Integrating Emerging Data Sources into Operational Practice

State of the Practice Review

www.its.dot.gov/index.htm
Final Report—December 2016
FHWA-JPO-16-424

U.S. Department of Transportation

Produced by Cambridge Systematics, Inc.
U.S. Department of Transportation
Office of the Assistant Secretary for Research and Technology

Notice

(Cover Image from iStockphoto.)

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA-JPO-16-424		

4. Title and Subtitle	5. Report Date
Integrating Emerging Data Sources into Operational Practice: State of the Practice Review	December 2016
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Doug Gettman, Kelsey Hales, Alison Voss, Alan Toppen, and Bheeshma Tumati	

9. Performing Organization Name And Address			10. Work Unit No. (TRAIS)
Kimley Horn and Associates 7740 N. 16th St. Suite 300 Phoenix, AZ 85020	and	Deloitte Consulting LLP 1919 North Lynn Street Arlington, VA 22209	
Under Contract to: Cambridge Systematics, Inc. 4800 Hampden Lane, Suite 800 Bethesda, MD 20814			11. Contract or Grant No. DTFH61-12-D-00042

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
U.S. Department of Transportation ITS Joint Program Office-HOIT 1200 New Jersey Avenue, SE Washington, DC 20590	Final Report
	14. Sponsoring Agency Code HOIT

15. Supplementary Notes

The GTM for the U.S. DOT is Jon Obenberger.

16. Abstract

The purpose of this report is provide agencies responsible for Transportation Systems Management and Operations (TSM&O) with an introduction to successful Big Data tools and technologies that can be used to aggregate, store, and analyze new forms of traveler-related data that may be useful for operations. While traditional sources of transportation data for TSM&O will remain, emerging data sources, largely those from Connected Travelers, Connected Vehicles, and Connected Infrastructure, will represent a significant opportunity for Departments of Transportation (DOTs) and localities to improve TSM&O practices. In addition, this report will identify ways these collection, storage, and analytics practices can be integrated into the next generation of transportation management systems. Big data techniques outside of the transportation field were considered, to identify practices that may be useful within the transportation field.

The first chapter reviews the common functions of TSM&O and provides a state of the practice summary of how data and information currently are acquired, processed, stored, and analyzed. The second chapter identifies emerging data sources from connected vehicles (CV), connected travelers, and other sources relevant to TSM&O and predicts the point(s) of access of these data to a DOT. Chapter 3 then characterizes each of the emerging sources by current and future volume and data velocity (the rate at which data is generated and the rate at which the data accumulates over time). The future data volumes are assessed at a national scale and at the scale of a "typical" agency. The data volumes for the national level are computed and presented only to assess the sheer scale. Chapter 4 provides an overview of Big Data tools and technologies. Chapter 5 then introduces the reader to the popular and common platforms for ingesting, processing, and analyzing large volumes of information. Lastly, chapter 6 introduces cost models for commercial tools and platforms.

17. Key Words	18. Distribution Statement
Transportation Systems Management and Operations (TSM&O), emerging data sources, connected travelers, connected vehicles, connected infrastructure, Big Data tools and techniques	No restrictions.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	132	N/A

I* (MODERN METRIC) CONVERSION FACTORS

APPROXIMATE CONVERSIONS TO SI UNITS				
SYMBOL	**WHEN YOU KNOW**	**MULTIPLY BY**	**TO FIND**	**SYMBOL**
in	inches	25.4	millimeters	mm
ft	feet	0.305	meters	m
yd	yards	0.914	meters	m
mi	miles	1.61	kilometers	km
in^2	square inches	645.2	square millimeters	mm^2
ft^2	square feet	0.093	square meters	m^2
yd^2	square yard	0.836	square meters	m^2
ac	acres	0.405	hectares	ha
mi^2	square miles	2.59	square kilometers	km^2
fl oz	fluid ounces	29.57	milliliters	mL
gal	gallons	3.785	liters	L
ft^3	cubic feet	0.028	cubic meters	m^3
yd^3	cubic yards	0.765	cubic meters	m^3
NOTE: volumes greater than 1000 L shall be shown in m^3				
oz	ounces	28.35	grams	g
lb	pounds	0.454	kilograms	kg
T	short tons (2000 lb)	0.907	megagrams (or "metric ton")	Mg (or "t")
°F	Fahrenheit	5 (F-32)/9 or (F-32)/1.8	Celsius	°C
fc	foot-candles	10.76	lux	lx
fl	foot-Lamberts	3.426	$candela/m^2$	cd/m^2
lbf	poundforce	4.45	newtons	N
lbf/in^2	poundforce per square inch	6.89	kilopascals	kPa

SYMBOL	**WHEN YOU KNOW**	**MULTIPLY BY**	**TO FIND**	**SYMBOL**
mm	millimeters	0.039	inches	in
m	meters	3.28	feet	ft
m	meters	1.09	yards	yd
km	kilometers	0.621	miles	mi
mm^2	square millimeters	0.0016	square inches	in^2
m^2	square meters	10.764	square feet	ft^2
m^2	square meters	1.195	square yards	yd^2
ha	hectares	2.47	acres	ac
km^2	square kilometers	0.386	square miles	mi^2
mL	milliliters	0.034	fluid ounces	fl oz
L	liters	0.264	gallons	gal
m^3	cubic meters	35.314	cubic feet	ft^3
m^3	cubic meters	1.307	cubic yards	yd^3
g	grams	0.035	ounces	oz
kg	kilograms	2.202	pounds	lb
Mg (or "t")	megagrams (or "metric ton")	1.103	short tons (2000 lb)	T
°C	Celsius	1.8C+32	Fahrenheit	°F
lx	lux	0.0929	foot-candles	fc
cd/m^2	$candela/m^2$	0.2919	foot-Lamberts	fl
N	newtons	0.225	poundforce	lbf
kPa	kilopascals	0.145	poundforce per square inch	lbf/in^2

* SI is the symbol for the International System of Units. Appropriate rounding should be made to comply with Section 4 of ASTM E380. (Revised March 2003).

Table of Contents

List of Tables

List of Figures

Chapter 1 Overview

The proliferation of data collected and stored from people and devices connected to the Internet is an important trend for businesses, individuals, and governments. Emerging data from travelers, vehicles, infrastructure, and other sources is expected to transform how agencies manage their transportation systems. The purpose of this project is provide agencies responsible for Transportation Systems Management and Operations (TSM&O) with an introduction to successful Big Data tools and technologies that can be used to aggregate, store, and analyze new forms of traveler-related data that may be useful for operations. In addition, this project will identify ways these collection, storage, and analytics practices can be integrated into the next generation of transportation management systems.

The project is divided into four reports, which the reader is encouraged to consider together as a complete set. The first report (this document) provides a review of the state of the practice in Big Data tools and technologies and characterizes the nature of emerging data sources that will need these tools to be effectively used. The second report identifies specific use cases for these Big Data approaches across common TSM&O practices in light of the availability of the new data sources. The third report details some proposed aggregation and edge-processing schemes to reduce the burden of the Department of Transportation's (DOT) Information Technology (IT) systems to consume and store all possible "raw" data, while retaining the maximum amount of information from the new sources. The fourth report then provides some recommendations on how these emerging sources and acquisition, processing, and analytics techniques can be integrated into future next generation transportation management systems.

The purpose of this report (report #1 of the 4 listed above) is as follows:

- Provide a summary of emerging data sources and their potential volumes relevant to TSM&O practices.
- Provide an overview of the current state of the practice in IT tools and technologies in the "Big Data" space.
- Raise awareness of TSM&O practitioners and IT professionals associated with TSM&O agencies of these coming trends and the order-of-magnitude challenges in data handling that this presents.

The intended audience will be those practitioners with some IT experience, and those that wish to gain better awareness of IT issues in Big Data. As more travelers and vehicles become connected and new sources of information relevant to TSM&O emerge, new ways of acquiring, processing, and storing data will be required if the data is to be transformed into information and used to improve operational practice.

Big Data is a term that typically is used loosely and in a nonspecific manner. This report provides some definitions and descriptions of current "big data" technologies and tools, which will help the reader to understand better what is "under the hood" of these Big Data systems. Information on the

predicted volumes of emerging sources is presented to indicate the scale of the data issues that could be expected for TSM&O agencies over the next 10 years. After reading this report, it will become clear to the reader that, if and when these data sources are made available for TSM&O purposes, the changes needed to DOT IT systems will be significant. The report also indicates the need for methods to preprocess much of the data before it is actually stored in DOT systems. Methodologies for this type of preprocessing will be discussed in report #3 of this project.

This report addresses five principal questions:

1. What are the categories of emerging data sources for TSM&O?
2. What will be the nature of these sources five and 10 years into the future (e.g., volume, velocity, cost, availability)?
3. What are the current industry trends in big data?
4. What are leading commercial tools and functionality of systems designed for big data?
5. What are the existing computational platforms and technology and their relative costs and capabilities?

After reading this report, it will become clear to the reader that, if and when these data sources are made available for TSM&O purposes, the changes needed to DOT IT systems will be significant.

The report is divided into six technical chapters. The first chapter reviews the common functions of TSM&O and provides a state-of-the-practice summary of how data and information currently are acquired, processed, stored, and analyzed. The second chapter identifies emerging data sources from connected vehicles (CV), connected travelers, and other sources relevant to TSM&O; and predicts the point(s) of access of these data to a DOT. Chapter 3 then characterizes each of the emerging sources by current and future volume and data velocity (the rate at which data is generated and the rate at which the data accumulates over time). The future data volumes are assessed at a national scale and at the scale of a "typical" agency. The data volumes for the national level are computed and presented only to assess the sheer scale. The remainder of the project will only consider requirements for solutions and technologies for management of the new information at regional scale.

Chapter 4 provides an overview of Big Data tools and technologies. "Big Data" as a buzzword means little without some concrete information describing the moving parts. Chapter 5 then introduces the reader to the popular and common platforms for ingesting, processing, and analyzing large volumes of information. Chapter 6 introduces cost models for commercial tools and platforms.

This project builds upon previous work in this area, which are listed in the References section. These sources provide an excellent introduction to the subject of big data and what it could mean for transportation operations. (U.S. Department of Transportation, Intelligent Transportation System (ITS) Joint Program Office, "Big Data's Implications for Transportation Operations: An Exploration," Publication No. FHWA-JPO-14-157, December 2014; McKinsey Global Institute, "Big Data: The next frontier for innovation, competition and productivity," May 2011. Accessed at: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.; Kimley-Horn and Associates, Inc., "Traffic Management Centers in a Connected Vehicle Environment," Transportation Management Center (TMC) Pooled Fund Study, March 2014.; U.S. Department of Transportation, ITS Joint Program Office, "Big Data and ITS," White

Paper, October 2013. Accessed at: http://connectedvehicle.itsa.wikispaces.net/file/detail/ITS+and+Big+Data+White+Paper+Final+Draft+10_2+%282%29.docm; International Transport Forum, "Big Data and Transport: Understanding and assessing options, 2015." Accessed at: http://www.itf-oecd.org/big-data-and-transport-understanding-and-assessing-options.) In addition, these references introduce the opportunities and challenges of using crowdsourced data for TSM&O, provide background details on the U.S. DOT Connected Vehicles program data architecture and standards, and identify Traffic Management Center (TMC)/TSM&O agency trends. The reader is encouraged to review these documents for additional background as this report seeks to expand on these sources with new material. (U.S. Department of Transportation, ITS Joint Program Office, "Estimate Benefits of Crowdsourced Data from Social Media," Publication No. FHWA-JPO-14-165, February 2015; American Association of State Highway and Transportation Officials (AASHTO), "National Connected Vehicle Field Infrastructure Footprint Analysis," Publication No. FHWA-JPO-14-125, June 2014.)

Chapter 2 State of the Practice of Data for Transportation Systems Management and Operations

The purpose of this chapter is to characterize the state of the practice in use of data for Transportation Systems Management and Operations (TSM&O). This chapter introduces categories of activities that are performed by TSM&O agencies and the types of data that feed these activities. After reading this chapter, the reader will understand the types of agency activities and systems that may be enhanced by the use of emerging data sources that will be available over the next 10 years. The next chapter identifies categories of emerging data sources and estimates their likely point(s) of access to the Departments of Transportation (DOT).

Chapter Objectives:

- Identify state of the practice in using data for TSM&O activities.
- Identify current limitations of TSM&O practices and how emerging sources may enhance operations.

The National Operations Center of Excellence (NOCoE) defines TSM&O as:

"An integrated program to optimize the performance of existing infrastructure through the implementation of systems, services, and projects designed to preserve capacity and improve security, safety, and reliability of the transportation system.

The term includes regional operations collaboration and coordination activities between transportation and public safety agencies; and improvements to the transportation system such as traffic detection and surveillance, arterial management, freeway management, demand management, work zone management, emergency management, electronic toll collection, automated enforcement, traffic incident management, roadway weather management, traveler information services, commercial vehicle operations, traffic control, freight management, and coordination of highway, rail, transit, bicycle, and pedestrian operations."

TSM&O is the active management of the multimodal transportation network by collecting data on system performance and making adjustments to real-time controls, information, and demand-management strategies. TSM&O is not the implementation or construction of new facilities or rebuilding existing facilities and a variety of other functions of DOTs. Other functions of DOTs could very well be affected by new emerging data sources, including the data sources identified in this report, but are not discussed here. In addition, those additional DOT functions may very well be improved by application of the same data acquisition, marshalling, and analysis tools and platforms.

TSM&O practices that will likely be affected by the availability of new data include, but are not limited to, the following:

- Incident and event management.
- Road hazard warnings.
- Speed warnings.
- Traffic signal timing.
- Freeway ramp metering.
- Variable speed limits/recommendations and lane-use control strategies.
- Dynamic message sign displays.
- Work zone implementation.
- Broadcasted and Personalized Traveler information.
- Congestion pricing, road user fees, and tolls.
- Performance measurement, including weather and emissions monitoring.
- Asset management and maintenance.

Data Sources of Today

TSM&O organizations have been connecting to infrastructure to obtain information for operations for more than 40 years. All of the core missions of TSM&O organizations are supported by the collection of device status and sensor data. Connected infrastructure devices that provide data for TSM&O include traffic signals, ramp meters, Closed-Circuit Television (CCTV), vehicle detection stations, Road Weather Information Systems (RWIS), flood warning sensors, high wind warning sensors, and a variety of other specialty devices. Incident data is the other major existing source of information used by TSM&O agencies. Each of these devices and data sources for TSM&O will be discussed briefly in the following sections to identify existing limitations and opportunities.

Traffic Signals

Traffic signals are capable of reporting second-by-second status of every controllable phase and every detector connected to the signal for local operation and conditions monitoring. General signal status includes the operating mode, such as coordination, free, or flash; and more detailed alarms and specialty conditions, such as stop time, local manual, preemption, transit priority, cycle time, plan parameters, and pedestrian and bicycle activity. Traffic detectors come in a variety of forms, including in-pavement inductive loops, video, and radar. Typical information from detectors is the presence of a large metallic object within a certain region of the pavement at a certain time. This information is used in real time by the signal to indicate the need to turn the light green for a particular direction of travel, and keep it green to service the demand. There is much more complexity to traffic signal operations than is necessary to discuss here.

The main limitation of current data collection of traffic signals is that the information contains no indications of individual vehicles' intent (where they came from or are going to) and vehicles that are not currently in a detector region are virtually invisible to the system. This becomes especially challenging in highly congested conditions when the signal, or signal system attempting to coordination operations of multiple intersections (on an arterial, for example), cannot know what the true demands are because the control algorithms cannot "see" past where their detectors are located. While there is a variety of sophisticated methods developed and used by current TSM&O organizations (generally characterized as Adaptive Signal Control Technology (ASCT)) to handle these situations, new data sources from connected vehicles and travelers can help to improve signal operations immensely. Pervasive availability of data from connected vehicles and travelers may reduce agency burden for maintenance of sensors and associated systems.

Currently, central systems typically collect and store this status information in a Relational Database Management System (RDBMS). Users of the central system can view real-time operations on maps and tabular displays, and many systems have detailed aggregation algorithms and tools for analyzing performance in various ways. The new emphasis on active performance management of signal operations has emerged over the past five to seven years through the Every Day Counts program, which focused on the adoption of ASCT. However, much of the data collected by traffic signal systems is simply deleted by the RDBMS after a specified number of days in order to maintain responsive database performance and for various other institutional reasons, such as the need to respond to records requests from accident-injury lawyers for data that is sufficiently "old." There certainly is an opportunity to use this rich historical performance information in new ways to improve practice using new tools and technologies that are designed to handle large volumes and a large variety of data, even without any introduction of new data from connected vehicles, connected travelers, or other sources.

There is certainly an opportunity to use this rich historical performance information in new ways to improve practice using new tools and technologies that are designed to handle large volumes and variety of data, even without any introduction of new data from connected vehicles, connected travelers, or other sources.

Ramp Meters

Like traffic signals, ramp meters are capable of reporting second-by-second status of every controllable lane and every detector connected to the signal for local operation and conditions monitoring. Ramp meters generally have detection zones on the freeway that are used to measure local conditions in order to set the rate at the ramp to a reasonable input level. When congestion is high on the freeway, metering rates are reduced; and when congestion is low, metering rates are increased (or the meter is turned off completely). Ramp meters also generally have queue detectors upstream of the stop line that are used to measure demand, as well as change operations to "flush the queue" if those detection zones are determined to have vehicles sitting in that location for extended periods of time. While ramp meters sometimes meet public opposition, they are proven to reduce crashes in the merging area and generally reduce congestion on the freeway.

The main limitation of current data collection of ramp meters is similar to that for traffic signals. The information contains no indications of individual vehicles' intent (where they came from or are going to), and vehicles that currently are not in a detector region are virtually invisible to the system. This becomes especially challenging in highly congested conditions when the ramp queue is large and the congestion on the freeway also is determined to be high. Sophisticated algorithms for setting ramp

metering rates in a corridor have existed for more than 30 years, but implementation (like ASCT on arterials) has not been widespread, typically due to the need to maintain the traditional detection assets that are needed to fuel the algorithms and traffic models. New data sources from connected vehicles and travelers can help to improve ramp metering operations immensely. Even existing sources from third-party link-speed providers would enable new methods of smarter ramp metering responses. Integrating ramp metering responses with signal timing on arterials is another opportunity area where new sources of data would improve practice.

Closed-Caption Television

Real-time streaming feeds from fixed-location cameras have played a prominent role in TSM&O operations for the past 30 years. Closed Caption Television (CCTV) are capable of reporting subsecond status of their field of view. Both fixed-aspect and pan-tilt-zoom (PTZ) cameras are used extensively for freeway, arterial, and transit management. Video analytics is frequently used by agencies for incident detection, typically in tunnels. Cameras on arterials also are capable of performing some vehicle counting and conditions monitoring in addition to their role in actively operating the traffic signal. Cameras have similar limitations to in-ground pavement loops and other similar technologies that they only know of the presence of a large object within a certain region of the pavement at a certain time. While some newer CCTV systems have capabilities to track objects in the field of view (and/or fusing views from multiple cameras or using fish-eye lenses) and store trajectories of these objects for analytics, this practice is not yet common in TSM&O. Similar to traffic signal status data, most TSM&O agencies do not retain CCTV images for any extended period of time, typically for institutional reasons, but also because no analytics methods are readily available to extract information from historical trends of what the camera saw at any particular time of day, day of week, or under certain conditions. There is certainly an opportunity to use this rich historical performance information in new ways to improve practice using new tools and technologies that are designed to handle large volumes and variety of data, even without any introduction of new data from connected vehicles, connected travelers, or other sources. New data sources from connected vehicles and travelers may reduce agency needs to invest in further CCTV coverage.

Vehicle Detection Stations

Vehicle detection stations are capable of reporting the second-by-second status of every detector connected to the collection station for local operation and conditions monitoring. Typically, the data is aggregated on the station controller and communicated to the central system in 20- or 30-second summaries of traffic volume, speed, and occupancy. This data is then typically used to color the link of a conditions map red for high congestion (low speed, high occupancy) and green for low congestion (high speed, low occupancy). Volumes are used for reporting and trend analysis. Some sophisticated algorithms are used for incident detection by correlating the data from multiple detection stations. These methods are common, but typically are less reliable than incident reports from 911 call centers (via the public). Similar to traffic signals, the typical information from detectors is the presence of a large metallic object within a certain region of the pavement at a certain time.

The main limitation of vehicle detection technology is that the information contains no indications of individual vehicles' intent (where they came from or are going to), and vehicles that currently are not in a detector region are virtually invisible to the system. This becomes especially challenging in highly congested conditions when the speed drops to virtually zero, and there is almost always a vehicle in the detection zone. These stations and their associated central systems have difficulty in reliably determining the "back of queue." New sources of third-party link-speed data (derived from trajectories

of probe vehicles) have improved this practice immensely. Most TSM&O agencies now find that Google maps with traffic is now much more accurate than their traffic conditions maps driven by vehicle detection stations. New data sources from connected vehicles and travelers will help to improve freeway and arterial conditions monitoring; and it is expected that over the next 10 years, many more TSM&O agencies will continue to abandon maintenance of vehicle detection stations (VDS), and take advantage of other more reliable sources from third-party providers.

Like traffic signal data, central systems typically collect and store VDS status information in an RDBMS. Users of the central system can view real-time operations on maps and tabular displays, and many systems have detailed aggregation algorithms and tools for analyzing performance in various ways. New emphasis on active performance management of freeway operations has emerged over the past 10 years, resulting in systems like the California Department of Transportation's (Caltrans) Freeway Performance Measurement System (PeMS) and Maryland's Regional Integrated Transportation Information System (RITIS). These systems come closest to managing "Big Data" with the Caltrans PeMS system, hosting over 10 years of freeway conditions data, incident data, and other information in a multi-Terabyte-distributed Oracle RDBMS. PeMS and RITIS represent the closest match of existing systems that perform data acquisition, marshalling, and analysis functions for TSM&O organizations. Other agency systems that perform similar functions include the Florida TSM&O tool and Nevada's Freeway and Arterial System of Transportation (FAST) performance dashboards. These systems will be discussed further in the next report of this project.

Other Sensors

Other sensors used by TSM&O agencies include environmental sensor stations, which are detection stations with a variety of weather sensors for road surface temperature, precipitation, high wind, visibility, water lever, etc. Other sensor types include truck escape ramp sensors and a variety of other specialty devices. Most of these sensors are capable of reporting second-by-second status, but typically the reporting from these sensors is on a much less frequent basis, such as 30 seconds or one minute. Status changes from these devices on more frequent increments typically are not useful. This data is then used to color the icon of the device on a conditions map to a certain status (cold or warm, raining or not, etc.). Historical data are stored and used for reporting and trend analysis.

The main limitation of current data collection from specialty sensors is that the information is only related to the specific location where the sensor is deployed. Weather sensors, for example, can show rain in that specific spot when the surrounding area is dry (if you have ever traveled in the mountains of Colorado, for example, microbursts are quite common). Agencies have stretched their limited resources to deploy such devices in key locations, such as in mountain passes or at freeway-to-freeway junctions. New sources of open data from the National Weather Service, for example, has helped to fill in many of these gaps in existing data collection. New data sources from connected vehicles and travelers will vastly help to improve freeway and arterial conditions monitoring with respect to road-weather conditions across the agency's network and reduce costs of agencies to deploy monitoring sensors.

Like the other sensor sources, central systems typically collect and store specialty sensor status information in a RDBMS. Users of the central system can view real-time operations on maps and tabular displays, and many systems have detailed aggregation algorithms and tools for analyzing conditions in various ways. However, much or most of the data collected by these sensors is simply deleted by the RDBMS after a specified number of days in order to maintain responsive database performance. There is certainly an opportunity to use this rich historical performance information in

new ways to improve practice using new tools and technologies that are designed to handle large volumes and variety of data, even without any introduction of new data from connected vehicles, connected travelers, or other sources.

Incident Data

Most State DOTs and regional management agencies use incident data feeds from law enforcement agencies for incident response and traveler information. Most larger agencies are now obtaining link-speed and incident indication information from third-party data providers (HERE, Inrix, TomTom, and others). A small and growing subset of agencies are obtaining somewhat more granular vehicle status data from crowdsourcing apps (i.e., Waze). While most of the information in an incident report is standardized, many incident reports from law enforcement agencies have textual descriptions of location and onsite activity status (officers onsite, officers en-route, number of injuries or vehicles involved, etc.), which must be manually deciphered by operators in a Traffic Management Center (TMC) to accurately locate the incident on a map and encode in incident management logs. Crowdsourcing data typically is automatically geolocated in the reported status.

Like the other sources, central systems typically collect and store incident status information in an RDBMS. Users of the central system can view real-time status on maps and tabular displays; and many systems have detailed algorithms and tools for response planning, such as automatically placing messages on multiple dynamic message system (DMS) signs at the same time, pointing CCTV cameras at the location, and posting information to 511. Incident data typically is not deleted by the RDBMS with the same veracity as other data sources after a specified number of days in order to maintain responsive database performance, since the tables do not grow at the same rate as the tables that store sensor information. However, there is still certainly an opportunity to use this rich historical performance information in new ways to improve practice using new tools and technologies that are designed to handle large volumes and variety of data. Systems, such as PeMS and RITIS, have tools for this purpose that have been proven useful for historical analysis.

Opportunities to Improve Practice

As discussed in the previous section, current sources of data for TSM&O are heavily focused on connected infrastructure and proven technologies. Database and data management practices tend to be conservative, simply because Information Technology (IT) and hardware investments must last for 10 years or more, given agency procurement cycles and limitations of funding. TSM&O agencies do not have complete situational awareness from spot-sensors that are geographically dispersed with limited capabilities. As the worldwide trend of mobile technology has advanced, data from emerging mobile sources can improve a wide range of TSM&O practices in the following ways:

- Incident and event management—improved incident response, onsite monitoring, and management.
- Road hazard warnings—higher fidelity location information, more accurate confirmation of hazard types, more timely warnings.
- Speed warnings—specific recommendations to different vehicle types based on roadway conditions, more timely warnings.

- Traffic signal timing—better operation in oversaturated conditions, more timely updates to fixed timings, broad-based adaptive controls, reduced reliance on physical sensor devices and maintenance, shift towards in-vehicle data delivery, performance monitoring of signals with no physical links to DOT communications infrastructure.
- Freeway ramp metering—more accurate and coordinated corridor metering algorithms.
- Variable speed limits/recommendations and lane-use control strategies—more accurate and coordinated responses, shift towards in-vehicle signage reducing needs for infrastructure investments.
- DMS displays—more accurate messaging, shift towards in-vehicle signage for more personalized recommendations, reduced need for infrastructure investments.
- Work zone implementation—higher safety for workers and drivers, higher resolution maps of work zone geometries, real-time information on new zone locations, less need to manually update locations.
- Broadcasted and Personalized Traveler information—higher fidelity information, more accurate and timely information, personalized recommendations.
- Congestion pricing, road user fees, and tolls—more granular toll rates, more accurate congestion prices, personalized tolls, and road user fees.
- Performance measurement, including weather and emissions monitoring—higher fidelity analysis, more comprehensive coverage of geography, reduced need for infrastructure investments.
- Asset management and maintenance—reduced need for infrastructure investments, faster detection and response to equipment failures.

Chapter 3 Categories of Emerging Data Sources

Chapter Objectives:

- Identify data sources that will be available for TSM&O over the next 10 years.
- Identify how TSM&O agencies will access these data.

The purpose of this chapter is to identify emerging data sources for Transportation Systems Management and Operations (TSM&O). An ***emerging data source*** is characterized by information that is potentially relevant for TSM&O, and that has not yet been widely capitalized on. New technologies are anticipated to accelerate the availability of this data over the next two decades. This chapter introduces categories of emerging data sources, and assesses the likely point(s) of access to the TSM&O agency. After reading this chapter, the reader will understand the types of emerging data that will be available over the next 10 years. The next chapter provides estimates of the volumes, velocity, and storage needs for these emerging data.

We have categorized Emerging Data Sources into four general classifications:

1. Connected Travelers.
2. Connected Vehicles (with three subcategories).
3. Connected Infrastructure.
4. Other Potential Sources (with three subcategories).

Each of these categories will be discussed separately in the following narrative.

While the information from these sources may be relevant to a wide variety of agency activities, the focus of this report is on the applicability of the information for TSM&O strategies. Each source also is characterized by the way in which the data will likely be consumed by TSM&O organizations (i.e., the ***point of access***). The point of access is important in identifying how acquisition, marshaling, and analysis tools and technologies can be applied appropriately.

The report will emphasize real-time data (for use in operations), but historical data may be equally important as data storage and processing capacities improve with new tools that will be presented later in this report. Historical data, while typically the purview of planning departments, can provide performance measure insights to TSM&O agencies, particularly as the ability to predict traffic based on current and historical data will likely become more reliable over time.

Connected Travelers

The rapid adoption of the smartphone has enabled hundreds of new ways that travelers exchange information with agencies and commercial entities related to travel and transportation. As of 2015,

more than 68 percent of adult Americans own a smartphone, and this number will only increase. Of those ages 18 to 29, 86 percent have a smartphone, as do 83 percent of those ages 30 to 49. (Pew Research Center, "U.S. Technology Device Ownership 2015," Accessed May 13, 2016, http://www.pewInternet.org/2015/10/29/technology-device-ownership-2015.) As of 2011, there were more cell phones and tablets in the U.S. than the entire American population—328 million connected devices compared to 312 million people. (CNN Money, "U.S. cell phones, tablets outnumber Americans—Oct. 12, 2011," Accessed May 13, 2016, http://money.cnn.com/2011/10/12/technology/cellphones_outnumber_americans/index.htm.)

Virtually ubiquitous 3G/4G cellular networks and prevalent open Wi-Fi networks allow travelers to achieve almost uninterrupted connectivity. Several studies have indicated that millennials (people who are between 18 to 34 years old as of 2016), in particular, would rather be without a car than a smartphone. (Zipcar, "Millennials & Technology: A Survey," Accessed May 13, 2016, http://www.slideshare.net/Zipcar_Inc/millennial-slide-share-final-16812323.) Young adult adoption rates of driver licenses is now lower than any point since 1980. (Frontier Group and U.S. Public Interest Research Group (PIRG) Education Fund, "Transportation and the New Generation: Why Young People Are Driving Less and What It Means for Transportation Policy," Accessed May 13, 2016, http://www.uspirg.org/sites/pirg/files/reports/Transportation%20%26%20the%20New%20Generation%20vUS_0.pdf.) While video conferencing and telecommuting may reduce work-related travel demand, and social media have connected groups of like-minded individuals without the need for physical meet ups, the majority of the workforce still needs to commute. Furthermore, people enjoy meeting face-to-face for recreation, entertainment, and meals. When people leave home, they almost always travel with their phone or Web-enabled tablet device. While there will always be a small segment of the population that is not connected to the network (e.g., does not own a smart device, or device currently is inoperable), the majority of travelers already are connected to a suite of apps and services (i.e., Wi-Fi, global positioning system (GPS), data, etc.) through a personal device that knows their physical location on a relatively granular level (accurate within several meters and updated every several seconds).

While traveler information once consisted solely of push notifications to the traveler, apps are now collecting information about the user's activities and location to provide content to the app itself (e.g., Google Maps, Waze, etc.), but also to personalize information for the user based on their current location. Crowdsourcing is a popular term for organically collecting data on field conditions from mobile devices. Examples of location-aware mobile applications that provide value to users include Moovel (formerly RideScout), Uber, or MaaS, which are "mobility as a service" companies that are mode agnostic and seek to provide the best available option to get from Point A to Point B, including options such as transit, bicycle, or taxi. These companies collect user information in order to provide individualized recommendations via user profiling. This ***"digital exhaust"*** or breadcrumbs of traveler location, activities, and status (e.g., riding in a car, riding on transit, walking, biking, etc.) hold significant promise for TSM&O activities. With the important caveat that privacy remains critical, these services present a new source of information on traveler behavior that was once only available via expensive and time-consuming travel surveys. In the past, rarely, if ever, have TSM&O strategies incorporated traveler demographics and behaviors, but now, this may be changing with the potential availability of this kind of data.

Some TSM&O use cases for connected traveler data include:

- Populating agency traffic condition maps.
- Populating regional or project-specific origin-destination matrices.

- Identifying transit ridership and usage patterns.
- Identifying bicyclist and pedestrian usage patterns.
- Determining the traffic impacts of proposed development and construction projects.
- Identifying locations of incidents and traffic-impacting events.
- Updating traffic signal timing.

TSM&O use cases of Emerging Data Sources will be explored in further detail in a subsequent report. Additional information is available in the References section [1-7].

Points of Access for Connected Traveler Data

TSM&O organizations currently can obtain connected traveler data from at least three different methods:

1. Directly through an agency-branded app.
2. Indirectly through a third-party source.
3. Indirectly though social media outlets (e.g., posts, tweets, feeds, etc.).

Almost every region with a 511 system has a branded 511 app (or suite of apps for relaying traffic conditions and transit schedules). Many agencies also have citizen reporting apps, which allow the public to report infrastructure issues such as potholes. Examples of agency-branded apps include the Utah Department of Transportation (DOT) Citizen Data program and the Los Angeles Metropolitan Transportation Authority (MTA) 511 app. (Google Play Store, "Utah Department of Transportation (UDOT) Citizen Reports—Android Apps on Google Play," Accessed May 13, 2016, https://play.google.com/store/apps/details?id=gov.utah.udot.citizenreport.; Los Angeles County Metropolitan Transportation Authority (Metro), "Metro Mobile App," Accessed May 16, 2016, https://www.metro.net/mobile/metro-mobile-app.) Current 511 apps typically are only data-push applications that do not record or store user-related information. Adoption rates generally are not significant compared to other commercial products (in major regions typically less than two to three percent of travelers). These apps could be leveraged, with appropriate privacy protocols, for collecting traveler data useful for TSM&O activities. It could be argued that adoption rates of such 511 apps in the future could be increased if new, location-based features and functionality were provided; in particular, features and functions that only TSM&O agencies can provide. The value proposition would be to provide such functions in exchange for traveler behavior data, which could enhance TSM&O activities.

Higher adoption rates would lead to a need for tools and technologies to store and process the information. If an agency-branded app becomes extremely popular, the data load on the agency's cloud or physical servers may require commercial capacity levels. If the app is not widely adopted, it may not warrant sophisticated tools and approaches for data management.

Indirect data collection by DOTs through a third-party source will almost always be via the Internet and stored in the Cloud or physical servers. Note that today many DOTs do not utilize cloud-based storage and applications, but in line with the general trends of Information Technology (IT) infrastructure, it will likely happen in the near future. Indirect data collection through a third-party source could include the provision of aggregated datasets or raw data on individual travelers that is appropriately anonymized.

A real-time example of connected traveler data collection is the partnerships that many agencies now have with Waze. Waze provides almost raw connected traveler information to agencies for free in exchange for agency data on traffic incidents and work zones, as well as closed-circuit television (CCTV), dynamic message sign (DMS), and other field device feeds for sharing on the Waze app. The primary delivery mechanism is a real-time Application Programming Interface (API) that provides individual traveler speeds on a segmented network of roadways relevant to the agency (e.g., within State boundaries). In one of the first Waze partnerships with Rio de Janeiro in Brazil, the app has generated only around 110,000 users among the City's 6.45 million total population. (World Population Review, "Rio De Janeiro Population 2016—World Population Review," Accessed May 13, 2016, http://worldpopulationreview.com/world-cities/rio-de-janeiro-population.) Note that penetration rates of Waze in other countries, such as Israel (where it was invented), are greater than 90 percent. In addition, considering the fact that Android users cannot uninstall Google Maps, the market penetration rate of Waze is roughly equivalent to the number of Android users that use location-based services (Waze is a separate traveler information app, but is owned by Google and uses the traffic conditions information from Google Maps.). Similar restrictions are true for Apple iOS devices and Apple map products. Individual user trips or trip history, even anonymized, currently are not shared by Waze, Google, or others (Waze shares link speeds as reported by individual users, but not their entire trip.). Although many app developers will note that they do keep databases of user trip histories, and the default is almost always set for the user to "opt in."

In a less real-time manner, many agencies also purchase travel data from third parties, such as AirSage and Cellint, which track users' cell phone movements throughout the cellular network. The movement of phones while engaged in a moving vehicle can provide a rough notion of origin-destination flows. Typically, this data is purchased on a one-time basis for a specific project or analysis, via a File Transfer Protocol (FTP) download, .ZIP file transfer, or database transfer. However, this type of traveler information does not provide a completely accurate representation of travel flows, since not all drivers are engaged in phone conversations, and not all phone conversations continue from origin to destination. The data also can grow to sizes that may require application of Big Data tools or technologies for robust analysis.

Parsing of open social media feeds is a popular trend in marketing and image management. In fact, there is an entire industry devoted to social media monitoring to help businesses identify the effectiveness of marketing campaigns and new product introductions, and to evaluate consumer sentiment regarding a personality, technology, or social topic. Applicability for TSM&O activities remains challenging as precise location-based data is not always included in user messages, and messages often include inconsistent, abbreviated spelling. Application is better suited for major emergency situations that evolve over an extended period of time than for everyday traffic incidents or congestion. The point of access for parsing open social media feeds is similar to the collection of individual traveler trips from third parties; it will most likely be collected from a third-party cloud, and stored in the TSM&O agency cloud or physical server.

Connected Vehicles

Connected vehicles are divided into three categories of emerging data sources:

1. "Proprietary" (commercial) connected vehicle systems.

2. "Open" connected vehicle systems (i.e., U.S. DOT-sponsored technologies; Dedicated Short-Range Communications (DSRC)).
3. Radio Frequency Identification (RFID), Wi-Fi, and other technologies.

Each type has a different point or points of access, relevant to the implications for application of advanced tools and technologies for use in TSM&O.

Commercial Connected Vehicle Systems

Commercial connected vehicles include cellular connections to a private cloud from the vehicle's infotainment system or third-party in-car systems for vehicle tracking and data collection. Currently, commercial in-vehicle systems primarily are used for the purpose of gaining Internet access for passengers' nomadic devices and infotainment systems.

Nearly every automaker now offers connected car options. In this rapidly evolving area, automakers, telecommunications providers, technology companies, and content producers are establishing strategic alliances to provide the full range of services necessary for new and emerging products. According to a 2015 report by PricewaterhouseCoopers (PricewaterhouseCoopers, "Connected Car Study 2015: Racing ahead with autonomous cars and digital innovation, 2015," Accessed at http://www.strategyand.pwc.com/reports/connected-car-2015-study.), the primary use cases for automakers and their partners are:

- Entertainment.
- Mobility management (e.g., navigation, traffic, incidents).
- Driver assistance/safety.
- Vehicle health monitoring/recalls/remote diagnostics.
- Driver health monitoring.
- Fleet management.
- Insurance premium evaluation.
- Autonomous driving.

These are potentially rich sources of data on driver and vehicular behavior, which can significantly benefit consumers and auto suppliers. However, there has been little discussion to date on sharing this type of data with DOTs, although momentum is building for using commercial aftermarket devices for setting insurance premiums and road mileage rates as an alternative to gasoline taxes. (Oregon DOT, "MyOReGO | A new way to fund roads for all Oregonians," Accessed May 13, 2016, http://www.myorego.org.)

Third-party vehicle tracking and data collection companies, such as INRIX, HERE, TomTom, Garmin, etc., have monetized vehicle status information (primarily location and speed) for sale to DOTs on a subscription basis, providing coverage in a DOT's area of influence (a State or region). These companies have sharing agreements with vehicle fleets (and in-dash navigation systems of some Original Equipment Manufacturers, or OEMs), as well as data sharing agreements with private owners of aftermarket navigation devices. In the U.S. alone, these companies collect data in excess of a terabyte per month. (Texas A&M Transportation Institute, "Strategic Research Program: Big Data Scan," Accessed May 13, 2016, http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/161505-1.pdf.) As of June 2016, HERE has published a connected vehicle data sharing standard, which may greatly accelerate the availability of trajectory-based commercial connected vehicle data to DOTs. (HERE, "HERE, automotive companies move forward on car-to-cloud data standard," Accessed July 1, 2016, https://lts.cms.here.com/static-cloud-content/Newsroom/290616_HERE_automotive_companies_move_forward_on_car_to_cloud_data_standard.pdf.)

As of June 2016, HERE has published a connected vehicle data sharing standard, which may greatly accelerate the availability of trajectory-based commercial connected vehicle data to DOTs. (HERE, "HERE, automotive companies move forward on car-to-cloud data standard," Accessed July 1, 2016, https://lts.cms.here.com/static-cloud-content/Newsroom/290616_HERE_automotive_companies_move_forward_on_car_to_cloud_data_standard.pdf.)

Many DOTs are now routinely purchasing this data to supplement their existing vehicle speed monitoring systems (i.e., in-pavement loop detectors, radar, and video) and meet section 1201 Federal requirements for the dissemination of real-time mobility information. These suppliers also have the ability to provide archived data to DOTs for analysis and planning applications. Note that there are sometimes restrictions on the use of this purchased data, and agencies must be aware of these restrictions before entering purchasing agreements.

There also is sizeable interest from the Cellular industry to provide DSRC-like services by upgrading existing infrastructure. Recent research and development indicates that low-latency vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications can be provided by use of Long Term Evolution (LTE) cellular phone towers by locating the switching hardware (and/or software) at the tower, instead of going to the Cloud and then back down to the other vehicle. This solution can compete with the low-latency abilities of DSRC if the frequency of vehicle status updates is reduced to 20Hz (5 updates per second versus 10 updates per second for DSRC). (AT&T Labs Research, "Enabling Vehicular Safety Applications over Long-Term Evolution (LTE) Networks," Accessed May 13, 2016, http://web2-clone.research.att.com/export/sites/att_labs/techdocs/TD_101260.pdf.) Similarly, the major communication system providers already are working towards 5G wireless systems, which are geared toward supporting the Internet of Things (IoT)—the network of physical objects (devices, vehicles, machines, etc.) embedded with electronics, software, sensors, and network connectivity that enables these objects to collect and exchange data—with expected widespread adoption in the 2019 to 2022 timeframe. (IEEE Spectrum, "Autonomous Driving Experts Weigh 5G Cellular Network Against Dedicated Short Range Communications," Accessed May 13, 2016, http://spectrum.ieee.org/cars-that-think/transportation/self-driving/autonomous-driving-experts-weigh-5g-cellular-network-against-shortrange-communications-to-connect-cars.) The implications of IoT will be discussed in more detail throughout this document.

5G technology is expected to double or triple 4G bandwidth, but continue to have the same limitations as existing cellular architectures—download bandwidth to the remote device is significantly higher compared to the upload bandwidth.

Point of Access: Private Cloud to DOT Cloud or Physical Point

In the case of third-party providers of segment-based speed data, this information is made available to DOTs from a private cloud to a DOT cloud (a cloud system owned, operated or leased by DOTs) or physical server(s) through a defined API. If vehicular condition data was made available directly from OEM systems, it would most likely use a similar data exchange method. Direct availability of vehicular condition data from the OEM or the devices installed for insurance premium settings remains elusive. Since there are existing business models for third-party providers of this type of information, it is perhaps more likely that these existing supply-chain portals would be expanded to include more and more commercial connected vehicles rather than the introduction of competing products provided directly by OEMs. The road user fee pilot, conducted by Oregon DOT, is perhaps the leading prototype model for how commercial Connected Vehicle (CV) technologies could be used for TSM&O, since the usage and travel data of individual road users already is being provided to the DOT via "opt in" data feeds. The road user currently has the option to either "opt-in" and share location data with Oregon DOT, or "opt-out" and only provide mileage information in order to compute the usage fee.

U.S. Department of Transportation/Public Connected Vehicles (Dedicated Short-Range Communications)

In contrast to the proprietary systems being developed by private companies seeking to compete for sales and revenue, open connected vehicle platforms are being developed by U.S. DOT for mobility and safety applications. This program has a long history; and background details are available from National Highway Traffic Safety Administration (NHTSA), U.S. DOT, Government Accountability Office (GAO), American Association of State Highway and Transportation Officials (AASHTO), and other related sources. Open connected vehicle systems rely on DSRC technology to send vehicle status data to other vehicles and infrastructure access points with very low latency (~20 times per second (50 ms) from transmission to receipt). The low latency requirement is necessary for crash prevention safety applications. NHTSA has announced a notice of proposed rulemaking, which mandates the use of DSRC radios in all new passenger vehicles sold in the United States. If this rule is officially enacted, a similar rule for commercial vehicles (including buses) will likely follow. The DSRC band (5.9GHz) currently is protected by the Federal Communications Commission (FCC) for licensed use of vehicle safety and mobility applications.

In the U.S. DOT connected vehicle platform, each vehicle broadcasts its "heartbeat" Basic Safety Message (BSM) information containing its speed, location, heading, etc. every 100ms (10ms). Any other vehicle or roadside infrastructure device that is equipped with the DSRC equipment and is within line-of-sight to the transmitting vehicle can receive the heartbeat information. Raw BSMs by themselves are snippets of a vehicle's status at the specific time and location where the message was broadcasted (i.e., a single point of digital exhaust). In addition to BSMs, a Probe Data Message (PDM) encapsulates a string of "snapshots" (a more comprehensive data element than the BSM) over a longer timeframe to provide a vehicle trajectory information that could be shared with a roadside unit (RSU).

The snapshots are added to the PDM at fixed time intervals, at certain events, when a vehicle starts or stops, or based on vehicle speed which means that slow-moving vehicles would have PDMs that cover shorter distances, and faster moving vehicles (say, on a freeway) would have PDMs that cover longer distances. The PDM currently is defined as having a fixed-size buffer, storing a maximum of 32 snapshots (although varying numbers for the snapshot storage limits currently are being discussed in the standards development process) before starting a new PDM. If the DSRC-equipped vehicle passes an RSU requesting PDMs, it transmits its current PDM which can cover up to 120 seconds or 1km whichever comes last. If no RSU is in range, the PDM snapshots are deleted from the 32 in the buffer in a defined manner preferentially to maintain as long as a trajectory as possible with interim points deleted.

Other messages that can be broadcast by the vehicle include the Signal Request Message (SRM). This message allows a transit, freight, emergency, or other authorized vehicle (which could in theory include any DSRC-equipped vehicle) to request priority at a traffic signal. This application has been successfully demonstrated in the Multimodal Intelligent Traffic Signal System (MMITSS) tests in Arizona and California. Transit priority requests are stored in DOT databases for the entire test duration. As a result, this project has successfully demonstrated the derivation of signal performance metrics from BSMs through simulation. (University of Arizona, Multimodal Intelligent Traffic Signal Systems (MMITSS) Concept of Operations, December 2012. Accessed at: http://www.cts.virginia.edu/wp-content/uploads/2014/05/Task2.3._CONOPS_6_Final_Revised.pdf.) Additional messages currently are being proposed, including a SpeedProfile message, and an enhancement to the BSM that includes modal information.

The Michigan DOT Data Use Analysis and Processing (DUAP) system is designed for a collection of PDMs from DSRC-equipped vehicles, as well as data feeds from commercial CVs, agency fleet GPS devices, and other sources. (Michigan DOT, "VII Data Use Analysis and Processing: System Requirements Specification," December 2007, Accessed May 13, 2016, http://www.michigan.gov/documents/mdot/MDOT_DUAP_SysReq_Final_220099_7.pdf.) Applications of the stored trajectory data currently are in development.

Demonstrations by U.S. DOT's Connected Vehicle Safety Pilot program indicated the ability of BSMs to plot high-resolution trajectories of equipped vehicles on Geographic Information Systems (GIS) maps. Other existing test beds in California, Florida, New York, and Virginia have tested the data exchange of PDMs and BSMs. The upcoming Connected Vehicle Pilot Deployment sites in New York, Wyoming, and Florida will demonstrate additional safety and mobility applications.

Point of Access: Roadside Unit to Agency Physical Server

In the case of U.S. DOT Public Connected Vehicles, the point of access of the PDM/BSM data is directly from the vehicle itself to the roadside unit. This requires the DSRC equipment to be within line-of-sight and within range (generally 500m without obstructions).

Radio Frequency Identification, Wi-Fi, and Bluetooth Data

Radio Frequency Identification (RFID) technology is commonly used to transmit information over very short distances (~10m). These applications typically are employed by private or public-private operators for purposes, such as tolling, parking, weigh-in-motion checks for commercial trucks, fuel dispensing, and fleet management (check-in and check-out). As such, a vehicle's location and required PII is read at the point of contact between the device reader and the mobile unit (tag, sticker,

or device that holds static information). Sharing of the information with DOTs for TSM&O purposes becomes challenging, due to the embedded PII, which is necessary for payment and account transactions, as well as the lack of open standards (although there are some common standards employed by multiple public-private operators such as E-Zpass).

Wi-Fi and Bluetooth technologies can extend the transmission range to approximately 30m and are commonly used for vehicle (or device) reidentification. The Media Access Control (MAC) address of the Bluetooth device or Wi-Fi radio is detected at one location, and then reidentified at a second location. Travel time between the two beacons can then be computed. No PII is exchanged between the units, because most solutions hash the MAC address to further prevent tracking of a specific device through a network of beacons.

Point of Access: Variable

Private and public-private RFID readers are not commonly shared with DOTs due to the embedded PII. If they were to be shared in the future, it would likely be a Cloud API connection from the private or public-private operator to a DOT Cloud (a cloud system owned, operated, or leased by DOTs) or physical server. Bluetooth and Wi-Fi travel time data collection systems are typically aggregated by an onsite or Cloud processor, and then shared with DOT Cloud or physical servers via API. Some vendors provide access to individual travel times (determined by matching a specific MAC addresses at two separate readers), while others only provide summaries of travel times by time of day (TOD), or day of week (DOW), etc. for each pair of readers. Travel times captured using Bluetooth technologies are typically more often used by local agency DOTs compared to subscription probe-based services, such as INRIX and HERE. For example, the City of Austin, Texas, maintains approximately 50 Bluetooth readers for travel time data collection on regional arterials. The agency chose to purchase these devices as capital equipment versus leasing the equipment using a SaaS subscription model that is promoted by several vendors. Data transferred from the Bluetooth provider cloud system is stored in the agency's Advanced Traffic Management System (ATMS), but is typically transferred to long-term storage after 60 to 90 days as per most agencies' ATMS policies. After this period, the agency rarely accesses this data.

Connected Infrastructure

While the Internet of Things is a relatively new term in the world of big data, TSM&O organizations have been connecting to infrastructure to obtain information for more than 40 years. One of the core missions of TSM&O divisions is to collect device status and sensor data, and provide command and control actions to field devices from the Traffic Management Center (TMC). Connected infrastructure devices include traffic signals, ramp meters, CCTV, DMS, vehicle detection, Road Weather Information System (RWIS), flood warning, high wind warning, and a variety of other devices. Emerging application devices include Active Traffic Management applications of Variable Speed Limits (VSL) and Lane Control System (LCS). Integrated Corridor Management (ICM), interconnections of arterial control and freeway control systems, as well as transit and other demand management systems, are becoming more popular. Center-to-Center connections between regional systems are common for regions with multiple agency partners to share information and improve operations.

While the Internet of Things is a relatively new term in the world of Big Data, TSM&O organizations have been connecting to infrastructure to obtain information for more than 40 years.

There are emerging data sources in connected infrastructure, however. Several high-profile bridge failures have instigated the need to monitor bridge health in quasi-real time. Deploying and connecting emissions sensors is in the research and development stage. While there is uncertainty in how data from connected vehicles and travelers will be obtained and processed, the DOTs have solid experience in connecting information infrastructure, though data storage and management using Big Data technologies and techniques are still in need of evaluation. Most agencies today simply purge detailed data after a certain archival period (e.g., 30 or 60 days) using standard features of Relational Database Management System (RDBMS), such as Oracle and MS SQL Server. Large-scale storage of CCTV video and analysis of that video by emerging techniques, such as machine-learning or pattern matching, is a key consideration for future TSM&O practices. Safety and efficiency trends relative to TSM&O actions could potentially be found by automated analysis of CCTV images that would be cumbersome to attempt manually.

Other Sources

Mobile Sensors

The popular media and many white papers on Big Data are fond of identifying the "Google car" as a significant generator of data as its spinning Light Detection and Ranging (LiDAR) scanner collects 2 to 20GB/s of information from the area surrounding the vehicle. According to https://ark-invest.com/research/googles-driverless-car-massive-data-request, Google's intent to LiDAR map every road in the U.S. will require up to 70 Petabytes of storage. In addition to LiDAR, 3D cameras for StreetView images can generate up to 60 MB/s. Detailed radar, sonar, and GPS could add an additional 160 KB/s. There are many use cases for LiDAR point clouds, particularly in construction management and asset management. However, it is challenging to imagine the value of real-time or quasi-real-time sharing of LiDAR point cloud data with a DOT for TSM&O purposes as the vehicle moves through a space. Perhaps, virtual reality headsets might be used by future TMC operators to navigate near-real-time 3D spaces shared by suitably equipped automated vehicles. Certainly 3D camera views (even 2D camera views) could be beneficial, particularly in incident, event, and security management situations when there is no existing view from a fixed camera location with or without pan-tilt-zoom (PTZ). Many TSM&O organizations already rely on StreetView for assessment of field locations without having to physically travel to the location. Real-time video from onsite vehicles is certainly of value.

Points of Access: Mobile Sensor Data

There are no known uses of mobile 3D point clouds and/or mobile 3D video by TSM&O organizations. If such data was available in 5 to 10 years, the source would likely be via the Internet from the mobile sensor vehicle directly to a DOT cloud (a cloud system owned, operated, or leased by DOTs) or physical server. Such information would likely be only available when requested or pushed and not transmitted at all times. Perhaps, initially it would only be available from agency-owned incident management vehicles. Since 3D static maps currently are available on StreetView and other free sources, there would not likely be a revenue market for comprehensive 3D visual maps on a paid basis as a complete data set. There is some emerging market for 3D LiDAR maps, particularly for construction and asset management. It is challenging to envision use cases for maps for an entire agency's region for TSM&O purposes, even within the next 5 to 10 years.

High-Resolution Maps

A number of vendors are now offering high-resolution digital map products (HERE, MapBox). These digital maps are focused on lane-level accuracy geometry, accurate placement of all traffic control signs and advisories, allowable traffic controls at intersection junctions, and major street furniture. Currently, the maps are being marketed to support automated driving. DOTs, however, could be a natural consumer of such detailed data, in particular, since some V2I applications require precise description of the geography at intersections. It is likely that in the next 0 to 5 years most agencies will generate these GIS files in-house or through contract mechanisms. Over the next 3 to 10 years, however, it is potentially likely that procurement of such data by DOTs for use in outward-facing V2I applications would be more common than in-house generation due to the labor involved to generate such information and the lack of staff in TSM&O organizations to perform these tasks.

Points of Access: High-Resolution Maps

Current vendors offer access to high-resolution digital maps on their cloud-based servers for direct use from the Cloud as a subscription service. Business models may evolve to allow the native files to be downloaded, clipped, and converted to GIDs and so on at some time in the future. The challenge with local storage, like any mapping products maintained by DOTs today, is keeping the information up to date. Perhaps, initially, it would only be available from agency-owned incident management vehicles. Combining high-resolution connected traveler and connected vehicle data with high-resolution digital maps seems a reasonable consideration for TSM&O activities in the next 5 to 10 years, and thus the local storage or cloud-based access to these assets seems forthcoming.

Aggregated and Nonaggregated Transactional Data

TSM&O encapsulates some components of longer-term trends and changes to travel behavior. New information regarding supply chain and logistics management, purchasing behaviors, real estate marketing and valuation, and other economic transactions can affect TSM&O decisionmaking in subtle and perhaps not so subtle ways, particularly when it involves freight. An example of the value of these type of data is being modeled by Quetica, which serves Iowa DOT, among other clients. Iowa DOT utilizes Quetica to apply commodity-specific, county-level, cross modal global freight flow data to supply chain optimization analysis as a courtesy to companies in, or considering locating to, Iowa. In addition to these economic transaction-based sources, there are comparable data sources associated with utility transactions, such as transportation-relevant energy (electrical, natural gas); and telecommunications (fixed and mobile data utilization) patterns that may affect the manner in which agencies respond to changes in freight flows. At an aggregated level, these data may not initially be considered big enough to warrant new ways of storing and analyzing the information. Certainly if nonaggregated information was available over the next 3 to 10 years, procurement of this information would be challenging for TSM&O agencies to store and analyze with existing technologies.

Points of Access: Aggregated and Nonaggregated Transactional Data

Current vendors offer access to aggregated transactional data through traditional download from a cloud server. Nonaggregated data would like be encapsulated similarly and shared with DOT through a cloud-to-cloud API.

Chapter 4 Projected Nature of Emerging Data Sources

The purpose of this chapter is to characterize the nature, volume, and velocity of relevant emerging data sources. A quantitative approach is applied to estimate the scale of the big data challenges facing Transportation Systems Management and Operations (TSM&O) organizations. While some past reports have estimated data volumes, particularly for the United States Department of Transportation (U.S. DOT) Connected Vehicle (CV) program, few other sources have consolidated various emerging sources in one report. After reading this chapter, the reader will have an appreciation for the scale of the information that may be available to TSM&O practitioners in 5- and 10-year time horizons in each of the categories. The reader will understand that new "Big Data" technologies, and tools will be needed for extracting information and value from the sources. In subsequent chapters, we survey the marketplace of big data solutions and technologies to manage and extract value from emerging data sources. These three chapters provide the reader with more indepth understanding of available tools and technology solutions.

Chapter Objectives:

- Estimate the volume and velocity of data from each emerging data source category in 5- and 10-year time horizons.
- Estimate the total data storage needed to store and retain all collected information in 5- and 10-year time horizons.

Emerging Data Sources will come from Connected Travelers, Commercial Connected Vehicles, U.S. DOT (Public) Connected Vehicles, Connected Infrastructure, and Other sources. In this chapter, we assess the projected nature of these data sources for three timeframes:

1. 2016 (current year).
2. 2021 (5-year time horizon).
3. 2026 (10-year time horizon).

The projections of each of the categories include estimates of the data availability, delivery method(s), volume, velocity, and total storage requirements. This chapter is organized as follows:

- Assumptions that apply across most of the emerging sources.
- Estimates of the current population, which will generate content for each data source and any additional data elements that are available from each source.
- Estimates of projected growth rates of each source.
- Comparisons of the data volume and velocity of the various sources across the U.S. for a typical agency on a typical day.
- Anticipated cumulative data growth assuming all received information is archived and processed using big data tools and technologies for TSM&O applications.

Assumptions

For comparison purposes across the various emerging data sources, we focus on the common elements across all modes (including transit, walking, biking, and of course riding in their personal, shared-use, or for-hire vehicle): ***trajectory data of travelers' position and speed while traversing the transportation network***. This ***"digital exhaust"*** or breadcrumbs of the travelers' location over time while traveling is the common denominator of the first three types of emerging sources, and is one of the cornerstones of the U.S. DOT CV program. At the moment, we will disregard additional data elements, such as trip purpose, motivation, vehicle condition information, and so on (that might be available from one source, but not another) for the purpose of estimating the comparative volumes and velocity of incoming information from the emerging sources. For the fourth category of "other sources," such as 3D Light Detection and Ranging (LiDAR) point clouds, high resolution maps, etc. photo snapshots, video, etc., we list additional assumptions in that subsection. A consolidated list of assumptions is provided here. Narrative discussion of the basic assumption of cumulative data from one traveler per day follows.

List of All Assumptions

1. The "digital exhaust" data of travelers' trajectories in the system is the primary data element used to estimate data volume and storage requirements.
2. One data point in a trajectory is 1KB in size (speed, location, acceleration, and various other status elements).
3. A typical day is a weekday where most adult persons are typically working and traveling for the purposes of commuting to and from a workplace or school or performing their work duties by traveling.
4. Travelers spend 60 minutes in a typical day in travel.
5. Security credentials on the U.S. DOT CV system change once every 5 minutes.
6. Data points in a trajectory are collected once every 10 seconds of travel.
7. Second by second updates of digital exhaust are probably sufficient for most TSM&O mobility applications.
8. A probe data message trajectory holds 32 individual data points.
9. Cumulative data from one traveler per day in Probe Data Messages (PDM), commercial connected vehicles, or connected traveler apps is 500KB.
10. There are approximately 345,000,000 personal devices in the U.S. that could provide trajectory data.
11. Just because an app is installed in a smartphone does not mean the data will be transmitted unless the user activates the app.
12. In 2016, 1 percent of those personal devices (3,500,000) currently are being used, and the data is available to DOTs.
13. There are 1,250,000 miles of roads in the U.S.; meaningful for TSM&O.

14. An average road segment is 0.25 mile, resulting in a segmented map of the U.S. with 5,000,000 segments.
15. Status updates of each segment are provided by a third-party data provider once per minute.
16. The status update for one segment is 1KB.
17. Dedicated Short-Range Communication (DSRC) range is 1,500ft.
18. Vehicles interacting with a DSRC Roadside Unit (RSE) travel at 50ft/s.
19. In 2016, interactions of vehicles with Radio Frequency Identification (RFID) readers occur two to four times per day and negligibly contribute to today's or future year's predictions of data loading.
20. In 2016, there are 300,000 traffic signals in the United States.
21. A National Transportation Communications for ITS Protocol (NTCIP) poll message for signal and detector status is 2KB.
22. In 2016, 60 percent of traffic signals in the U.S. are polled once per second.
23. Other connected infrastructure devices are polled two times per minute with a 2KB status message per poll.
24. In 2016, there are 60,000 connected closed circuit television (CCTV) in the U.S. and 90,000 other devices.
25. CCTV streams are 350Kbps.
26. A high-definition (HD) map of a 0.25m segment is 5MB.
27. 3D video streams at 60 frames per second is 32Mbps at 1080p resolution.
28. A 3D video of a 0.25m segment has a video file size of 318MB.
29. A 3D point cloud of a 0.25m segment is the same size as a 3D video.
30. A 3D movie or 3D point cloud cannot in the foreseeable future be streamed wirelessly over existing or foreseeable technologies.
31. 3D image snapshots would be transmitted once per 10 seconds during an incident.
32. 3D images are 30MB each.
33. Incidents during which 3D images are broadcast from the site last 30 minutes.
34. Aggregated transactional data is not of appreciable volume or velocity to be considered further.
35. A typical agency has 1,000,000 travelers.
36. A typical agency has 1,000,000 registered vehicles.
37. A typical agency has a travel network comprising 50,000 segments.
38. A typical agency has 1,000 traffic signals, 300 CCTV, and 200 other connected devices.
39. In 2016, a typical agency manages 30 incidents in a typical day. The number of incidents in 2021 and 2026 does not change.

40. In 2016, 1 percent of travelers in the region of a typical agency use a connected traveler app.
41. In 2021, 20 percent of vehicles are DSRC connected vehicles, and 20 percent of all signals are DSRC signals.
42. In 2026, 50 percent of vehicles are DSRC connected vehicles, and 50 percent of all signals are DSRC signals.
43. In 2021, 15 percent of travelers use connected traveler apps, and 15 percent of vehicles return commercial connected vehicle data.
44. In 2026, 50 percent of travelers use connected traveler apps, and 50 percent of vehicles return commercial connected vehicle data.
45. In 2021, 70 percent of traffic signals are connected and in 2026, 80 percent of traffic signals are connected.
46. In 2021, there are 15,000 additional CCTV and 15,000 additional other devices connected to DOT networks.
47. In 2026, there are 50,000 additional CCTV and 50,000 additional other devices connected to DOT networks.
48. In 2021, 6 percent of incidents are covered by 3D video feeds.
49. In 2026, 16 percent of incidents are covered by 3D video feeds.
50. In 2021, all regions of the U.S. have a comprehensive HD map.
51. In 2026, all regions of the U.S. have a comprehensive 3D HD digital map.

Cumulative Data Load for One Traveler for One Day

Each Basic Safety Message is roughly 320 bytes in size. (Federal Communications Commission (FCC), Accessed May 13, 2016, https://apps.fcc.gov/.) The Vehicle Infrastructure Integration (VII) Wireless Access in Vehicular Environment (WAVE) Short Message (WSM) was estimated at 1.5KB per message. We split the difference here for this analysis and assume that the basic breadcrumb/snapshot is **1KB** in size. The Probe Data Message holds 32 snapshots and Security Credential Management System (SCMS) certificates are changed once per five minutes, so we surmise that one PDM will be collected for each 5-minute interval (assuming proximity to a Roadside Unit (RSU) during that time), or 9.375 seconds between breadcrumbs. We assume that most travelers spend ~60 minutes per day (or 3600s transmitting this information) traveling, whether they use taxi, bike, walk, transit, carpool or drive alone. This results in tracking approximately 385 breadcrumbs (one snapshot for each ~10s) per traveler per day. Rounding up for overhead and to simplify the calculations, we estimate the data load per traveler across the sources that provide general traveler trajectory information to be **500KB/traveler/day**. Regardless of source (commercial CV, connected traveler app, RSU/ On Board Unit (OBU), etc.), it seems reasonable for the purpose of volume estimation to assume that once-per-10-seconds tracking of individual user breadcrumbs is sufficient for many if not most TSM&O applications.

Discussions of additional assumptions and calculations of other data loads are provided in each section. Note if these assumptions or calculations are off by 25 to 50 percent in either direction (high or low), the conclusions reached regarding the need for advanced tools and technologies remain unchanged.

Current Year

In this section, we estimate the levels of data that currently are consumed/available from the four categories and subcategorizations of emerging data sources:

- Connected travelers.
- Commercial connected vehicles.
- Open connected vehicles (U.S. DOT/DSRC).
- RFID/Wi-Fi.
- Other sources.

These current year (2016) estimates will then be extrapolated to future years based on estimated growth rates of each technology, data source, and/or applications. The estimates provided here are intended to illustrate the relative scope and size of the emerging sources. Note also that all sources may not be needed if focusing especially on the digital exhaust of individual vehicles and travelers. Some agencies may opt to obtain data from one source, obviating the need to obtain it from another. For example, if an agency developed a regional connected traveler app, provided the app to citizens for free, and the app was adopted regionally by enough users; it might not be necessary to also purchase connected vehicles' digital exhaust data from third-party providers. Similarly, investments in applications to obtain data from connected vehicles may reduce the need for agencies to invest in other traditional technologies for TSM&O such as connected infrastructure. For this exercise to estimate the total data volumes, however, we neglect these cannibalization issues with respect to an agency selecting one or another to provide similar information.

Connected Travelers

As of 2015, Pew research indicates that 68 percent of all Americans have smartphones, and 45 percent own tablet computers. Upwards of 50,000,000 additional smart watches and athletic wearables may be available to provide traveler-related data as well. (Pew Research Center, "U.S. Technology Device Ownership 2015," Accessed May 13, 2016, http://www.pewInternet.org/2015/10/29/technology-device-ownership-2015.) However, since many of them do not actually include global positioning system (GPS) tracking capabilities, we will discount the number of wearables by 50 percent.

Using round numbers and a U.S. population estimate of 320,000,000, this equates to:

- 218,000,000 smartphones.
- 144,000,000 tablets.
- 25,000,000 wearables with GPS.

Tech-focused people may carry a smartphone, have a connected tablet in their backpack, and wear a smartwatch all at the same time. For the sake of simplicity, we estimate that approximately 25 percent of travelers are carrying multiple smart devices at any given moment, which reduces the number of tablets and wearables by 25 percent and results in **345,000,000 potential devices** that could provide traveler data. Currently, smartphone Operating System (OS) developers (Google, Apple, Microsoft, RIM, etc.) have not attempted and will probably not attempt to monetize users' travel data for the DOT market. So the potential number of individual devices for Connected Traveler information is limited to

the number of people who have opted to share their mobility data through existing commercial traveler information and mobility apps, future potential "smart city" apps, or "My511" public or public-private-partnership (PPP) apps. While Waze may have more than 90 percent market penetration in Israel, the adoption rate in the U.S. is much lower. We surmise that ~1 percent of existing smartphone, tablet, and wearable users have "opted in" for data sharing or are using an existing app that currently can provide data related to travelers to an agency DOT, or in these round numbers, **3,500,000 current connected travelers**. At 500KB per traveler per day, the total data load becomes approximately **1.75TB per day**. Penetration levels of new users of connected traveler applications will be used to estimate growth of data from this source for 5- and 10-year horizons.

Commercial Connected Vehicles

Commercial connected vehicles include 3G and 4G connections from the vehicle's infotainment system or embedded/aftermarket in-car systems installed by a third-party for vehicle tracking and data collection. There currently are roughly 250,000,000 registered vehicles in the United States. (U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology, Bureau of Transportation Statistics, "Number of U.S. Aircraft, Vehicles, Vessels, and Other Conveyances," Accessed May 13, 2016, http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_11.html.) Approximately 16,000,000 vehicles were sold in the U.S. in 2013. We estimate perhaps 5,000,000 vehicles today have 3G/4G connected features. But none of these sources currently are being marketed to DOT as data sources for TSM&O uses. The road mileage fee as an alternative to gasoline taxes offers the best potential option for data to be shared with DOTs on a large-scale basis, but requires significant legislative changes to enable this policy in every State. (Oregon DOT, "MyOReGO | A new way to fund roads for all Oregonians," Accessed May 13, 2016, http://www.myorego.org.) We are thus left with third-party vehicle tracking and data collection systems such as INRIX, HERE, TomTom, Garmin, etc. which have monetized the vehicle status information (primarily location and speed) for sale to DOTs on a subscription basis covering their area of influence (a State or region). Lacking a reliable estimate of existing commercial connected vehicles, we use the 5,000,000 estimate of currently enabled 3G/4G privately owned vehicles as the estimate of vehicles used by these aggregation services, combined and available across all of the services. These services do not provide all vehicle trajectories but rather **aggregate the information into lists of vehicle speeds per segment across a network of roads to the DOT**.

As of 2013, there are over 4.1 million miles of roads in the United States. (U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology, Bureau of Transportation Statistics, "Public Road and Street Mileage in the United States by Type of Surface(a) (Thousands of miles)," Accessed May 13, 2016, http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_04.html.) If we exclude local roads, there are over 1.25 million miles of freeway, arterial and collector roads. (U.S. Department of Transportation, Federal Highway Administration (FHWA), Policy and Governmental Affairs, Office of Highway Policy Information, "Highway Statistics 2013: User's Guide," Accessed June 27, 2016, http://www.fhwa.dot.gov/policyinformation/statistics/2013.) Based on a sample of segment data from a leading traffic data provider for a medium sized MPO, we assume an average segment length of ¼-mile. A segmented list of road sections for the entire U.S. road network would therefore total more than **5,000,000 segments**. If each segment is updated by these services to the DOT on a minute-by-minute basis, and each segment status update is roughly the size of a single Basic Safety Message BSM (~1KB), the total data load from aggregation services

(assuming a DOT buys only one service and not multiple that cover the same road segment), is thus:

1440 minutes / day * 5,000,000 segments * 1KB = **7.2TB per day**

These data rates will grow marginally as new road segments are added to the road network. Penetration rates of new services for commercial CV data will be estimated then for projections of data from this source for 5- and 10-year horizons.

U.S. Department of Transportation/Public Connected Vehicles

As of 2016, there are fewer than 10,000 vehicles equipped with U.S. DOT connected vehicle technology across the U.S. and less than 200 RSUs. With the Connected Vehicle Pilot Deployment projects scheduled for 2018, these numbers should roughly double. There are two components to the emerging data supplied by U.S. DOT connected vehicles: 1) the PDMs; and 2) the detailed BSM digital exhaust when in range of an RSU. The numbers of equipped vehicles that exchange information with existing RSUs today is not relevant in comparison to the other data sources. For the purpose of future estimation, we estimate again that a single vehicle will generate approximately 500KB of data in PDMs during the ~60 minutes of travel per day. The penetration rates of RSUs will be used to determine the actual (lower) predicted data load for the 2021 and 2026 time horizons. For burst transmissions of detailed BSM exhaust when in range of an RSU we estimate the range of the DSRC coverage at ~500m or 1,500ft and an average vehicle travel speed of 35mph or 50ft/s. The total range of the vehicle approaching and departing the view of the RSU, would be a distance of 3000ft, or 60s in range. At 10KB per second (assuming each BSM is 1KB), the data transmission for **one interaction** would be **600KB/interaction**. The penetration rates of RSUs will be used to determine the actual predicted data load for the 2021 and 2026 time horizons. In 2016, the data from public connected vehicles is ~zero for a typical agency.

Radio Frequency Identification Data

Private and public-private RFID readers are not commonly shared with DOTs due to the embedded Personally Identifiable Information (PII) issues. If they were to be shared in the future, it would likely be a Cloud Application Programming Interface (API) connection from the private or public-private operator to a DOT Cloud (a cloud system owned, operated, or leased by DOTs) or physical server. For the purpose of growth rate estimation, we surmise that RFID information in the form of Bluetooth and Wi-Fi travel time data collection systems are considered connected infrastructure and captured in that category of data. Other commercial CV applications based upon RFID information would be included in the data rates estimated for commercial CV sharing and thus are not captured here under a separate category. Furthermore, the number of interactions of a Connected Vehicle or Traveler with RFID readers is perhaps two to four times per day. At a similar per-transaction data transfer as a single BSM (~1KB, even at 10KB) the impact on the data loading analysis is negligible when compared with the other sources and is thus ignored.

Connected Infrastructure

As of 2016, there are more than 300,000 traffic signals in the United States. (Institute of Transportation Engineers, “National Traffic Signal Report Card, 2012,” Accessed at http://library.ite.org/pub/e265477a-2354-d714-5147-870dfac0e294.) Additional infrastructure devices include CCTV cameras, Dynamic Message Signs (DMS), ramp meters, fog warning, wind warning, detector stations, weigh-in-motion stations, lane control systems, dynamic speed limit signs, school

zone flashers, Bluetooth/Wi-Fi readers, and more. Nationally, these additional devices combined might total 100,000 additional devices. Traffic signal status is typically monitored on a second-by-second basis. The status message used in most Standard protocols (NTCIP, AB3418E) is around 2KB per poll. For 24-hour per day connectivity, the data load for an individual traffic signal is 24 x 60 x 60 x 2KB = **173MB/day**. Not all of the traffic signals in the U.S. are constantly connected as there are still many dial-up systems and unconnected devices. For estimation purposes we surmise based on our experience with State and Local agencies that perhaps 60 percent of the 300,000 current traffic signals currently are polled 24x7. Thus the national load in 2016 for traffic signal data is around **31TB/day**. Other devices such as system detection, DMS, weather stations, and so on are typically polled on a 20 to 30 second basis. With a similar 2KB poll message for each device, data load for an individual device would be estimated at 24 x 120 (60 minutes x 2 polls per minute) x 2KB = **5.8MB/day**. Thus the national load for 100,000 of these devices totals **0.58TB/day**. The total of the two categories (traffic signals and other devices) is roughly then **32TB/day**. Streaming video is a popular infrastructure technology for TSM&O and amasses significant storage requirements if stored. According to the American Association of State Highway and Transportation Officials (AASHTO) Footprint analysis report, there are over 10,500 CCTV for freeway monitoring in 2014. Arterial monitoring cameras are deployed by local and regional agencies typically between 1:10 and 1:1 ratios of traffic signals to cameras. As an average, we assume a 1:5 deployment of signals to cameras for an additional 60,000 CCTV deployed at intersections. A 350Kbps video stream is approximately **4GB/day** if recorded second-by-second, resulting in a national potential daily load of streaming video of 70,500 x 4GB/day = **282TB/day**.

Other Emerging Sources

As of 2016, there is essentially no sharing or storing of 3D LiDAR or mobile video feeds from connected or automated vehicles for TSM&O. HD maps also are not consumed by TSM&O agencies yet. These products would likely be consumed based on segmentation of a network similar to commercial connected vehicle data or traditional digital maps. A segmented list of road sections for the entire road network of the U.S. (excluding local roads) is estimated at **5,000,000 segments**. HD maps and 3D resources are not streaming services, and the content describing each segment of an HD map is estimated at **5MB (**this is the typical size of most detailed CAD drawings of quarter-mile roadway sections)**.** The entire database for the U.S. would encompass approximately **25TB**. A 3D point cloud or 3D stereoscopic feed for a 0.25km segment might require a total video file size of **318MB** (3D video is approximately 32Mbps for 1080p/60fps for one field of view; six cameras are typically needed at minimum for a 360-degree 3D movie. A 0.25km segment is traversed by a vehicle at 35mph, 50ft/s in 16.5s). A 3D point cloud would be estimated to be roughly the size of a 3D stereoscopic video at **318MB** per segment. Thus, a 3D map of the U.S. road network would total **1.6PB** in data storage requirements. These assets, like HD maps, would not be updated on a daily basis (and no agency would reasonably have use for or need the entire coverage area. Perhaps 3D feeds for specific segments or areas would be obtained from probes during incidents or events. Real-time delivery of streaming high-resolution 3D content might require more than 192Mbps of bandwidth, which is *way* outside of the estimates for upload support for 5G wireless (current upload speeds of 4G wireless are approximately 5Mbps). Thus a real-time feed might consist of a sequence of some 3D "snapshots" being delivered from the probes to a Traffic Management Center (TMC) instead of a continuous stream. We assume an average incident duration of 30 minutes which would result in a series of 180 3D images (one every 10 seconds). Each 3D image is approximately 30MB (6 1920x1080 24bit color images), for a total data feed per incident of **5.4GB/incident**.

Aggregated transactional data such as the cross modal freight flows stored and analyzed by Iowa DOT, for example, are on the order of 69MB per year for the Freight Analysis Framework file for 2013. (U.S. Department of Transportation, Federal Highway Administration, Freight Management and Operations, Office of Operations, "Freight Analysis Framework," Accessed June 27, 2016, http://ops.fhwa.dot.gov/freight/freight_analysis/faf/index.htm#faf4). Disaggregated information might increase this level of information by a factor of 100,000 to **6.9TB/year**. If all *disaggregated* flows were shared with a DOT on a daily basis, this would comprise **19GB/day.** Aggregated data is not of considerable size to be relevant in the conversation of additional resources and disaggregated transactions are unlikely to be shared with DOT/TSM&O agencies in the foreseeable future.

Aggregated transactional data such as the cross modal freight flows stored and analyzed by Iowa DOT, for example, are on the order of 69MB per year for the Freight Analysis Framework file for 2013. (U.S. Department of Transportation, Federal Highway Administration, Freight Management and Operations, Office of Operations, "Freight Analysis Framework," Accessed June 27, 2016, http://ops.fhwa.dot.gov/freight/freight_analysis/faf/index.htm#faf4.) Disaggregated information might increase this level of information by a factor of 100,000 to 6.9TB/year. If all disaggregated flows were shared with a DOT on a daily basis, this would comprise 19GB/day. Aggregated data is not of considerable size to be relevant in the conversation of additional resources and disaggregated transactions are unlikely to be shared with DOT/TSM&O agencies in the foreseeable future.

Complications of the Confluence of Connected Travelers with Commercial Connected Vehicles

The current "battle for the dashboard" between Original Equipment Manufacturers (OEM), major mobile operating system providers (Google, Apple, Microsoft, RIM, etc.) complicates the analysis of data growth rates of connected travelers and connected vehicles. Almost all current OEM vehicles have tethering capability for drivers to use their Bluetooth-enabled cellular phones hands-free in their vehicle. Google, Apple, and others are planning for more comprehensive integration with drivers' smartphone apps to be available directly on the dashboard via Apple Carplay and Android Auto products, to name a few. This significantly blurs the lines between a Commercial CV and a connected traveler. Perhaps in 10 to 15 years the delineation will be between Commercial CVs being truly commercial (i.e., taxis, delivery vehicles, and other business-owned vehicles) where the driver is an employee, and all other data will be considered as originating from the connected traveler applications whether they are driving in their own vehicle or a rented one. For the purpose of the data loading analysis in this report, we overlook this confluence issue since there is such uncertainty in predicting future events.

Typical Agency Data Loads

One agency does not manage the entire U.S. for TSM&O purposes, so it is less relevant to evaluate the data loading for the entire U.S. than for a single system which would have to ingest, process, and analyze all of this information at once. Agencies come in a variety of organization and jurisdictional control structures. There are perhaps six canonical types of agency organizations that are responsible for existing data ingestion and will be responsible for ingestion of future connected traveler and connected vehicle data:

- State DOT, focused on freeways only (e.g., Tennessee DOT).

- State DOT, freeway/arterial responsibilities (e.g., Virginia Department of Transportation (VDOT), Caltrans, Utah Department of Transportation (UDOT)).
- Combined State DOT and City/County (e.g., Austin Combined Transportation, Emergency, and Communications Center (CTECC)).
- Multi-State coalition (e.g., I 95 corridor).
- Local city/county or other municipal (isolated or "rural," e.g., Lubbock, Texas).
- Local city/county or other municipal (urban/suburban, e.g., Seattle, Miami/Dade County).

For the purpose of projecting data loads for a "typical" agency in the current year, we assume the following:

Table 1. Data loading assumptions for a "typical" agency (2016).

Agency Characteristic	Projected Data Load
Regional population under jurisdiction	1,000,000 adult travelers
Regional population of vehicles	1,000,000 registered vehicles
Regional population of connected infrastructure	1,000 traffic signals, 200 closed-circuit television, 300 additional devices
Regional road network	50,000 segments
Connected vehicles	~zero
Roadside units	~zero
Regional incidents per day	30

Projected Growth Rates

The American Association of State Highway and Transportation Officials (AASHTO) footprint analysis provides the basis for growth rate predictions for public connected vehicles as shown in figure 1.

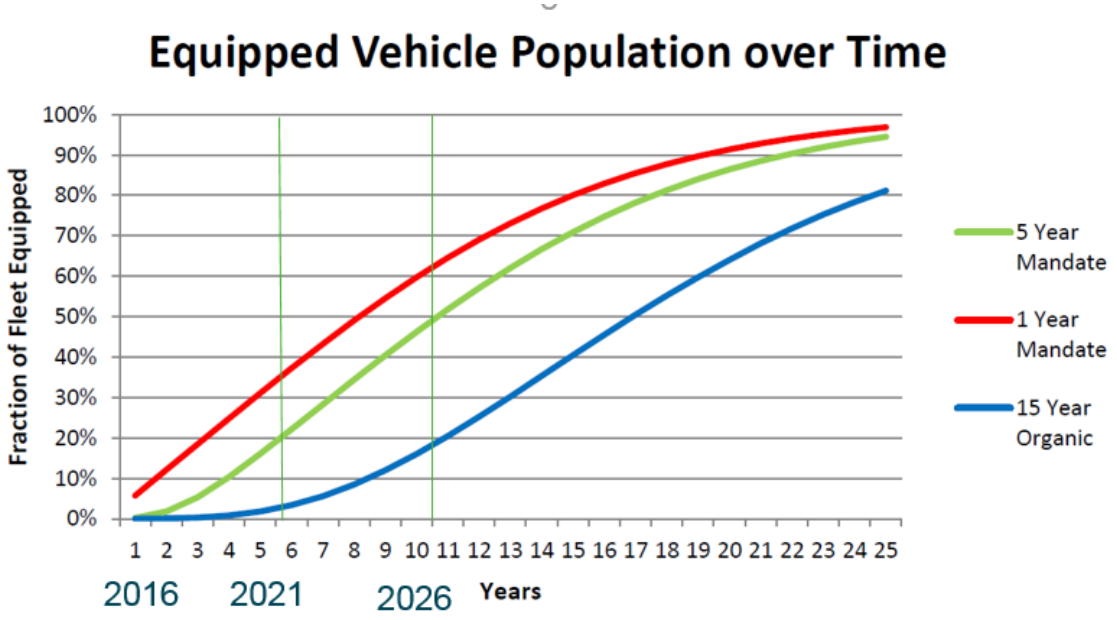

Figure 1. Graph. Predicted growth rates of public connected vehicles.
(Source: American Association of State Highway and Transportation Officials, 2014.)

As shown in the figure, the AASHTO analysis provides three potential growth rates assuming a "strong" mandate of immediate adoption of the DSRC technology (1-year mandate), a "medium" mandate of adoption of the technology over a five-year timeframe, and "weak" adoption of the technology with only a recommendation by National Highway Traffic Safety Administration (NHTSA) but no mandate ("15 year organic"). For the purpose of our data loading estimation, we use the "medium" growth trajectory which results in 20 percent penetration of the technology in 2021 and 50 percent penetration of the technology by 2026. For vehicle-to-infrastructure (V2I) applications and the resulting potential value to TSM&O, the infrastructure deployment must keep pace with the fleet penetration, so we use the same numbers (20 percent and 50 percent) for the penetration levels of RSUs for the Nation and for our typical agency. Using the assumption of 250,000,000 vehicles in the U.S., this leads to 50,000,000 connected vehicles in 2021 and 125,000,000 connected vehicles in 2026 and thus per-day data rates nationally of **25TB/day** for the probe messages (500KB) in 2021 and **62.5TB/day** in 2026. For the RSU-vehicle interaction data, we assume a 20 percent penetration rate of RSUs in 2021, in which a vehicle passes an RSU on 20 percent of their travel or 12 minutes of their 60 minutes of travel. Using the average in-range estimate of one minute, this results in 12 RSU interactions for a total of 600KB*12 = 7.2MB/day/vehicle or a national total of **360TB/day** in 2021 and **900TB/day** in 2026 for RSU-vehicle interactions. Adding together the daily probe data PDMs and the RSU interactions, the total load is **385TB/day** in 2021 and **963TB/day** in 2026. Notably the burst transmissions from CVs to RSUs will be shown to vastly outpace the other sources. This implication is important for the design and implementation of an appropriate PDM, and compression and data aggregation approaches for the RSU-OBU interactions. These topics will be addressed in a subsequent report.

Growth Rates of Connected Travelers

Figure 2 indicates projected growth rates of use and availability of connected traveler applications with approximately 15 percent penetration in 2021 and 50 percent penetration by 2026. Strong adoption (i.e., "killer apps" that provide significant value to the user and return on investment (ROI) from developers and data providers) and removal of PII barriers could raise adoption to perhaps 70 percent by 2026. Continual issues with PII and lack of interest from data providers (low ROI) could see significantly lower (20 percent) availability of connected traveler data in 10 years. Obviously these are approximations, but for the purposes of sizing potential Big Data tools and technologies for a typical TSM&O agency, these values are reasonable assumptions for estimating the potential data loads.

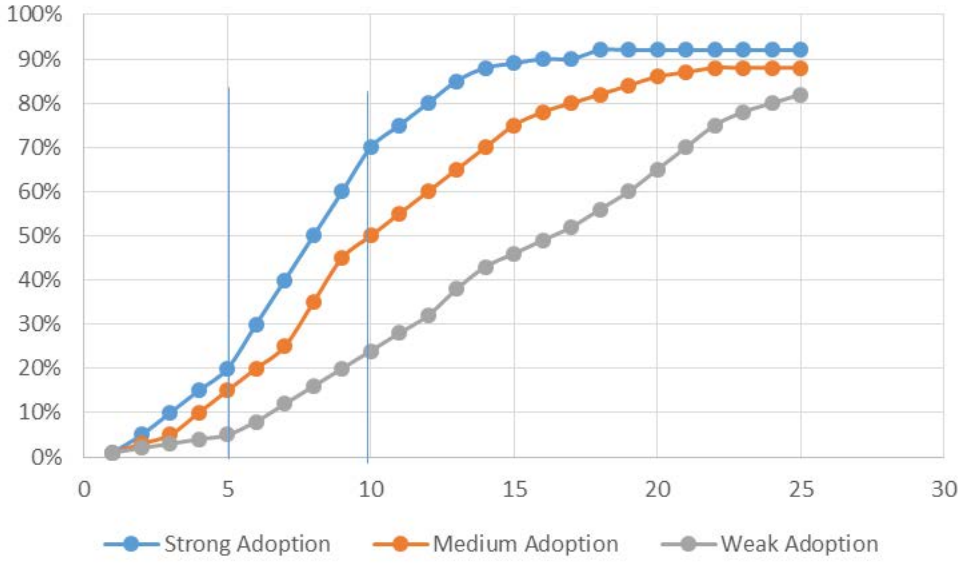

Figure 2. Graph. Connected traveler population growth rate over time.
(Source: Kimley-Horn and Associates, Inc., 2016)

Using the assumption of 345,000,000 connected travelers and the 500KB/day estimate of the data load per traveler at 15 percent penetration in 2021 and 50 percent penetration in 2026 results in per day national data loading of **25.9TB/day** and **86.25TB/day**, respectively. This is roughly 10 percent of the loads from CV-RSU interactions.

Growth Rates of Connected Commercial Vehicle Services

Figure 3 indicates projected growth rates of use and availability of connected commercial vehicle data and applications with approximately 15 percent penetration in 2021 and 50 percent penetration by 2026. It is likely that third-party providers will partner with OEMs to provide new data products (i.e., anonymized high-resolution trajectories, origin-destination tables, etc.) assuming the PII issues related to vehicle usage can be suitably addressed. Similarly to connected travelers, strong adoption and removal of PII barriers could raise adoption rates to perhaps 70 percent by 2026. Continual issues with PII and lack of interest from data providers (low-ROI willingness of DOTs to pay for data) could see significantly lower (20 percent) availability of connected commercial vehicles data in 10 years. Obviously, these are approximations, but for the purposes of sizing potential Big Data tools and

technologies for a typical TSM&O agency, these values are reasonable assumptions for estimating the potential future data loads.

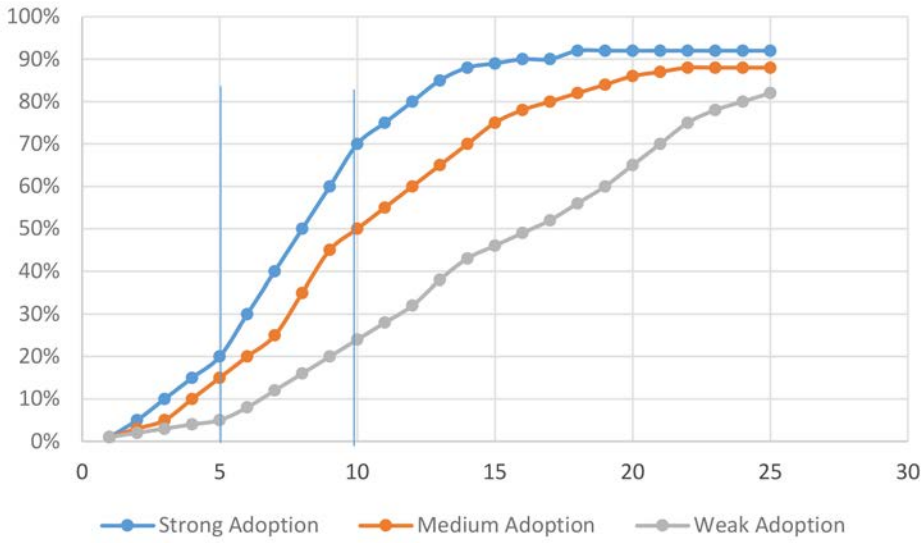

Figure 3. Graph. Connected commercial vehicle population growth rate over time.
(Source: Kimley-Horn and Associates, Inc., 2016)

Using the assumption of 250,000,000 vehicles this leads to 37,500,000 commercial connected vehicles in 2021 and 125,000,000 commercial connected vehicles in 2026. At 500KB per CV per day, the resulting data loads are **18.75TB/day** in 2021 and **62.5TB/day** in 2026.

Growth Rates of Connected Infrastructure

Figure 4 indicates projected growth rates of use and availability of connected infrastructure data with approximately 70 percent penetration in 2021 and 80 percent penetration by 2026. This is dominated by ongoing efforts by agencies to connect their traffic signals with high-speed, agency-owned or leased IP networks. Reduction in costs, availability of funds, and renewed emphasis on connectivity, particularly related to the benefits potential for V2I applications, could raise penetration to perhaps 90 percent by 2026. Obviously, these are approximations, but for the purposes of sizing potential big data tools and technologies for a typical TSM&O agency, these values are reasonable assumptions for estimating the potential future data loads.

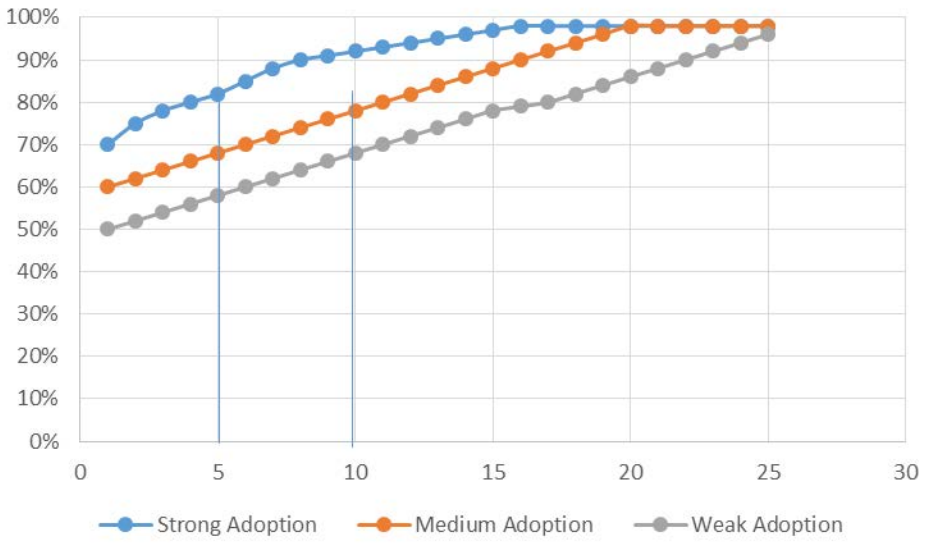

Figure 4. Graph. Growth in connected infrastructure (existing) population over time.
(Source: Kimley-Horn and Associates, Inc., 2016)

In addition to connecting existing devices, agencies are continually purchasing new devices and new sensor systems such as emissions sensors, bridge monitoring sensors, radiation detectors, and the like. We estimate based on "medium" levels of adoption of new devices that 15,000 new devices would be deployed in 5 years and 50,000 new devices in 10 years nationwide. For CCTV, we assume the same growth with 15,000 new CCTV in 2021 and 50,000 in 2026.

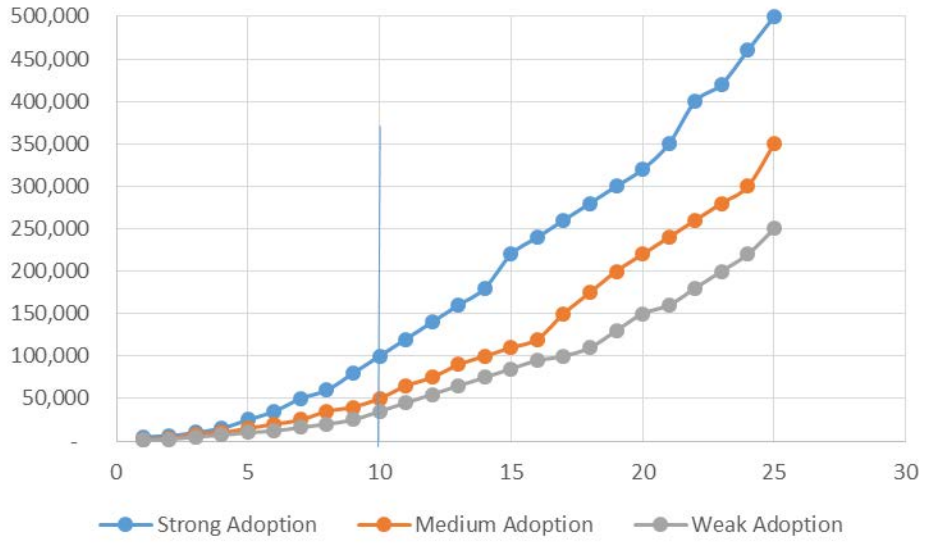

Figure 5. Graph. Growth in connected infrastructure (new devices) population over time.
(Source: Kimley-Horn and Associates, Inc., 2016)

Using these assumptions, 70 percent of the 300,000 existing traffic signals at 173MB/day and an additional 115,000 other devices (100,000 connected devices today plus an additional 15,000 devices) at 5.8MB per day, and 85,500 CCTV at 4GB per day results in a daily national load of **379TB per day** in 2021 (dominated 90 percent by CCTV). In 2026, with 80 percent of the 300,000 signals and 150,000 additional devices connected, and 120,500 CCTV the daily national load is thus **524TB per day (92 percent video).**

Growth Rates of Other Sources

Figure 6 indicates projected growth rates of use of 3D video devices for incident response management as a percentage of all incidents that occur. This is primarily driven by agency investment in the technology and the payoff of such onsite telepresence on management effectiveness. With such uncertainty regarding the technology and the usefulness of the application to TSM&O agencies, we forecast weak adoption of the technology in 2021 at 6 percent coverage of incident occurrence and 16 percent coverage of incident occurrence by 2026. For other sources such as HD maps, we estimate that by 2021 all regions of the U.S. have available a comprehensive HD map. By 2026, all regions of the U.S. are assumed to have available a comprehensive 3D HD digital map.

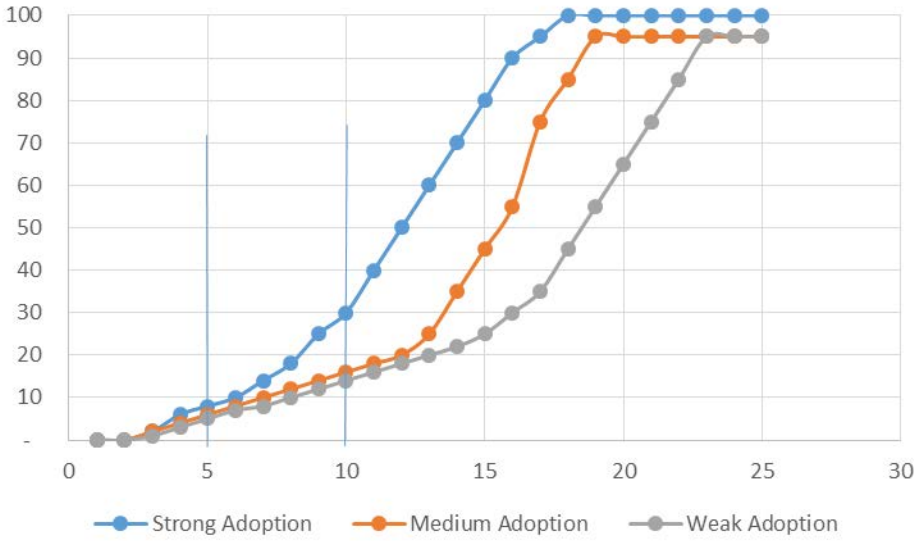

Figure 6. Graph. Growth in percentage of incident coverage by 3D mobile cameras.
(Source: Kimley-Horn and Associates, Inc., 2016)

Summary of Growth Rates of Emerging Data Sources

Table 2 summarizes the daily data storage loading expected for the entire United States during year 1 (2016), cumulative through year 5 (2021), and cumulative through year 10 (2026).

Table 2. Summary of daily data volume for the entire United States.

Source	Data Volume per day per device	Today (2016)	5 Years (2021)	10 Years (2026)
Connected Travelers via 3rd party	~500KB	1.75TB	26TB	86TB
Connected vehicles (commercial)	~1KB per segment	7.2TB	7.5TB	8TB
Connected vehicles (commercial—future)	~500KB (nonaggregated)	~zero	18.75TB	62.5TB
Connected vehicles (public/dedicated short-range communications)	~600KB per interaction + 500KB/day (probe)	~zero	385TB	900TB
Connected infrastructure	~173MB (signal) and ~5.8MB (other device) ~4GB (video)	314TB	379TB	524TB
Other sources	~5.6GB/incident	~zero	1.68TB/day	28GB/day
Total		323TB	820TB	1609TB

Calculation methods for the total daily data loading of each data source category for 2021 and 2026 are discussed in the previous sections for each type of emerging data source.

Typical Agency Data Growth Rates

As indicated earlier, one agency does not (and will not) manage TSM&O activities for the entire U.S., so it is less relevant to evaluate the data loading for the entire United States as if one system would have to ingest, process, and analyze all of this information at once.

For the purpose of projecting data loads for a "typical" agency we assume the following:

Table 3. Data loading assumptions for a "typical" agency.

Agency Type	Projected Data Load
Regional population under jurisdiction	1,000,000 adult travelers
Regional population of vehicles	1,000,000 registered vehicles
Regional population of connected infrastructure	1,000 traffic signals, 300 additional devices, 200 closed-circuit television
Regional road network	100,000 segments
Connected vehicles	~zero
Roadside units	~zero

With these round number assumptions the data loads of "today" (2016) for a typical agency are estimated to be approximately:

Table 4. Data loading estimations for a "typical" agency.

Data Type	Density	Data Load
Connected travelers	1% of 1,000,000 (10,000)	5GB/day (0.005TB)
Commercial vehicles	1% of 5,000,000 segments	72GB/day (0.072TB)
Commercial vehicles (future)	0% of 1,000,000 vehicles	~zero
Connected vehicles	~zero	~zero
Connected infrastructure	1000 signals (60%), 300 other, 200 closed-circuit television	975GB/day (0.975TB)
Other sources	~zero	~zero

Using the projected penetration rates for each of the Emerging Data Sources listed previously and using the same calculations detailed in the national data loading analysis in the previous section, Table 5 summarizes the data loading per day for a typical agency. In 2021, the data loads from connected traveler and commercial connected vehicles emerging sources are on par with the current data loading from connected infrastructure. The data loading from public connected vehicles would be roughly triple the other sources combined in 2021 and treble in 2026.

Table 5. Summary of daily data storage loading for a typical agency.

Source	Data Volume per day per device	Today (2016)	5 Years (2021)	10 Years (2026)
Connected travelers via 3rd party/opt-in app	~500KB (Waze)	0.005TB	0.075TB	0.25TB
Connected vehicles (commercial)	~1KB per segment	0.072TB	0.075TB	0.1TB
Connected vehicles (commercial—future)	~500KB (nonaggregated)	~Zero	0.075TB	0.25TB
Connected vehicles (public/dedicated short-range communications)	~650KB per interaction + 500KB (probe)	~Zero	1.53TB	3.6TB

Table 5. Summary of daily data storage loading for a typical agency (continuation).

Source	Data Volume per day per device	Today (2016)	5 Years (2021)	10 Years (2026)
Connected infrastructure	~173MB (signal) and ~5.8MB (other device) ~4GB (closed-circuit television)	0.975TB	1.083TB	1.51TB
Other sources	3D Video	~zero	11.2GB/day	28GB/day
Total		1.05TB	2.84TB	5.61TB

Calculation methods for the total daily data loading of each data source category for 2021 and 2026 are discussed in the previous sections for each type of emerging data source.

Data Velocity

In this section we discuss the data velocity and cumulative storage needs for a typical agency. Data Velocity is the rate at which data is generated and the rate at which the data needs to be processed. There are primarily two categories of data processing, batch, and streaming. Batch processing is for analysis done after-the-fact. Data that does not require immediate action can be analyzed independently from the real-time performance of the system. Streaming processing enables real-time decisionmaking and alerts. Streaming and batch analyses each have their pros and cons, and the appropriate method depends largely on the organization's particular use case and business need. Transportation management centers have needs for both streaming and batch data processing. Use cases for both will be explored further in a subsequent report.

Connected Travelers

Each of the Emerging Data Sources are not transmitted to the DOT at the same rates. This section summarizes the data rates for each source or how often each dataset is refreshed. Connected traveler information from an existing source such as Waze is updated in the data feed each minute, providing new status of anonymized app users. We project that the same delivery mechanism would likely be used in 5- and 10-year time horizons. As the volume of connected travelers increases, this velocity may become an issue for an individual agency and need big data techniques to adequately capture the volume. However, there are specific traveler to infrastructure applications where second-by-second data from connected travelers is warranted and useful for TSM&O. Pedestrian and bicycle detection in particular is a challenge for many agencies across the U.S. and second-by-second delivery of user location is critical for effective enabling of traffic signal operations. Some pilot projects already have started to undertake this obstacle, such as in Austin, Texas, but with very low numbers of cyclists or pedestrians. Scalability of such solutions are likely to require big data treatments to work on a citywide scale with hundreds of thousands of road users.

Commercial Connected Vehicles

Similarly, commercial aggregated connected vehicle data based on map-segments is typically updated each minute. We anticipate that the existing update methodology will continue in five and 10 years. This data is not accelerating at the same rate that connected vehicles and connected traveler volumes will increase because it is solely dependent on the DOT adding new roads or the third-party increasing the density of their map-segmentation. Neither of these are rather likely to have significant implications for data velocity. Additional types of products such as origin-destination tables and the like would increase the volume, but not the velocity. Future commercial connected vehicles data might be delivered in a similar manner to DOT with snippets of trajectories delivered each approximately every minute.

U.S. Department of Transportation/Public Connected Vehicles

The velocity of U.S. DOT/DSRC connected vehicle data is very different. The vehicle transfers 10Hz status updates and probe data messages when in range of an RSU. We denote this as “burst transfer.”.” It is unlikely that the velocity would increase even faster than this and it is probably more likely that the velocity will decrease as the deployed density of connected vehicles climbs higher and higher. Safety applications, such as red-light-running warnings require 10Hz updates, but second-by-second updates are probably sufficient for TSM&O mobility applications. Some data processing will likely take place at the RSU, further reducing the velocity into the TMC.

Connected Infrastructure

Status data from connected infrastructure is today either delivered on a second-by-second basis or on a 20- to 30-second update interval. In the future, we predict more use of “report on change” (sometimes referred to as Simple Network Management Protocol (SNMP) “traps”) methods which reduce bandwidth requirements. Although traffic signal status does vary on a 10Hz basis, there are many periods of the day (i.e., the middle of the night) when the status remains largely the same and there is no need to tell the TMC again that the light is still green and no one is driving by.

Table 6 summarizes the data velocity expected for each emerging data source for a typical agency during year 1 (2016), cumulative through year 5 (2021), and cumulative through year 10 (2026).

Table 6. Summary of data velocity for a typical agency.

Source	Data Volume per day per device	Today (2016)	5 Years (2021)	10 Years (2026)
Connected Travelers via 3rd party/opt-in app	~500KB (Waze)	Minute-by-minute	Minute-by-minute	Minute-by-minute
Connected vehicles (commercial)	~1KB per segment	Minute-by-minute	Minute-by-minute	Minute-by-minute
Connected vehicles (commercial—future)	~500KB (nonaggregated)	~none	Minute-by-minute?	Minute-by-minute?

Table 6. Summary of data velocity for a typical agency (continuation).

Source	Data Volume per day per device	Today (2016)	5 Years (2021)	10 Years (2026)
Connected vehicles (public/dedicated short-range communications)	~600KB per interaction + 500KB probe	Burst transfer	Burst transfer	Burst transfer
Connected infrastructure	~173MB (signal) and ~5.8MB (other device), ~4GB (closed-circuit television)	Second by second, 30-second report	Second by second, report on change	Second by second, report on change
Other sources	~5.4GB (3D Video)	N/A	~10 second snapshots	~10 second snapshots

Data Storage

Data storage for Emerging Data Sources becomes overwhelming if the agency considers storing all data ever collected. This currently is being done now by systems such as Regional Integrated Transportation Information System (RITIS) and Caltrans Performance Measurement System (PeMS) with connected infrastructure and some probe and commercial connected vehicle feeds. The data requirements for existing systems are incomparable to the data stores required for the five- and 10-year time horizons when including the emerging sources; particularly if an agency sees the need to store all of the public connected vehicle Basic Safety Messages (BSM), which is unlikely. Potential approaches to reducing the data footprint of the BSMs at the TMC through edge processing and data aggregation will be discussed in more detail in a subsequent report. We assume for this analysis that the penetration rates climb incrementally each year and remain steady for the entire year.

Table 7 summarizes the total data storage required for each emerging data source for a typical agency during year 1 (2016), cumulative through year 5 (2021), and cumulative through year 10 (2026).

Table 7. Summary of data storage for a typical agency.

Source	Data Volume per day per device	Today (2016)	5 Years (2021)	10 Years (2026)
Connected Travelers via 3rd party/opt-in app	~500KB	1.8TB	73TB	370TB
Connected vehicles (commercial)	~1KB per segment	5TB	159TB	294TB
Connected vehicles (commercial—future)	~500KB (nonaggregated)	~Zero	68TB	365TB
Connected vehicles (public/dedicated short-range communications)	~600KB per interaction + 500KB (probe)	~Zero	1594TB	6546TB
Connected infrastructure	~173MB (signal) and ~5.8MB (other device), ~4GB (closed-circuit television)	38TB	1878TB	4244TB
Other sources	~5.4GB (3D Video)	1TB	3.6TB	4.6TB
Total		45.8TB	3776TB	11823TB

Assuming a monthly storage cost of $0.03/GB, which is what Amazon currently charges for its enterprise data storage (there are several factors in how costs are calculated for various tiers of storage and capacities, but this is a planning level estimate), this equates to a cost of approximately **$1,500/mo** for data storage to store all data in 2016 in the Cloud. This grows to a staggering **$110,000/mo to store more than 3,700TB in 2021**, which makes it clear that the raw BSMs and raw video simply cannot be stored "forever." The value of individual BSMs for TSM&O is marginal, at best, so aggregation and storage of derived performance measures or summary statistics **must** be applied. Similarly for raw video, analytics will be needed to extract out performance data, or store only anomalies or special events. Reducing the BSM and video storage by a factor of 100 by aggregation, still results in a cost-prohibitive **$40,000/mo for storage of all ~1,300TB from 2016 to 2021 in the Cloud**. Considering the cost per gigabyte of hard drive storage has dropped by a factor of 46 in the past 10 years, the growth in the amount of data (2026 projections are 70 times 2016 estimates) will be partially offset by reduced costs for storage, though we may not see the same decreases with the growth in higher performance (and higher cost) solid state storage. Multiple strategies will need to be applied to determine what data to keep and for how long, similar to the ways in which TSM&O agencies manage connected infrastructure information in traditional databases today.

Density of Emerging Data Sources in a Typical City

These growth rates can be summarized visually as shown in the following figures. Figure 7 shows a typical city dominated by connected infrastructure, with a few connected travelers and connected vehicles. In 2021 (figure 8), connected vehicles and Travelers' penetration rates begin to overwhelm the connected infrastructure and in 2026 (figure 9), the density of connected infrastructure is close to complete and connected travelers and Vehicles are widespread throughout the jurisdiction. In each of the figures, one icon indicates multiple devices or travelers. RSUs and Infrastructure are represented by a scale of 10, connected vehicles by a scale of 100 and Travelers by a scale of 1000. Rough approximations of the growth rates for each are shown in each figure for 2021.

Figure 7. Illustration. A typical agency's connectedness in 2016.
(Source: Kimley-Horn and Associates, Inc., 2016.)

Figure 8. Illustration. A typical agency's connectedness in 2021.
(Source: Kimley-Horn and Associates, Inc., 2016.)

Figure 9. Illustration. A typical agency's connectedness in 2026.
(Source: Kimley-Horn and Associates, Inc., 2016.)

This chapter laid out a set of estimates of the volume and velocity of data that may be possible as connected vehicles, travelers, and other sources emerge. Even if the assumptions are 25 percent too high or too low, the conclusions remain unchanged. **Data management and data analysis will require Big Data tools and methods if a TSM&O agency seeks to realize its value**. Significant strategies will be needed to address what to store and for how long in order to maximize cost effectiveness for already budget-constrained TSM&O agencies. The next three sections transition to a discussion of currently available big data technologies and tools. Chapter 5 discusses trends in the use of big data in government and industry. Chapters 6 and 7 identify and compare vendor solutions that currently are available in the marketplace.

Chapter 5 Industry and Government Trends in Big Data

What is Big Data?

"In pioneer days they used oxen for heavy pulling, and when one ox couldn't budge a log, they didn't try to grow a larger ox. We shouldn't be trying for bigger computers, but for more systems of computers."—Grace Hopper. (Schieber, Philip, "The Wit and Wisdom of Grace Hopper," Accessed May 9, 2016, http://www.cs.yale.edu/homes/tap/Files/hopper-wit.html.)

The purpose of this chapter is to raise awareness of the reader in common terms and technologies in the Big Data ecosystem. The reader will gain a better appreciation for some of the component tools and technologies and understand, at an entry level, of what pieces do what. Given the predictions of the volume and velocity of emerging data sources for Transportation Systems Management and Operations (TSM&O) from the previous chapter, the reader will understand that current information technology (IT) systems are not capable to handle such volumes. After reading this chapter, the subsequent chapters introduce the readers to common industry systems that implement tools and technologies for handling massive data sets. The terms and concepts introduced in this chapter will be used in subsequent chapters.

Chapter Objectives:

- Provide definitions and descriptions of the moving parts of Big Data.
- Introduce Gartner's Hype Cycle for categorization of technology readiness.
- Categorize tools and technologies.
- Identify trends in each of the process model steps.
- Introduce the key elements of Internet of things and explain the potential relevance to TSM&O.

"Big Data" is the buzzword given to the ongoing phenomenon of data production and consumption on a massive scale, and it generally means very different things to different organizations. A Big Data problem typically occurs when an organization generates and/or consumes more data than their IT infrastructure can handle. However, at what point that organization reaches the stage of Big Data depends on the circumstances and capabilities of the organization. For example, a regional grocery chain might feel overwhelmed at receiving 100GB per day, where **Facebook with over a billion users handles approximately 500TB per day**. (Gigaom, "Facebook is collecting your data—500 terabytes a day," Accessed April 4, 2016," https://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/.) Alternatively, an operations center overseeing a large rural district can anticipate a volume of data from emerging data sources that may to them seem significant (less than 1TB per day), but would pale in the face of the data volume expected from the same data sources at a large urban traffic management center (greater than 5TB per day).

Data production and consumption have been rapidly growing for far longer than the term Big Data has existed, and past solutions have focused on building bigger, better, and faster machines. (Domo, "Data Never Sleeps 3.0," Accessed April 4, 2016, https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/.) However, with the understanding that data is only going to continue growing at an exponential rate (as demonstrated in the forecasts in chapter 4 of this report), it has come to be accepted that Big Data is an enduring problem that requires a new paradigm shift. Similarly to Intel's shift from its focus on increasing the clock speed of its microprocessors to multicore architectures, the big data paradigm shift is about using clusters of networked computers to distribute tasks in parallel, which is far more efficient and scalable than relying on bigger and faster hardware.

Google is a noteworthy benchmark for embracing this shift. In 2003, Google released a report titled "The Google File System" describing their solution to the big data problem applied to the problem of searching the vastness of the Internet for meaningful and relevant information, something we take for granted today. (Ghemawat, Sanjay, Gobioff, Howard, and Leung, Shun-Tak, "The Google File System," Accessed April 4, 2016, http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf.) Because their search indexes were so large and because search results from the index needed to be nearly instantaneous, they devised the Google File System (GFS) is a highly reliable, distributed file system that runs on a cluster of inexpensive, commodity hardware. The cluster can be easily scaled up by adding individual servers to quickly accommodate rapid changes in data storage and processing needs. In addition, it distributed keyword queries to more rapidly return results from a rapidly growing user base. The GFS laid the foundation for the design of the **Hadoop Distributed File System (HDFS)—**an open source software framework managed by the Apache Software Foundation (Apache) to store and process massive amounts of data quickly and efficiently across multiple nodes. (Apache, "What is Apache Hadoop?" Accessed April 4, 2016, http://hadoop.apache.org/.)

> The Big Data paradigm shift is about using clusters of networked computers to distribute tasks in parallel, which is far more efficient and scalable than relying on bigger and faster hardware.

Google released a second report titled "MapReduce: Simplified Data Processing on Large Clusters" in 2004 that described Google's cost effective approach to quickly process large volumes of data. (Dean, Jeffrey and Ghemawat, Sanjay, "MapReduce: Simplified Data Processing on Large Clusters," Accessed April 4, 2016, http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf.) **MapReduce is a batch data processing technique that breaks large tasks into smaller ones to be concurrently executed in a distributed fashion across the cluster**. The results of each small task are aggregated once processing is complete. Google recognized that crawling the Internet and indexing keywords lent itself to parallel processing and by devising a framework that did not rely on expensive mainframes or super computers, they could scale much more efficiently and cost effectively.

Together, HDFS and the MapReduce computing framework have become the foundation of the Hadoop ecosystem. While other solutions, which will be discussed in this report, gained popularity for a time, Hadoop expanded to become an entire ecosystem and has adopted nearly every other big data concept within its open source suite of tools. The leading big data companies are building their offerings upon this open source ecosystem and offering enterprise services for initial set-up, configuration, and ongoing

> While the Big Data landscape is rapidly evolving, there is as yet no successor on the horizon for Hadoop.

support, similar to companies that sell distributions of the open source Linux operating system. While the big data landscape is rapidly evolving, there is as yet no successor on the horizon for Hadoop, though new projects are continuously adding new functionality and improving the usability of the Hadoop ecosystem.

Characteristics of Big Data

The five V's are a good place to start when trying to describe big data. The first four V's (volume, velocity, variety, and veracity) are attributes of the data itself, each requiring additional considerations to supplement and modernize traditional systems. While the fifth V (value) is the business benefit that can be created using Big Data.

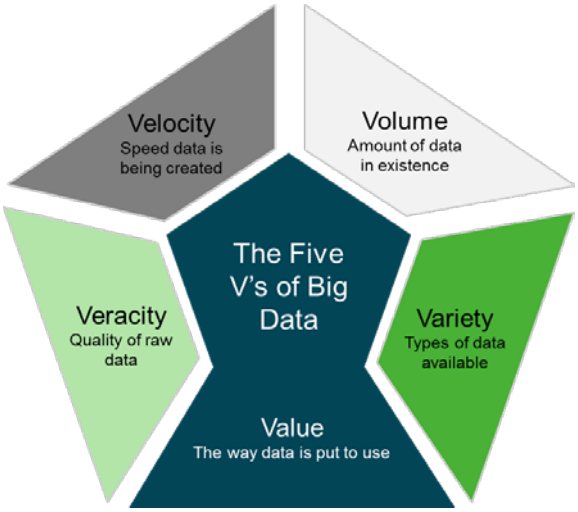

Figure 10. Illustration. The five V's of big data.
(Source: Deloitte, 2016.)

Descriptions of the first four V's are as follows:

- Volume: is the total amount of data in existence. Data is constantly being generated at faster speeds and organizations are interested in collecting more of it for analysis. Infrastructure scalability becomes paramount as data volumes increase exponentially and storage priorities must be adapted.

- Velocity: is the rate at which data is generated and the rate at which the data needs to be processed. There are primarily two categories of data processing, batch and streaming. Batch processing is for analysis done after-the-fact and in large chucks at a time. Data that does not require immediate action can be analyzed independently from the real-time performance of the system. Streaming processing enables real-time decisionmaking and alerts by analyzing the data as soon as it arrives. Streaming and batch analyses each have their pros and cons, and the appropriate method depends largely on the organization's particular use case and business need. Transportation management centers have needs for both streaming and batch data processing, and use cases for both will be explored further in a subsequent report.

- Variety: refers to the different data sources (e.g., connected vehicles, connected travelers, and connected infrastructure) and types of data being generated (e.g., operational and relational data, video data, etc.). When an organization is interested in collecting as much data as possible for analysis, the ability to store and analyze a wide variety is an important factor to consider. For example, the characteristics of the data may require more complex data governance (policies describing proper handling and management of data, frequently related to security and privacy concerns). (Thomas, Gwen, "Defining Data Governance," Accessed May 12, 2016, http://www.datagovernance.com/defining-data-governance/.) Other considerations include storage capabilities for structured, unstructured, and semistructured data and advanced analysis techniques to make use of complex data (e.g., unstructured image files).

- Veracity: refers to the quality of the raw data being received. This includes challenges organizations face collecting information they can trust with data free of biases, noise (background data, impossible for machines to understand), abnormalities, or general inaccuracies. Veracity also can refer to the collection of unwanted data. Organizations may want to collect as much data as possible, but may not know what to do with it all once they have it. This is particularly true in our use cases for TSM&O specifically related to the collection of Basic Safety Messages (BSM). This will be explored further in subsequent reports. Veracity can play a particularly important role in automated decisionmaking without human interaction and intervention (e.g., adaptive traffic control, automated incident alerts, or future concepts for regional congestion pricing or road user charging).

- Value: refers to the potential that big data offers to unlock new insights, make faster and smarter decisions, and improve practice in TSM&O. Volume, velocity, variety, and veracity make big data into the beast that it is to manage. However, to create value out of the Emerging Data Sources, TSM&O organizations will need to find ways to manage the four V's in a way that maximizes the return on investment (ROI) of data as an asset.

The Explosion of Big Data

The rise of social media and the Internet of Things (IoT) gave birth to the data explosion that introduced big data to the general population. (Nammari, Brian, "IoT, Social Media and their Monster Child called Big Data, What is next?" Accessed April 4, 2016, https://medium.com/@bnammari/iot-social-media-and-their-monster-child-called-big-data-what-is-next-899eba9f6b7b#.svyjd5wmg.) The Internet and large volumes of data certainly existed before social media and IoT, but the exponential growth of data and Internet use that can be seen today requires a new understanding of the changes in the data landscape and places more importance on data as an asset than ever before.

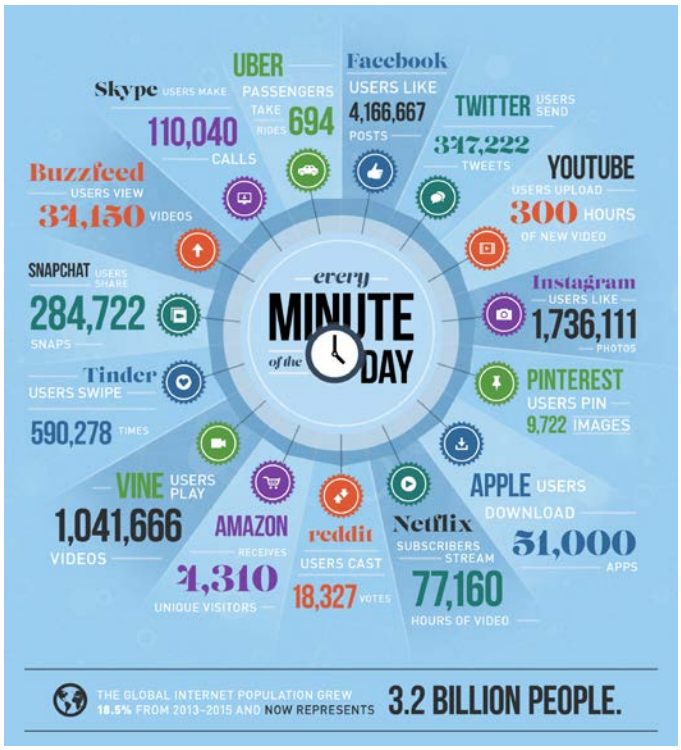

Figure 11. Infographic. Data never sleeps 3.0: how much data is generated every minute?
(Source: Domo, 2015.)

Social media has led to a steady increase in Internet traffic and data generation with no sign of slowing momentum. Domo is a business intelligence software company that released an infographic and report called "Data Never Sleeps" in 2011 showing per minute statistics on many of the most popular social media platforms, including Facebook, YouTube, Twitter, and Instagram. (Tepper, Allegra, "How Much Data is Created Every Minute?" Accessed April 5, 2016, http://mashable.com/2012/06/22/data-created-every-minute/#SAV6YUMJSmq7.) According to their research, the global Internet population grew 5.6 percent in the previous year and at the time was 2.1 billion people strong. They released a second report ("Data Never Sleeps 2.0") in 2013 showing an increase in per minute statistics across the board and an increase in global Internet users to approximately 2.4 billion people. (Domo, "Data Never Sleeps 2.0," Accessed April 5, 2016, https://www.domo.com/learn/data-never-sleeps-2.) However, most recently in 2015, Domo released "Data Never Sleeps 3.0" to show significant increases in per minute statistics of the most popular social media platforms and a jump of 18.5 percent to reach approximately 3.2 billion global Internet users. (Domo, "Data Never Sleeps 3.0," Accessed April 4, 2016, https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/.)

The Internet of Things is the interconnected system of sensors, wearable tech, phones, smart appliances, connected vehicles, and other devices (or things) broadcasting data via the Internet. IoT is the overarching concept for innovative ideas like connected cities in which connected travelers (and all the devices they carry), connected vehicles, and connected infrastructure will all be capable of communicating and interacting with each other to the benefit of organizing citywide system operations. IoT data in a connected city would support TSM&O's **Operations Capability Improvement Process** by providing rich data for planning and budgeting decisions and performance measures. (Federal

Highway Administration, "Organizing for Operations," Accessed May 12, 2016, http://www.ops.fhwa.dot.gov/plan4ops/focus_areas/organizing_for_op.htm.)

The Importance of Big Data in Transportation Systems Management and Operations

The emergence of the transportation data sources identified earlier indicates that Departments of Transportation (DOT) will soon be facing significant big data challenges of their own. Connected travelers, vehicles, and infrastructure will drive growth in data that will enhance TSM&O; but this new data requires an IT infrastructure, processes, and skills capable of handling data acquisition, marshalling and analysis. The remainder of this section discusses the big data process model, industry and government developments in big data, and emerging data analysis techniques.

Making Sense of Big Data

TSM&O organizations will face a number of hurdles when turning their big data opportunities into meaningful actions. Without proper planning and consideration, a big data solution can turn into an inefficient system with latency and performance issues. Just like a transportation solution for congestion in a growing city, a big data solution should be thoughtfully designed before any physical construction takes place. The process model depicted in the figure below is a useful way of thinking through a big data solution by breaking it into four distinct steps. TSM&O organizations will need to evaluate and design a big data solution that holistically considers each of these steps.

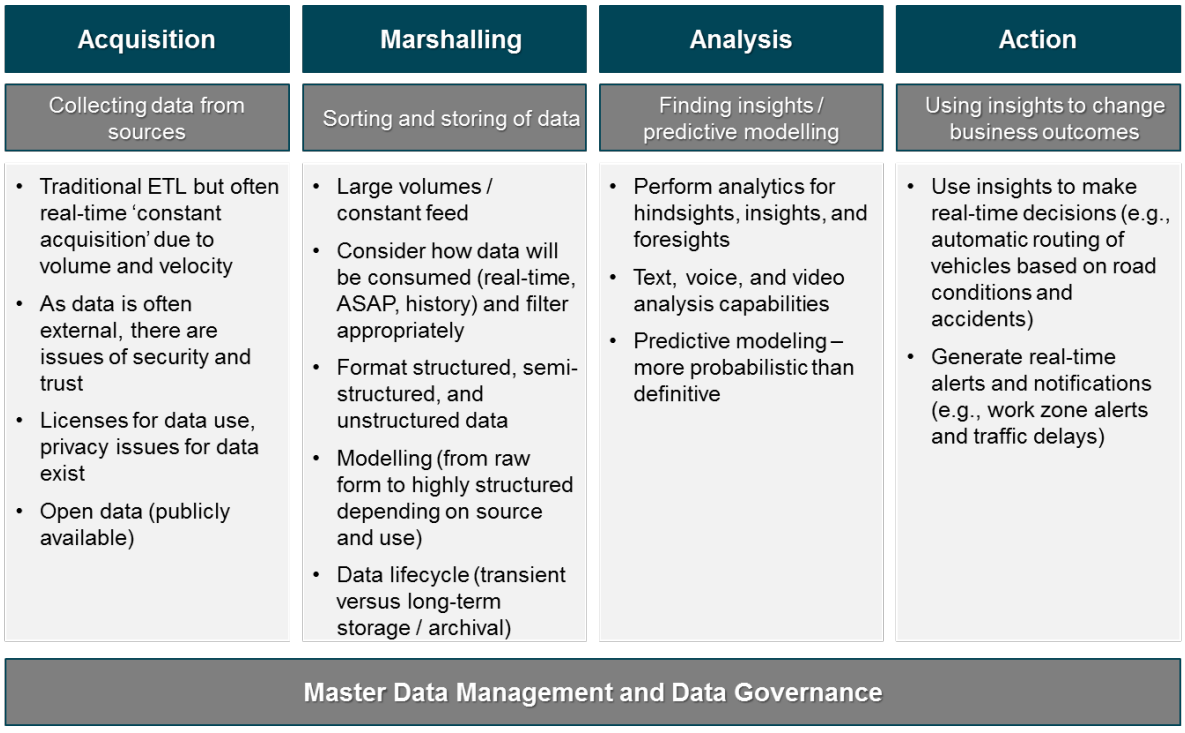

Figure 12. Chart. A big data process model: acquisition, marshalling, analysis, and action.
(Source: Adapted directly from the capgemini big data process model, 2012.)

Acquisition

Acquisition refers to the collection and preprocessing of data from a variety of systems within the organization (internal sources) and systems outside the organization (external sources). External sources of data frequently require more consideration due to differing data formats, additional privacy, governance, and security concerns (e.g., rights to use and distribute, sensitivity of the data, corrupt or malicious files, etc.). Additionally, methods for ingesting data are continuing to evolve and depend on many factors: 1) the volume of data coming in; 2) the data source; 3) how quickly data is needed; and 4) how much preprocessing of data is necessary (e.g., extract/transform/load (ETL)) before being ready for analysis. The methods and characteristics of how data is acquired are important to consider because they can directly affect the capabilities and considerations for how data is marshalled, analyzed, and acted upon. Currently, data acquisition (or 'constant acquisition') for TSM&O agencies is in the form of polling field devices using National Transportation Communications for ITS Protocol (NTCIP) or proprietary protocols, accessing data feeds from third parties using Web services and sharing traveler conditions data with other systems.

Marshalling

Marshalling refers to the **sorting and storing of data**. The five V's are of particular importance when it comes to marshalling the data. The high volume, fast velocity, diverse variety, and questionable veracity of big data requires a robust and adaptable storage solution to harness its value. A big data solution must be capable of storing all types of data an organization is interested in collecting at the speed needed to collect and process it for actionable insights. This includes the ability to compress and archive legacy data as well as newly collected data that isn't necessary for immediate or frequent analysis.

Analysis

Analysis refers to how an organization wants to use their data, including the ability to find insights and inform decisions through advanced analytical techniques and visualization. Analysis can be performed at many different speeds and can use a wide variety of tools and techniques. One set of methods uses statistical, descriptive, and predictive models to provide hindsight, insight, and foresight, respectively. For example, predictive models may one day be used to forecast traffic conditions based on weather, incidents, historical traffic data, and other factors. Additionally, new techniques are rapidly emerging to analyze data previously considered too difficult, including unstructured data like text, audio, and video. Improvements in video analytics technologies may make streaming video from closed-circuit television (CCTV) much more valuable than it is today. With properly designed acquisition, marshalling, and analysis methods, tools such as live, **interactive "dashboards" can be designed to minimize the time from data ingestion to actionable insights**. This is the kind of thing that Performance Measurement System (PeMS) and Transportation Information System (RITIS) are doing for TSM&O agencies today with traditional databases and acquisition methods.

Action

Action refers to the use of insights gained during the analysis stage to change business outcomes. For example, making real-time decisions based on traffic conditions and crash reports to generate alerts and notifications informing in-vehicle devices or nomadic devices of potential traffic delays. Insights may yield different actions depending on the priorities of decisionmakers from operations engineers, planners, first responders, departmental leadership, and regional coordinators. The effect

of actions based on insight is the potential value (the fifth V of big data) that TSM&O organizations can achieve through big data. Because action is typically a result of the analyses being performed and dependent on the business outcome desired, it is the last consideration of a big data solution, sequentially. In other words, a complete solution requires additional consideration specific to the business problem to be solved, but it is not considered in detail in this report.

Big Data Trends

Given big data's recent introduction as a serious technology disrupter, it is important to understand the current trends of the market overall, as well as the trends across industries and government agencies. The information contained in this report is a high-level "snapshot" of the current big data landscape as of spring 2016. Many big data tools are quickly gaining traction, and **new technologies are constantly emerging and maturing.** (Olavsrud, Thor, "21 data and analytics trends that will dominate 2016," Accessed April 5, 2016, http://www.cio.com/article/3023838/analytics/21-data-and-analytics-trends-that-will-dominate-2016.html.) Technology is always expanding and pushing the limits of the next great innovation, and, with that perspective in mind, this report is intended to provide an introduction to the technologies, tools, and practices anticipated to be at the forefront of the big data movement in the next 5 to 10 years.

Increased Adoption

The biggest trend in the realm of big data is increased adoption. Gartner's "Hype Cycle for Emerging Technologies" **very noticeably left Big Data out of its 2015 edition**. While it had been a staple of the famous Hype Cycle in past years and was trending down the Slope of Disillusionment as expected in 2014, the Gartner author commented on its absence saying "We've retired the big data hype cycle. I know some clients may be really surprised by that because the big data hype cycle was a really important one for many years. But what's happening is that big data has quickly moved over the Peak of Inflated Expectations and has become prevalent in our lives." (Woodie, Alex, "Why Gartner Dropped Big Data Off the Hype Curve," Accessed April 6, 2016, https://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/.) Essentially, the big data movement has become commonly accepted and, in some cases, specific analysis and action concepts have been introduced in its stead (e.g., Machine Learning and Natural Language Processing). The Gartner Hype Cycle will be discussed in more detail in subsequent sections. Several key areas of increased adoption of big data tools, technologies, and related concepts are described below.

Big Data in Strategic and Tactical Operations

Up to this point, big data has been focused on accomplishing very specific, operational business objectives (e.g., Google Maps and other Navigation systems use traffic data to provide basic route guidance to individual drivers). However, the new wealth of available data is opening opportunities in high-level strategy work as well. According to a report, "Big Data—Moving from the operational to the strategic," released by Wipro, **one of the next big stepping stones for big data will be in strategic and tactical business endeavors.** (Sanjiv, K.R., "Big Data—Moving from the operational to the strategic," Accessed April 6, 2016, http://www.wipro.com/documents/Wipro-analytics-big-data-moving-from-the-operational-to-the-strategic.pdf.) Data from connected vehicles could provide insight into crash hotspot locations leading to a reduction in incidents at these locations. Likewise, connected traveler data could tell us how readily individuals in a particular corridor will change their travel

patterns in response to congestion, incidents, weather and other factors, leading to better traveler information and more targeted Integrated Corridor Management (ICM) strategies.

Open Source Software in the Government

Agencies across the Federal Government are recognizing the importance of open source software (OSS) as is evident by several reports released by government agencies. One report released in 2013 by the Department of Homeland Security (DHS) titled, *Open Source Software in Government: Challenges and Opportunities*, describes the current use of OSS, the barriers they see to its mainstream adoption, and the next steps that need to be taken. (U.S. Department of Homeland Security, "Open Source Software in Government: Challenges and Opportunities," Accessed April 6, 2016, https://www.dhs.gov/sites/default/files/publications/Open%20Source%20Software%20in%20Government%20%E2%80%93%20Challenges%20and%20Opportunities_Final.pdf.) Another report released by the White House Office of Management and Budget (OMB) in 2014, titled *The Open Government Partnership* details the Government's belief in the value of and their commitment to OSS both across government agencies and with the general public. (The Open Government Partnership, "Announcing New Open Government Initiatives," Accessed June 27, 2016, https://www.whitehouse.gov/sites/default/files/microsites/ostp/new_nap_commitments_report_092314.pdf.)

The price tag (i.e., "free") is tempting for any organization, but perceived risks to security, lack of a controlled release roadmap and potential lack of commercial support has many holding back and continuing to pursue more traditional solutions. In response to this, companies that sell support for OSS and/or additional proprietary software (e.g., Red Hat Enterprise Linux) have made OSS a viable option for customers who require stability and support. For example, the United States Census Bureau released a request for quotation (RFQ) to explore open source Hadoop technology through vendors that support the OSS tools. Cloudera, Hortonworks, and MapR are three examples of companies that sell support for their own distributions of Hadoop (each of these will be discussed in subsequent sections of this report). As vendors like these work to provide comprehensive services for open-source big data technologies, the draw of these solutions becomes the commodity hardware they are built on providing high scalability with significantly reduced cost; just the "do more with less" that government agencies require.

Open Data in the Government

An executive order was issued in 2013 demonstrating the Government's position on the Open Data Policy advocating greater transparency of data and interoperability between government agencies through the use of open and shared machine-readable data. (The White House, Office of the Press Secretary, "Executive Order—Making Open and Machine Readable the New Default for Government Information," Accessed June 27, 2016, https://www.whitehouse.gov/the-press-

office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.) Consistent with this mission, United States Department of Transportation (U.S. DOT) has recognized the importance of open data with the creation of the Open Source Application Development Portal (OSADP), Research Data Exchange (RDE), and Operational Data Environment (ODE) for the purposes of "promot[ing] open source development of software applications that use connected vehicle technology and data to help travelers avoid delays." (U.S. Department of Transportation, Federal Highway Administration, "Open Source Application Development Portal," Accessed June 27, 2016, http://www.itsforge.net.; U.S. Department of Transportation, Office of the CIO, "Privacy Impact Assessment—Federal Highway Administration (FHWA) Open Source Application Development Portal," Accessed June 27, 2016, https://cms.dot.gov/sites/dot.gov/files/docs/OSADP__FHWA_PIA_Adjudicated_082514.pdf.)

Big Data at the Census Bureau

The U.S. Census Bureau (USCB) is designing and implementing a big data analytics platform as part of an initiative to implement "Adaptive Survey Design" across the entire organization. Adaptive design is an innovative approach to collect survey and census data in more efficient ways to dynamically assign data collection methods, save the USCB both time and money, and increase the quality of the data collected.

USCB is positioning their solution architecture at the leading edge of big data solutions poised to adapt more quickly to future technologies and trends.

http://dc-aapor.org/SystemsInfrastructureMathurThieme.pdf.

Machine Learning and Cognitive Analytics in Industry

Machine learning is the technology behind designing machines capable of pattern recognition to "learn" without being explicitly programmed to do so. Cognitive computing uses machine learning, natural language processing, and other technologies to simulate the abilities of the human brain. **Cognitive analytics** is the combination of cognitive computing and analytic techniques to make sense of data in a smarter and more efficient way. (Deloitte, "Cognitive analytics: The three-minute guide," Accessed April 6, 2016, http://public.deloitte.com/media/analytics/pdfs/us_da_3min_guide_cognitive_analytics.pdf.) Each of these techniques becomes more accurate and more useful as they process more data making them an attractive tool to be adopted into a big data solution. These tools may become critical to fully utilize the potential of big data over the next 10 years.

For example, a report by Forbes describes how cognitive analytics has the potential to offer unimaginable insight into the world of financial fraud prevention and protection, where small improvements in performance can turn into hundreds of millions in savings. With the ability to continuously learn from the data being analyzed, financial analytical systems will be capable of uncovering insights previously unseen, automatically recognizing and managing user patterns, and aiding in evidence-based decisionmaking. (Drury, Nicholas and Sarkar, Sandipan, "How Cognitive Computing Impacts Banks and Financial Markets," Accessed April 6, 2016, http://www.forbes.com/sites/ibm/2015/11/09/how-cognitive-computing-impacts-banks-and-financial-markets/#31e622e525e5.) Many of the same principles may be able to be applied to understanding traffic patterns more thoroughly based on weather conditions, traffic congestion, visibility, and other factors to make accurate short-term predictions and provide appropriate preventative measures if enough data can be quickly and reliably assimilated into a prediction model.

Security and Governance in Big Data

Security and data governance are topics frequently heard in the debate surrounding big data and other naturally related topics (e.g., cloud and OSS), and this conversation is not any less important for TSM&O agencies. A **trend we see as particularly important for agencies in big data governance is the added consideration of shared data**, whether that's shared across a single agency, shared between local agencies, or shared across entirely disparate groups (e.g., local TSM&O agencies, auto manufacturers, private ride sharing companies like Uber and Lyft, etc.). Shared data requires all data consumers (both individual users and organizations as a whole) to consider the appropriate governance controls and regulations required across the data lifecycle.

The inherently shared data of connected travelers, vehicles, and infrastructure will likely be capable of painting a rather clear picture of U.S. citizens as they move about their daily lives, and could be used for dishonest purposes if not properly protected and regulated. A Corporate Partnership Board report from the International Transport Forum titled, "Big Data and Transport: Understanding and assessing options" explores the work still to come in understanding the privacy implications of big data in transportation. The report comments that a "New Deal on Data" may be necessary in the near future to redefine and regulate data ownership between data producers, consumers, and subjects. (International Transport Forum Corporate Partnership Board, "Big Data and Transport: Understanding and assessing options," Accessed April 6, 2016, http://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf.) Many State DOTs already have formed **data governance committees and initiatives** specific to their organizational structures (including Strategic Highway Safety Plan (SHSP), Highway Safety Improvement Program (HSIP), Traffic Records Coordinating Committees (TRCC), etc.); however, these committees will need to consider new and unique challenges posed by the emerging data sources discussed in previous sections as well as how they wish to be interacting with other transportation organizations in the coming years. (Transportation Research Circular, "Improving Safety Programs Through Data Governance and Data Business Planning," Accessed June 27, 2016, http://onlinepubs.trb.org/onlinepubs/circulars/ec196.pdf.) The Federal Highway Administration (FHWA) currently is in the process of creating a Data Governance Plan to advise agencies on topics such as these. As of spring 2016, the first of six volumes has been published for dissemination across DOTs, and the full plan is estimated to take several years to complete and publish. (Federal Highway Administration, "Data Governance Plan Volume 1: Data Governance Primer," Accessed May 14, 2016, https://www.fhwa.dot.gov/datagov/.) The security concerns and proper use of these emerging data sources should be considered during the creation of this report and, consequently, TSM&O agencies should consider these reports and others like them when acquiring, marshalling, and acting on emerging data.

Emerging Technologies and Concepts

As discussed previously, **big data is officially considered on the Plateau of Productivity** according to the Gartner Hype Cycle and into recognized and established territory. However, there are a number of specific technologies that are still a part of the Hype Cycle and may have some distance to travel before they can be considered firmly established technologies. A few key emerging technologies and concepts are discussed below.

Hadoop as a Leader

Hadoop is likely the most established technology on this list, but it also is a rapidly changing ecosystem that many of its subcomponents can still be considered emerging technologies. Hadoop is an enterprise big data storage and computational platform, the core components of which include **HDFS** (distributed file system), **MapReduce** (computing framework), and **Yet Another Resource Negotiator (YARN)** (resource management tool). Hadoop is an open source initiative founded by the nonprofit Apache that has been divided into subprojects to be worked on by volunteers from Apache, other vendors, and the open-source community at large. These subprojects make up the total ecosystem of wide-ranging capabilities that Hadoop offers and enable the fast-paced adaptability that Hadoop is known for. Just a few of the more commonly known subprojects include a relational database (Hive), a nonrelational database (HBase), a distributed coordination service (Zookeeper), and an in-memory, distributed computational engine (Spark).

Due in large part to its open-sourced nature, Hadoop has quickly become the foremost big data solution available in the market today. New components can be added as needed by opening a new subproject and contributing directly. While Apache provides all OSS completely free of charge, several vendors have chosen to package their own distributions of Hadoop and sell their support services as well as their own proprietary additions. More detail will be provided on the most popular Hadoop vendors and the services they offer as well as how they compare to other big data solutions in subsequent sections; however, it is worth understanding from the outset the wide reach Hadoop has in the big data field. Quite a few variations and support packages are available through vendors, but the adaptability and speed with which new technology and tools can be developed make Hadoop a significant leader in big data technologies at this time.

Geographic Information Systems Tools for Hadoop

A technology that already is pervasive in TSM&O is Geographic Information Systems (GIS). GIS is designed to store, manage, analyze, and visualize geographic and spatial data. There are several GIS tools available for Hadoop on the market currently, however, some are more mature than others. ESRI (an international GIS software company) has developed several open source advanced GIS tools for Hadoop. Additionally, there are GIS tools emerging that run on the Spark in-memory engine, including Magellan and SpatialSpark providing a new class of high performance geospatial computing for big data. (GetInData, "Geospatial analytics on Hadoop," Accessed April 7, 2016, http://getindata.com/blog/post/geospatial-analytics-on-hadoop/.) Since transportation data is integrally based on location, consideration of GIS tools and technologies will be an important element in subsequent reports.

Big Data on the Cloud

Many organizations have given their predictions for big data in the coming years and mention **an inevitable move to the cloud.** (Olavsrud, Thor, "21 data and analytics trends that will dominate 2016," Accessed April 5, 2016, http://www.cio.com/article/3023838/analytics/21-data-and-analytics-trends-that-will-dominate-2016.html.; IDC, "IDC Reveals Worldwide Big Data and Analytics Predictions for 2015," Accessed April 7, 2016, http://www.idc.com/getdoc.jsp; Rossi, Ben, "Top 8 trends for big data in 2016," Accessed April 7, 2016, http://www.information-

> Since transportation data is integrally based on location, consideration of GIS tools in Big Data systems and technologies will be an important element.

age.com/technology/information-management/123460615/top-8-trends-big-data-2016.) The inherent scalability and flexibility required to support a big data solution make cloud an attractive option from the start. Cloud technology is no longer considered an emerging trend; however, big data solutions on the cloud using open source cloud technologies and tools is a recent evolution.

There are primarily two cloud options to be considered for a big data solution: **platform-as-a-service (PaaS)** and **infrastructure-as-a-service (IaaS)**. IaaS solutions set up and manage the foundation of a system (network, storage, servers, and virtualization); PaaS solutions set up and manage the foundation as well as the operating system, middleware, and runtimes required. OpenStack and Docker are IaaS and PaaS OSS cloud options, respectively. OpenStack and Docker both offer free, interoperable, and adaptable options for organizations to move into the cloud, and both solutions are gaining popularity exponentially with OpenStack being referred to as "the next Linux". (IT Business Edge, "Ten Reasons Why OpenStack Will Rule the Enterprise," Accessed April 7, 2016, http://www.itbusinessedge.com/slideshows/ten-reasons-why-openstack-will-rule-the-enterprise.html.) Additionally, many traditional cloud providers (e.g., Amazon Web Services (AWS), IBM, Microsoft, etc.) offer their own Hadoop PaaS solutions. A detailed discussion of the relative costs and capabilities of those platforms will be elaborated more in subsequent sections.

Government organizations, including many TSM&O agencies also are starting to join the cloud conversation. While government organizations tend to be risk-adverse in nature and maintain tight control of their IT assets, they have recognized the potential benefits of deploying in the cloud, including instant and cost efficient scaling. Several early adopters already have made the leap, including the General Services Administration (GSA), Department of the Interior, Department of Agriculture, National Aeronautics and Space Administration (NASA), and the National Oceanic and Atmospheric Administration. (InformationWeek, "5 Early Cloud Adopters in Federal Government," Accessed April 7, 2016, http://www.informationweek.com/government/cloud-computing/5-early-cloud-adopters-in-Federal-government/d/d-id/1315911.) **One of the biggest obstacles lies in a lack of certification by the Federal Risk and Authorization Management Program (FedRAMP) of most major cloud providers**. This continues to create lingering doubts as to the security and privacy that cloud currently can offer. However, as one of the early adopters of the cloud, GSA has been a big push in the Federal Government's "Cloud First" initiative that requires "agencies take full advantage of cloud computing benefits to maximize capacity utilization, improve IT flexibility and responsiveness, and minimize cost." (General Services Administration, "Cloud IT Services," Accessed April 7, 2016, http://www.gsa.gov/portal/content/190333.)

Internet of Things (IoT)

With data streaming in from every direction (from connected travelers, vehicles, and infrastructure, for example), IoT platforms are being developed to connect to these devices and ingest the data with use-cases that cut across a variety of industries and markets in very similar ways.

In 2015 a number of predictions were made for the Internet of Things in a report produced by the International Data Corporation (IDC), an organization that provides insight and strategy on emerging market opportunities. In this report, IDC predicts that:

- **"Within 5 years, all industries will have rolled out IoT initiatives with more than 90 percent of all IoT data hosted on PaaS."**
    - This prediction gives insight into the significance that businesses are placing in the future of IoT, the versatility that it must have to cross into every industry, and the fact that

agencies such as TSM&O organizations will only manage the applications and the information, with the hardware, software, and data hosted on a Platform.

- **"By 2017, 90 percent of enterprise system practices will adopt new business models to manage the service-oriented, nontraditional infrastructure."**
  - This prediction is a technical way of saying that IoT is coming, but businesses don't necessarily know how to use it yet and their business currently can't handle it. This is particularly true for most DOTs that are primarily reliant on "on premise" operations today. With the speed at which DOTs can react to market changes, the 2017 prediction is likely premature for TSM&O practices.
- **"By 2018, 40 percent of IoT-created data will be stored, processed, and analyzed at the edge of the network."**
  - The "edge of the network" refers to the devices collecting and distributing the data in an IoT system. These are sometimes referred to as "edge devices" and include routers, switches, and processors closest to where the data is delivered to the organization. This prediction indicates that instead of bringing all data into a central location, there will be a trend to manage data closer to the devices that provide it, reducing data movement and eliminating massive accumulation of data. This is particularly true for the emerging sources evaluated in this report; the collection, processing, and transmittal of raw BSMs all the way back to the Traffic Management Center (TMC) will be intractable. Strategies for handling these issues will be explored in a subsequent report.
- **"The movement of large quantities of data produced will cause roughly 50 percent of IT networks to be constrained and 10 percent overwhelmed by IoT devices."**
  - This prediction assumes that (based on the second prediction in which IoT is expected to significantly affect the normal operating procedure of businesses) IoT will overwhelm the businesses that don't react or don't react quickly enough. This will likely be the case for TSM&O organizations given the analysis in this report in chapter 3.

It is clear from the analysis in chapters 3 and 4 that TSM&O organizations will need to address how IoT tools and technologies intersect with big data tools and technologies in some form or fashion as more and more data becomes available from more and more devices. While the general concepts of IoT have been part of TSM&O for more than 40 years, the scale of connected infrastructure (100s-1000s) is pale in comparison to the data from the millions of travelers and millions of vehicles that will be available to the DOT in the next 10 years. The growth of generalized technologies in IoT for a variety of other industries will likely lead to new innovations that apply to TSM&O just as easily as other application sectors.

Every year, Gartner releases an updated version of the Hype Cycle for Emerging Technologies. The report shows at-a-glance the current state and projected progress of many of the most popular emerging technologies in the market today. Each Hype Cycle drills down into five key phases of a technology's complete lifecycle. According to Gartner, those phases are:

While the general concepts of IoT have been part of TSM&O for more than 40 years, the scale of connected infrastructure (100s-1000s) is pale in comparison to the data from the millions of travelers and millions of vehicles that will be available to the DOT in the next 10 years.

- **Innovation Trigger:** A potential technology breakthrough kicks things off.
- **Peak of Inflated Expectations:** Early publicity produces a number of success stories—often accompanied by scores of failures.
- **Trough of Disillusionment:** Interest wanes as experiments and implementations fail to deliver. Producers of the technology shake out or fail.
- **Slope of Enlightenment:** More instances of how the technology can benefit the enterprise start to crystallize and become more widely understood.
- **Plateau of Productivity:** Mainstream adoption starts to take off.

Based on a user's appetite for risk, they can use this report to assess their willingness to invest in a given technology.

In July of 2015, Gartner released its newest version of the Emerging Technology Hype Cycle. The report estimates that IoT has just crossed the "Peak of Inflated Expectations" and is beginning its trend down the "Trough of Disillusionment," where implementations fail and curiosity fades. The phase is predicted to last 5 to 10 years before mainstream adoption will takeover in the "Plateau of Productivity" phase. The IoT platform also appears on the Hype Cycle just before the "Peak of Inflated Expectations."." Further in this report, we summarize some available platforms, but at this stage in their development very little details are known about their effectiveness or real-world utility, as astutely noted by Gartner.

Figure 13. Graph. Gartner's Hype Cycle for emerging technologies.
(Source: Gartner, 2015.)

While these terms might have jarringly negative language, movement through the Hype Cycle is an important path every new technology travels to eventually reach the "Plateau of Productivity." As discussed previously in this report, big data was previously on the Hype Cycle, but was left off the most recent edition due in part to its increasing legitimacy and in part to its broad scope.

The IoT reference model provides a standardized and simplified model of IoT systems and applications. The reference model is comprised of seven levels; however, it is important to note that data may flow in both directions. The following reference model shows how big data fits within IoT:

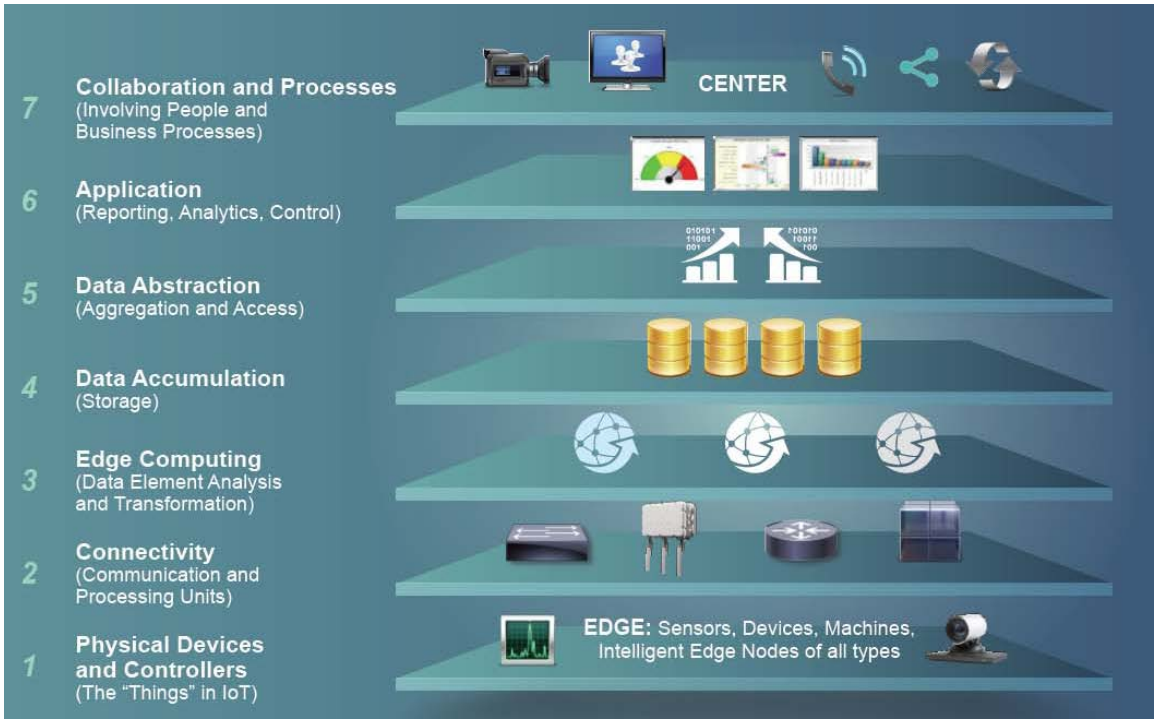

Figure 14. Illustration. Internet of things reference architecture.
(Source: Cisco, 2013.)

Physical devices and controllers are considered the "things" in the Internet of things. They include a wide range of endpoint devices that send and receive data. In the context of this report, connected travelers, vehicles, and infrastructure are the "Things."

Connectivity represents the communication and processing by the existing networks. The most important function of this level is reliable, timely information transmission. In the context of this report, this is described by the Point of Access for each data source in the first section.

In **edge computing**, network data flows are converted into information that is suitable for Level 4. On this level, data is evaluated, formatted, and assessed on a packet-by-packet basis before sending the data to storage. **This is a critical issue for TSM&O organizations**. Design and evaluation of

computing architectures for edge processing of the Emerging Data Sources (particularly public connected vehicles) will be the subject of a subsequent report.

During **data accumulation**, data in motion is converted to data at rest. As the data is put to rest, it becomes usable by applications on a nonreal-time basis. Applications can access the data when necessary. **Data Accumulation is a critical issue for TSM&O organizations**.

Data abstraction focuses on rendering data and its storage to provide faster querying capabilities by applications. The Data Abstraction process includes reconciling data formats, assuring data consistency, and consolidating data into one place. **Data abstraction is a critical issue for TSM&O organizations**.

The **application** level is where information interpretation occurs. Software interacts with the data abstraction level to view the data stored. At this level, business intelligence reports are utilized to understand the data for operational decisions. **How big data tools and applications integrate with legacy TSM&O applications and systems is a critical issue for TSM&O organizations**.

Collaboration and processes is the human interaction and business process layer. This level empowers people to do their work better by providing the right data from the application level, at the right time, to make meaningful decisions.

The general structure of IoT describes the component layers of the information processing system that will be needed as connected vehicles and travelers join the existing connected infrastructure of the TSM&O agency in becoming a reality over the next 5 to 10 years. Because these commercial and open source platforms are right at the top of the Gartner Hype Cycle, precious little is known about the details. Certainly since the data sources from connected travelers and vehicles are still emerging, it is not expected that an IoT platform would be able to acquire, marshal, and analyze the data without significant investment in time and funding by a particular agency, a consortium, FHWA/U.S. DOT, or even third-party providers intending to sell TSM&O agency these capabilities as a Service. Subsequent reports will explore the requirements of TSM&O applications that may serve to drive features and functionality of IoT offerings so that, in 5 to 10 years' time, they can closer to "plug and play" for TSM&O-related applications.

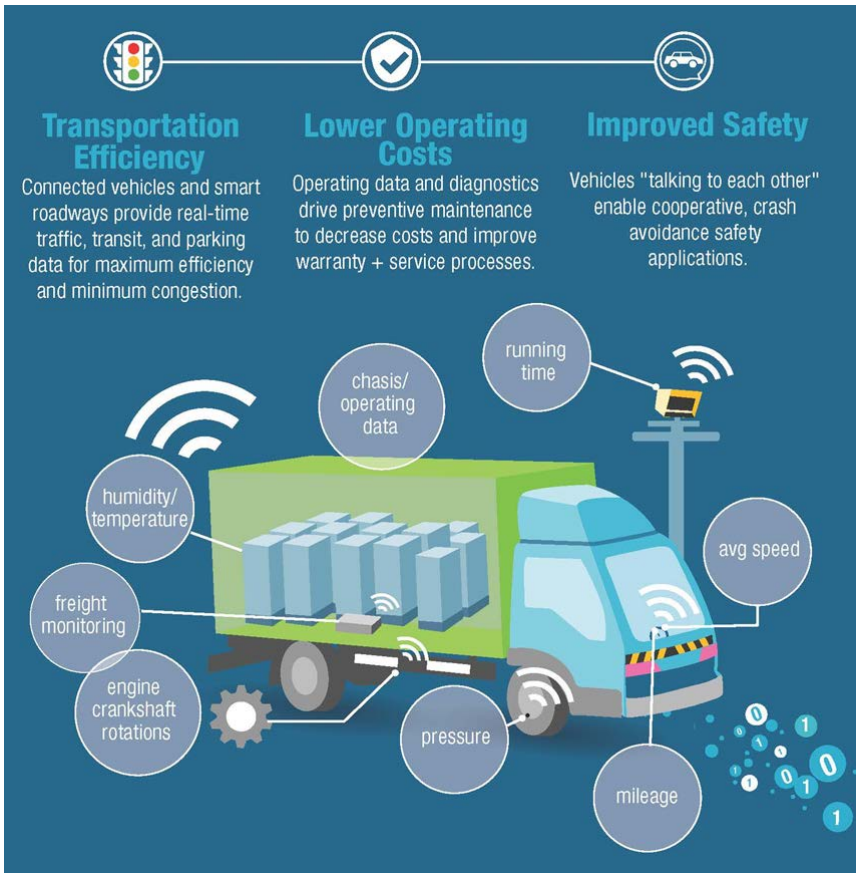

Figure 15. Illustration. Potential impact of Internet of Things in the freight shipping industry.
(Source: Datastax Infographic https://jaxenter.com/Internet-things-data-go-112274.html, 2014.)

Acquisition Trends

In the immediately preceding sections, several overall trends seen in the big data arena across industries and government agencies are discussed. In the next few sections, some trends specific to each step of the big data process model will be discussed (except Action). Many of the trends discussed may seem easily generalizable and relevant to any system implementation. However, each is made more complicated by the five V's of big data discussed earlier. In the case of data acquisition, several trends worth watching include organizations' desires to collect the "right" data and save time, money, and resources on unnecessary data; to collect more data in real-time to make faster decisions; and to connect to legacy databases and leverage their investments in their current IT infrastructure. These issues are all relevant to TSM&O and will be discussed in more detail in subsequent reports.

Collecting the Right Data

One of the most daunting tasks in implementing a big data solution is sifting through the new wealth of available data for the data that really matters. Big data solutions are designed around inexpensive and scalable storage resources; however, as cheap as storage can be, it still has a cost and should be managed as an asset. In this case, the less extraneous, unused data stored, the better ROI an organization will realize on its big data solution. This is a problem TSM&O agencies likely already

face, but will be magnified to an incredible scale if the currently available (but uncollected) data and the emerging data sources discussed previously become part of agencies' regular intake. Wipro's report regarding big data's move into strategic initiatives showed that organizations overwhelmingly fear data "quality issues" and "difficulty in assessing which data is truly useful (data overload)." (Sanjiv, K.R., "Big Data—Moving from the operational to the strategic," Accessed April 6, 2016, http://www.wipro.com/documents/Wipro-analytics-big-data-moving-from-the-operational-to-the-strategic.pdf.) This will be critically true for the collection of detailed connected vehicles data identified earlier. TSM&O organizations new to big data will need to develop methods of choosing what data to collect, sifting through the available data to understand what should be kept, and storing and archiving based on frequent and infrequent use, respectively. All TSM&O organizations, for example, struggle with maintenance of fixed assets (in pavement loop detectors, radar, and video cameras) for traffic detection. The emerging data sources of connected travelers and vehicles may be able to replace these assets and reduce the expense of maintaining and replacing the equipment. The desire to ingest, store, and analyze every piece of data that could provide business insights is dependent on avoiding irrelevant data that wastes resources.

Collecting Real-Time Data

While collecting data in real-time may seem to be in direct opposition with the previous trend and almost guarantee data overload, streaming data in real-time or near real-time is essential to enable the agility organizations will need to make informed and evidence-based decisions just as quickly as the data comes in. The speed with which agencies are capable of ingesting and processing the data will directly affect the speed with which they can ultimately use the data (i.e., perform analyses and execute actions based on those analyses). Once again, this may be a problem agencies currently deal with, but the magnitude and frequency with which we expect streaming data to increase is likely to be staggering. Many tools and technologies enabling streaming data acquisition speeds will be discussed in detail in subsequent sections.

Coexistence of Big Data Tools and Legacy Databases

Every big data solution should consider and incorporate in its design an organization's legacy IT infrastructure. This may include both integrating systems still relevant to business processes and retiring outdated or newly redundant systems. An excellent example of systems likely to remain relevant are the many sources of data that organizations are built around. A complete big data solution will still require integration with these data sources. Deciding to implement a big data solution is an important business decision for any organization, and examination of legacy investments is an incredibly important first step in that decision that requires careful consideration, technical, financial, and otherwise. For example, if an agency currently uses Oracle for their Freeway Management System and MSSQL for their real-time Arterial Management System, these legacy databases will need to be integrated with or entirely consumed by the Linux-based Hadoop. This would have to be a business decision based on the technical skills requirements, potential ROI, total investment already made in Oracle and MSSQL, and other factors.

If an agency currently uses Oracle for their Freeway Management System and MSSQL for their real-time Arterial Management System, these legacy databases will need to be integrated with or entirely consumed by the Linux-based Hadoop.

Marshalling Trends

The most recognizable aspect of the big data solution is data marshalling. A marshalling solution must be capable of storing the total volume and variety of data an organization collects at the velocity needed for timely action. Additionally, many marshalling solutions require replication factors to ensure fault tolerance of the data being stored and the processes being executed, and would require at a bare minimum the storage resources for the total replicated data volume. **Software-defined storage** and the use of **commodity hardware and virtualization** are just a few of the popular trends in data marshalling.

Software-Defined Storage

Software-defined storage (SDS) is the decoupling of the physical hardware and the specific software that defines how the physical hardware works. There is often confusion in the distinctions between SDS, storage virtualization, and traditional storage, and there is little standardization of terms and definitions at this stage. In a traditional system, hardware is controlled by a layer of built-in software. This can limit flexible configuration capabilities and introduce significant vendor lock-in based on the requirements an organization has. Storage virtualization is when several storage resources (e.g., multiple physical servers) are virtually combined as one complete set and then logically partitioned into separate resources, similar to a redundant array of independent disks (RAID). This option can lighten some of the restrictions introduced by traditional storage alternatives. SDS, however, sets up a brand new software layer responsible for data replication, snapshots, and other hardware management capabilities that would typically be the responsibility of the hardware if it had those capabilities at all. SDS can be implemented on commodity hardware or hardware that already exists to **offer a more interoperable solution for organizations concerned with vendor lock-in or limitations due to their hardware specifications**. SDS can eliminate many of these concerns and provide a more cost efficient, flexible solution with automated management capabilities. (Rouse, Margaret, "software-defined storage," Accessed April 7, 2016, http://searchsdn.techtarget.com/definition/software-defined-storage.)

However, as with any new technology trend, SDS solutions require careful consideration of the business needs and IT capabilities. Several concerns a former Dell engineer shared include the vagueness in SDS definitions, a lack of understanding of performance needs, and the risky desire to manage every aspect of the organization. **SDS is still a growing concept**, and the definition and understanding of the technologies available today are unclear at best. Organizations should be careful to understand what their storage performance needs are, what SDS services are available, and what SDS services can actually provide. Additionally, organizations should be careful to understand the complexity required to manage an SDS solution and the potential risk they are taking on if the solution fails. (Vekiarides, Laz, "5 bitter truths about software-defined storage," Accessed April 7, 2016, http://www.infoworld.com/article/2997239/storage/5-bitter-truths-about-software-defined-storage.html.)

Commodity Hardware

One of the most frequently mentioned traits of big data marshalling is the use of **commodity hardware as a means of reducing storage costs**. The use of commodity hardware applies to Hadoop and SDS solutions, but may not apply to a massively parallel processing (MPP) database which may use high performance hardware. MPPs and Hadoop platforms are both discussed in more detail in further sections. Hadoop and many other technologies have embraced the fact that servers will eventually fail. Based on this fundamental principle, many big data technologies are limiting the dependency on hardware investment and designing solutions that are extremely fault tolerant, robust, and highly available systems largely unaffected by server failures.

For example, Hadoop builds in a **default data replication factor of three** that breaks data into chunks, copies those chunks twice, and then stores them in distributed locations across storage nodes, (a "node" is a storage server computer). This method of data storage ensures the fault tolerance of the data; in the case that one node fails, a master node makes a copy of all the data stored on the lost node and redistributes the **new copies around the cluster again to maintain the replication factor of three**.

This has serious implications for storage capacity planning; TSM&O organizations will need to perform extensive sizing and planning exercises to understand the total storage anticipated and then consider the replication factor they require for fault tolerance. Using commodity hardware ensures that an organization loses as little as possible when a server inevitably fails because organizations can choose to either repair the failed server (without having the rest of the platform dependent on server recovery) or simply remove and replace with another for minimal loss. Deploying a solution in the cloud produces many of the same benefits by removing all concerns of physical servers; however, it has its own set of challenges as mentioned earlier and expanded in future sections.

Virtualization

Virtualization is another option to decrease hardware costs, increase customization and interoperability, and increase system availability. Virtualization is the act of creating a virtual version of the network, operating system, application, and/or other levels of a system. There are many gradations of virtualization; hardware virtualization being one of the most common. Hardware virtualization is the creation of one or several virtual machines (VM) within the physical hardware that behave as a completely independent computing platform. VMs have their own operating system, middleware, and applications, independent from that of the physical hardware where the VM lives. Many TSM&O organizations already are embracing the concepts of virtualization for traditional applications and databases. The use of VMs has been shown to reduce the energy footprint of data centers, decrease investments in hardware, increase availability, reduce vendor lock-in, and reduce the time to provision, transfer, and manage servers. **Proper configuration of VMs is critical** to maintaining application performance requirements and, in particular, the read/write performance of database access and retrieval (both legacy and big data).

TSM&O organizations will need to perform extensive sizing and planning exercises to understand the total storage anticipated and then consider the replication factor they require for fault tolerance.

Data Analysis Trends

Once data is acquired, prepared, and stored, it needs to be used. Data analysis is the final, crucial step in the data process model before achieving the end goal of actionable insights. The benefit of developing many ways to acquire and store high volumes and diverse varieties of data is to derive business value, and data analysis trends are becoming more sophisticated and robust as the ability to collect, store, and process rich and diverse data improves. The list below describes some of the biggest trends in data analysis. This topic is discussed in more detail specifically for TSM&O practices in a subsequent report.

Increased Investment in Skills

"Data scientists are the people who understand how to fish out answers to important business questions from today's tidal wave of structured and unstructured information. A good data scientist has to be able to speak the language of business as well—which is what separates data scientists from great analysts or data management experts. Data scientists want to build things—not just give advice." (Deloitte, "Data scientists: The three-minute guide," Accessed April 6, 2016, http://public.deloitte.com/media/analytics/pdfs/us_ba_Deloitte3minDatascientist_021813.pdf.) You may have to spend a lot to retain a data scientist suited to your organization. But when you consider that person's ability to influence smarter, more focused investments in other areas such as technology, it's a premium worth paying." (Deloitte, "Cognitive analytics: The three-minute guide," Accessed April 6, 2016, http://public.deloitte.com/media/analytics/pdfs/us_da_3min_guide_cognitive_analytics.pdf.)

As more organizations shift to a data-driven mindset, **the demand for data scientists is outpacing supply**. The result is a highly competitive landscape for attracting and keeping top talent. This demand makes data science an attractive field of study in academia, particularly at the Master's level (where students are more likely to go into industry jobs rather than stay in academia). The first such program was started at North Carolina State University (NCSU) in 2007. Since then similar programs have been developing exponentially. NCSU's Institute of Advanced Analytics performs a survey of data science Master's programs each year and categorizes them into four types: Master of Science (MS) in Analytics (MSA), MS in Data Science (MSDS), MS in Business Analytics (MSBA), and other MS programs, tracks and concentrations. The following figure demonstrates the exponential growth of these programs at NCSU alone. (North Carolina State University Institute for Advanced Analytics, "Degree Programs in Analytics and Data Science," Accessed May 14, 2016, http://analytics.ncsu.edu/?page_id=4184.)

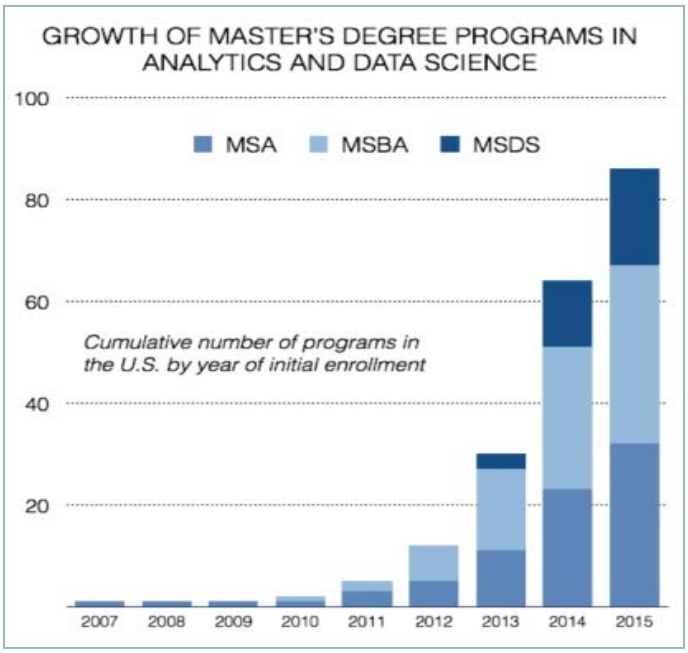

Figure 16. Graph. Growth of Master's degree programs in Analytics and Data Science.
(Source: North Carolina State University Institute for Advanced Analytics, 2016.)

As academia adapts to supply fresh data science talent, **organizations will be capable of promoting a more robust data science program**. But academia is not the only pipeline TSM&O should look to tap when building up data analysis talent, as it could take decades for supply to catch up with demand in this field. Partnerships, internal recruitment and training, and outsourcing are go-to strategies for building a solid data science team.

Datafication

"***Datafication***" is the buzzword used to describe how our daily life's activities are being turned into computerized data and this is particularly true for the Emerging Data Sources for TSM&O. With more data available, the transportation network's operations and health can be further quantified, monitored, and optimized. Behavioral patterns that were previously unusable are now being turned into data that can improve processes and operations, including the transportation industry. Subsequent reports will explore further how integrated transportation networks utilizing smartphones, sensors, and connected vehicles will improve TSM&O operations.

Predictive Analytics Driving Efficiency

Predictive models exploit patterns found in transactional, operational, maintenance, usage, historical, and other types of data to identify risks and make predictions about the future. For example, car insurance companies use predictive analytics to calculate the likelihood that a policy holder will experience or cause a collision, as well as the extent to which the insurance company would be liable. Similarly, predictive models are commonly applied to determine future traffic patterns due to new land uses and crash rates of new infrastructure elements. As more data becomes available from the emerging data sources, TSM&O organizations can likely predict a variety of safety and mobility-related metrics with better accuracy and refine TSM&O processes accordingly. These use cases will be discussed further in subsequent reports.

Machine Learning and Cognitive Analytics

As discussed in a previous section, machine learning is a scientific discipline combining computer science and statistics to use algorithms that make predictions about future data points based on what algorithms have "learned" from an historical dataset. These algorithms are powerful because they **enable computers to perform actions that weren't explicitly programmed**, which in turn allows data scientists to uncover insights that would have been lost otherwise. Some popular techniques include **Neural Networks** and **Nonparametric Statistics**. While adaptive traffic control systems are in widespread use today, most if not all do not learn from their past actions. Using machine learning algorithms, traffic signal timing parameters could be updated automatically based on the abundance of information available from emerging data sources based on current traffic conditions, typical peak traffic hours, crashes reported, etc.

As a reminder, **cognitive analytics** uses natural language processing, machine learning, and analytical techniques to simulate the abilities of the human brain. Cognitive analytics has the potential to be a valuable addition in TSM&O. Transportation networks are expanding to include even more data for analysis from connected vehicles, Travelers, and Infrastructure, much of it unstructured and semistructured data. Cognitive analytics can be used to sift through this data and find connections otherwise lost using traditional analytical techniques. For example, the Tennessee Highway Patrol (THP) implemented cognitive analytics to combat traffic accidents and fatalities by building a "real-time probability heat map that suggests where incidents will likely occur." The Tennessee Integrated Traffic

Analysis Network (TITAN) has shifted the THP mindset "from 'patrol and respond' to 'anticipation and prevention." (Huddleston, Greg, "Cognitive Analytics Is Helping To Reduce Roadway Fatalities in Tennessee," Accessed June 27, 2016, http://www.forbes.com/sites/ibm/2016/04/28/cognitive-analytics-is-helping-to-reduce-roadway-fatalities-in-tennessee/#204a37673dad.)

By employing cognitive analytics to the emerging data sources, decisions based on scenarios, such as determining the traffic impacts of proposed development and construction projects, can be made initially by an algorithm; then augmented/approved/overturned by a TSM&O decisionmaker allowing one decisionmaker to become significantly more productive. Additionally, the machine learning component of cognitive analytics allows the augment/approve/overturn interaction of seasoned decisionmakers to train the algorithm, imbedding their experience into the system. This is a powerful way to create systemic knowledge transfer that is not restrained by physical location. The experience of the best decisionmakers can be leveraged by new employees to immediately improve their performance. Furthermore, cognitive analytics with natural language processing can be leveraged to understand and incorporate public sentiment through analysis of social media and other sources of public opinion into the design and construction of a new infrastructure, such as a metro line. (Holecy, Miro, "Cognitive Computing Can Help to Meet Citizens' Expectations from Transportation Services," Accessed April 7, 2016, https://www.ibm.com/blogs/insights-on-business/government/cognitive-computing-transportation/.) **Cognitive Analytics Technologies, like IBM Watson, are still on the Gartner Hype Cycle**.

Real-Time Analytics and the Internet of Things

IoT has created an explosion of data that is accelerating opportunities to use real-time, streaming data. Social media platforms, connected vehicles, roadside units (RSUs), GPS applications, and other sensors and devices are streaming constant information that can be gathered and analyzed to make real-time decisions. For example, users are becoming active participants in the dissemination of traffic data by using Waze to report crashes, traffic congestion, police activity, gas prices, and other hazards and information for drivers. This data is then immediately available for other drivers to adapt their routes on-the-fly and avoid potential slowdowns. More details for both real-time analytics and IoT will be provided in subsequent sections.

Geospatial Analysis

Geospatial analysis is the gathering, display, and manipulation of imagery, GPS, satellite photography, and other data. The data is rendered in terms of geographic coordinates or street addresses and postal codes. Geospatial data enables an organization to supplement traditional data with time and location, as well as spatial and surface analyses, and plays particular importance for TSM&O. The emerging data sources analyzed in this project will enable new forms of network analyses of traffic patterns, origin-destination matrix synthesis (vehicles, pedestrian, etc.), construction project prioritization, and other use cases. Geospatial analysis using big data will be a key component of the content in subsequent reports.

Chapter 6 Leading Commercial Practices and Tools

This chapter briefly describes the leading commercial practices and tools used to aggregate, store and use large amounts of data. The chapter is organized according to the steps of the big data process model described in chapter 5 through the topics of data acquisition, marshalling, and analysis. After reading this chapter, the reader will have a better understanding of some of the key leading practices and commercial tools for handling extremely large data sets. These tools and practices are not common today in Transportation Systems Management and Operations (TSM&O); so it is important to understand the general landscape before discussing the application of certain tools and technologies. Chapter 6 then provides some cost and capability comparisons of commercially available tools that are expressly designed for handling large data sets.

Chapter 6 Objectives

- Introduce leading commercial practices and tools in each of the process model steps.
- Introduce key terminology, product names, and capabilities of tools prevalent in the industry.

A common organizational theme of the following sections is to first discuss leading and/or emerging practices in this landscape and then discuss a few of the leading tools and technologies used to accomplish these practices. Any examples provided are only intended to be representative; and are not considered recommendations. Additionally, the information contained below, should be considered from the perspective of a snapshot in time (spring 2016) and is likely to change rapidly. Every effort has been made to provide the most leading edge information to ensure maximum continued relevancy.

Data Acquisition

Leading Practices

In 2012, it was estimated that over 400 million tweets were sent on a typical day. (Farber, Dan, “Twitter hits 400 million tweets per day, mostly mobile,” Accessed May 9, 2016, https://www.cnet.com/news/twitter-hits-400-million-tweets-per-day-mostly-mobile/.) In terms of data analysis, special tools are required to effectively analyze data produced in this quantity and velocity. An API, or Application Programming Interface, is the term used to refer to the interface between one software system and another for the purpose of exchanging data. For example, Twitter has access to a massive amount of data that many TSM&O agencies may be interested in harnessing, and they have developed several APIs to allow the public to access their data, including:

- **Firehose APIs:** deliver data as it becomes available, but without limitations on the number of searches—the only way to see 100 percent of the data in real time. Using a Firehose API is like fishing with a net stretched across the river: 100 percent of the fish will be caught.

- **Streaming APIs:** use a persistent connection (i.e., one that doesn't close) to deliver data immediately as it becomes available. This can be compared to fishing with a pole in a river. There is a limit to the number of fish that can be caught.
- **Search APIs:** use a temporary connection to search for matching criteria within a pool of existing data at that given time. Using a Search API is like picking out a certain fish for dinner, based on a set of criteria (price, type of fish, weight, color, etc.) and choices are limited to the current selection of fish in the market.
- **RESTful APIs:** use a temporary connection to a Web service to initiate a single request for data. Using this type of API is like calling a marketplace vendor to ask if a certain type of fish is available or in stock.

A software system is typically described as **exposing** an API for other systems to connect to. The consideration of the types of APIs for the different emerging data sources is a key consideration for the use cases for TSM&O applications to be developed further in subsequent reports.

Emerging Practices

IoT is proliferating across all industries and market sectors enabled by the Internet and Internet Protocol (IP) communications. Very similar to the way that ruggedized Ethernet switches have revolutionized the communication networks of DOT to existing and new infrastructure over the last 15 years, the general trend of IoT technologies and platforms for use by a variety of industries may enable TSM&O organizations to acquire data from connected travelers, vehicles, and infrastructure at lower cost and easier integration than "rolling our own" in the next 10 years. The following components make up the fundamental infrastructure required to support IoT:

- **Edge devices:** the "things" including wearable technology, vehicles, industrial equipment, phones, and anything that can be connected to a network, fitted with sensors, actuators, or embedded computers.
- **Gateways:** close the gap between devices in the field and the cloud, where data is collected, stored, and manipulated by business applications. Gateways will use secure connectivity solutions via Dedicated Short-Range Communications (DSRC), cellular, Wi-Fi, Bluetooth, and ZigBee networks.
- **Cloud platforms:** provide APIs and tools that enable developers to build real-time IoT applications to connect devices with the cloud.

The use of these components for data collection has been happening for generations, in particular at Departments of Transportation (DOT) for connecting infrastructure for TMS&O. Existing systems use specific protocols and data collection architectures, and have limited scalability to connected vehicles and travelers as they currently exist. The massive global push for standardized IoT platforms is incredibly well understood at this stage, but is an important trend that may reduce cost and complexity to DOTs in implementing solutions for TSM&O applications since the need for such technologies extends far wider into industrial and business needs. How IoT platforms may be leveraged to seamlessly connect existing infrastructure and connected devices, how to integrate or combine with a big data solution, and other questions will all need to be answered in the coming years. Several new platforms for IoT are discussed in the next chapter.

Data Acquisition Tools

As with all sections about technology, the tools and technologies discussed are only examples that are intended to represent the current landscape rather than a recommendation or suggested approach. Additionally, the information provided is considered highly susceptible to rapid innovation and changes and is only provided as a "snapshot" in time of many of these tools.

As mentioned previously, Hadoop is seen as an open-source market leader, but frequently it also is the only option to be found for specific tasks. For example, getting large-scale data into a distributed environment has led to several Apache products developed for Hadoop to make data ingestion faster and easier. While each of these tools were developed for the Hadoop platform, some may be capable of working with other solutions (e.g., massively parallel processing (MPP) databases) with minor modifications (e.g., Sqoop and Storm).

1. **Apache Flume:** is a distributed and highly available tool for collecting, aggregating, and moving large amounts of fast moving data (e.g., transactional log data or a Twitter feed), into a central repository. It is tightly integrated with the Hadoop ecosystem and is often used to get data into Hadoop Distributed File System (HDFS).
2. **Apache Kafka:** can handle real-time feeds and is considered a general purpose tool not necessarily designed for Hadoop. Hadoop is just one of the possible systems capable of receiving data streamed through Kafka.
3. **Apache Sqoop:** allows for efficient bulk transfer between Hadoop and a more traditional structured data store such as a relational database. This characteristic makes Sqoop particularly important for organizations looking to update and/or transfer their traditional relational databases.

A few other important data acquisition tools in the Hadoop ecosystem include Apache Storm and Apache Chukwa. Storm is a real time event processing framework in the Hadoop ecosystem. Like Spark (which will be discussed more in subsequent sections) it is the underlying layer for many other tools to distribute jobs and consolidate their results in a Hadoop cluster. Finally, Chukwa is a data collection system used to monitor large distributed systems by collecting system metrics and log files.

Data Marshalling

Leading Practices

The automated steps needed to clean, organize, store, and manage data become vitally important when too much data is acquired in a given timeframe to perform these tasks manually. Often, these processes are more time consuming in central processing units (CPU) cycles than the queries and reports that run on the data during the analysis phase. These steps are referred to as data marshalling, and due to the challenges associated with the veracity, velocity, volume, and variety of big data, new considerations for traditional data marshalling practices are becoming necessary. For example, when a problem of volume comes up, which it frequently would in a big data solution (e.g., scalability, replication, archival, backups, etc.), one general solution is to use inexpensive, commodity hardware to offset the cost of excess storage. The next few sections will discuss many leading practices in data marshalling beyond the use of commodity hardware and how each has had to adapt

to for scalability, compression and archival, fault tolerance and high availability, replication, security, and backup and disaster recovery.

Scalability

Scalability refers to a system's ability to increase and/or decrease storage resources as needed. Based on the predictions mentioned earlier, the connected city of 2021 is forecasted to grow 400 percent in connected travelers, 300 percent in connected vehicles, and 25 percent in connected infrastructure. As the connected city continues to grow, the amount of data stored within the system will need to grow from terabytes to petabytes in 10 years—explaining the importance of scalability.

Drastic increases in volume and velocity from emerging data sources cannot be ignored, yet frequently, organizations do not know (and perhaps cannot know) the peak scale of their data. Being able to scale quickly and affordably is a must. To that end, many solutions recommend using low-cost storage systems that allow them to achieve quick and affordable scalability. For instance, Hadoop and some MPP databases recommend the use of commodity hardware for storage.

Scalability is an important factor for many of the subsequent practices discussed. The sheer volume that emerging data sources, such as connected vehicle data, bring introduces the most obvious layer of complexity. From there, the other four Vs of big data introduce new complexities and inefficiencies to the way that data needs to be handled.

Compression/Archival

> Often, data marshalling processes are more time consuming in CPU cycles than the queries and reports that run on the data during the analysis phase.

For example, one method of managing the size of data being accumulated (particularly when considering the volume of historical data that may not be needed often, but either can't or shouldn't be destroyed) is compression and/or archival.

Compression and archival refer to the ability of a system to reduce data storage needs by decreasing the resources required and consolidating them for long-term storage, respectively. For example, as old information becomes outdated or isn't used frequently, the cost of storing the information may outweigh its value. This will particularly be true for "raw" BSM data if the TSM&O organization is storing this natively since more than 1.3PB of the 2.1PB total storage needed for a typical agency in 2021 will be Basic Safety Messages (BSM). However, both data compression and archival increase latency when the data is needed again (through methods of decompressing or retrieving data), making the decision of what to compress and/or archive a strategic business decision. However, as always, big data is complicated by the other four Vs, and the increased speed with which data is brought into organizations makes fast and accurate business decisions a necessity.

The ability to provide fast and efficient data compression options to an organization are key. Hadoop, for example, inherently provides multiple methods for compressing/decompressing data at varying speeds and efficiencies. Additionally, many solutions exist to provide the massive storage required to eliminate the need for archival procedures.

Fault Tolerance/High Availability

Planned downtime activities for a typical system can include kernel switch, hardware maintenance, and operating system maintenance. Unplanned downtime can occur due to unforeseen issues which lead to application nonavailability.

The combination of fault tolerance and high availability make up the ability of the system to remain operational in the case of system failure (e.g., failure of a single node). The two concepts are very closely related, and frequently confused; however, there are slight differences. A **highly available system provides methods (typically involving software and hardware) for maximizing the time in operation for the system** (e.g., through shared services and resources), **and a fault tolerant system provides methods (typically purely hardware) of continuous service** (e.g., through failover nodes). For instance, if a traffic signal's power source is lost, a highly available system would immediately dispatch maintenance, in the meantime, however, a fault tolerant system would have a battery backup system failover in place to immediately replace the power loss without system downtime. Performance may be affected in the fault tolerant scenario (perhaps the signal can only perform in flash mode), but it ensures that the system will remain active rather than having an intersection with dark signal heads. The two concepts are not mutually exclusive and are frequently considered together to provide maximum availability of the system.

The more complex a system becomes, such as an increase in the number of nodes or distributed organization of the data, the more difficult to manage the solutions becomes. As a result, there are critical considerations that organizations must understand when designing their solution, including the **cost of downtime**, the **recovery time objective** (the maximum time an information technology (IT) process can be down before suffering unacceptable consequences), and the **recovery point objective** (the maximum amount of data an IT process can afford to lose while down before suffering unacceptable consequences).

Luckily, because most people need to sleep and most do so at night, there is little travel activity during the night. **Planned downtime** of tools and systems for TSM&O can be managed during periods of low travel activity with minimal impact. **Unplanned downtime** of tools and technologies during the day impacts operations only to the point that the systems are used for mission-critical use cases. These issues will be analyzed further during subsequent reports.

Replication

Replication is the system's ability to copy data to provide fault tolerance of system processes. The major advantages of replication are to:

- Improve data availability and scalability.
- Provide a failsafe backup.
- Provide load balancing.

These advantages occur because replication of data provides redundancy and distribution of the data so that computations and other processes can occur in duplication and across resources. For a solution built to handle the data volumes projected for the next decade, the replication of data becomes significantly larger and more complex to manage, and understanding exactly where the data

is (and all copies of it) becomes a more challenging task. The solutions discussed in this report rely on a master node to track all data stored on the system.

Security

Security provides the ability of a system to meet legal and regulatory compliance standards through user access controls, data encryption, and so on. According to CNN Money, 47 percent of U.S. adults had their personal information exposed by hackers in 2014. (Pagliery, Jose, "Half of American adults hacked this year," Accessed May 14, 2016, http://money.cnn.com/2014/05/28/technology/security/hack-data-breach/.) With major security breaches and fraud incidents making international headlines, organizations are taking steps to address the growing problems of advanced persistent threats, fraud, and insider attacks.

Big data requires the same security standards as traditional Relational Database Management System (RDBMS) and networks, but introduces several problems at scale. These problems include challenges such as diverse data sources and formats, speed at which it's being received, granular access controls, and others. Security is constantly being refined to meet the standards of the Federal Government. The main approach suggested for TSM&O organizations relative to Personally Identifiable Information (PII) and data security is not to store any PII at all, such as is being baked in to the public connected vehicle system.

When determining security solutions, organizations will need to determine the security requirements of their specific Agencies. For examples, if the recommended approach of not ingesting PII data is followed, the system will have lower security standards to meet. Otherwise, Agencies will need to explore and implement the latest security protocols and standards (e.g., attribute-based encryption). In addition, organizations should establish comprehensive controls for configuration and management of a multinode environment which may far exceed the security provisions in a typical agency DOT environment today. This will be explored further in subsequent reports.

Backup/Disaster Recovery

Many agencies mistakenly believe that routine backup operations will have them covered in the event of an outage or a disaster; however, **data backup and disaster recovery are not the same**.

Data backup refers to the ability of the system to preserve all necessary content by systematically storing data, while data recovery refers to the process of how data will be recovered in the case of serious failures. **Disaster recovery** is intended for the most extreme types of system failure, such as an earthquake, flood, fire, and requires geographically distant data centers and disaster recovery centers to prevent large data clusters and their back-ups from suffering effects of one large disaster. Many and most TSM&O organizations currently implement backup strategies; far fewer consider disaster recovery strategies.

Both of these safeguards present a serious problem when the size of the solution is considered. Each solution will likely have their recommended approach; however, Hadoop's inherently fault tolerant design and data replication process make standard backups unnecessary (although Agencies will have standards they need to follow or get approved through appropriate channels). The need for disaster recovery of emerging data source storage is an open question and would depend on the how the information is needed for mission-critical use cases. This will be explored further in subsequent reports.

Data Marshalling Tools

As noted in the sections above, there are many tools that can be used for data marshalling. Several tools and technologies used to store, manage and prepare data that will be discussed in subsequent sections include **Massively Parallel Processing (MPP) databases**, **NoSQL databases**, and the **Hadoop ecosystem**. This section will briefly explain these tools and technologies.

As with all sections about big data tools, the tools and technologies discussed are only examples that we feel represent the current landscape and do not represent a recommendation or suggested approach. Additionally, the information provided is considered highly susceptible to rapid innovation and changes and is only provided as a "snapshot" in time as of spring 2016.

Massively Parallel Processing Databases

MPP databases use a "shared-nothing" architecture, where neither memory nor disk storage is shared among processors, to isolated resources across independent compute nodes within a system eliminating single points of failure. (Stonebraker, Michael, "The Case for Shared Nothing," Accessed June 27, 2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.5370&rep=rep1&type=pdf.) Each node is linked to a master node (or multiple masters) that manages data loading, storage, preparation, and processing resources across each compute node. This architecture allows jobs to be split into smaller processes to be executed across the system concurrently, which is referred to as **parallel processing**. MPP databases are extremely scalable due to the simplicity of adding additional identical compute nodes (or master nodes) to provide more resources. The following figure demonstrates what an MPP database might look like and how it can easily scale horizontally:

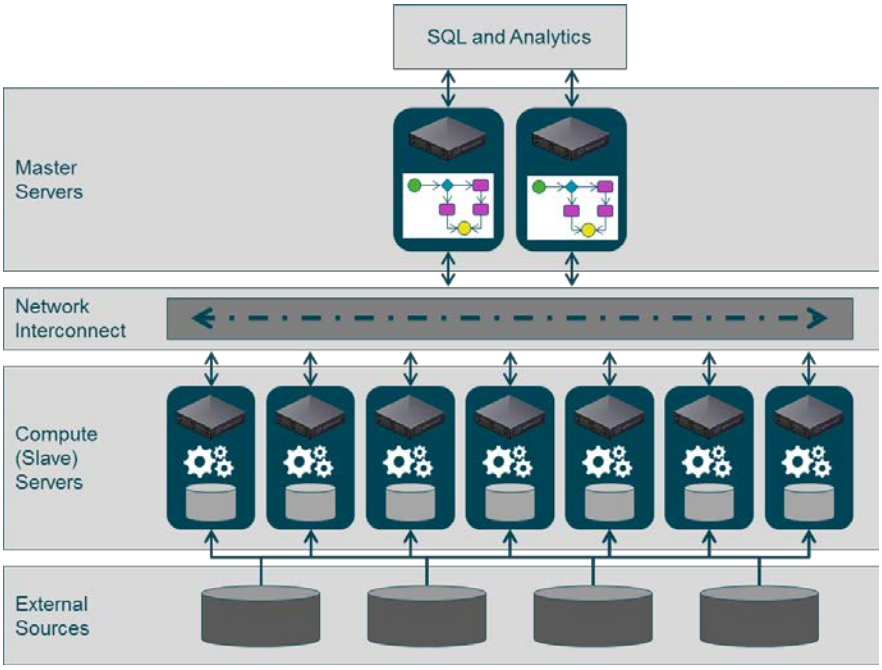

Figure 17. Illustration. Massively parallel processing shared-nothing architecture.
(Source: Adapted directly from an EMC Corporation InFocus article, "The Data Warehouse Modernization Act," 2016.)

Several examples of MPP databases include Teradata, Greenplum, Netezza, and Exadata and are very briefly described here. (DB-Engines, "System Properties Comparison Netezza versus Oracle versus Teradata," Accessed May 12, 2016, http://db-engines.com/en/system/Netezza%3Boracle%3Bteradata.) Additional details about the costs and capabilities will be discussed in subsequent sections.

- **Teradata** provides a market-leading MPP database that relies on a tightly integrated commercial-off-the-shelf (COTS) hardware and software solution. Teradata is typically considered the most expensive option due in part to its high support costs, required software and hardware package, and significant maintenance efforts.
- **Greenplum** is an MPP database developed by Pivotal that is recommended to be implemented on commodity hardware for significant cost savings. However, Greenplum only provides software licenses and support; hardware is not an intrinsic part of the solution.
- **PureData** is an IBM developed MPP solution that is much more recent to the market and lags in adaptability (e.g., programming languages and operating systems (OS) supported), but may provide cost savings to make up for this.
- **Exadata** is Oracle's MPP database on the market, and provides significantly more adaptability (e.g., programming languages and OSs supported) than other solutions, has been in the market almost as long as Teradata, and has a significant market share.

For all their differences, each of the MPP databases discussed offers a market-tested solution with what appear to be converging cost models. However, they will each also have similar limitations, including market trends and unstructured data capabilities. **MPP databases are Relational Database Management Systems (RDBMSs) and do not ingest, store, and process unstructured data and NoSQL (discussed in the following section) tasks very efficiently**. Part of the reason for this ties into the first limitation, which is that MPPs are established technologies (and have many benefits because of this). Hadoop and many of the tools it enables are forward-looking technologies. Much of this report is intended to focus on the trends of big data, hence the focus on Hadoop.

NoSQL (Not Only SQL) Databases

Another tool worth mentioning for data marshalling is the use of **Not Only SQL (NoSQL) databases**. NoSQL databases are large, scalable databases that organize data in different ways to use unstructured, semi-structured, and complex data together. One of the methods by which NoSQL improves performance above traditional SQL-based RDBMS is to sacrifice select aspects of ACID (Atomicity, Consistency, Isolation, and Durability) processing. This introduces potential losses of data in the case of failure, but saves time and resources during transactions and uses mitigating strategies (e.g., data replication techniques) to minimize data loss.

Examples of NoSQL databases include the following:

- **Key Value Store:** database records are stored and retrieved using a key that uniquely identifies the record. It is commonly used for large volume and high velocity transactional applications, such as Amazon. Examples include Redis, and Dynamo.
- **Document Store:** is a subclass of Key Value Store, where metadata (data about data; e.g., column headers) is extracted for further optimization. Frequently used for applications that

use text-heavy, digital context, such as Kindle. Metadata is the most heavily accessed data and needs low response time. One example includes MongoDB.

- **Column Store:** stores data in primary keys by columns to significantly speed up search functions. This is used for Customer Relationship Management (CRM) systems, library catalogs, and other ad hoc inquiry systems. One example includes Cassandra.

- **Graph Store:** stores data in "nodes" (single data points) and "edges" (relationships between nodes). This is used in relationship-heavy applications, such as social networking Web sites or a telecommunications provider's networks. One example includes Neo4J.

A NoSQL database is generally not a standalone product that will meet all of the needs of TSM&O applications, and they will not be discussed independently during subsequent sections analyzing cost and capabilities. Traditional SQL databases can't handle the unstructured, semistructured, and complex data and rely heavily on ACID processes. NoSQL provides options for organizations to break out of those constraints and are an important consideration for agencies that must handle significant amounts of unstructured, semistructured, and otherwise complex data types. Several of the examples provided are actually Apache research projects that can be implemented as a component of the Hadoop ecosystem. For TSM&O use cases of the emerging data sources, the graph store concept may hold significant promise. This may be discussed further in subsequent reports.

The Hadoop Ecosystem

Hadoop is an open source, enterprise big data file system designed to provide reliable and scalable distributed storage and computing. It is referred to as an "ecosystem" because it is built on a distributed file system called the Hadoop Distributed File System (HDFS) with a computing framework called MapReduce and customized through the use of add-on components as shown in figure 18. **Hadoop is a generic processing framework designed to execute queries and other batch read operations against massive datasets that can scale from tens of terabytes to petabytes in size**.

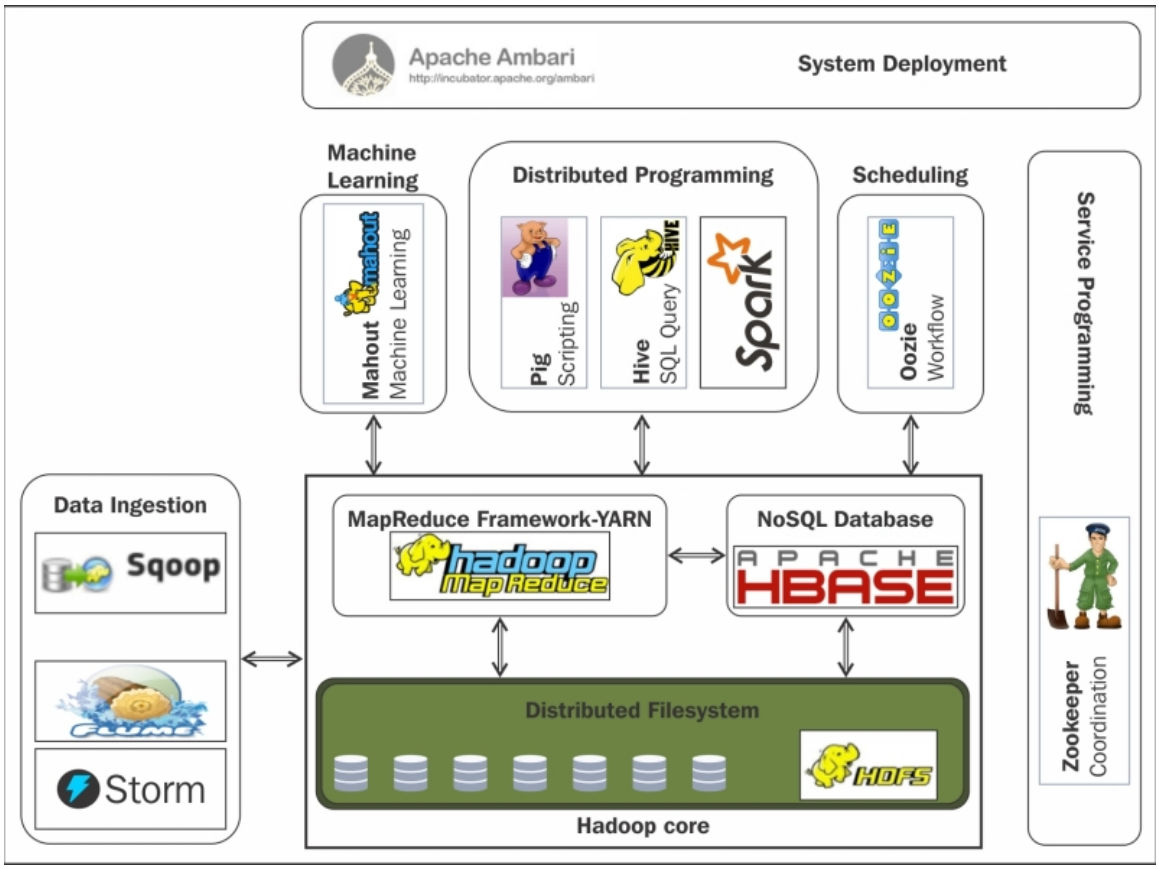

Figure 18. Illustration. A typical Apache Hadoop ecosystem.
(Source: *Hadoop Essentials* by Swizec Teller, 2015.)

Originally developed and published by Google, the distributed file system concept grew into the open source big data solution that it is today. Hadoop's popularity continues to grow because it continues to meet the needs of many organizations for flexible data storage and analysis capabilities with the goal of also maintaining costs.

Hadoop has been particularly useful in environments where massive server farms (large collections of servers) are used to collect data from a variety of sources. Hadoop is able to process parallel queries as big, background batch jobs on the same server farm. This saves the user from having to acquire additional hardware for a traditional database system to process the data, assuming such a traditional system can even scale to the required size. Hadoop also reduces the effort and time required to load data into another system, allowing the user to process it directly within Hadoop. This overhead would become impractical with very large datasets.

Several vendors have taken the freely available, Apache open source code and developed their own distributions (some with proprietary add-ons and adaptations). Several vendors that provide Hadoop-bases platforms include Cloudera, Hortonworks, Map R, Greenplum, IBM, and Amazon; however, Cloudera, Hortonworks, and MapR tend to lead the pack as of spring 2016 and will be discussed further in subsequent sections.

- **Cloudera** is the vendor that has been in the market longest and historically holds the biggest market share. Cloudera provides a few proprietary tools to add capabilities of the cluster, including a cluster management tool and an in-memory query engine; however, they also don't support certain Apache projects in pursuit of others (e.g., Cloudera supports Spark instead of Storm). They provide a free version with several limitations (e.g., not all proprietary capabilities are enabled) as well as an "enterprise" version that unlocks all capabilities, and includes support.
- **Hortonworks** provides 100 percent open source products and contributes everything that it develops back to the open source community. Hortonworks is capable of quickly adapting and growing to new Apache releases due to its lack of incompatibility with proprietary work. This may be particularly important for Agencies that are interested in using leading edge tools (e.g., geospatial tools).
- **MapR** provides the most proprietary product, including the file system that it's built on (MapRFS). Its ability to upgrade with Apache releases is much more limited; however, it has a reputation as the fastest and most efficient Hadoop platform. Also provides a free version as well as its "enterprise" edition.

Data Analysis

Leading Practices

Descriptive Analytics

The bread and butter of data analysis using basic statistics such as sums, averages, and percent changes is called **descriptive statistics**. Descriptive statistics are often the first step on the journey of data discovery, but can only give limited insights. What's more, descriptive statistics can be deceptive when applied without regard for the shape (or distribution) of the data being described. For example, averages are a widely used descriptive statics that can be misleading to use in the case of outliers. (Paret, Michelle, "Using the mean in Data Analysis: It's Not Always a Slam-Dunk," Accessed April 5, 2016, http://blog.minitab.com/blog/michelle-paret/using-the-mean-its-not-always-a-slam-dunk.) To avoid making decisions on improperly applied summary statistics, a thorough exploration, and understanding of the data is necessary.

A massive volume of data can affect an analyst's ability to thoroughly explore and understand their data particularly when it's rapidly changing and perhaps from questionable sources (the crux of the big data problem). These problems are equally as true at the introductory level of statistics as they are for more advanced analyses, in part because these descriptive analytical techniques are the foundation of all analytics to come.

Querying and Reporting

Querying and reporting are basic functions of virtually any analytically oriented group or function across an enterprise. Querying typically seeks to answer a specific question and is accomplished by isolating a subset of data based on the query criteria (i.e., licensed drivers, aged 35 to 50, owning a midsized sedan that is more than 5 years old). Reporting typically encompasses summary statistics

and aggregations of historical data and is thereby limited to gaining insights on the past with a blind spot to the future.

While there exist many proprietary, commercial reporting tools in the this vendor space, statistical programming languages such as R and Python have developed their own open source reporting tools and packages that allow for reproducibility (in other words, accuracy of the analytical model due to its ability to be reproduced) and streamlining of basic reporting. These tools and many, many others (commercial and open source, both) are rapidly being added as important aspects of a big data ecosystem with an eye towards analytical flexibility (e.g., the in-memory engine Spark is providing the ability to write Spark code in R, Python, and Scala with more languages to be supported in the future).

Interactive Visualization

Interactive visualization includes real-time searching for patterns to discover new information For example, the image below shows a Google Maps screenshot of the aftermath of gridlock in Texas after a serious collision of two semitrailers. Visualizations open up a new way for people, from experienced data analysts to inexperienced laymen, to understand the data and draw meaning.

Figure 19. Google maps image of Texas gridlock.
(Source: http://www.cnbc.com/2015/02/27/15-cars-2-semis-gridlock-us-75-in-texas.html, 2015.)

The advent and proliferation of powerful statistical and analytical tools that can be applied to the big data ecosystem has given rise to the need for visualization tools. The increasing complexity and sophistication of statistical data analysis methods require equally powerful methodologies to visualize and interpret output results, especially for TSM&O use cases that are highly geographic in nature. In addition, due to the increasing size of data, “dense graphical representations are more effective for exploration than spreadsheets and charts. Furthermore, because of the exploratory nature of the analysis, it must be possible for the analysts to change visualizations rapidly as they pursue a cycle involving first hypothesis and then experimentation.” (Polaris Web site)

Emerging Practices

Each type of analysis discussed above (descriptive, querying and reporting, and visualizations) can be considered the foundational building blocks (in the order discussed) that are explicitly necessary to enable the use of **advanced analytics**, which can be used as a blanket term for the latest emerging practices in data analysis. The following sections briefly describe several of the most popular emerging practices in analysis.

Forecasting

Forecasting is a hallmark of advanced analytics because it is the first opportunity to use data in terms of the future. All previous skills discussed are only capable of providing insights into past and occasionally real-time experiences. Forecasting analytics generally exploit properties of data measured over time to provide these insights. A sequence of numbers measured over a continuous time interval is referred to as a time-series; this can include data like Road Weather Information System (RWIS) and traffic flows.

Time series data is an excellent example of data that will likely be streaming in real-time and has the ability to grow rapidly with the expected increase in connected vehicles, travelers, and infrastructure. If every connected traveler and vehicle is providing real-time data (e.g., speed, route being taken), significant and accurate predictions of future actions may be able to be realized (e.g., expected traffic delays, predicted route improvements, etc.).

Regression

Regression analysis is one of the more commonly used analyses that may not generally be considered an "emerging" practice; however, it is an important tool in any analytics process. Regression is used to understand the correlation between a dependent variable and one or many independent variables. For example, regression could be used to understand the relationship between gas prices and crime rates on the number of commuters using metro systems in large metropolitan areas. In this example, the price of gas and the city crime rate are the predictors, or independent variables, and the number of commuters is the dependent variable. The effect (i.e., increase or decrease) on the number of commuters when gas prices rise and fall and when crime rises and falls can be estimated, or inferred, by regression. Specific inferences are possible in regression because of mathematical assumptions made in regression analysis; understanding underlying assumptions of regression are important to be able to make sound inferences. As the volume of data used in a regression increases, the model is able to converge on more precise estimations of true parameter of interest. As the variety of data available to consider in an analysis increases, so do the options data scientists have to meet mathematical assumptions needed to build a useful regression model.

Each of the most popular statistical programming languages (R, Python, SAS, etc.) will have their own regression models/tools that will require less technical coding (for example, instead of using MapReduce directly); however, they will require at least minimal statistical programming skills.

Optimization

Optimization uses mathematical models to find the best course of action according to some objective function, in a situation based on constraints and possible alternatives. While in simple situations, people do this intuitively; however, there are many complex systems where the most beneficial set of

control actions or Demand Management actions are nonobvious. For example, determining ramp metering rates in a freeway corridor or traffic signal timings on an arterial. The Emerging Data Sources from connected travelers and vehicles can enhance the abilities of TSM&O organizations to provide better service to the public. **Python**, **R**, and **SAS** (a nonopen source statistical programming language) each provide several advanced optimization methods such as linear and nonlinear programming, genetic algorithms, and other search methods.

Machine Learning

Machine learning is the science of enabling computers to act without being explicitly programmed. Examples include self-driving cars and effective Web searching. Machine learning combines aspects of computer science and statistics to provide powerful analysis of data. Specifically, machine learning concerns the construction of algorithms which learn from past data to make predictions about future data points.

Clustering is one of the most intuitive and fundamental machine learning algorithms in use today across a wide range of domains. Clustering is about finding similar subgroups of data points in a dataset. For example, a clustering of traffic incidents at a specific intersection or road segment may be affected by specific variables, such as visibility, inclement weather, pavement conditions, etc. Additionally, clustering could be used for targeted marketing campaigns aimed at increasing public transportation use.

Network Analysis

Network analysis is used to visualize complex networks with graphical tools. It is a powerful way to mathematically represent complex systems across many different domains and inherently the configuration of transportation systems.

Today, there is a surge of interest in visualizing networks for analysis purposes. The advent of statistical computing and the development of new techniques and models for their analysis and interpretation have accelerated interest as well. For example, network analysis could be used to assess transportation networks for pain points or vulnerabilities. These would be crucial insights that could be leveraged for future expansion of highways and bridges, etc. Emerging tools include the **NoSQL graph stores** discussed in a previous section. **Neo4j** is one of the most popular options with a significant market share and an open source version available for anyone to use.

Data Analysis Tools

A variety of tools and techniques are available today to analyze large volumes of data in batch, near real-time, and real-time speeds; however, many of these analytical tools are dependent on the statistical programming languages used. These tools can be nonstatistical languages like Java in addition to more traditional statistical languages like **R**, **Python**, and **SAS**. Many of the leading and emerging practices discussed above are simply models and methods that each language develops independently (e.g., optimization models and regression models). Hadoop and other solutions provide many ways to employ these methods as well as a few ways to employ them directly (e.g., MapReduce code). The following sections identify what we've referred to as the "tools" of data analysis; however, this is slightly misleading because these are not specific implementation recommendations, but rather generic methodologies that give consideration to what the business needs of the Agencies are. For example, SAS provides several options for enhancing the big data

ecosystem through in-memory, in-database, high-performance computing products. Each product provides tradeoffs between performance and statistical needs.

Disk-Based Analytics

Disk-based analytics is the traditional method used for reading, writing, and processing data where applications query the data stored directly on physical disks. A disk-based database reads and writes the data directly from the disk and brings the data to the code. When dealing with large amounts of data, the data movement between physical disks and memory can create latency issues. It is typically the slowest option available, but has been around longer and is more frequently used and understood by users.

MapReduce and **grid computing** are leading examples of nontraditional, disk-based analytics that use processing on distributed storage to improve performance.

In-Database Analytics

Compared to traditional disk-based, **in-database analytics** are generally a faster, more flexible and efficient way to process increasingly large data. They use a distributed architecture (e.g., MPP or Hadoop) to process large datasets in blocks across a cluster of servers; however, to reduce data movement and latency, the code is brought to the data. It takes significantly fewer resources to move the code to the data rather than the other way around allowing for a more scalable solution and faster processing speed.

Key products that utilize this model include SAS In-Database and SPSS In-Database; however, most languages will have some in-database options available. SAS has significantly adapted their standard (proprietary) language (which is a disadvantage in terms of user adoptability) to provide in-database tools that are critical for performing large merges and sorts (e.g., merging traffic and incident reports with map locations).

In-Memory Analytics

In-memory analytics is the fastest method of data analysis and is best suited for solving complex and time-sensitive business scenarios. The key advantage is the lift of data into memory to reduce data movement and increase performance speeds. With the cost of memory (and storage in general) always decreasing, in-memory analytics are becoming a more cost effective approach. However, even with a general decrease in the cost of memory, it is still significantly more expensive than disk storage, and in-memory analysis will require more expensive, memory-intensive machines because the entire dataset needs to be lifted into memory simultaneously.

Products like **Spark**, **SAP Hana**, **SAS LASR**, and SAS In-Memory Statistics are all prime example of in-memory analytic tools. Additionally, utilizing in-memory analytics provides the necessary speed to enable visual analytics and streaming analytics.

Visual analytics is the use of interactive visual interfaces for the purpose of analytical reasoning. This methodology can be seen as an approach combining visualization, data analysis, and human factors (cognition, perception, etc.). Using visual analytics, users may directly interact with data analysis capabilities to produce meaningful information and develop insight from dynamic, ambiguous, and

often conflicting data. **Zoomdata** and **Databricks** are both market analysis tools that allow customers to explore, analyze, and communicate data in a visual manner.

Streaming analytics is the ability to analyze data as soon as it arrives to predict an outcome; it is what enables real-time analysis and, perhaps more importantly, real-time action. One of the most common examples of streaming data is Twitter and other social media platforms. If an Agency was interested in understanding the current public sentiment as important news was being announced, data could be streamed in real-time from Twitter, Facebook, and other platforms based on keyword queries. Analyses could be performed using in-memory tools, and impact the news being provided.

Real-time insights can be gained from streaming analytics with connected cell phones, vehicles, infrastructure, and other devices. Many IoT scenarios will be extremely applicable for streaming analytics, such as the detection of nonrecurrent queues, bottlenecks, or changes in origin-destination patterns.

Apache Spark Streaming and **Apache Storm** are two leading examples on the market for this capability.

Several benchmarking and performance tests have been described here to demonstrate the leaders in streaming analytic capabilities. The results of benchmarking tests can be seen in the table below:

Table 8. Benchmarking tests comparing speed and performance of Apache products.

Benchmarking Test	Summary
Spark tied for the Daytona Graysort Competition	• Spark competed and tied in the Daytona Graysort Competition where they processed 100TB of data (1 trillion records) on disk with 206 EC2 machines in 23 minutes. • Previous record by Hadoop's MapReduce was 72 minutes using 2100 machines.
Comparison between Spark Streaming, Storm, and Flint	• Yahoo performed a benchmark comparison of three of the top streaming Structured Query Language platforms. • Found that Spark, while having higher latency, is capable of handling higher throughput.
Benchmark of SAS In-Memory for Hadoop and Revolution Analytics	• SAS performed a comparison of Revolution Analytics' RRE7 and their own In-Memory Statistics for Hadoop in response to previous benchmarks performed by Revolution. • SAS disputes the original Revolution benchmark, and shows faster performances.

Big Data Deployment Options

Given the many tools that could all combine and overlap to produce one comprehensive big data solution, several deployment options are available based on Agencies' big data capability maturity and their IT policy on data security and privacy.

The figure below shows the gradient of implementation options available, and each option is described briefly below the figure.

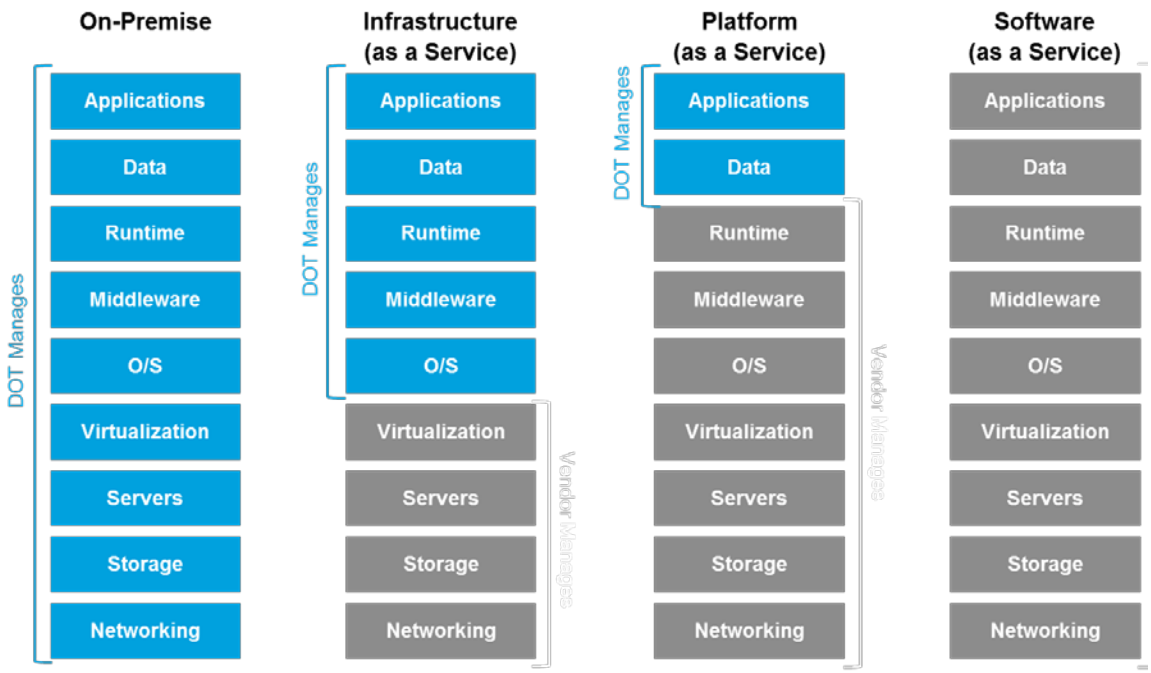

Figure 20. Diagram. Internet Technology considerations for on-premise, Infrastructure-as-a-Service, Platform-as-a-Service, and Software-as-a-Service implementations.
(Source: Deloitte, 2016.)

The "as a service" (aaS) models generally assume a cloud-based deployment (gray boxes) for the aspects of control an organization is willing to sacrifice for simplicity and possibly cost savings.

- **On-Premise:** Deployment offers users the ability to install, manage, and maintain every aspect of a big data deployment. Typical on-premise deployments require significant up-front costs (hardware, software licensing, etc.) but allow for greater control of the system.
- **Infrastructure-as-a-Service (IaaS):** Deployment provides scalability needs and minimizes responsibility for the DOT. Users are responsible for managing applications, data, runtime, middleware, and operating system. Instead of having to purchase hardware outright, users can purchase IaaS based on consumption, similar to electricity or other utility billing.
- **Platform-as-a-Service (PaaS):** Deployment allows users to develop, test, and deploy applications quickly and efficiently. With PaaS, users are only responsible for data and application tiers. Similar to IaaS, users can purchase PaaS on a subscription basis ultimately paying just for what they use.
- **Software-as-a-Service (SaaS):** Deployment uses the Web to deliver applications. Most SaaS applications can be easily accessed directly from a Web browser on the client's side. This model is maintained entirely by the vendor. Like the other service models, users purchase a subscription to access the application.

Chapter 7 Cost and Capabilities of Computational Platforms

Chapter 7 Objectives:

- Introduce gartner's magic quadrant for assessing commercial tools.
- Describe differences and similarities of cost models for commercial tools and systems.
- Describe capabilities and characteristics of commercial tools in MPP, Hadoop, and Internet of Things (IoT) categories.

This chapter addresses current computational platforms in the market and their relative costs and capabilities. In this chapter, the reader will gain increased awareness of the commercial tools and systems currently available in the marketplace for handling massive data sets. Terminology introduced in chapters 5 and 6 is not typically redefined in this chapter. After reading this chapter, the reader will have entry level understanding the similarities and differences of commercial providers' cost models. Because the cost models vary so widely it is challenging to put an apples-to-apples pricing comparison together of one versus another. Note also that the use of vendor names and specific technology descriptions are not recommendations of these tools and systems by U.S. Department of Transportation (DOT).

There are three broad categories of big data computational platforms discussed in this report. One type is the **massively parallel processing (MPP) database**. As a reminder, an MPP database processes large volumes of data via multiple node processors (connection points in a network), which segment the data into time efficient, manageable quantities. (http://searchnetworking.techtarget.com/definition/node.) In an MPP database, each node processor has its own operating system and memory resulting in a "shared-nothing" architecture. (http://whatis.techtarget.com/definition/MPP-massively-parallel-processing.) An MPP database quickly processes large volumes of data, but it cannot easily share data between the nodes, which makes dynamic analysis and communication difficult.

Hadoop platforms are the second category of big data platforms and are quickly becoming synonymous with big data. Hadoop is a "Java-based programming framework that supports the processing of large datasets in a distributed computing environment". (http://searchcloudcomputing.techtarget.com/definition/Hadoop.) The Hadoop distributed file system (HDFS) facilitates the rapid transfer of data across thousands of nodes and can be implemented in-house on commodity servers or as a cloud-based platform-as-a-service solution (both of which will be discussed as implementation options).

The third category of big data platforms can be summarized as a **cloud-based IoT platform**. Cloud-based IoT platforms facilitate the secure connection and management of devices (e.g., cars, phones, etc.) to support the acquisition, marshalling, and analysis of nontraditional and semi-structured data sources.

An aside worth noting before mentioning several solutions that Departments of Transportation can pursue, is the degree of change management and skills investment that is incredibly important to understand. While some solutions are sure to provide an easier transition or a less steep learning curve (e.g., Cloudera's user-friendly console, Microsoft's interoperability with other Microsoft products, etc.), each of these will present significant new hurdles for Information Technology (IT) staff, data analysts, and other personnel to manage and effectively use. However, these tools also are cutting edge and innovative solutions that may draw in new, excited staff and partners. Most platform providers will provide IT support (for a cost) and will have a wealth of free online resources as well as paid training sessions that can be attended in-person.

Gartner's Magic Quadrant

Gartner's **Magic Quadrant** is the industry gold standard for information technology market research and comparing vendor products. The quadrant provides a qualitative analysis into a market through its direction, maturity, and participants. Gartner's Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics is shown below in figure 21 and identifies the leaders, challengers, niche players, and visionaries in data warehousing and data management. Several key **vendors of both MPP databases** and **Hadoop distributions** are labeled in **purple** and **green**, respectively; however, each vendor is analyzed for its holistic abilities in this specific category. In other words, some vendors may provide additional data warehousing services that are being evaluated by this Magic Quadrant in addition to the specific capabilities discussed in this document (i.e., their rankings would be inflated) or some vendors might have additional capabilities relevant to this report not considered in Gartner's analysis (their rankings are undervalued). Gartner's analysis should only be used as a starting point to begin to understand the vendor; detailed understanding of the products they offer and the business requirements they can or can't meet will be more important. This section will focus on the leading vendor technologies in each previously described category of big data platforms and does not constitute a recommendation of these products.

Before comparing the capabilities and costs of these solutions, it is worth understanding how they are similar. Both pure big data solutions (MPP and Hadoop platforms) offer easily scalable, large volumes of data storage, distributed computing capabilities for increased performance and analytical capabilities, and a robust, fault tolerant system. Each of the IoT platforms will provide a cutting edge opportunity to connect to more devices and bring in more data than ever before in a potentially cost effective and efficient way. The vast majority of solutions discussed recommend the use of **inexpensive commodity hardware** (with the exception of perhaps Teradata) or **cloud-based implementations** in which the desired resources can be customized to fit an Agency's needs. **Open source is generally embraced with each vendor**, but flexibility for independent capabilities, tools, and languages is a minimum (e.g., Python and R code, NoSQL database options, etc.). Additionally, with a focus on open source and flexible solutions, enterprise-level support capabilities are an important differentiator for each vendor. While most offer analytical capabilities automatically, it is still important to have capabilities within the Transportation Systems Management and Operations (TSM&O) agency because outside big data vendors won't have the domain expertise in transportation operations required to provide the full benefit of the tools.

Figure 21. Graph. Gartner's magic quadrant for data warehouse and data management solutions for analytics.
(Source: Gartner, 2015.)

Engineered Massively Parallel Processing Platforms

In MPP databases, data is partitioned across multiple database servers (or nodes) with each node having separate memory and processors to process data locally. All communication is via a network interconnect. There is no disk-level sharing of data between the processors (i.e., the processors do not access information on the same hard drive; explained previously as the "shared-nothing architecture").

Teradata

Gartner ranked Teradata in the *Leaders Quadrant*. Teradata is designed to accommodate large data warehouse implementations (e.g., multi-PB solutions) for its customers, which include Apple, Walmart, and eBay. (Harris, Derrick, "Why Apple, eBay, and Walmart have some of the biggest data warehouses you've ever seen," Accessed May 14, 2016, https://gigaom.com/2013/03/27/why-apple-ebay-and-walmart-have-some-of-the-biggest-data-warehouses-you've-ever-seen/) The figure below demonstrates the communication of resources where parsing engines (Pes) manage and optimize queries, a Bynet enables internode communication, access module processes (AMP) execute queries and manage the database, and disk resources hold the data. Each AMP has full control of its own memory, disk, and central processing unit (CPU), demonstrating the shard-nothing architecture. Teradata's solution includes a fully integrated, complete, out-of-the-box solution, including hardware, software, network, operating system, and enterprise support.

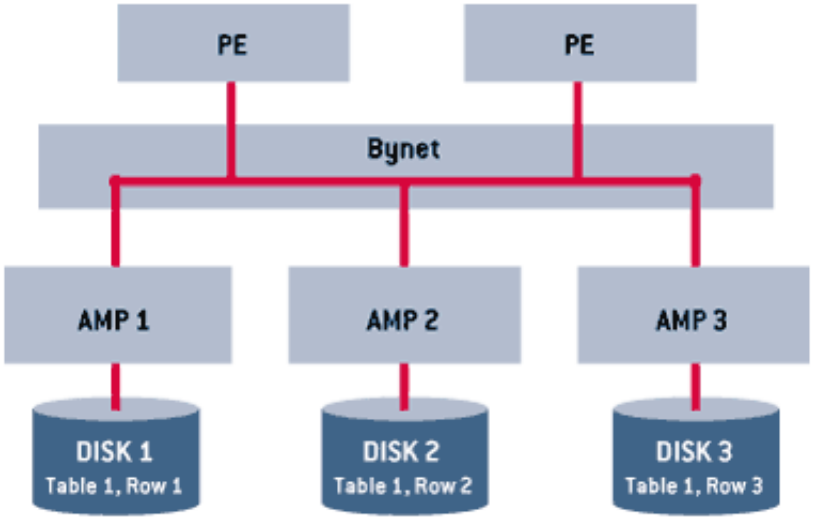

Figure 22. Illustration. Teradata sample architecture.
(Source: Windows IT Pro, 2000.)

In a two-year old study, the International Technology Group (ITG) compared the cost of an IBM PureData solution to a Teradata solution, and concluded that the three-year total cost of ownership for Teradata was on average 1.5 times more expensive than PureData. (International Technology Group, "Cost/Benefit Case for IBM PureData System for Analytics: Comparing Costs and Time to Value with Teradata Data Warehouse Appliance," Accessed May 13, 2016, https://tdwi.org/~/media/5BE30CAF543C4820A7139AAE81DA590F.PDF.) Acquisition of the products themselves (this initial upfront costs) were virtually identical (which has converged over the years), but maintenance and support, deployment, and personnel cost significantly less for PureData. Deployment costs averaged 3.8 times higher for Teradata, and deployment times averaged in the 10-20-day range for PureData and 100 days-6 months range for Teradata. Related to this, the "lost opportunity costs" (money lost due to delays in getting to production) were between 2.9 to 5.3 times higher for a Teradata solution. These results are nearly identical to the same analysis published the year prior.

Key Capabilities

- Teradata has been in the market the longest, has historically been the market leader, and provides a honed product capable of scaling significantly.
- Provides a fully integrated product, including hardware, software, network, operating system for a lump sum price with additional yearly support costs added on.
- Incorporates fully integrated analytical tools, including in-database capabilities to reduce unnecessary data movement as well as geospatial, big data tools, visualization, and other tools.

- Employs a completely shared-nothing architecture (all computing resources are isolated from each other and controlled only by the master node) to enable more significant utilization of cluster resources (essentially, compute nodes aren't wasting resources in interactions with other nodes).

Pricing

Teradata provides a fully integrated solution, including hardware, software, network, and operating system and charges for an initial, upfront cost with yearly maintenance and support costs added on. Teradata also does not typically recommend commodity hardware options for their solution which will increase the initial price substantially, but also needs to be factored into long-term scalability constraints.

Direct information regarding costs are unavailable from this vendor; however, based on experience, Teradata has historically been the highest in both initial, upfront costs and in total cost of ownership. The information from the study mentioned above is two years old and the gap is likely still converging; however, the differences are significant enough to consider fully when analyzing options.

IBM PureData System for Analytics

Gartner ranked IBM in the *Leaders Quadrant*. IBM's Netezza has been rebranded as a part of their PureSystems Suite, and is now called PureData System for Analytics—Powered by Netezza. PureData still provides the same fully integrated (coming with hardware, software, network, etc.) MPP database and still uses the Netezza name for continuity and brand recognition. The figure below shows the PureData system in conjunction with Revolution Analytics (a vendor for the open source analytical programming language, R), and provides an example of how PureData works. The traditional Netezza servers make up the analytical backbone of the solution (hence, "Powered by Netezza") and communicate independently with each shared-nothing disk storage unit and SQL processing host (for query management). The Netezza servers contain a snippet processing unit (SPU) that functions very similarly to the Teradata AMP to isolate resources and ensure dedicated processing power.

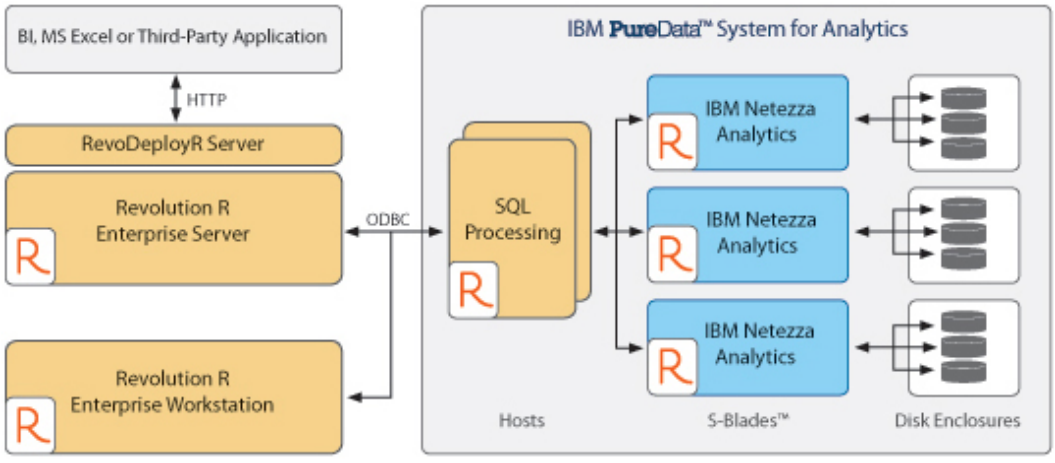

Figure 23. Illustration. IBM PureData sample architecture.
(Source: Personal Blog of Dr. Albert Spijkers, 2015.)

PureData frequently advertises its speed to market and ability to adapt quickly as primary differentiators between competitors. Two studies discussed below comparing IBM's PureData to Teradata and Oracle's Exadata appear to back these claims up.

As described above, in a two-year old study, the ITG compared the cost of an IBM PureData solution to a Teradata solution, and concluded that the three-year total cost of ownership for Teradata was on average 1.5 times more expensive than PureData. (International Technology Group, "Cost/Benefit Case for IBM PureData System for Analytics: Comparing Costs and Time to Value with Teradata Data Warehouse Appliance," Accessed May 13, 2016, https://tdwi.org/~/media/5BE30CAF543C4820A7139AAE81DA590F.PDF.) Acquisition of the products themselves (this initial upfront costs) were virtually identical (which has converged over the years), but maintenance and support, deployment, and personnel cost significantly less for PureData. Deployment costs averaged 3.8 times higher for Teradata, and deployment times averaged in the 10 to 20-day range for PureData and 100 days to 6 months range for Teradata. Related to this, the "lost opportunity costs" (money lost due to delays in getting to production) were between 2.9 to 5.3 times higher for a Teradata solution. These results are nearly identical to the same analysis published the year prior.

In another two-year old study, the ITG compared the cost benefits of an IBM PureData solution to an Oracle Exadata solution, and concluded that the three-year total cost of ownership for Exadata was on average 1.8 times more expensive than PureData. (International Technology Group, "Cost/Benefit Case for IBM PureData System for Analytics: Comparing Costs and Time to Value with Teradata Data Warehouse Appliance," Accessed May 13, 2016, https://tdwi.org/~/media/5BE30CAF543C4820A7139AAE81DA590F.PDF.) Every type of cost considered (acquisition, maintenance and support, deployment, and personnel) were significantly less for PureData. Deployment times ranged between 4 days-3 months for PureData and between 2 weeks-12 months for Exadata. The "lost opportunity costs" (money lost due to delays in getting to production) were on average 3 times higher for an Exadata solution.

Key Capabilities

- Provides a fully integrated product, including hardware, software, network, operating system for a lump sum price with additional yearly support costs added on.
- Is known for its speed to deployment and manageable pricing structure (as shown in the above reports).
- Provides the benefits associated with having a wide range of other IBM technologies to supplement and possibly customize their solutions in a similar way to how they adapted the Netezza database into their new PureData System for Analytics.

Pricing

Direct information regarding costs are unavailable from this vendor; however, the two studies mentioned above provided insights into the pricing difference between PureData and its competitors two years ago. The gap is likely converging; however, they both provided significant enough differences to be worth considering when analyzing options.

Pivotal Greenplum

Gartner ranked Pivotal in the *Visionaries Quadrant*. Pivotal's Greenplum solution is the only MPP database discussed here to provide the option of inexpensive, commodity hardware as a means of keeping down costs. The use of commodity hardware (which is more prevalent in Hadoop architectures) assumes that hardware failure is inevitable, particularly in big data solutions with up to thousands of nodes, and it's not worth spending the time and money to fix the hardware at the cost of lost data and resources. Commodity hardware is therefore easier and cheaper to both scale out with and replace when inevitable failures occur.

Greenplum is one of the smaller vendors for MPP appliances, and the only *Visionary* discussed here. However, they were responsible in 2009 for one of the largest data warehouses in the world with eBay. According to a blogger who sat down with eBay executives, eBay implemented two massive data warehouses using Teradata for one and Greenplum for the other. The Teradata warehouse was slightly smaller, but responsible for more varied data and more complex workload management tasks. **The Greenplum warehouse was intended to hold 6.5PB of user data and 17 trillion records with an ingest rate of approximately 50TB per day (**Recall from section 1 the total storage of a typical agency is expected to require roughly 2PB in 2021 at less than 2TB per day in 2021). The following figure shows a sample physical architecture for Greenplum that demonstrates the shared-nothing architecture discussed previously.

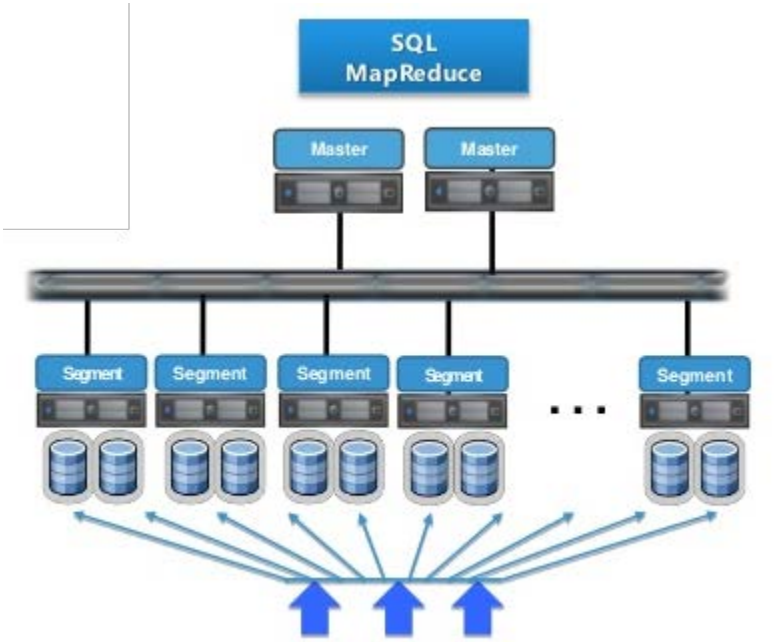

Figure 24. Illustration. Pivotal Greenplum sample architecture.
(Source: TekSlate, 2014.)

Key Capabilities

- Offers multiple deployment options, including the use of user-provided commodity hardware, the more typical tightly integrated, optimized for high performance solution required by other vendors to bring down costs, or even a virtualized Infrastructure-as-a-Service (IaaS) environment (the use of commodity hardware is a clear cost benefit, but may introduce interfacing issues due to a lack of hard standards and repeatable implementations).
- Tends to be more innovative and adaptable given its smaller size and lack of a long-standing history and customer base (e.g., use of commodity hardware, use of open source, etc.); this could make Pivotal an interesting choice for a DOT looking to stay adaptable or a risky option and introduce too many unknowns.

Pricing

Greenplum can provide the standard fully integrated solution, including hardware, software, network, and operating system and charges for an initial, upfront cost with yearly maintenance and support costs added on. Or the customer can purchase commodity hardware to be integrated with the purchased software solution and support from Greenplum.

Direct information regarding costs are unavailable from this vendor; however, based on experience, Greenplum has historically worked to maintain their low-cost reputation.

Oracle Exadata

Gartner ranked Oracle in the *Leaders Quadrant*. Oracle's Exadata solution is depicted in the figure below with a detailed breakdown of the Oracle software components that comprise it. Worth noting is that the complete Oracle solution is composed of both their traditional Real Application Cluster (RAC) servers (the two servers above the InfiniBand) and their Exadata solution composed of the Exadata cells (the three servers below the InfiniBand) which practice the shared-nothing architecture. Also worth noting is that Oracle may not recommend inexpensive, commodity hardware directly, but because they offer such a wide range of hardware appliances, the choice is up to the organization as to how expensive and resource optimized they want to be.

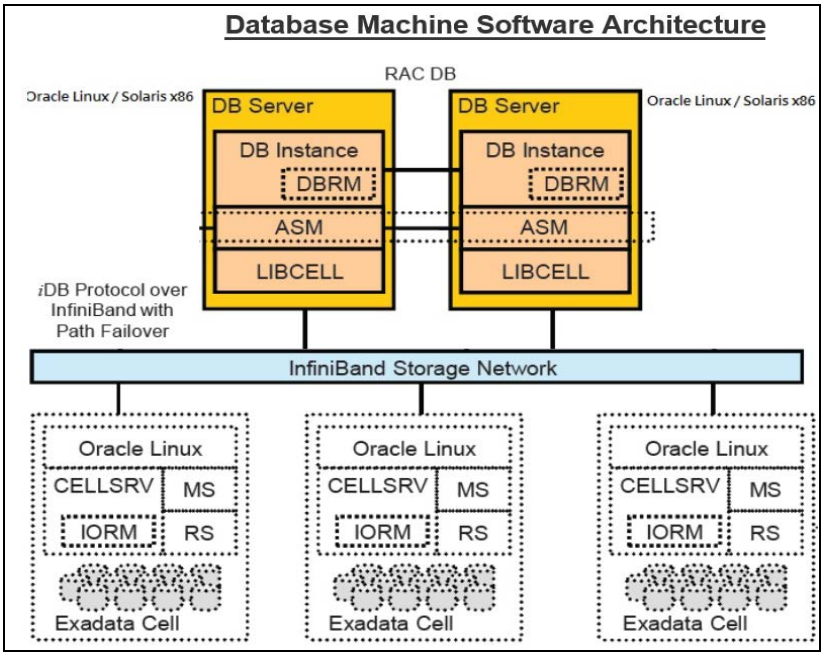

Figure 25. Illustration. Oracle Exadata sample architecture.
(Source: UnixArena Blog, "Architecture of Exadata Database Machine—Part 2" 2014.)

As described above, in a two-year old study, the ITG compared the cost benefits of an IBM PureData solution to an Oracle Exadata solution, and concluded that the three-year total cost of ownership for Exadata was on average 1.8 times more expensive than PureData. (International Technology Group, "Cost/Benefit Case for IBM PureData System for Analytics: Comparing Costs and Time to Value with Teradata Data Warehouse Appliance," Accessed May 13, 2016, https://tdwi.org/~/media/5BE30CAF543C4820A7139AAE81DA590F.PDF.) Every type of cost considered (acquisition, maintenance and support, deployment, and personnel) were significantly less for PureData. Deployment times ranged between 4 days to 3 months for PureData and between 2 weeks-12 months for Exadata. The "lost opportunity costs" (money lost due to delays in getting to production) were on average 3 times higher for an Exadata solution.

Key Capabilities

- Provides a fully integrated product, including hardware, software, network, operating system for a lump sum price.
- Provides a full range of database, data warehousing, and data management technologies to supplement and possibly customize their solutions.
- Provides incredibly transparent pricing structures with a massive suite of options to sort through (prices are likely to decrease with discounts).

Pricing

Oracle provides an incredibly detailed and transparent pricing structure for organizations to customize their solution based on hardware resources (e.g., RAM, disk, cores, etc.), software products, and support costs.

> Based on the data volume and velocity estimates provided in chapters 2 and 3, $2,000,000 is a very rough initial estimate for the initial cost of an MPP solution.

A total cost estimate is not possible given the lack of capacity planning and sizing exercises and extremely comprehensive pricing structure Oracle provides; however, **based on the data volume and velocity estimates provided in chapters 2 and 3, $2,000,000 is a very rough initial estimate for the initial cost**. This would not include the annual support fees (generally about 20 percent of the initial, upfront cost, per year). Additionally, the study mentioned above provided insights into the pricing difference between Exadata and PureData two years ago. The gap is likely converging; however, they both provided significant enough differences to be worth considering when analyzing options.

Additional pricing details can be found through the following links for hardware and software, respectively:

- http://www.oracle.com/us/corporate/pricing/exadata-pricelist-070598.pdf.
- http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf.

Distributed Hadoop Platforms

There are three market-leading Hadoop distributions: Cloudera, Hortonworks, and MapR. These distributions can be implemented on-premise or in the cloud as Infrastructure-as-a-Service (IaaS); Platform-as-a-Service (PaaS) options will be discussed in the following section (these options were briefly discussed in a previous section). An on-premise implementation of Hadoop will be discussed in this section. Each distribution provides a mature solution with significant market share, and a very similar set of capabilities. Organizations will need to determine exactly what their business needs are to properly compare.

Cloudera Distributed Hadoop

Gartner ranked Cloudera in the *Visionaries Quadrant*. Cloudera has been in the Hadoop market longer than any other vendor (which explains its high market share), and includes proprietary tools like Impala, Cloudera Navigator, Cloudera Manager, and Cloudera Director for providing a more efficient query engine and cluster and data management tools. Cloudera does not support several Apache products and tends to be slower to upgrade Apache releases; however, they have a more focused, user-friendly solution with well-established support capabilities. For example, Cloudera Manager is a proprietary cluster management tool that can launch a cluster with a few clicks and allows configurations to be managed and changed quickly and easily (both of which take time and more technical skills using the command line).

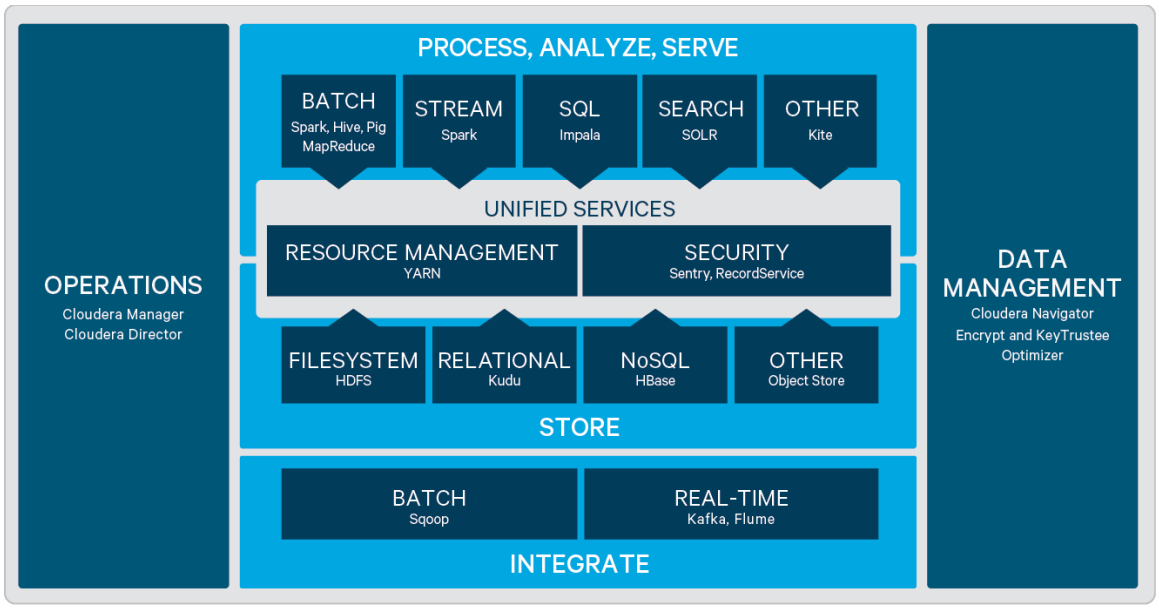

Figure 26. Illustration. Cloudera sample architecture.
(Source: Cloudera, 2016.)

Key Capabilities

- Includes proprietary tools like Impala, Cloudera Manager, Cloudera Navigator, and Cloudera Director that tend to be more focused on accomplishing specific customer needs (may be incredibly useful for DOTs if they decide to develop specifically relevant tools like geospatial capabilities and graph stores more heavily over the next few years).
- Leads the current market share (many third-party vendors would be interested in collaborating with a vendor that clearly leads the market share enabling faster development and integration with new products).

Pricing

A number of editions are available, including a free Express version. The Express version provides limited management functionality (which is a key attraction for many customers) and extremely limited support capabilities. Several Enterprise versions are available for yearly subscription costs per node based on business needs (Basic, Data Engineering, Analytical, etc.).

Direct information regarding costs are unavailable; however, based on experience, $2,600 per node per year for base software and support is an initial estimation.

Additional pricing details can be found through the following link:
http://www.cloudera.com/content/dam/www/static/documents/datasheets/cloudera-enterprise-datasheet.pdf.

Hortonworks Data Platform

Gartner ranked Hortonworks in the *Visionaries Quadrant*. The Hortonworks Data Platform (HDP), is the only vendor among the three distributions that is built using entirely open source components and gives everything they develop back to the open source community. This enables Hortonworks to have faster upgrades with Apache-developed products. HDP also provides a broad range of deployment options for Hadoop (e.g., Windows Server to Linux to virtualized Cloud deployments) making it a more versatile Hadoop distribution. Ambari is Apache's and Hortonworks' response to Cloudera Manager.

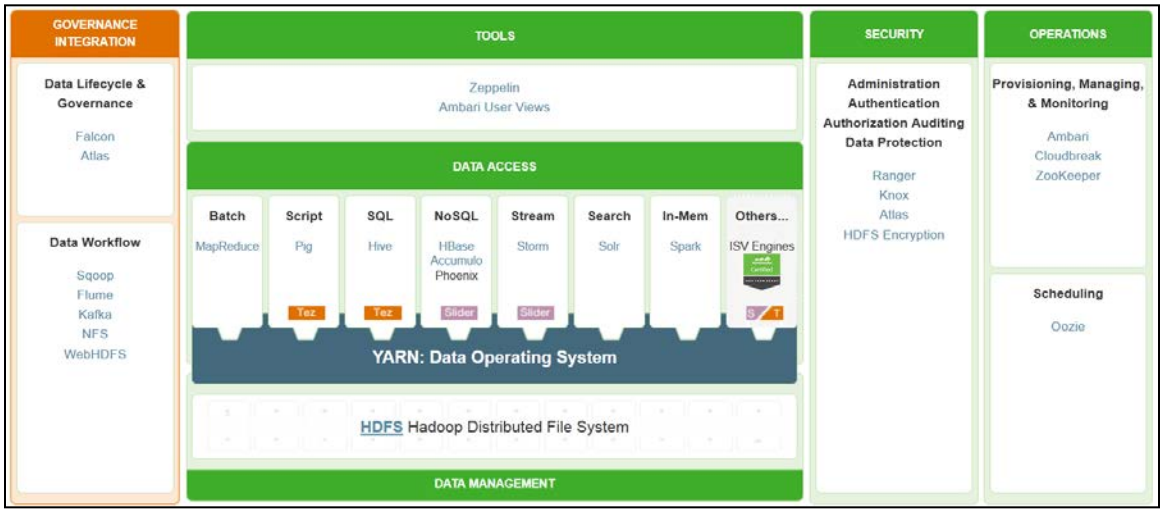

Figure 27. Illustration. Hortonworks sample architecture.
(Source: Hortonworks, 2014.)

Key Capabilities

- Provides shorter times between upgrades of Apache products and can innovate faster due to their 100 percent open source commitment (if Apache or Hortonworks decides to pursue a specific new capability or tool, a significant pool of people can work towards the project leading to faster capabilities; may be important to DOTs if there are significant gains to be made with new capabilities that look like they're approaching quickly).
- Provides a wealth of experience-backed expertise due to its creation as a Yahoo! Spin-off; Yahoo! Has the largest cluster known in the world at around 40,000 nodes and was one of the most innovative and massive contributors to the Apache projects.

Pricing

Hortonworks provides three subscription options: Jumpstart, Enterprise, and Enterprise Plus. Jumpstart is a 6-month subscription; Enterprise and Enterprise Plus are yearly subscriptions. Various details are provided regarding the level of support that can be expected, including support contacts, time until response based on the severity of the problem, hours of direct support, etc.
Direct information regarding costs are unavailable; however, based on experience, $1,200 per node for the basic software, service, and support is an initial estimation.

Additional pricing details can be found through the following link:
http://hortonworks.com/services/support/.

MapR Converged Data Platform

Gartner ranked MapR Technologies in the *Visionaries Quadrant*. MapR employs the most significant percentage of proprietary products, including the foundational file system they call MapRFS. MapR integrates and expands on many of the most popular Apache tools (e.g., Spark, Drill, and others) in addition to commercial tools (e.g., SAS, SAP, and others) to provide a comprehensive and flexible solution.

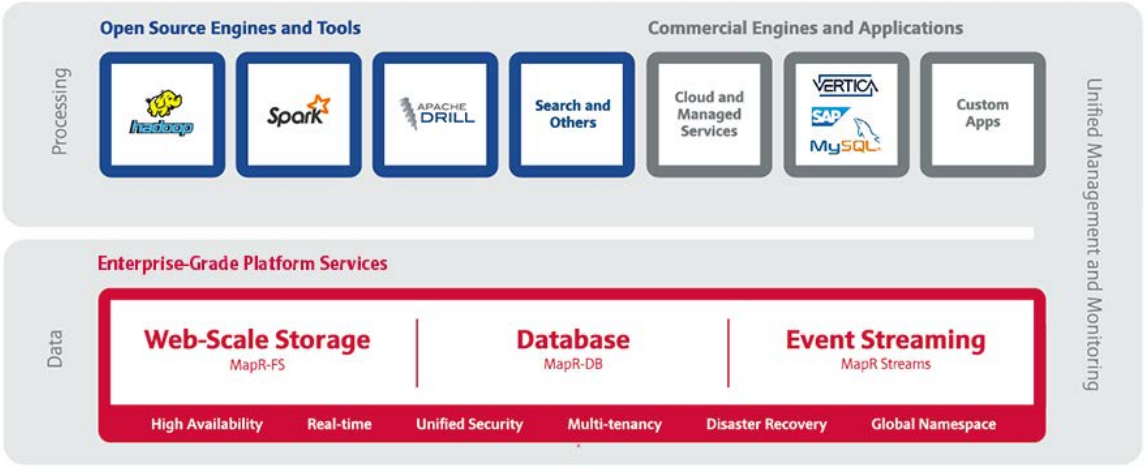

Figure 28. Illustration. MapR sample architecture.
(Source: MapR, 2015.)

Key Capabilities

- Provides extensive proprietary products to supplement the open source components and commercial tools that they integrate with; tends to force the issue of vendor lock-in.
- Considered to have the objectively fastest and most efficient platform, which can become extremely important if DOTs are interested in pursuing significant real-time applications (every second counts).

Pricing

MapR provides two subscription options for their distribution: Community and Enterprise. The community edition provides unlimited, free production use; however, it has limited features and capabilities (e.g., high availability, fault tolerance, consistent snapshots, etc.) and offers no commercial support. The Enterprise edition provides each of the capabilities they offer, round-the-clock commercial support, and support for specific products (e.g., Spark, Hbase, Solr, etc.) for an additional subscription cost.

Direct information regarding costs are unavailable; however, based on experience, $2,600 per node per year for the base software and support is an initial estimation.

Additional pricing details can be found through the following link: https://www.mapr.com/products/mapr-distribution-editions.

Cloud-Based Hadoop Platforms

The cloud-based Hadoop platform implementations discussed in this section are Platform-as-a-Service (PaaS) models where the cloud vendor provides a Hadoop computing platform preintegrated with existing cloud services. The time it takes to stand up the Hadoop platform is dramatically reduced, provisioning can be done in a straightforward manner, and the solution can be scaled easily and on an as-needed basis. For the convenience of integration with cloud services, faster provisioning, and simple scalability, organizations will sacrifice some flexibility in customization options.

The following four platforms discussed are some of the most well-known cloud vendors in the market, and they have used each of the Hadoop distributions previously discussed to build a cloud-based Hadoop platform. Pricing can be difficult to discuss because there is no standard yardstick from which to measure, each vendor has a different process, and the cloud infrastructure is generally sold separate from the Hadoop platform. For example, an organization will choose their cloud infrastructure base with the resources and at the cost-point they need, and then they will choose the Hadoop PaaS component that meets their needs.

Additionally, the Magic Quadrant may not be as applicable for each of these vendors. Several, like Microsoft and IBM, can provide services at every level of data warehousing (e.g., on-premises, traditional RDBMS, and cloud offerings), while Amazon and Google are limited to the diverse cloud offerings they provide. However, if they are on the Magic Quadrant, their ranking will be mentioned briefly.

Amazon Web Services Elastic MapReduce

Gartner ranked Amazon Web Services (AWS) in the *Challengers Quadrant*. Amazon is widely known for its cloud options and understood to be a market leader, and while the number of nuanced choices they have can seem overwhelming for an organization to consider, the takeaway message is that AWS will not lack customizable options when the business needs are determined. For example, Amazon's Elastic MapReduce (EMR) clusters can use EC2 instances as virtual Linux servers for the master and slave nodes, Amazon S3 for bulk storage of input and output data, and CloudWatch to monitor cluster performance. Additionally, EMR uses the open source Apache distribution of Hadoop by default, but allows organizations to use a MapR distribution if that's what they desire (they do not offer Cloudera or Hortonworks distributions). (Amazon Web Services, "Amazon EMR," Accessed May 9, 2016, https://aws.amazon.com/elasticmapreduce/.)

Key Capabilities

- Provides interoperability with other Amazon products, including the ability to use different offerings for optimized performance of specific functions (e.g., Amazon S3 for storage, Amazon EC2 for compute resources, etc.).
- Provides optimized and transparent pricing models and options available for underlying cloud services as well as platform offerings (e.g., high memory options).

- Is simple and efficient to scale up or down based on current storage needs (e.g., on an hourly basis or a seasonal basis given peak driving times of day and year).

Pricing

The price of EMR can range from $0.011/hour/node to $0.270/hour/node based on the instance used for the Hadoop compute platform (e.g., optimized for memory, general purpose, etc.). However, additional cost must be considered for computation and storage resources (e.g., EC2 and S3, respectively) and any additional AWS services the business requires.

Additional pricing details can be found through the following link: https://aws.amazon.com/elasticmapreduce/pricing/.

Microsoft Azure HDInsight

Gartner ranked Microsoft in the *Leaders Quadrant*. Microsoft's Azure HDInsight is a PaaS model that runs Microsoft's Hadoop platform, called HDInsight, on their Azure cloud platform. HDInsight uses a Hortonworks distribution of Hadoop, and does not provide for the use of any other distributions. Using Microsoft's platform brings increased interoperability with other Microsoft products (which may be important for Microsoft-heavy organizations), but lacks flexibility elsewhere (i.e., vendor lock-in). (Microsoft Azure, "What is Hadoop in the cloud? An introduction to Hadoop components in HDInsight for big data analysis," Accessed May 9, 2016, https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/.)

Key Capabilities

- Provides integration and interoperability with Microsoft products, tools, and other services, including Excel and BI tools.
- Provides organizations the option of letting Microsoft manage more aspects of the infrastructure (valuable for organizations particularly concerned about a lack of existing skills; less valuable for organizations that rely on maximum flexible control over their environment).
- Allows the creation of hybrid applications (applications that combine on-premises and cloud-based tools) as a way of connecting legacy datacenters to the Hadoop computing platform.

Pricing

Microsoft's Azure HDInsight is offered as Standard (an enterprise solution) and Premium (an enterprise solution with additional advanced analytical capabilities). The price of a Standard cluster can range from $.008/hour/node to $3.04/hour/node, and the price of a Premium cluster can range from $.010/hour/node to $3.36/hour/node. The wide range in price is dependent on the computing resources each node provides, including primarily RAM, disk, and number of cores. However, the prices provided include both the Hadoop platform instances as well as the storage/computation resources all-in-one.

Additional pricing details can be found through the following link: https://azure.microsoft.com/en-us/pricing/details/hdinsight/.

IBM SoftLayer BigInsights

Gartner ranked IBM in the *Leaders Quadrant*. IBM's SoftLayer BigInsights provides a PaaS solution using IBM's SoftLayer cloud infrastructure and their proprietary BigInsights Hadoop distribution. IBM provides several additional proprietary tools, including the browser-based analytics tool called BigSheets and the MPP SQL engine for Hadoop called BigSQL. However, it also lacks flexibility due to its restrictive use of IBM proprietary software (i.e., vendor lock-in). (IBM, "Hadoop-as-a-service, big data analytics in the cloud," Accessed May 9, 2016, http://www-03.ibm.com/software/products/en/ibm-biginsights-on-cloud.)

Key Capabilities

- Includes additional proprietary analytical accelerators such as text analytics, machine learning, and geospatial analysis, and data mining.
- Includes additional proprietary Hadoop analytical tools such as Big SQL, BigSheets, and Big R.

Pricing

Pricing information is not available from this vendor. They primarily sell their services through third-party agreements and enterprise licenses, and do not provide publically available cost breakdowns.

Google Cloud Dataproc

Gartner has not formally ranked Google in the Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics. Google was the original creator of the Google File System for which Hadoop was founded and continues to be a leader in the field of advanced big data applications. Google's Cloud Dataproc provides increased interoperability across other Google Cloud Platform products, including the well-known BigTable, BigQuery, and others. (Google Cloud Platform, "Cloud Dataproc," Accessed May 9, 2016, https://cloud.google.com/dataproc/.)

Key Capabilities

- Provides built-in integration with Google's proprietary products, including Cloud Storage, BigQuery, BigTable, Cloud Logging, and Cloud Monitoring.
- Provides "image versioning" to allow users to choose between bundled versions of Hadoop products, including Spark, Pig, Hive, and others (useful when compatibility issues come into play; simplifies deployments).
- Provides tools for developers to manage a cluster multiple ways, including a Web user interface (UI), a Google Cloud software development kit (SDK), RESTful Application Programming Interfaces (API), and secure shell (SSH) access (managing the cluster includes the technical details of managing resources, configuring the platform, and monitoring and maintaining system health).

Pricing

Google's Dataproc solution is billed based on the underlying resources used. The cost of Dataproc itself is $0.01/hour/vCPU used. Total cost of ownership would also need to include the underlying cost of the cloud infrastructure (e.g., Google's compute engine, persistent disk storage, and cloud monitoring services).

Additional pricing details can be found in the following location:
https://cloud.google.com/dataproc/pricing.

Cloud-Based Internet of Things Platforms

The cloud-based Internet of Things (IoT) platform is a special type of computational and integration platform supporting message collection from a variety of devices and the execution of big data analytics to discover patterns and trends. While many companies are beginning to develop IoT platforms, the following five were determined to be a representative set of the capabilities on the market in spring 2016.

IoT is still a new concept, and the IoT platforms are still evolving to fill the right business and technical needs. The platforms listed below provide a very similar set of capabilities at this time, and will likely evolve and distance themselves from one another with specific vendors becoming clear market leaders in the very near future. Because this report is only a snapshot in time and not intended to provide any recommendations at this time, each platform will only briefly be mentioned with a short understanding of the most distinguishing factors and price comparison. Additionally, Gartner's Magic Quadrant is not explicitly relevant for these platforms and will not be mentioned here.

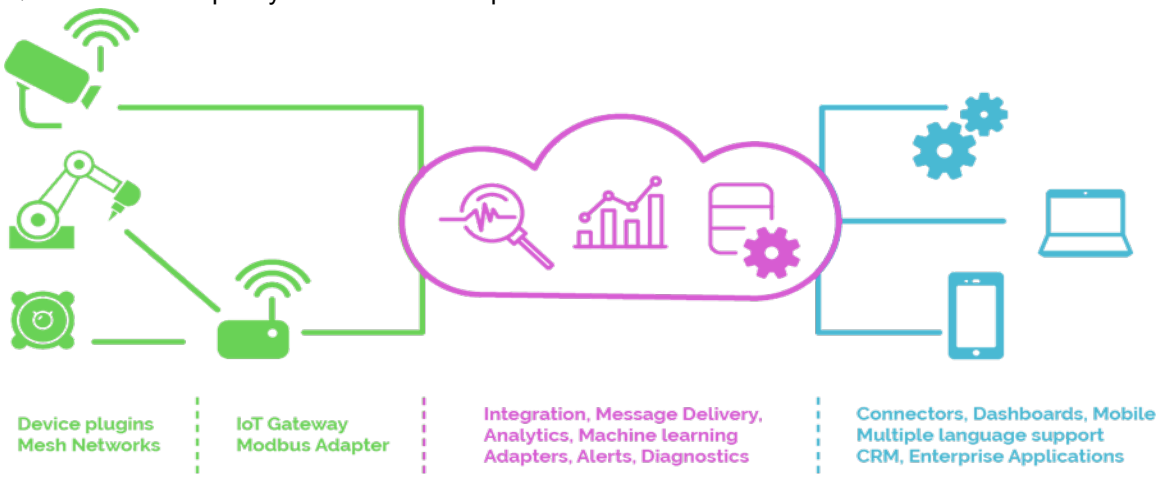

Figure 29. Illustration. Representation of the Internet of Things.
(Source: https://www.robomq.io/, 2016.)

The image above shows a simplified representation of how connected devices interface with IoT gateways to eventually reach a cloud platform for storage, management, and analysis. For the purposes of this report, Agencies would be able to connect roadside units (RSUs) and other acceptable connected devices such as traffic equipment with sensors, phones through user buy-in

apps (e.g., a Boston-based app that detects bad roads through connected travelers' phone movements), etc. (City of Boston, "Street Bump: Help Improve Your Streets," Accessed May 13, 2016, http://www.cityofboston.gov/DoIT/apps/streetbump.asp.)

Amazon Web Services Internet of Things

Amazon Web Services (AWS) provides an IoT managed cloud platform to deliver secure and efficient communication between edge devices (e.g., sensors, actuators, embedded devices, or smart appliances) and the underlying AWS cloud foundation. AWS also has partnered with hardware manufacturers like Intel, Texas Instruments, Broadcom and Qualcomm to create starter kits compatible with their platform. These starter kits are physical kits with sensors, actuators, and other devices designed to help users begin using the IoT platform. (Amazon Web Services, "AWS IoT," Accessed May 9, 2016, https://aws.amazon.com/iot/.)

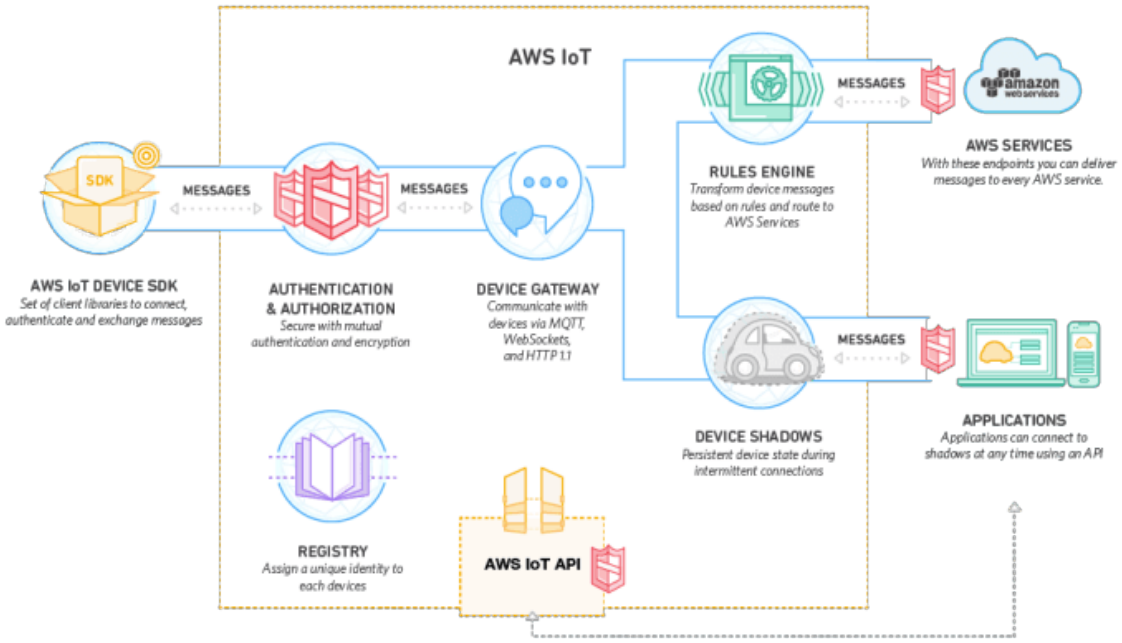

Figure 30. Illustration. Amazon Web Services Internet of Things platform sample architecture. (Source: https://paolopatierno.wordpress.com/2015/10/13/an-iot-platforms-match-microsoft-azure-iot-vs-amazon-aws-iot/, 2015.)

Pricing

Amazon charges on a pay-as-you-go model (typical for cloud deployments) with no minimum fees and free delivery to other Amazon services (e.g., Amazon S3, Amazon DynamoDB, etc.), for example, a big data platform that uses Amazon's S3 model. Their free tier provides up to 250,000 messages per month, and their paid tier provides $5 per 1,000,000 messages. A message is considered bidirectional (meaning it costs to both send and receive messages), and a message is in 512 byte increments (meaning a 1024 byte message would be billed as two messages). In terms of the emerging sources identified in this report, a connected traveler's daily load of 500KB/day would represent perhaps 500 1KB messages, or 1,000 500-byte chunks in Amazon's pricing structure. **The**

150,000 connected travelers of a typical agency in 2021 would thus cost $750/day for the 150,000,000 total messages ($274,000 per year). Total cost of ownership also would need to include the underlying cost of the cloud infrastructure using Amazon S3 or other offerings.

Additional pricing details can be found through the following link: https://aws.amazon.com/iot/pricing/.

Microsoft Azure Internet of Things

Microsoft Azure's IoT platform can be used for same general purpose as the AWS; however, in a similar vein as Microsoft's HDInsight platform (though perhaps even more so), Microsoft is capable of ensuring an extra degree of interoperability with other Microsoft products. Which is particularly important for organizations that rely heavily on Microsoft products.

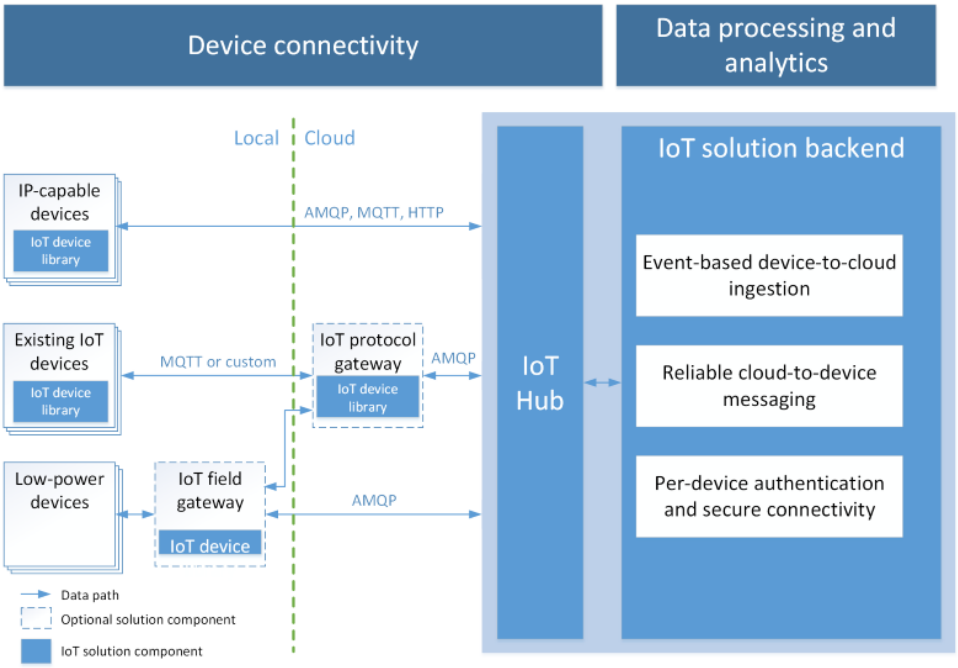

Figure 31. Illustration. Microsoft Azure Internet of Things sample architecture.
(Source: Microsoft, 2016.)

Pricing

Microsoft charges on a pay-as-you-go model (typical for cloud deployments) with no cancellation fees or upfront costs. They offer three subscription models: Free, S1, and S2. The Free model offers up to 8,000 messages at 0.5 KB each per day; the S1 model offers up to 400,000 messages at 4 KB each per day for $50 per month, and the S1 model offers up to 6,000,000 messages at 4 KB each per day for $500 per month. In terms of the emerging sources identified in this report, a connected traveler's daily load of 500KB/day would represent perhaps 500 1KB messages, or 1,000 4KB chunks in Microsoft's pricing structure. **The 150,000 connected travelers of a typical agency in 2021 would thus cost $333/day (using the straight-line pricing of 6,000,000 messages for $50/month) for the 150,000,000 total messages ($122,000 per year)**.

Additional pricing details can be found through the following link: https://azure.microsoft.com/en-us/pricing/details/iot-hub/.

IBM Watson Internet of Things

IBM Watson's IoT platform may distinguish itself from other platforms for Agencies with several focused set of automobile industry-related use cases. (IBM, "IBM Watson Internet of Things," Accessed May 9, 2016, http://www.ibm.com/Internet-of-things/.) If IBM can separate itself from the rest of the pack in the transportation industry, they may be a vendor worth keeping an eye on in the future.

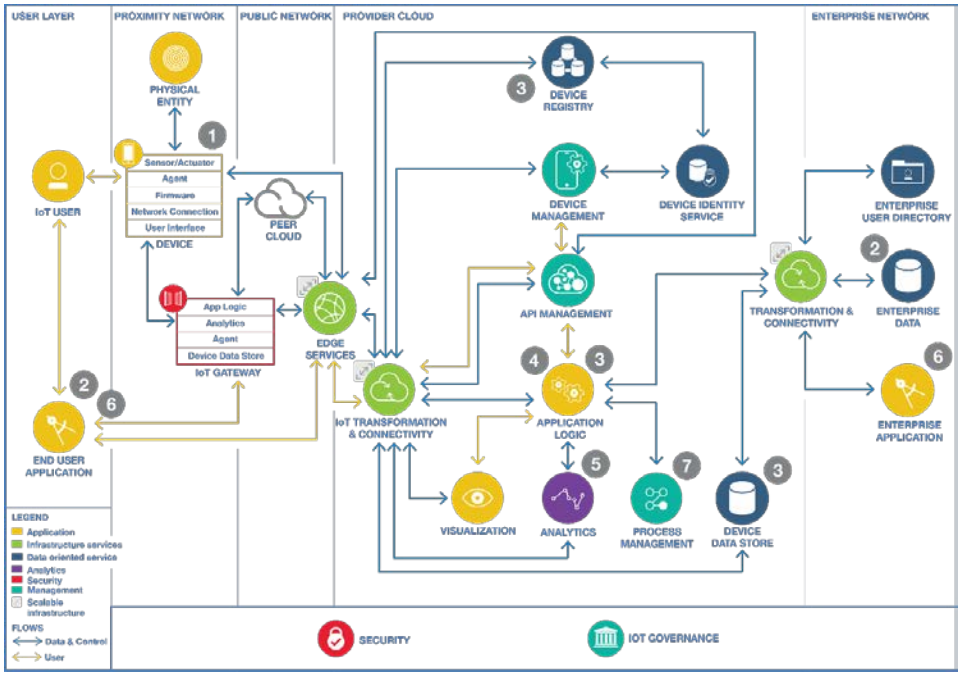

Figure 32. Illustration. IBM Watson sample architecture.
(Source: IBM, 2016.)

Pricing

IBM charges based on the number of devices connected per month, the amount of data exchanged by those devices, and the amount of data stored in historical databases. Up to 20 devices per month, 100MB of data traffic (equivalent to 50,000 messages) per month, and 1GB of storage per month are free with any plan. Total cost of ownership also would need to include the underlying cost of the cloud infrastructure using IBM Bluemix.

Direct information regarding costs are unavailable from this vendor.

Cisco Internet of Things Cloud Connect

Cisco and Intel both distinguish themselves from other vendors through their deep market experience in the "things" of IoT as opposed to their work with cloud offerings. (Cisco, "Internet of Things (IoT)," Accessed May 9, 2016, http://www.cisco.com/c/en/us/solutions/Internet-of-things/iot-products.html.) This strength could become a serious advantage in the future, or their lack of expertise in cloud solutions may be too much for them to overcome.

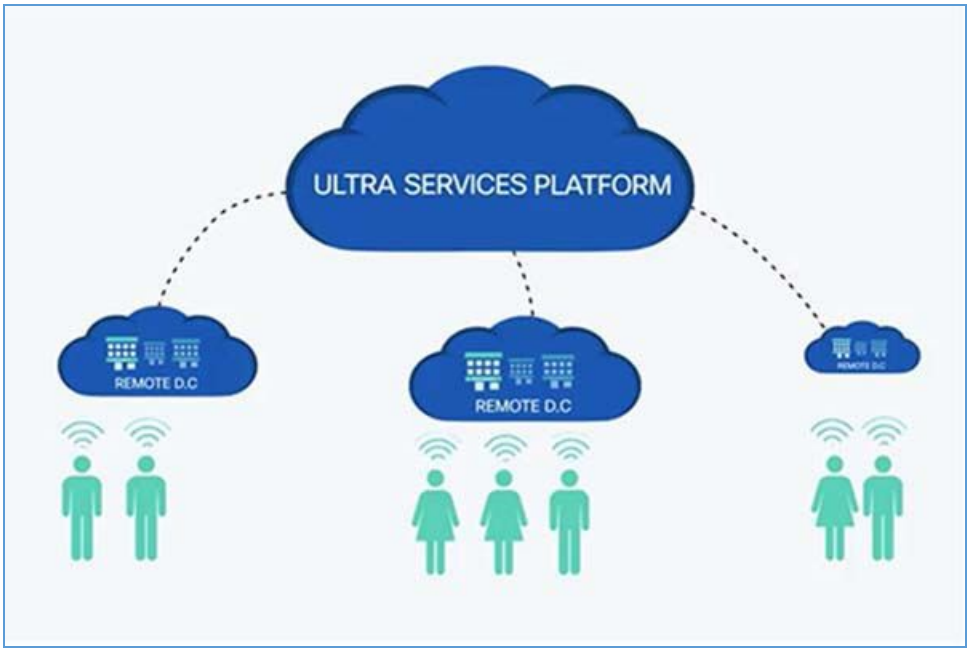

Figure 33. Illustration. Cisco Internet of Things sample architecture.
(Source: Cisco, 2016.)

Pricing

Pricing information is not available from this vendor. They primarily sell their services through third-party agreements and enterprise licenses, and do not provide publically available cost breakdowns.

Intel Internet of Things

As stated before, Intel and Cisco both distinguish themselves from other vendors through their deep market experience in the "things" of IoT as opposed to their work with cloud offerings. This strength could become a serious advantage in the future, or their lack of expertise in cloud solutions may be too much for them to overcome.

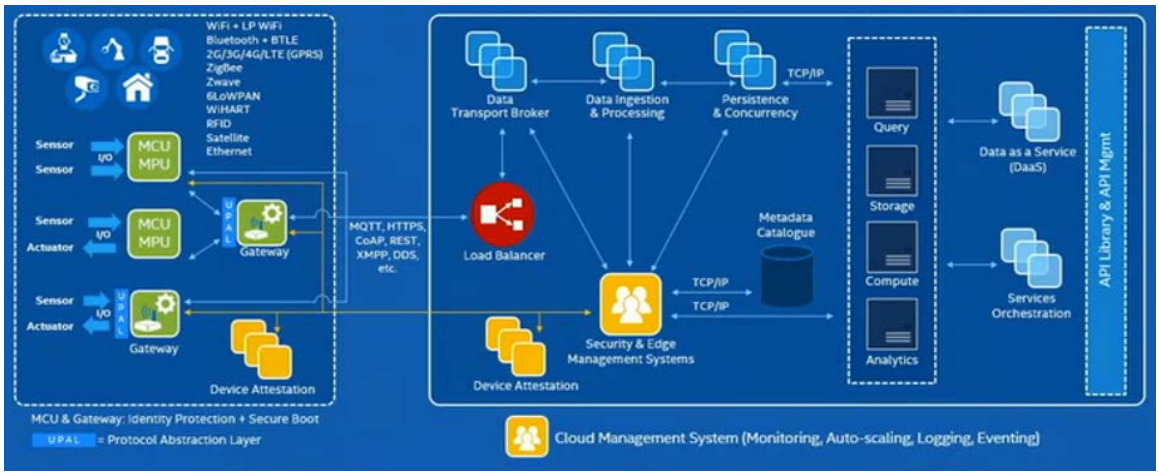

Figure 34. Illustration. Intel Internet of Things sample architecture.
(Source: https://theiotlearninginitiative.gitbooks.io/internetofthings101/content/documentation/Intel.html, 2016.)

Pricing

Pricing information is not available from this vendor. They primarily sell their services through third-party agreements and enterprise licenses, and do not provide publically available cost breakdowns.

Pricing Comments

Many of these products either have very limited information readily and publically available about their specific pricing structures or have incredibly detailed and challenging to decipher information regarding every possible approach that can be taken. While the transparency of the latter is commendable, it unfortunately still doesn't make it easy to make a direct apples-to-apples comparison of any products.

As much detail as could be found, reasonably interpreted, and assumed to be helpful was included about each vendor's product above. However, additional they generally lacked consistency and occasionally relevance and interpretability. To that end, an appendix titled "Big Data Tools and Technologies Implementation Considerations" has been attached with a list of questions that are extensive though not by any means exhaustive to 1) demonstrate the difficulties in finding the information desired; and 2) provide a prelude to the work anticipated in subsequent reports when a detailed analysis of alternatives may be performed.

Chapter 8 Summary

This report is intended to give the reader an understanding of what emerging data sources relevant to Transportation Systems Management and Operations (TSM&O) could be available in the next 10 years and what the data acquisition and storage implications will be. The second objective of the report in the context of the overall project, is to raise awareness of the TSM&O practitioner to some of the moving parts and terminology of the generic buzz-word "big data" and provide some examples of currently available systems and platforms available in the marketplace. Subsequent reports will explore further into the implications of these tools to address opportunities and challenges for TSM&O applications, architectures for collecting the volumes of data for public Connected Vehicle (CV) systems, and directions for integrating Big Data tools and technologies with existing Transportation Management Systems.

While traditional sources of transportation data for TSM&O will remain, emerging data sources, largely those from connected travelers, connected vehicles, and connected infrastructure, will represent a significant opportunity for Departments of Transportation (DOT) and localities to improve TSM&O practices. However, the volume of data will create challenges for DOTs to manage and use it. A single traveler will likely contribute 500KB of trajectory-related travel behavior data per day; assuming that the Personally Identifiable Information (PII) issues can be suitably addressed for the connected traveler and commercial connected vehicles sources. This is a very strong assumption. However, it is an issue that existing third-parties that currently provide data to DOTs understand and will likely be working towards resolving as they are well aware that this higher-fidelity data source can be of considerable value for TSM&O.

In rough terms, if all data available to an agency through the emerging data sources was consumed and stored, **the volume of data per day consumed by an agency today of 1TB will more than double by 2021 to 2.5TB and be a more than five times higher daily consumption rate of 5.2TB by 2026**. 5.2TB per day is approximately 60GB per second, or 600Gbps. For comparison, a top-tier business access plan from commercial providers in 2016 is typically 150-200 Mbps per connection. Many businesses utilize multiple connections to increase total download capacity. In 10 years, it would be likely that commercial access plans will offer this level of download capacity to businesses or agencies. Acquisition of all of this data will require new methodologies predicated by components of the Big Data ecosystem.

If all of this emerging data is *stored*, the cumulative storage of a typical agency would be more than **3 Petabytes in 2021 and 10 Petabytes by 2026**. More than 80 percent of the data will be raw Basic Safety Messages (BSM) and compressing closed-circuit television (CCTV) images. Even compressing CCTV and BSM data sources 100:1 and/or only storing derived analytics or summaries will still require on the order of **1 Petabyte of storage in 2021 and 3 Petabytes by 2026**. Additional strategies to compress and consolidate information from the raw data will be needed as any current agency would struggle to justify the costs for this level of storage for the return on investment (ROI) that it provides. Some calculations of the comparative benefits versus costs would be of value to quantify the value proposition. DOTs also are on a much different procurement cycle than commercial businesses, and investments of this nature must last 10+ years, or the yearly subscription costs must

be justified and planned for years in advance. Investing in enhanced software capabilities for a typical DOT almost certainly means *not* investing in some other activity, capital project, or equipment purchase.

The data availability and volumes are educated projections with the important caveat that PII issues are resolved sufficiently to unlock the sharing of transformative information on travel behaviors with public agencies for the benefit of the public. Nonetheless, by any measure, the growth rate of these data sources will generate volumes of data that will require new methods and tools to realize their value. **The volume, variety, velocity, and veracity of connected traveler, connected vehicle, and connected infrastructure data will put TSM&O agencies firmly into the realm of "big data."** Even without the connected vehicles and connected travelers data, most agencies are not utilizing the information they currently collect from infrastructure alone, particularly traffic signal systems, in meaningful ways due to limitations in current Information Technology (IT) infrastructure (RDBMS) and lack of Big Data tools, experience, and expertise. Perhaps a good first step would be for agencies to gain experience with these tools and techniques as an expansion to their current tools and legacy systems.

There is a significant variety of tools and methods available now for data acquisition, marshalling, and analysis that are proven in a variety of use cases and markets. Data acquisition and marshalling technologies are the most mature of the three primary components; "out of the box" or "plug and play" analysis components are continuing to mature in 2016. There is no single best tool, technology, or provider for a particular agency or a particular TSM&O application. Rather, the tools and methods must be appropriate for the data, sufficiently mature, stable and supported, and within the ability of the data analysts to properly use. **The two primary categories of currently available commercial systems are the massively parallel processing (MPP) and Hadoop-based ecosystem solutions.** Both approaches are well represented in the marketplace and appear to have the capabilities to handle the size and velocity of data for TSM&O purposes by 2021. For example, the Greenplum MPP appliance is marketed in 2016 to hold up to 6.5PB at ingestion rates of up to 10TB per day; clearly within the requirements of 2PB total storage and 2.5TB/day ingestion rates estimated in chapter 3. Costs for these tools and technologies (along with storage) are substantial, relative to their value, and this tradeoff will need to be strongly considered by budget-constrained TSM&O agencies. Hadoop is mentioned many times in this report as it is such a strong central component of the Big Data world and the foundation of almost all commercially available systems outside of the MPP market. The MPP also is considered an aging technology since it simply (as far "simply" describes a massively complex system of interprocess communications and sophisticated software components) distributes an RDBMS across tens, hundreds, or thousands of processors. The Hadoop ecosystem essentially builds on top of the MPP concept of parallelization with new ways of organizing information that improve speed and responsiveness for processing unstructured information of massive sizes; particularly data that is not well suited for the strict ACID rules of RDBMS, such as images, documents, music, and so on.

As time passes, there is no doubt that new and increasingly sophisticated methods and tools will be developed to deal with increasingly bigger datasets and analysis suites will become more mature. The methods to make sense of these large datasets will require individuals and teams with skill sets that span software development, database administration, IT, statistical analysis and modeling, and interpersonal communication. Just as importantly, these individuals also must have domain knowledge in TSM&O to understand the information, perform meaningful analyses, and effectively communicate results. These individuals are now known as data scientists. Just as TSM&O has emerged over the

last 20 years as a functional area within DOTs and local agencies, data science will be an important area of focus over the next 10.

As hardware technology becomes ever more powerful and the data sources grow, the *cluster* of commodity processors and data storage devices is replacing the *server* as the primary unit of computing (per the prescient Grace Hopper). DOTs need to understand this paradigm shift as they work with their IT departments to plan for new applications of these emerging data source. However, it is unrealistic to expect DOTs to have capabilities in these new tools because they are evolving so quickly; rather they will need to partner with IT staff, data scientists, and system providers to engineer solutions.

References

1. U.S. Department of Transportation, ITS Joint Program Office, "Big Data's Implications for Transportation Operations: An Exploration," Publication No. FHWA-JPO-14-157, December 2014.

2. McKinsey Global Institute, "Big Data: The next frontier for innovation, competition and productivity," May 2011. Accessed at: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.

3. Kimley-Horn and Associates, Inc., "Traffic Management Centers in a Connected Vehicle Environment," TMC Pooled Fund Study, March 2014.

4. U.S. Department of Transportation, ITS Joint Program Office, "Big Data and ITS," White Paper, October 2013. Accessed at: http://connectedvehicle.itsa.wikispaces.net/file/detail/ITS+and+Big+Data+White+Paper+Final+Draft+10_2+%282%29.docm.

5. International Transport Forum, "Big Data and Transport: Understanding and assessing options, 2015." Accessed at: http://www.itf-oecd.org/big-data-and-transport-understanding-and-assessing-options.

6. U.S. Department of Transportation, ITS Joint Program Office, "Estimate Benefits of Crowdsourced Data from Social Media," Publication No. FHWA-JPO-14-165, February 2015.

7. AASHTO, "National Connected Vehicle Field Infrastructure Footprint Analysis," Publication No. FHWA-JPO-14-125, June 2014.

8. Pew Research Center, "U.S. Technology Device Ownership 2015," Accessed May 13, 2016, http://www.pewInternet.org/2015/10/29/technology-device-ownership-2015.

9. CNN Money, "U.S. cell phones, tablets outnumber Americans—Oct. 12, 2011," Accessed May 13, 2016, http://money.cnn.com/2011/10/12/technology/cellphones_outnumber_americans/index.htm.

10. Zipcar, "Millennials & Technology: A Survey," Accessed May 13, 2016, http://www.slideshare.net/Zipcar_Inc/millennial-slide-share-final-16812323.

11. Frontier Group and U.S. PIRG Education Fund, "Transportation and the New Generation: Why Young People Are Driving Less and What It Means for Transportation Policy," Accessed May 13, 2016, http://www.uspirg.org/sites/pirg/files/reports/Transportation%20%26%20the%20New%20Generation%20vUS_0.pdf.

12. Google Play Store, "UDOT Citizen Reports—Android Apps on Google Play," Accessed May 13, 2016, https://play.google.com/store/apps/details?id=gov.utah.udot.citizenreport.

13. Los Angeles County Metropolitan Transportation Authority (Metro), "Metro Mobile App," Accessed May 16, 2016, https://www.metro.net/mobile/metro-mobile-app.

14. World Population Review, “Rio De Janeiro Population 2016—World Population Review,” Accessed May 13, 2016, http://worldpopulationreview.com/world-cities/rio-de-janeiro-population.

15. PricewaterhouseCoopers, “Connected Car Study 2015: Racing ahead with autonomous cars and digital innovation, 2015,” Accessed at http://www.strategyand.pwc.com/reports/connected-car-2015-study.

16. Oregon DOT, “MyOReGO | A new way to fund roads for all Oregonians,” Accessed May 13, 2016, http://www.myorego.org.

17. Texas A&M Transportation Institute, “Strategic Research Program: Big Data Scan,” Accessed May 13, 2016, http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/161505-1.pdf.

18. HERE, “HERE, automotive companies move forward on car-to-cloud data standard,” Accessed July 1, 2016, https://lts.cms.here.com/static-cloud-content/Newsroom/290616_HERE_automotive_companies_move_forward_on_car_to_cloud_data_standard.pdf.

19. AT&T Labs Research, “Enabling Vehicular Safety Applications over LTE Networks,” Accessed May 13, 2016, http://web2-clone.research.att.com/export/sites/att_labs/techdocs/TD_101260.pdf.

20. IEEE Spectrum, “Autonomous Driving Experts Weigh 5G Cellular Network Against Dedicated Short Range Communications,” Accessed May 13, 2016, http://spectrum.ieee.org/cars-that-think/transportation/self-driving/autonomous-driving-experts-weigh-5g-cellular-network-against-shortrange-communications-to-connect-cars.

21. University of Arizona, Multimodal Intelligent Traffic Signal Systems (MMITSS) Concept of Operations, December 2012. Accessed at: http://www.cts.virginia.edu/wp-content/uploads/2014/05/Task2.3._CONOPS_6_Final_Revised.pdf.

22. Michigan DOT, “VII Data Use Analysis and Processing: System Requirements Specification,” December 2007, Accessed May 13, 2016, http://www.michigan.gov/documents/mdot/MDOT_DUAP_SysReq_Final_220099_7.pdf.

23. Federal Communications Comission, Accessed May 13, 2016, https://apps.fcc.gov/.

24. U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology, Bureau of Transportation Statistics, “Number of U.S. Aircraft, Vehicles, Vessels, and Other Conveyances,” Accessed May 13, 2016, http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_11.html.

25. U.S. Department of Transportation, Office of the Assistant Secretary for Research and Technology, Bureau of Transportation Statistics, “Public Road and Street Mileage in the United States by Type of Surface(a) (Thousands of miles),” Accessed May 13, 2016, http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/national_transportation_statistics/html/table_01_04.html.

26. U.S. Department of Transportation, Federal Highway Administration, Policy and Governmental Affairs, Office of Highway Policy Information, “Highway Statistics 2013: User’s Guide,” Accessed June 27, 2016, http://www.fhwa.dot.gov/policyinformation/statistics/2013.

27. Institute of Transportation Engineers, “National Traffic Signal Report Card, 2012,” Accessed at http://library.ite.org/pub/e265477a-2354-d714-5147-870dfac0e294.

28. U.S. Department of Transportation, Federal Highway Administration, Freight Management and Operations, Office of Operations, “Freight Analysis Framework,” Accessed June 27, 2016, http://ops.fhwa.dot.gov/freight/freight_analysis/faf/index.htm#faf4.

29. Schieber, Philip, “The Wit and Wisdom of Grace Hopper,” Accessed May 9, 2016, http://www.cs.yale.edu/homes/tap/Files/hopper-wit.html.

30. Gigaom, “Facebook is collecting your data—500 terabytes a day,” Accessed April 4, 2016,” https://gigaom.com/2012/08/22/facebook-is-collecting-your-data-500-terabytes-a-day/.

31. Domo, “Data Never Sleeps 3.0,” Accessed April 4, 2016, https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/.

32. Ghemawat, Sanjay, Gobioff, Howard, and Leung, Shun-Tak, “The Google File System,” Accessed April 4, 2016, http://static.googleusercontent.com/media/research.google.com/en//archive/gfs-sosp2003.pdf.

33. Apache, “What is Apache Hadoop?” Accessed April 4, 2016, http://hadoop.apache.org/.

34. Dean, Jeffrey and Ghemawat, Sanjay, “MapReduce: Simplified Data Processing on Large Clusters,” Accessed April 4, 2016, http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf.

35. Thomas, Gwen, “Defining Data Governance,” Accessed May 12, 2016, http://www.datagovernance.com/defining-data-governance/.

36. Nammari, Brian, “IoT, Social Media and their Monster Child called Big Data, What is next?” Accessed April 4, 2016, https://medium.com/@bnammari/iot-social-media-and-their-monster-child-called-big-data-what-is-next-899eba9f6b7b#.8rleu1q43.

37. Tepper, Allegra, “How Much Data is Created Every Minute?” Accessed April 5, 2016, http://mashable.com/2012/06/22/data-created-every-minute/#SAV6YUMJSmq7.

38. Domo, “Data Never Sleeps 2.0,” Accessed April 5, 2016, https://www.domo.com/learn/data-never-sleeps-2.

39. Federal Highway Administration, “Organizing for Operations,” Accessed May 12, 2016, http://www.ops.fhwa.dot.gov/plan4ops/focus_areas/organizing_for_op.htm.

40. Olavsrud, Thor, “21 data and analytics trends that will dominate 2016,” Accessed April 5, 2016, http://www.cio.com/article/3023838/analytics/21-data-and-analytics-trends-that-will-dominate-2016.html.

41. Woodie, Alex, “Why Gartner Dropped Big Data Off the Hype Curve,” Accessed April 6, 2016, http://www.datanami.com/2015/08/26/why-gartner-dropped-big-data-off-the-hype-curve/.

42. Sanjiv, K.R., “Big Data—Moving from the operational to the strategic,” Accessed April 6, 2016, http://www.wipro.com/documents/Wipro-analytics-big-data-moving-from-the-operational-to-the-strategic.pdf.

43. U.S. Department of Homeland Security, "Open Source Software in Government: Challenges and Opportunities," Accessed April 6, 2016, https://www.dhs.gov/sites/default/files/publications/Open%20Source%20Software%20in%20Government%20%E2%80%93%20Challenges%20and%20Opportunities_Final.pdf.

44. The Open Government Partnership, "Announcing New Open Government Initiatives," Accessed June 27, 2016, https://www.whitehouse.gov/sites/default/files/microsites/ostp/new_nap_commitments_report_092314.pdf.

45. The White House, Office of the Press Secretary, "Executive Order—Making Open and Machine Readable the New Default for Government Information," Accessed June 27, 2016, https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.

46. U.S. Department of Transportation, Federal Highway Administration, "Open Source Application Development Portal," Accessed June 27, 2016, http://www.itsforge.net.

47. U.S. Department of Transportation, Office of the CIO, "Privacy Impact Assessment—Federal Highway Administration (FHWA) Open Source Application Development Portal," Accessed June 27, 2016, https://cms.dot.gov/sites/dot.gov/files/docs/OSADP__FHWA_PIA_Adjudicated_082514.pdf.

48. Deloitte, "Cognitive analytics: The three-minute guide," Accessed April 6, 2016, http://public.deloitte.com/media/analytics/pdfs/us_da_3min_guide_cognitive_analytics.pdf.

49. Drury, Nicholas and Sarkar, Sandipan, "How Cognitive Computing Impacts Banks and Financial Markets," Accessed April 6, 2016, http://www.forbes.com/sites/ibm/2015/11/09/how-cognitive-computing-impacts-banks-and-financial-markets/#31e622e525e5.

50. International Transport Forum Corporate Partnership Board, "Big Data and Transport: Understanding and assessing options," Accessed April 6, 2016, http://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf.

51. Transportation Research Circular, "Improving Safety Programs Through Data Governance and Data Business Planning," Accessed June 27, 2016, http://onlinepubs.trb.org/onlinepubs/circulars/ec196.pdf.

52. Federal Highway Administration, "Data Governance Plan Volume 1: Data Governance Primer," Accessed May 14, 2016, https://www.fhwa.dot.gov/datagov/.

53. GetInData, "Geospatial analytics on Hadoop," Accessed April 7, 2016, http://getindata.com/blog/post/geospatial-analytics-on-hadoop/.

54. IDC, "IDC Reveals Worldwide Big Data and Analytics Predictions for 2015," Accessed April 7, 2016, http://www.idc.com/home.jsp.

55. Rossi, Ben, "Top 8 trends for big data in 2016," Accessed April 7, 2016, http://www.information-age.com/technology/information-management/123460615/top-8-trends-big-data-2016.

56. IT Business Edge, "Ten Reasons Why OpenStack Will Rule the Enterprise," Accessed April 7, 2016, http://www.itbusinessedge.com/slideshows/ten-reasons-why-openstack-will-rule-the-enterprise.html.

57. InformationWeek, “5 Early Cloud Adopters in Federal Government,” Accessed April 7, 2016, http://www.informationweek.com/government/cloud-computing/5-early-cloud-adopters-in-Federal-government/d/d-id/1315911.

58. General Services Administration, “Cloud IT Services,” Accessed April 7, 2016, http://www.gsa.gov/portal/content/190333.

59. Rouse, Margaret, “software-defined storage,” Accessed April 7, 2016, http://searchsdn.techtarget.com/definition/software-defined-storage.

60. Vekiarides, Laz, “5 bitter truths about software-defined storage,” Accessed April 7, 2016, http://www.infoworld.com/article/2997239/storage/5-bitter-truths-about-software-defined-storage.html.

61. Deloitte, “Data scientists: The three-minute guide,” Accessed April 6, 2016, http://public.deloitte.com/media/analytics/pdfs/us_ba_Deloitte3minDatascientist_021813.pdf.

62. North Carolina State University Institute for Advanced Analytics, “Degree Programs in Analytics and Data Science,” Accessed May 14, 2016, http://analytics.ncsu.edu/?page_id=4184.

63. Huddleston, Greg, “Cognitive Analytics Is Helping To Reduce Roadway Fatalities in Tennessee,” Accessed June 27, 2016, http://www.forbes.com/sites/ibm/2016/04/28/cognitive-analytics-is-helping-to-reduce-roadway-fatalities-in-tennessee/#18d7aded3dad.

64. Holecy, Miro, “Cognitive Computing Can Help to Meet Citizens’ Expectations from Transportation Services,” Accessed April 7, 2016, http://insights-on-business.com/government/cognitive-computing-transportation/.

65. Farber, Dan, “Twitter hits 400 million tweets per day, mostly mobile,” Accessed May 9, 2016, http://www.cnet.com/news/twitter-hits-400-million-tweets-per-day-mostly-mobile/.

66. Pagliery, Jose, “Half of American adults hacked this year,” Accessed May 14, 2016, http://money.cnn.com/2014/05/28/technology/security/hack-data-breach/.

67. Stonebraker, Michael, “The Case for Shared Nothing,” Accessed June 27, 2016, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.5370&rep=rep1&type=pdf.

68. DB-Engines, “System Properties Comparison Netezza vs. Oracle vs. Teradata,” Accessed May 12, 2016, http://db-engines.com/en/system/Netezza%3Boracle%3Bteradata.

69. Paret, Michelle, “Using the mean in Data Analysis: It’s Not Always a Slam-Dunk,” Accessed April 5, 2016, http://blog.minitab.com/blog/michelle-paret/using-the-mean-its-not-always-a-slam-dunk.

70. Harris, Derrick, “Why Apple, eBay, and Walmart have some of the biggest data warehouses you’ve ever seen,” Accessed May 14, 2016, https://gigaom.com/2013/03/27/why-apple-ebay-and-walmart-have-some-of-the-biggest-data-warehouses-you’ve-ever-seen/.

71. International Technology Group, “Cost/Benefit Case for IBM PureData System for Analytics: Comparing Costs and Time to Value with Teradata Data Warehouse Appliance,” Accessed May 13, 2016, https://tdwi.org/~/media/5BE30CAF543C4820A7139AAE81DA590F.PDF.

72. Amazon Web Services, “Amazon EMR,” Accessed May 9, 2016, https://aws.amazon.com/elasticmapreduce/.

73. Microsoft Azure, “What is Hadoop in the cloud? An introduction to Hadoop components in HDInsight for big data analysis,” Accessed May 9, 2016, https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-introduction/.

74. IBM, “Hadoop-as-a-service, big data analytics in the cloud,” Accessed May 9, 2016, http://www-03.ibm.com/software/products/en/ibm-biginsights-on-cloud.

75. Google Cloud Platform, “Cloud Dataproc,” Accessed May 9, 2016, https://cloud.google.com/dataproc/.

76. City of Boston, “Street Bump: Help Improve Your Streets,” Accessed May 13, 2016, http://www.cityofboston.gov/DoIT/apps/streetbump.asp.

77. Amazon Web Services, “AWS IoT,” Accessed May 9, 2016, https://aws.amazon.com/iot/.

78. IBM, “IBM Watson Internet of Things,” Accessed May 9, 2016, http://www.ibm.com/Internet-of-things/.

79. Cisco, “Internet of Things (IoT),” Accessed May 9, 2016, http://www.cisco.com/c/en/us/solutions/Internet-of-things/iot-products.html.

APPENDIX A. List of Acronyms

511	DOT Branded Traveler Information Service
aaS	As-a-service
ACID	Atomicity, Consistency, Isolation, and Durability
AMP	Access Module Process
ASCT	Adaptive Signal Control Technology
AWS	Amazon Web Services
BSM	Basic Safety Message
C2F	Center-to-field
CCTV	Closed-Circuit Television
COTS	Commercial-off-the-shelf
CRM	Customer Relationship Management
CV	Connected Vehicle
CVPD	Connected Vehicle Pilot Deployment
DHS	Department of Homeland Security
DMS	Dynamic Message Sign
DOT	Department of Transportation
DSRC	Dedicated Short Range Communications
ESS	Environmental Sensor Station
ETL	Extract/Transform/Load
FedRAMP	Federal Risk and Authorization Management Program
GFS	Google File System
GIS	Geographic Information Systems
GSA	General Services Administration
HDFS	Hadoop Distributed File System
HDP	Hortonworks Data Platform
IaaS	Infrastructure-as-a-Service
ICM	Integrated Corridor Management
IDC	International Data Corporation
IoT	Internet of Things
IP	Internet Protocol
IT	Internet Technology
ITG	International Technology Group
LCS	Lane Control Signal
LiDAR	Light Detection and Ranging
MAC	Media Access Control
MPP	Massively Parallel Processing
MS	Master of Science
MSA	Master of Science in Analytics
MSDS	Master of Science in Data Science
NCSU	North Carolina State University
NHTSA	National Highway Traffic Safety Administration
NOCoE	National Operations Center of Excellence
NoSQL	Not Only SQL
NTCIP	National Transportation Communications for ITS Protocol
OBU	On Board Unit
OEM	Original Equipment Manufacturer

OMB	Office of Management and Budget
OSADP	Open Source Application Development Portal
OSS	Open Source Software
PaaS	Platform-as-a-Service
PDM	Probe Data Message
PE	Parsing Engine
PeMS	Performance Monitoring System
PII	Personally Identifiable Information
PPP	Public-Private-Partnership
PTZ	Pan-tilt-zoom
RAID	Redundant Array of Independent Disks
RDBMS	Relational Database Management System
RFID	Radio Frequency Identification
RITIS	Regional Integrated Transportation Information System
RSU	Roadside Unit
RWIS	Road Weather Information Systems
SaaS	Software-as-a-Service
SCMS	Security Credential Management System
SDK	Software Development Kit
SDS	Software-Defined Storage
SNMP	Simple Network Management Protocol
SPU	Snippet Processing Unit
SRM	Signal Request Message
SSH	Secure Shell
THP	Tennessee Highway Patrol
TITAN	Tennessee Integrated Traffic Analysis Network
TMC	Traffic Management Center
TSM&O	Transportation Systems Management and Operations
UDOT	Utah Department of Transportation
U.S. DOT	U.S. Department of Transportation
V2I	Vehicle to Infrastructure
V2V	Vehicle to Vehicle
VDOT	Virginia Department of Transportation
VDS	Vehicle Detection Stations
VII	Vehicle Infrastructure Integration
VM	Virtual Machine
VSL	Variable Speed Limits
WAVE	Wireless Access in Vehicular Environment
WSM	WAVE Short Message
YARN	Yet Another Resource Negotiator

APPENDIX B. Solution Implementation Considerations

Given the wide range of functional capabilities of big data solutions currently on the market, it is very difficult to make concrete, scalable, and universal recommendations or assessments on these solutions in such a manner that will address the needs of the majority of Transportation Systems Management and Operations (TSM&O) organizations. With such overarching considerations as budget, scale, user needs, and talent to grapple with, the process of determining which big data solution should be deployed is a process and not an event. The questions below were drafted with the intention of guiding the decisionmaking process an Agency may have in order to define its big data needs before it works to solicit vendors.

These questions are intended to be representative, thought-provoking, and extensive, but not exhaustive. More will likely be developed and adapted based on the results of the Task orders to follow.

- Preplanning
  - What is/are the problem(s) you seek to address?
  - What is your target State?
  - What is your preliminary budget?
  - Are you comfortable with open source products, or do you prefer commercial-off-the-shelf (COTS) products?
  - What are your current barriers to adoption?
  - At what level of maturity would you classify your organization's technical capabilities?
  - Will your solution be built, managed, and maintained by internal staff, contractors, or a combination of the two?
  - What are you interested in doing with your current hardware, software, etc.?
- Acquisition
  - Where does your data currently come from?
  - Are you interested in acquiring any new data sources?
  - How much data will you be ingesting?
  - What format(s) is/are this data in?
  - What tools will you ingest it?
- Marshalling
  - Where do you plan on hosting your data, locally or in the cloud?
  - Will the data need to be preprocessed in a specific way (e.g., extract/transform/load (ETL), normalization, merging, imputation, etc.)?

- How often do you anticipate needing to scale your solution up or down?
- What level of latency are you comfortable with in accessing the data?
- Will your data sources provide you with structured, unstructured, and/or semi-structured data?
- How long will the data reside in your solution?
- What are your organization's security and data governance standards for the data collected, for example, where does the data need to go when you're finished with it?
- What is your archival procedure and governance strategy?
- What is the acceptable loss in functionality/availability of your system?
- What is the criticality of data loss, for example will your solution be a system of record for any data?

- Analysis
  - What are the current languages, analyses, and tools/technologies are you currently using?
  - How complex are your current analyses and how frequently do you currently process data?
  - What analytical and programming languages are you interested in using (e.g., Java, SAS, R, Python, Scala, etc.)?
  - What analytical skills are you interested in expanding or willing to expand to?
  - How advanced will your data analysis procedures be (e.g., machine learning, natural language processing, etc.)?
  - Do you intend to perform media (e.g., audio, video, imagery, etc.) analysis, text analysis, or a combination of both?
  - Do you want to maintain manual or automated control over your analytical algorithms and procedures?
  - How quickly do you need to perform certain analyses (e.g., real-time, 24-hour cycles, etc.)?
  - What is the criticality for interruption of analyses, for example how mission critical are the analyses you are performing?

- Action
  - What do you know about your end users (e.g., their challenges, technical skills, etc.)?
  - What is the intended experience for your end user?
  - What are the desired outcomes of the end user's tasks (e.g., will they impact other aspects of the organization or other users)?
  - What questions do they seek to answer, and how frequently do they seek to answer them?

- Where will your users access the data (e.g., local client or through the Web)?
- What new software or tools will your users need?
- What is the criticality if your end users can't perform their tasks?
- To what extent would you enable the end user to manipulate the data, if at all?

- General maintenance and security
  - What is your enterprise's current inventory of hardware and software?
  - Would you prefer a platform with onsite technical support, or are you interested in outsourcing much of the support capabilities?
  - How sensitive is the data you're interested in using and how will you control and restrict access?
  - What regulations and standards does your solution need to meet?
  - How frequently will you need data and product lifecycles to reset?
  - How rigorous of a backup and recovery process will you require?

U.S. Department of Transportation
ITS Joint Program Office-HOIT
1200 New Jersey Avenue, SE
Washington, DC 20590

Toll-Free "Help Line" 866-367-7487
www.its.dot.gov

FHWA-JPO-16-424

U.S. Department of Transportation