DOT/FAA/TCTN-23/55

Federal Aviation Administration
William J. Hughes Technical Center
Strategic Initiatives Division
Atlantic City International Airport
New Jersey 08405

Literature Survey of Big Data

Technical Note

July 27, 2023

The research described in this report was funded by the FAA as part of its mission to improve aircraft safety. The views and opinions expressed are those of the author alone and do not necessarily represent the views of the FAA. The FAA assumes no liability for the contents or use thereof. The FAA has not edited or modified the contents of the report in any manner.

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's Full-Text Technical Reports page: actlibrary.tc.faa.gov in Adobe Acrobat portable document format (PDF).

Technical Report Documentation Page

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

1. Report No. DOT/FAA/TCTN-23/55	2. Government Accession No.	3. Recipient's Catalog No.

4. Title and Subtitle Literature Survey of Big Data	5. Report Date July 2023
	6. Performing Organization Code N/A
7. Author(s) Jason McGlynn	8. Performing Organization Report No. N/A
9. Performing Organization Name and Address U.S. Department of Transportation Federal Aviation Administration, William J. Hughes Technical Center Atlantic City International Airport, NJ 08405	10. Work Unit No. (TRAIS)
	11. Contract or Grant No. N/A
12. Sponsoring Agency Name and Address U.S. Department of Transportation Federal Aviation Administration NAS Systems Engineering and Integration Office Strategic Initiatives Division ANG-B6 Atlantic City International Airport, NJ 08405	13. Type of Report and Period Covered N/A
	14. Sponsoring Agency Code N/A

15. Supplementary Notes

16. Abstract

Mention the topic of big data, and a person is bound to experience information overload. Indeed, it is so complex with so many terms and details that people want to run away from it.

When used right, big data (BD) will help people access data they need in in real time and help managers make better decisions.

The purpose of this paper is to evaluate methods, procedures, and architectures for the storage and retrieval of all Federal Aviation Administration (FAA) research, engineering, and development (RE&D) data sets, to leverage on the technology innovation and advancement opportunities in the field of BD analytics. The paper also discusses all relevant Executive Orders (EOs), laws, and Office of Management and Budget (OMB) memorandums that were written to address what federal agencies under the OMB's jurisdiction must do to comply with various aspects of BD.

17. Key Words Big Data, storage, ingestion, ETL, security, infrastructure, policies, laws, metadata, batch processing, stream processing, tools, services, best practices, recommendations, architecture, analytics, machine learning	18. Distribution Statement This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc.faa.gov.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 108	22. Price

Contents

Appendices:

Figures

Tables

Executive summary

Big data (BD) is a new technology that is revolutionizing the way the world does business; it is evolving and it is responsible for giving people new insights. Those who understand and use it wisely will be able to make informed business decisions. However, because it is happening so quickly, it is hard to keep up with it all. Because the Federal Aviation Administration (FAA) collects so much data and needs to analyze it in real time to make informed life-saving decisions, it is vital for it to be apprised of the best and latest BD practices and methods.

This paper provides a literature review that examines multiple areas of BD. It is divided into two distinct parts.

1. <u>Best Practices and Methods in the BD Lifecycle.</u>

The first part examines the BD lifecycle and seeks to help users understand the best practices and methods used in each phase of the BD lifecycle. The BD lifecycle begins by collecting data from various sources and ingesting the raw data into a centralized data repository on the cloud. This is a time-consuming process known as Extract, Transform, and Load (ETL). ETL is a data ingestion tool that is responsible for not only producing clean, error-free data, but for automatically generating metadata which is vital for conducting searches. About 60-80% of scheduled time in any analytics project is spent on data ingestion, which consists of collecting and cleansing data. Once data is cleaned, users can access data in real time from the centralized data repository, depending on user credentials.

Before the advent of the cloud, data was stored in a data warehouse. This was fine, but it got very expensive because of maintenance and a constant need to upgrade servers. There was also a fear of running out of space and if recovery efforts would work. Now that data can be stored in the cloud, addressing these issues is the responsibility of the Cloud Service Providers (CSPs).

Once the data is in the cloud, security and privacy become major concerns. While it is not possible to eliminate all security and privacy risks, organizations must now develop policies to address them. In a cloud environment, it is vital that all sensitive data files be encrypted at every stage of the life cycle. Encryption is the main tool used to store and query BD safeguards. Because of confidentiality policies and procedures, it is also imperative to restrict data access to authorized users only. This enables data integrity with inherent mechanisms for access control. Since this can be a headache to manage, some organizations ask a CSP to manage it for them. Before going down this route, organizations should clearly identify the responsibility and expectations for each party.

With all the background work done, this paper discusses three types of BD analytics: 1) data mining, 2) machine learning (ML), and 3) data visualization. Accordingly, this paper describes the basic idea of each of these three types and how the FAA can use each.

2. Evolution of BD Laws Governing the FAA

Part 2 examines the many Executive Orders (EOs), laws, and Office of Management and Budget (OMB) memorandums that have been written over the years to address what federal agencies under the OMB's jurisdiction must do to comply with various aspects of BD. To help the FAA get a better understanding of what actions need to be taken to comply with these laws, a timeline of all EOs, laws, and OMB memorandums written concerning the various aspects of BD has been developed. After constructing this timeline, Table 10 (in Section 6) was created to summarize the EOs, laws, and OMB memorandums with which the FAA needs to comply.

This BD literature review closes by making recommendations to the FAA on the best practices it can utilize at each phase of the BD life cycle. Table 10 provides a summary of the EOs, laws, and OMB memorandums required for FAA compliance.

1 Introduction

1.1 Purpose

Mention the topic of big data, and a person is bound to experience information overload. Indeed, it is so complex with so many terms and details that people want to run away from it.

When used correctly, big data (BD) will help people access data they need in in real time and help managers make better decisions.

The purpose of this paper is to evaluate methods, procedures, and architectures for the storage and retrieval of all Federal Aviation Administration (FAA) research, engineering, and development (RE&D) data sets, to leverage the technology innovation and advancement opportunities in the field of BD analytics. The paper will also discuss all relevant government policies and laws that govern the management of BD.

1.2 Background

The book, “Big data: A Revolution That Will Transform How We Live, Work, and Think,” by Kenneth Cukier and Viktor Mayer-Schonberger describes a phenomenon that is literally taking the world by storm with its endless possibilities. This revolution, which is impacting the way industry, academia, and government agencies do business, is happening at such a rapid rate that it is hard to keep up with it.

It is a fact that data are power. The government collects a plethora of it. There are many benefits to collecting and analyzing data effectively including: improving decision-making and productivity, identifying and reducing inefficiencies, and reducing costs. We live in an age where budgets are tight and we need to do more with less. Accordingly, this paper will seek to identify ways the government can use all FAA generated or used RE&D data sets in the most efficient and effective manner possible.

1.3 Scope

This technical note addresses the following topics:

- Survey the variety, complexity, and characteristics of FAA research and engineering data in both its production and use.
- Compile, synthesize, and collate the methods, procedures, processes, technologies, and architectures that other groups within the FAA and external institutions (i.e., other

government agencies, industry, and academia) are currently using or developing to store and retrieve their research, engineering, and development or similar data.

- The Department of Transportation (DOT) and FAA guidance and policy, law, and executive orders that are relevant for data management of research data.
- The results of the information compiled, references utilized, and conclude with recommendations to meet FAA research data management technical and policy requirements.

2 FAA research and engineering data

The FAA collects a plethora of data in many formats for many reasons. Some data are structured (i.e., flight track, National Airspace System [NAS], mission support, weather, air traffic management, system, simulation [from studies and experiments], research, field, and operational data) and can fit nicely into a database.

The Fire Safety Research Branch collects structured data that is usually kept in excel files, scattered about the various personal computers (AIT) of our researchers or stored on our LabNet servers (we have multiple Redundant Array of Inexpensive Disks (RAID) drives dedicated to data). The data files are typically output from data acquisition modules that acquire signals from measurement devices in our experiments (temperature, pressure, velocity, heat flux, flow, etc.). Some experiments can have very small data files (~10 kb). Others have very large data files, for instance output from forward-looking infrared (FLIR) cameras and particle image velocimetry (PIV) files, on the order of several gigabytes per test.

Much of the data from fire experiments is in the form of video output from high-definition (HD) cameras. The raw video files from tests are stored on RAID drives, and this is only the video that we acquire ourselves, often times we have Advanced Imaging providing video support on top of our own cameras.

The FAA also collects unstructured data (i.e., video, websites, images, emails, pdfs, office docs, and research). Because data resides on servers, hard drives, knowledge sharing networks (KSN), shared drives, filing cabinets, workstations, and in silos (some of which are deemed proprietary), it is very difficult to manage, find, and access.

Much of our data is unstructured. If you look at our website, https://www.fire.tc.faa.gov/ you will see that the entire website contains various data elements – reports, presentations, PDFs, and several databases.

The MITRE Corporation was hired to assess the FAA's current and planned data provisioning capabilities. They interviewed stakeholders in the areas of air traffic management, post operational analysis, and investment analysis to identify and gain an understanding of data shortfalls. Four key shortfalls were identified:

1. Data are not always systematically collected, stored, and disseminated in a manner that allows for accessibility.
2. Data may be in its native use format and not in standardized ways that facilitate interpretation, understanding, and consistent use by others.
3. Adequate capabilities to process, analyze, and extract meaningful information from data are not universally available.
4. A lack of governance and organizational challenges may result in inadequate management of data."[1]

A key issue this report seeks to address is how to get all of this data both structured and unstructured in different formats into a centralized database so that employees can perform analytics and leverage the data to more informed decisions.

3 Data management life cycle

Given the various types of FAA data, the central question we need to answer is how to use BD to our advantage and to use the right tools to analyze the data to help managers make the best-informed decisions possible?

Table 1 shows the BD management life cycle or "…the sequence of stages that a particular unit of data goes through from its initial generation or capture to its eventual archival and/or deletion at the end of its useful life."[2] This paper discusses the BD management life cycle stages and presents the various tools that are available today to help external institutions (i.e., other government agencies, industry, and academia) store, retrieve, and analyze their RE&D data sets.

Table 1. Modernized data management life cycle

DATA INGESTION	**EXTRACT, TRANSFORM, AND LOAD (ETL)**	**STORAGE**	**DATA ANALYTICS**
- Batch Processing - Stream Processing	Cleans data and puts it in the right format	- EDW - RDBMS - Cloud	- Data mining - Machine Learning - Data Visualization

3.1 Data ingestion

Data ingestion is the process of dumping raw data from sources into a data lake, data hub, or storage repository. The ingested data comes from many file types, real-time data ingestion, and from batches. The four data types that can be ingested include:

1. Structured data – data that has a defined length and format with clearly defined data types whose pattern makes them easily searchable, such as relational databases, logs, and financial data.
2. Unstructured data – the Internet of Things (IoT), social media, video/audio recordings, customer channels, and external sources, such as partners and data aggregators. It is stored as different file types and is difficult to search.
3. Semi-structured data – contains semantic tags, but does not conform to the structure associated with typical relational databases, i.e., an MS Word file with metatags (see document properties in a Word file). Files are typically HTML, XML, RDF, or CSV [3].
4. Quasi-structured data – textual data with erratic data formats.

Once the data gets into the data lake, it is time to clean the data. This involves ensuring the accuracy of data (correcting or deleting incomplete, incorrect, inaccurate or irrelevant parts of the data from a record set, table, or database) and putting it into a common format so it can be processed and analyzed. Data ingestion tools perform transforming, mapping, and cleansing data, and they can be integrated with data governance and data quality tools [4]. The tools can also be used to help companies modify and format data for analytics and storage purposes [5].

Data ingestion consists of collecting and cleansing data. It takes 60-80% of scheduled time in any analytics project. Data scientists spend most of their time wrangling with this process of producing clean data. Dirty data can lead to inaccurate data analytics results and drive misguided decision making. Many enterprises begin data analytics projects without understanding how important it is to ensure the accuracy of data [6].

It can be quite challenging to choose the right file format when ingesting data because it can cause a data drift. “Data drift is a natural consequence of the diversity of big data sources. The operation, maintenance and modernization of these systems causes unpredictable, unannounced and unending mutations of data characteristics.” [7] Data drift is said to be the killer of data integrity and fidelity. “Data drift creates serious challenges for businesses looking to fully harness the insights available from big data.” [8] Furthermore, if data drifts are not addressed, it will cripple data analytics.

Before ingesting data into the data lake, it is a good practice to ensure that data are in the right format so there are not problems. Having data in the wrong format can be expensive to fix and can lead to data losses or bottlenecks in the overall architecture.

Therefore, the solution must address data drifts.

There are three types of data drifts:

1. Structural drift – when the data schema changes at the source. Common examples of structural drift are fields being added, deleted, re-ordered, or the type of field being changed.
2. Semantic drift – when the meaning of the data changes, even when the same structure.
3. Infrastructure drift – when changes to the underlying software or systems create incompatibilities [9].

If drift issues are not addressed, it can negatively affect your data lake. The following three questions may help:

- What is the most frequent way in which your data will be accessed?
- How is your data being stored at source (structured or unstructured)?
- What kind of partitioning needs do you have to take into account?[10]

While it is impossible to avoid data drift altogether, one should do the following to reduce the risks:

- Prototype your models before scaling them.
- Identify decision steps where your model will face multiple outcomes.
- Evaluate both the objective of the model and its assumptions.
- Cross-validate your data to optimize your solution.
- Ensure that all data come from the same time period. [11].

Batch processing and real-time (or stream) processing are the two ways to ingest data. These processes are described below.

3.1.1 Batch processing

"The ingestion layer periodically collects and groups source data and sends it to the destination system." [12] Although batch processing is cheaper and used when having near-real-time data is not important, it should be adequate to handle FAA RE&D data sets. Table 2 shows a list of batch processing tools that can be used to ingest data into a data lake.

Table 2. Batch processing ingestion tools for structured & unstructured data

Solution	**Developer**
Hadoop MapReduce	Apache
Apache Spark	UC Berkeley AMPLab
DISCO	Nokia
HPCC Systems	LexisNexis

3.1.2 Stream processing

Real-time (or stream) data ingestion means importing the data as it is produced by the source. It is more expensive because systems must constantly monitor sources and accept new information. Table 3 shows a list of stream processing tools that can be used to ingest data into a data lake.

Table 3. Stream processing ingestion tools

Solution	**Developer**
*Apache Nifi	Apache
Apache Kafka	Confluent
Apache Flume	--
Apache Storm	Twitter

*Used in Enterprise Information Management (EIM)

To determine the best data ingestion tools for your environments and needs, you need to answer the following questions:

- What kind of data will you be dealing with (internal/external, structured/unstructured, operational, etc.)?
- Who is going to be the key stakeholder of the data?
- What is your existing data management architecture?
- Who is going to be the data steward?
- Main question – Do we have the in-house skill set to carry out this migration successfully? [13]

The challenges of the data ingestion process include:

- Compromised compliance and data security regulations, making it extremely complex and costly. Verifying data access and usage can be problematic and time-consuming.

- Detecting and capturing data are a mammoth task owing to the semi-structured or unstructured nature of data and low latency.
- Improper data ingestion can give rise to unreliable connectivity that disturbs communication outages and result in data loss.
- Enterprises ingest large streams of data by investing in large servers and storage systems or increasing capacity of hardware along with bandwidth that increases the overhead costs [14].
- It takes a long time to ingest large tables with billions of rows and can come in many formats.
- Different data formats.
- Changes to the data sources schema.
- Incomplete and inaccurate data.

Since data ingestion is a laborious process, there are many reasons that businesses want to automate it as much as possible. These reasons include:

1. Improves time-to-market goals.
2. Increases scalability.
3. Frees up people's time to do analysis.
4. Mitigates risks.
5. Reduces errors.
6. Saves time and money [15].

3.2 Extract, Transform, and Load (ETL)

ETL is a data ingestion tool that includes the following three features:

1. Extracting – retrieving raw data from sources.
2. Transforming – cleansing and normalizing data and putting it into a format so it can be used by different applications. Tools do not work well with unstructured data.
3. Loading – placing the data in a database to be analyzed.

ETL serves two main functions, as follows:

1. Main functionality – act as a tool that is data cleansing oriented or data transformation oriented, or performs both functions.
2. Metadata support – "Metadata are structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource."[16] Metadata are data in context. An ETL is also responsible of using metadata to map source

data to destination. Thus choosing a tool that conforms to organizations metadata strategy is very important [17].

3.2.1 ETL main functionality

“ETL extracts raw data from disparate source systems (e.g., CRM software, inventory software, e-commerce applications, and web analytics), transforms all this data into a usable format suitable for querying and analysis, before finally loading it into a target system, which is typically a data warehouse, but could be any data repository.” [18]

ETL reduces the data ingestion process and produces faster results. While the current system works based on rules instead of code, it is often outdated. The ETL process performs a number of important functions to better organize and understand data. When selecting a cloud-based ETL tool, the buyer should consider the following factors:

- Use cases – the variety of use cases
- Data source – will the ETL work with different types of data (structured, unstructured, and high-dimensionality)
- Capabilities – flexibility to read and write data regardless of where it is and if the tool will let you quickly switch providers easily
- Integration – consider scope and frequency of integration efforts
- Business user – consider if the user understands about transforming data
- Budget – ETL reduces costs
- Business goals – consider the business needs when selecting an ETL tool

When selecting an ETL tool, an agency should consider the following:

- The ability to connect to and extract data from a variety of sources -- databases of all stripes (relational, NoSQL, etc.), BD systems built around technologies such as Hadoop and Spark, flat file repositories, application-to-application message queues, and more.
- A graphical user interface (GUI)-based design environment that supports drag and drop development of source to target mappings and ETL workflows.
- Team-based development capabilities for collaborating on integration projects, with associated version control and release management features.
- Basic data transformation functions, such as data type conversion, date reformatting and string handling, plus data mapping and workflow orchestration capabilities.
- Built-in data profiling software that can analyze source data for consistency, dependencies and other attributes before beginning the ETL process.

- Data quality and cleansing functionality for identifying and fixing errors in data sets, plus data synchronization for keeping data consistent in source and target systems.
- Metadata management support for synchronizing integration processes and documenting data transformation and business rules.
- A job scheduler, along with process management controls that provide things like runtime monitoring and error alerting, handling, and logging [19].
- The type of tasks, as the importance of certain functionalities over others will vary.
- The type of connections, as the ETL solution must be able to connect to other applications. Without this functionality, the processing power of the tool is useless [20].

"With traditional databases and batch ETL processes, performing analytical queries on huge volumes of data are time-consuming, complex, and cost intensive."[21] The main problem is that ETL is better suited for importing data from structured files or source relational databases into another similarly structured format in batches. It does not work for real time, streaming data. Batch ETL tools would work well for R&D data.

3.2.2 ETL metadata support

Extracting and loading data are easy. However, once the data gets into the data lake, the real work begins. When a user wants to find data in the data lake, data needs to have metadata that is keyword driven. Metadata are critical for conducting searching, for accurate analytics, and is the key to answering many questions. It "… identifies the attributes, properties and tags that will describe and classify information. It would be more appropriately defined as 'information about data.' It is represented in the form of any number of characteristics associated with the data information asset such as type of asset, author, date originated, workflow state, and usage within the Enterprise, among numerous others." [22] "Without business metadata, nontechnical users cannot search for data using business terms, work quickly, independently, and collaboratively, and get full value from a data lake."[23] Thus, it becomes a data swamp—an undocumented and disorganized data clutter that is nearly impossible to navigate, trust, or leverage for organizational advantage. Thus, "…an essential capability for developing and maturing BD processing services is to establish a comprehensive enterprise metadata management program." [22]

As part of M-13-13 Memorandum for the Heads of Executive Departments and Agencies on how the make data public, Federal agencies are required to publish metadata for all new information and collection efforts. It requires federal agencies to publish their information as machine-readable data, using searchable, open formats. It requires every agency to maintain a centralized

Enterprise Data Inventory that lists all data sets, and also mandates a centralized inventory for the whole government on data.gov [24].

The article titled, “Department of the Interior Metadata Implementation Guide—Framework for Developing the Metadata Component for Data Resource Management,” lays out a comprehensive plan to manage metadata effectively. The article says that a metadata implementation plan should identify key roles and responsibilities associated with metadata management processes and procedures. Following the principles below will help a business maximize the value of their data by documenting an integral business asset, sharing data, and improving the ability to search, discover, and use data. The guide will also lead to a better understanding of data quality, provide clarity of data relations, and identify redundancies in data, support data lifecycle management, and promote interoperability.

These goals are achieved through a series of actions defined in three major metadata implementation phases:

Phase 1: Getting started—Plan and organize

The article suggests that before data and metadata can be incorporated into any information system, metadata needs to be managed. Accomplishing this requires: 1) prioritizing an inventory of data assets accompanied by a listing of corresponding metadata records that meets bureau or office organizational needs, as well as its level of completion and 2) a gap analysis report that describes the status of each dataset, its associated metadata record and metadata type(s), and recommendations on any metadata improvements.

To determine what metadata are needed to adequately document data to meet organizational needs, an assessment of the current level of documentation that exists for the metadata types is needed. The information contained in metadata elements supports different needs and requirements depending on the audience. There are three types of metadata and many questions need to be asked:

1. Business Metadata – “categorizes the definition of business rules that apply to attribute properties.”
2. Technical Metadata – database’s physical characteristics – the actual definitions of fields, such as table and column names, in a database system.
3. Operational Metadata – administrative metadata. It’s about data file size, time of last load, and references to procedural scripts.

Phase 2. Create, Maintain, and Publish Metadata

Establish a repeatable approach to operationalize enterprise metadata management and to meet requirements for publishing metadata. This consists of the following tasks.

- Creating standards-based metadata including geographic location, keyword tagging, etc. for discovery, access, and use.
- Completing business, operational, and technical metadata (Digital Object Identifiers [DOIs]).
- Writing consistent data definitions (data dictionary) and data domains across the organization (even expanded to communities of interest).

Phase 3. Improving Metadata Management

Evaluate metadata implementation, metadata quality, and data architecture for data hosting, data services, metadata catalog, and overall metadata management across all functional areas. Develop user statistics to measure progress and success of metadata implementation through metrics.

Data governance has a data steward whose job it is to be "concerned with the meaning of data and the correct usage of data." https://blog.idatainc.com/data-governance-roles They are responsible to do the following tasks: define the schema and cleansing rules, decide which data should be ingested into each data source, and manage the treatment of dirty data.

Metadata should be developed continuously throughout the entire data lifecycle. This means that metadata should be generated along side the data.

The following attributes contribute to making a high quality data record:

- There needs to be consistency in certain fields. Some metadata fields require values that are written the same way to aid in machine-readability and discovery.
- Theme keywords should be drawn from thesaurus or keyword lists.
- Spell out acronyms with multiple meanings.
- Use appropriate profiles and extensions.
- Critical information for Content Standard for Digital Geospatial Metadata (CSDGM) records includes identification information, entities and attributes, data quality, access, use and liability constraints, distribution, and spatial references.
- Fields (i.e., attribute accuracy, logical consistency, completeness, and horizontal and vertical positional accuracy) describe the quality of your data. Provide as much detail in these sections as possible.

- Citations for all source data used in the dataset should be included [25].

3.2.2.1 Mistakes associated with Metadata

Twenty metadata experts revealed the five most common mistakes organizations make when it comes to metadata management, which are described below.

1. Not analyzing metadata requirements: Defining what objectives and goals need to be accomplished is the first step, and a very vital one, to have a clear picture in mind about the future scope and applications. Having measurable and definable objectives is imperative, and it will guide the course of the metadata management process.
2. Selecting metadata tools without conducting a thorough analysis of the options available: This requires an in-depth analysis in order to understand the different offerings and their pros and cons.
3. Not having a repository team responsible for metadata control: A repository team performs two very essential tasks: it collects and collates the data, and it provides access to the existing data. The absence of a dedicated repository team results in data mismanagement.
4. Not going for automation: Maintenance of the metadata also becomes a struggle over time. As the organization grows, so does the data inflow and outflow, making it very difficult to maintain manually.
5. Difficulty in accessing the metadata repository: If data access and retrieval requires too much effort, it defeats the purpose of metadata management [26].

The following are tips for selecting the right application performance monitoring tool:

- Conduct a thorough analysis of an organization's requirements.
- Consider different solutions and understand their strengths and weaknesses.
- Account for licensing costs, consulting fees, training, security and ancillary factors [27].

3.2.2.2 Metadata standards

Federal agencies are mandated to follow metadata standards outlined in Executive Order (EO) 12906, Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure endorsed by the Federal Geographic Data Committee (FGDC). These metadata records must comply with one of the FGDC-approved standards described below:

- CSDGM is a consensus limited to U.S. Federal agencies. Metadata are written in a plain text document.
- International Organization for Standardization (ISO) series of standards (19115). There is an ongoing effort to move towards adopting the ISO metadata standard, which will offer

updated tools, training materials and guidance documents to support ISO metadata implementation. Metadata will be using the Unified Modeling Language (UML), a standardized general purpose modeling language often used in software engineering [28].

3.2.3 Metadata strategy

A metadata strategy helps improve metadata governance across your agency. It documents current and future-state practices, as well as how your agency manages its metadata.

Moreover, it links to your broader data and information governance environment, including your information governance framework.

It is important to have a metadata strategy because it details which metadata are important to an organization, why it is important, and how the metadata will be captured, stored and used (by technical and business staff) – all at a strategic level [29].

A metadata strategy helps achieve the data orientation goals focusing on sharing data assets of an organization. It recognizes the value of data and its components and usage within and throughout the organization.

A metadata strategy serves as a map for managing the expanding requirements for information that the business places upon the IT environment. It also highlights how important a central data administration department is for organizations who are concerned about data quality, integrity, and reuse.

Developing and implementing a metadata strategy helps an organization measure the value of their information assets. Purchasing or developing a metadata storage facility can be an impetus for creating and implementing a Metadata Strategy [30].

A sound metadata strategy ensures seamless sharing, exchange and integration of tools and repositories and consists of the following building blocks:

- Identify the technical members most knowledgeable about the data.
- Identify the business area members most knowledgeable about the business data and business processes.
- Have business users identify the data stewards and data advocates for data and processes.
- Determine sources of record for data within an enterprise.
- Determine the usage of metadata.
- Identify all sources of metadata, including data models, rose models, CASE tools, silo repositories, data dictionaries, glossaries, third party dictionaries, business rules, data

mapping documents, business process mappings, data flow diagrams, corporate abbreviations, Microsoft Excel spreadsheets, catalogs etc.

- Determine the overall architecture for metadata storage. A metadata storage architecture defines an organization's overall strategy for metadata storage and retrieval of metadata. A centralized repository pulls in all the metadata in one location, a distributed architecture creates many small focused repositories and a hybrid architecture combines the two for best results.
- Determine integration points and processes necessary to consolidate and integrate the metadata.
- Determine metadata reporting and dissemination strategies.
- Evaluate full lifecycle metadata management tools.
- Standardize and document metadata sourcing processes.
- Determine how to keep the metadata up-to-date. This has always been a challenge in organizations. It is important to keep the metadata refreshed as obsolete or stale metadata can result in wrong decision-making.
- Standardize and document the metadata change management process and procedure.
- Define metadata security needs.
- Define components of the meta model.
- Identify issues and constraints.
- Form a metadata steering committee.
- Discuss and confirm the metadata management strategy with the management [31].

3.2.4 How Amazon Web Services (AWS) performs ETL

Since the FAA uses Amazon Web Services (AWS) for the cloud, it is vital to understand how AWS performs ETL. AWS has a managed ETL service called AWS Glue. The service generates ETL jobs on data and handles potential errors, to move data from source to destination using Apache Spark. ETL scripts from Glue can handle both semi-structured and structured data but not unstructured data.

AWS Glue service has three components, as follows:

1. Data Catalog – a managed service that lets you store, annotate, and share metadata in the AWS Cloud in the same way you would in an Apache Hive metastore. Each AWS account has one AWS Glue Data Catalog.
2. Job Authoring – this enables AWS Glue to generate Python/Scala code to move data from its source to its destination using the underlying Spark implementation. With the latest updates, Glue now supports running Scala Spark code.

3. Job Execution – serverless job executions. Developers do not need to deploy, configure, or provision servers for AWS Glue. Jobs automatically run in a Spark environment [32].

3.3 Storage

Since 2016, the FAA complied with Office of Management and Budget (OMB) M-19-19 – Federal Data Center Consolidation Initiative (FDCCI) and fulfilled the data center requirements of the Federal Information Technology Acquisition Reform Act (FITARA). M-19-19 rescinds and replaces the M-16-19 Data Center Optimization Initiative (DCOI). DCOI is part of the federal cloud computing strategy known as Cloud Smart that seeks to cut costs by closing and consolidating data centers and preventing new ones from being built.

BD consists of terabytes (or even petabytes) of information that can be a combination of structured data (databases, logs, SQL etc.) and unstructured (social media posts, sensors, and multimedia, etc.) data.

The ideal BD storage system allows an unlimited amount of data to be stored. It would feature high rates of random write and read access, and the flexibility and efficiency to deal with a range of different data models. It would also be able to store both structured and unstructured data, and for privacy reasons, only work on encrypted data.

Storage solutions need to address the velocity and variety of data. Velocity is important in the sense of query latencies, i.e., how long does it take to get a reply for a query?

Relational database management systems (RDBMS) were good at performing data management operations on structured data. The database processes data over an auxiliary server and uses specialized data manipulation languages and schema. Unfortunately, because of the characteristics of BD, relational databases cannot handle the extensive volume of data for storage with dynamic growth, frequent changing schema, complex data structures, more concurrent access needs, frequent input/output (I/O) needs, real-time processing needs, and consistency of a large number of storage servers. Currently, the FAA and EIM are using Oracle and PostgreSQL (aka Postgres) as their RDBMS.

Since RDBMS cannot handle unbounded volumes of data, there was a scale-up to scale-out that led to an explosion of new BD storage systems. The new technologies are not as stringent on data consistency in an effort to maintain fast query responses with increasing amounts of data. Additionally, there are problems storing metadata in RDBMS such as rolling back a transaction, recovering from a data corruption event or getting a reasonable-sized answer set for a given query [33].

Before the cloud existed, the Enterprise Data Warehouse (EDW) was comparable to a data lake. Employees with the proper user access could obtain the data needed to do the job. The purpose of the EDW was to integrate data across operational systems, which may be in operational silos and geographically distributed [34].

3.3.1 Big data storage technologies

Currently, there are a variety of BD storage technologies, as follows:

1. Distributed File Systems can store large amounts of unstructured data on commodity hardware. There are file systems with better performance, but Hadoop Distributed File System (HDFS) is an integral part of the Hadoop framework and has already reached the level of a de-facto standard.
2. Not only SQL (NoSQL) databases are the most important family of BD storage technologies. Compared to relational databases, they can handle all kinds of data sets and perform queries. They offer continuous availability and provide strong replication abilities along with read and write-anywhere with full location-independence support. It also doesn't need a specific caching layer to store data [35]. Therefore, NoSQL databases are better for storing metadata.

   NoSQL databases emerged as an enterprise solution to handle real-time data, BD, and a variety of data structures and models (i.e., document databases – MongoDB and CouchDB, graph databases –Apache Cassandra, key-value stores – Redis and Couchbase Server, Cache Systems– Redis and Memcache, Graph databases – Neo4) [36]. NoSQL databases are open sources. They are horizontally scalable with more clusters at lower costs, flexible schemas for unstructured data, high availability, accessibility, and are useful for large ever-changing data sets [37]. Redis and Neo4j are examples of NoSQL databases.

   BD consumes unbounded volumes of data. Hadoop-based solutions are offered by vendors such as Cloudera (2014a), Hortonworks (2014), and MapR (2014), as well as various NoSQL2 database vendors, in particular those that use in-memory and columnar storage technologies. Unlike RDBMS, that rely on row-based storage and expensive caching strategies, these BD storage technologies offer better scalability at lower operational complexity and costs.

3. NewSQL databases are modern relational databases that aim to provide the scalability of NoSQL databases while maintaining the transactional guarantees made by traditional

database systems. Stay away from locking records and process transactions in timestamp order and have concurrency control. Rely on memory heavily [38].

New SQL databases are helpful from a practical standpoint, how can an organization pick the best NoSQL database to meet their needs? In 2000, Eric Brewer introduced the Consistency, Availability, and Partition Tolerance (CAP) Theorem, which is essential in BD. To evaluate the performance of different storage architectures, organizations must compare and analyze the existing approaches using Brewer's CAP Theorem. A networked shared-data system can only have two out of the following three properties:

- Consistency – Asks if it is okay to have old data for a few seconds.
- Availability – Asks if it is okay if the system goes down for a few seconds or minutes.
- Partition Tolerance – Asks if one server is down, whether another server can handle transactions [39].

Dark data refers to the challenges companies face when dealing with most unstructured data. Unstructured data are said to comprise 80% of enterprise storage, and that percentage will increase. Many cloud storage vendors have emerged with technologies that combine virtually limitless storage repositories along with a varied set of metadata-enrichment capabilities that organizations need to understand and consider during the product-selection process.

3.3.2 Cloud storage types

Along with the different BD technologies, organizations should consider which of the following three types of cloud storage would best suit their needs.

1. Public Cloud Storage – a multi-tenant storage environment that is most suited for unstructured data on a subscription basis. Data is stored in the service providers' data centers with storage data spread across multiple regions or continents. The FAA Cloud Services (FCS) public cloud is operated by AWS.
2. Private Cloud Storage – in-house storage resources deployed as a dedicated environment protected behind an organization's firewall. Internally hosted private cloud storage implementations emulate some of the features of commercially available public cloud services, providing easy access and allocation of storage resources for business users, as well as object storage protocols. Private clouds are appropriate for users who need customization and more control over their data, or who have stringent data security or regulatory requirements.

3. Hybrid Cloud Storage – a mix of private and public cloud storage services with a layer of orchestration management to integrate operationally the two platforms. The model offers businesses flexibility and more data deployment options [40].

3.3.3 Block storage vs. Object-based storage

"Block storage systems are used to host databases, support random read/write operations, and keep system files of the running virtual machines. Data is stored in volumes and blocks where files are split into evenly sized blocks." [41]

Object-based storage is ideal for storing unstructured data. Object-based storage essentially clumps data, metadata tags, and a unique ID together to make one object. The metadata are customizable, which means you can input a lot more identifying information for each piece of data (i.e., content, retention, data protection, security, and other types of information the storage system accesses). The metadata of each object is used for indexing and searching.

When object-based storage is used, users can quickly find objects [42]. There is also a lot of data protection because it uses erasure coding, multi-copy mirroring, or a combination of the two. Objects are stored in a flat address space, which makes it easier to locate and retrieve data across regions.

There are four major benefits to using object-based storage:

1. Scalability – Object storage is known for its compatibility with cloud computing because of its unlimited scalability. Thanks to its flat structure, object storage does not have the same limitations as file or block storage. With object storage, scaling out and adding nodes is not a problem.
2. Faster data retrieval and better recovery – Unrestricted metadata allows storage administrators to implement their own policies for data preservation, retention and deletion. This, along with the way storage nodes are distributed across the structure, makes it easier to reinforce data and create better "disaster recovery" strategies.
3. Fewer limitations – Due to its flat data environment, object storage provides access that other storage systems cannot allow. Unlike metadata in file systems, which has limited file attributes, in object storage, metadata can be customized by any number of attributes. Additionally, data can be accessed on many systems using multiple protocols at the same time. Compared to other data storage architectures, object storage is less complex and requires fewer administrative resources.

4. Cost-effective – Because object storage scales out easier than other storage environments, it is less costly to store all your data. If users have a private cloud space, costs can be even lower [43].

3.3.4 Selecting a storage platform

Before deciding which storage technology to use, an organization must consider the following factors: security capabilities to protect sensitive data, data storage locations, service level agreements, technical support, how much space is needed, how frequently analytics will be performed, budget constraints, what features are available on mobile apps, and what types of data are to be processed [44] [45].

Experts were asked about what organizations need to do to have a sound data storage management strategy. They suggested that organizations need to know the data, understand compliance issues, have a data retention policy, look for solutions that fit the data, not be intimidated by upfront cost, and vet providers [46].

Since its release in 2006, AWS Simple Storage Service (S3) has been a leader and the unofficial industry standard a scalable, high-speed, web-based cloud storage service [47]. AWS S3 offers an extremely durable, highly available, and infinitely scalable data storage infrastructure at a very low cost with an easy-to-use application interface. Best of all, AWS S3 is available to the FAA and is able to build a data lake using both structured and unstructured data.

Object Storage Use Case

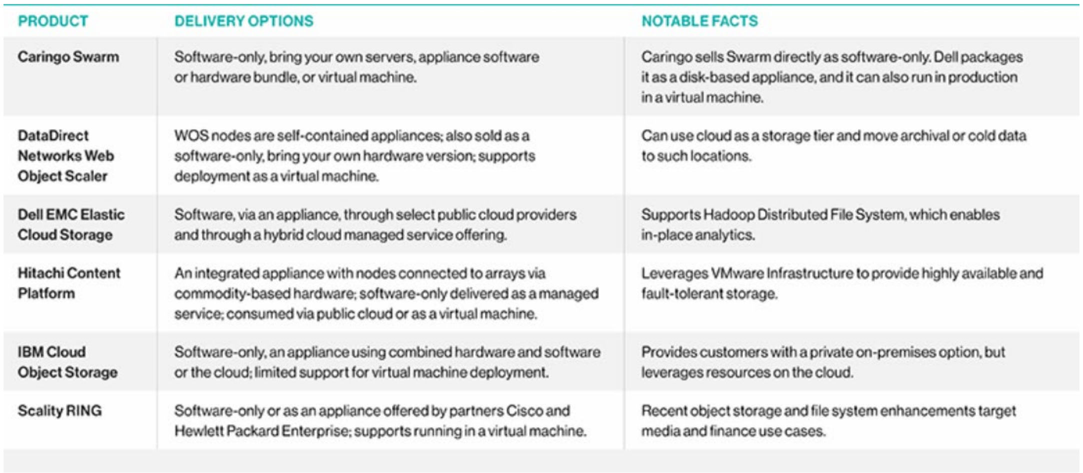
Healthcare organizations struggle with data storage as the volume of data produced by connected devices and EHRs continues to grow. Object storage is an option that will allow organizations to retrieve data from the data store quickly and securely.

According to Key Information Systems Director of Cloud Service, Clayton Weise, "there is potential in object storage to benefit the healthcare industry and move organizations away from dated legacy solutions, such as tapes."

"Object storage provides an inexpensive way to store vast pools of data, multiple petabytes up to exabyte scale within a single space. The data stored using object is always accessible, unlike tape where I have to know the serial number, track the tape, and physically retrieve it. . . . Data in object stores is always accessible. Object storage makes data retrieval far more convenient." [48]

Table 4 shows the broad range of options that the leading object storage vendors offer customers [49].

Table 4. Leading object storage systems



PRODUCT	DELIVERY OPTIONS	NOTABLE FACTS
Caringo Swarm	Software-only, bring your own servers, appliance software or hardware bundle, or virtual machine.	Caringo sells Swarm directly as software-only. Dell packages it as a disk-based appliance, and it can also run in production in a virtual machine.
DataDirect Networks Web Object Scaler	WOS nodes are self-contained appliances; also sold as a software-only, bring your own hardware version; supports deployment as a virtual machine.	Can use cloud as a storage tier and move archival or cold data to such locations.
Dell EMC Elastic Cloud Storage	Software, via an appliance, through select public cloud providers and through a hybrid cloud managed service offering.	Supports Hadoop Distributed File System, which enables in-place analytics.
Hitachi Content Platform	An integrated appliance with nodes connected to arrays via commodity-based hardware; software-only delivered as a managed service; consumed via public cloud or as a virtual machine.	Leverages VMware Infrastructure to provide highly available and fault-tolerant storage.
IBM Cloud Object Storage	Software-only, an appliance using combined hardware and software or the cloud; limited support for virtual machine deployment.	Provides customers with a private on-premises option, but leverages resources on the cloud.
Scality RING	Software-only or as an appliance offered by partners Cisco and Hewlett Packard Enterprise; supports running in a virtual machine.	Recent object storage and file system enhancements target media and finance use cases.

3.4 Security

“BD security is the processing of guarding data and analytics processes, both in the cloud and on-premise, from any number of factors that could compromise their confidentiality.” [50]

Many tools associated with BD and smart analytics are open source. Unfortunately, the tools can lead to BD security issues because they were not designed with security in mind.

Security mechanisms in BD technology are generally weak. Traditional security mechanisms are inadequate in the cloud because networks are so broad that a perimeter can no longer be clearly defined.

The top security and privacy risks that are specific to BD include the following:

1. Attackers use untrusted computational programs to extract and turnout sensitive information from data sources. Information leaks can corrupt your data and lead to incorrect results in prediction or analysis. To prevent malicious or fake data from infiltrating your data lake, implement access control, code signing, and dynamic analysis. Strategies are needed to control the impact of untrusted code.
2. Confirm that the data from sources put into the data lakes are valid and can filter malicious data from a good data source.
3. Some organizations cannot – or do not – implement access controls to divide the level of confidentiality within the company. Managing the role-based settings of different BD

users (admin, knowledge workers, end users, developers etc.) is a core part of implementing granular access control.

4. Activity monitoring – who is accessing what type of information and do they comply with a data security policy.
5. It is challenging to store data across thousands of nodes in terms of authentication, authorization, and encryption. Encrypting real-time data can have an impact on performance [51].

"Cloud computing experts believe that the most reasonable way to improve the security of BD are through the continual expansion of the antivirus industry." Some other ways to strengthen BD security include the following:

- "Focus on application security, rather than device security.
- Isolate devices and servers containing critical data.
- Introduce real-time security information and event management.
- Provide reactive and proactive protection." [52]

3.4.1 Solution and practices for cloud security issues

1. Encryption can be used to store and query BD safeguards confidentiality, restricting data access to authorized users while enabling data integrity with inherent mechanisms for access control [53]. Encryption was found to reduce the likelihood as well as the total cost of a data breach [54]. In a cloud environment, it is recommended that all files with sensitive data be encrypted at every stage of their life cycle. While encryption can protect sensitive and personal identifiable information (PII), it is important to be aware that it can impact performance. Another idea is to encrypt the sensitive data before sending it to a cloud service provider (CSP).
2. While logs are useful, they do not tell an organization what attackers are doing once inside. Logs have a limited ability to identify behavior leading up to an attack. AWS offers a host-based intrusion detection system (HIDS). It is better because it provides knowledge of the what, when, and where, before, during, and after an attack.
3. User access control – "Strong user access control requires a policy-based approach that automates access based on user and role-based settings. Policy driven automation manages complex user control levels, such as multiple administrator settings that protect the BD platform against inside attack." [55]
4. Intrusion detection systems (IDS) and intrusion prevention systems (IPS) are security workhorses and protect the BD platform from intrusion. If an intrusion succeeds, IDS quarantine the intrusion before it does significant damage.

5. Use a trusted CSP and ensure that they cannot access data. Ask the CSP if they can protect your data against a data loss by providing proper redundancy, backup, and disaster recovery solutions.
6. Understand the data storage regulations that the CSP follows [56].

A key concern for the FAA is how to protect sensitive data, particularly how it relates to RE&D data. A data breach exposes confidential, sensitive, or protected information to an unauthorized person. Files in a data breach are viewed and/or shared without permission.

The following is a smorgasbord of feedback that 34 experts provided as to the biggest mistakes companies make when it comes to securing sensitive data:

1. Companies do not understand “where their sensitive data resides because they have not set policies to systematically and consistently categorize their data, and consequently, they do not have controls in place to ensure that all categories of data are handled appropriately.” To protect sensitive data, companies must “set rules for handling it, implement technical controls to ensure it is actually handled properly, and educate your users about their role in keeping it safe.”
2. Not having a Data Classification Policy. Divide data into categories and label them – restricted (most sensitive), confidential or private (moderately sensitive), and public.
3. Businesses need to use the tools that vendors provide to pro-actively establish, monitor, and enforce security protocols and to limit internal access to sensitive content.
4. Not managing software vulnerabilities [57].
5. Consider a hybrid environment that combines public cloud services and a hosted IT infrastructure.
6. If your data contains sensitive government information, ensure that your cloud solutions offer strict controls to help you comply with country-specific requirements, such as General Data Protection Regulations (GDPRs) [58].
7. Create a firewall and do periodic backup and recovery.

3.4.2 Threats and vulnerabilities

“A threat is a potential cause of an incident that may result in harm to a system or an organization. A vulnerability is a weakness in the asset or system which is exploited by a threat. A threat agent carries out threats by exploiting one or more vulnerabilities.” The following are threats and vulnerabilities in the cloud environment: [59]

Data Breach

A data breach is a security incident of unauthorized release of private and sensitive information. It is considered the worst problem that the cloud computing service faces. It often happens when a cybercriminal infiltrates a database and compromises sensitive data, whether it is just merely that data or copying, transmitting or using it in any way. The average time between detection and containment of a data breach is 69 days. The costs associated with a breach were 25% lower for organizations that contained the incident within 30 days.

The mean time to contain (MTTC) is a crucial metric in any emergency response plan template. It depends on the organization's level of preparedness to rapidly switch into emergency response mode and execute the right tasks in the right order [60].

Some common reasons why data breaches occur include weak passwords, unfixed, old system vulnerabilities, malware, and human error.

Data Breaches occur in the following ways:

- Accidental Insider – An employee using a co-worker's computer and reading files without proper authorization permissions. The access is unintentional, and no information is shared. However, because it was viewed by an unauthorized person, the data are considered breached.
- Malicious Insider – Person accesses and/or shares data hoping to cause harm to an individual or company. The malicious insider may have legitimate authorization to use the data, but intends to use the information in nefarious ways.
- Lost or Stolen Devices – An unencrypted and unlocked laptop or external hard drive—anything that contains sensitive information—goes missing.
- Malicious Outside Actors – Hackers who use various attack vectors to gather information from a network or an individual.

To avoid a data breach, organizations should take the following steps:

- Patch and update software as soon as options are available.
- Encrypt sensitive data.
- Upgrade when software is no longer supported by the manufacturer.
- Enforce Bring Your Own Device (BYOD) security policies.
- Enforce strong credentials and multi-factor authentication.
- Educate employees on best security practices and ways to avoid socially engineered attacks [61].

- Use the latest antivirus protection.
- Give a user the privileges required for that user to fulfil his/her responsibilities.
- Protect Authentication Gateways [62].

Other Security Risks

Denial of Service (DoS) is an attack that effects a system's availability. In a DoS attack, there is only one source machine from which the attack originates and it is susceptible to mitigate. DoS attacks prevent legitimate users of a service from being able to access their data or applications.

Ransomware is malware where an attacker locks files or other resources in the victim's system generally through encryption. To decrypt the files, ransom is required.

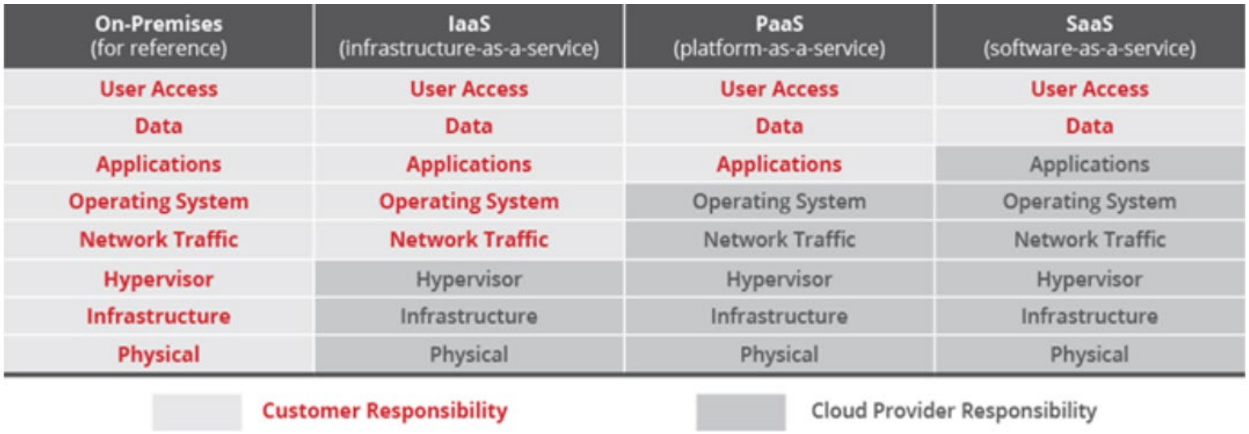
An Advanced Persistent Threat (APT) happens when an attacker infiltrates systems to establish a foothold in the infrastructure of an organization in the cloud in an attempt to steal data. "APTs get into cloud services through techniques like spear phishing, direct hacking, attack code on universal serial bus (USB) devices, penetration through the network and the use of an unsecured third-party Application Programming Interface (API). Advanced security controls, frequent infrastructure monitoring, and rigid process management are key to defending against this threat to cloud infrastructure." [59]

Within the cloud computing environment, there are four categories of security challenges:

1. Network level – issues involving network protocols and distributed nodes, distributed data, Internode communication. In the cloud environment, it is extremely difficult to ensure security since pieces of a file are stored in different locations.
2. User authentication level – encryption/decryption techniques, authentication methods such as administrative rights for nodes, authentication of applications and nodes, and logging.
3. Data level – data integrity and availability (i.e., data protection and distributed data)
4. Generic issues – traditional security tools and different technologies. A small security weakness in one technology (database, computing power, or network) can affect the whole system [63].

In the cloud computing environment, the CSP treats cloud security risks as a shared responsibility. The CSP covers security of the cloud. However, customers are responsible for securing the data they put in the cloud, as well as who can access it [64].

Figure 1 shows the shared responsibility of the CSP security and the customer in the cloud.



On-Premises (for reference)	IaaS (infrastructure-as-a-service)	PaaS (platform-as-a-service)	SaaS (software-as-a-service)
User Access	User Access	User Access	User Access
Data	Data	Data	Data
Applications	Applications	Applications	Applications
Operating System	Operating System	Operating System	Operating System
Network Traffic	Network Traffic	Network Traffic	Network Traffic
Hypervisor	Hypervisor	Hypervisor	Hypervisor
Infrastructure	Infrastructure	Infrastructure	Infrastructure
Physical	Physical	Physical	Physical

▢ Customer Responsibility ▨ Cloud Provider Responsibility

Figure 1. Shared responsibility model for security in the cloud

3.5 Architecture

BD architecture is the overarching system used to ingest and process data so it can be analyzed for business purposes. To ensure which people need to make smart business decisions, an organization must construct the architecture needed to perform in an efficient and cost-effective manner. Like an architect drawing blueprints for a building, data architects must design a model that considers many technological design elements.

BD architectures will vary based on a company's infrastructure and needs. It is a good practice for a company to base the design of a company's architecture on the functional and infrastructure requirements.

Without a good architecture in place, your BD systems will not function properly. If an organization does not have a good architecture, it may lead to frequent outages, inefficient resource utilization or saturation, and the lack of security, elasticity, scalability, and governance [65].

There are many advantages to using a cloud computing architecture. It can reduce IT costs because of unlimited data storage. When an organization stores data on a cloud server, there is no loss or damage to data because it can be backed up and recovered at any time or anywhere. Furthermore, it is integrated with any available software automatically and you can access it anywhere.

The architectural design must support the functional requirements of BD: ability to capture, organize, integrate, and analyze data as well as to act on the results of the analysis. A framework of BD technologies, known as a technology stack, addresses the functional requirements. The technology stack consists of eight architectural layers, as shown in Figure 2 [66]. Organizations will use different stack elements depending on the problem it is addressing.

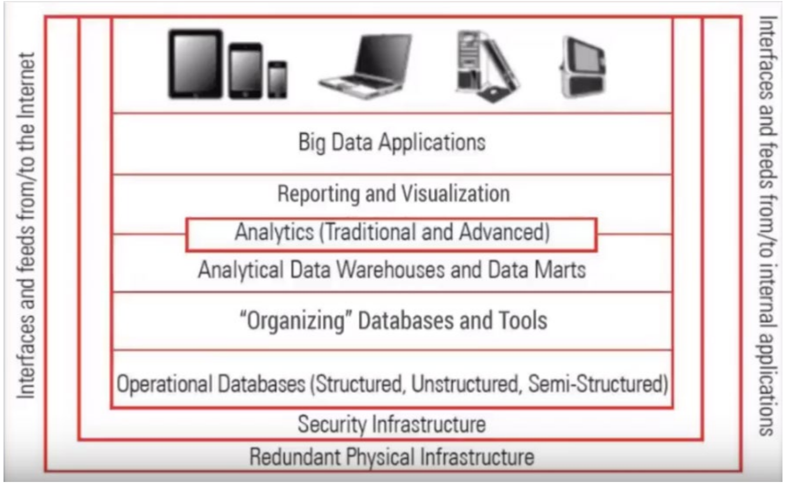

Interfaces and feeds from/to the Internet
Big Data Applications
Reporting and Visualization
Analytics (Traditional and Advanced)
Analytical Data Warehouses and Data Marts
"Organizing" Databases and Tools
Operational Databases (Structured, Unstructured, Semi-Structured)
Security Infrastructure
Redundant Physical Infrastructure
Interfaces and feeds from/to internal applications

Figure 2. Big data technology stack

The architectural design must be able to support:

- Infrastructure, management, and operational software
- High data storage
- Software development tools
- High power, high speed computation
- Redundancy

A BD architecture consists of the following four logical horizontal layers:

1. Various data sources.
2. Data storage – This layer stores data from the sources. Must be able to read data at various frequencies, formats, and sizes and store it in distributed file storage (DFS), cloud, structured data sources, NoSQL, etc.
3. Data processing/analysis layer – This layer interacts with stored data to extract business intelligence. This is where queries happen.
4. Data Output layer – This layer receives analysis results and presents them to the appropriate output layer [67].

BD architecture varies based on a company's infrastructure and needs, but it usually contains the following four logical vertical layers:

1. Information integration – Responsible for connecting to various data sources. The layer is used by components to store information in BD stores and to retrieve information from BD stores for processing.
2. BD governance – Guidelines that help enterprises make the right decisions about the data.
3. Quality of service – Layer responsible for defining data quality, policies around privacy and security, frequency of data, size per fetch, and data filters.
4. Systems management – Monitoring the health of the overall BD ecosystem [68].

Data architectures address data in storage, in use, and in motion; descriptions of data stores, data groups and data items; and mappings of those data artifacts to data qualities, applications, locations etc.

A data architecture is the system used to ingest and process huge amounts of data. It is the foundation for BD analytics. It provides criteria for data processing operations to make it possible to design data flows and control the flow of data in the system.

There are five architectural principles for building BD systems, as follows:

1. Building decoupled systems – Decoupled systems in BD are important because they allow you to alter one aspect of the system without affecting the other.
2. In choosing the right tool, you should consider data structure, latency, throughput, and access patterns.
3. Leverage managed services – Businesses should use CPS managed services because they are scalable/elastic, available, reliable, secure, and low maintenance. The CPS should perform management tasks so customers can focus on their core jobs.
4. Use log-centric design patterns – Since storage is cheap, most organizations do not delete any of their data. Organizations should build their BD system in a log-centric fashion. Immutable log files, which are protected from tampering, are copies of the original data in case anything happens to their system.
5. Be cost conscious – If the bill will be too high, the company may need to consider some different products. As a rule of thumb, the lower cost tools are often the most used [69].

A cloud computing architecture is comprised of the components and subcomponents. Components typically consist of two parts:

1. A front-end platform is what the user sees. It consists of interface applications required to access the cloud platform e.g., web browser.
2. A back-end platform is what the user cannot see (servers and storage). Users access the back end via the Internet, Intranet, or Intercloud [70].

The following seven best practices should be implemented when designing a solid cloud-architecture:

1. Proactive Planning – Plan as if your hardware will fail, and have a contingency mechanism in place to deal with unknown catastrophes.
2. Security – Security is more than a firewall. Building security layers will you protect your cloud, data from unauthorized actions, destructive forces, or breaches. Classify Data in different segments (Public, Private, and Shared), encrypt data, create a log of all details, enact policies to prevent Accidental overwrites/deletes/changes.
3. Reliability – Build a system that can to recover from outages and dynamically meet demands. Your system should be robust and work no matter what.
4. Design a system so you can efficiently manage your computer resources to meet your requirements.
5. Use cloud cost optimization strategies, including managed cloud services, to reduce cost to a minimum and use savings to serve your business.
6. Follow effective practices and procedures to manage workloads to drive operational excellence.
7. Take a multi-cloud strategy allowing you to migrate to other clouds and/or run the services balanced between two or more clouds. A multi-cloud strategy gives you the flexibility to achieve the best price to performance ratio without compromising the functionality and interoperability [71].

There are many types of cloud computing architectures. The following is a list of some examples:

1. Single "All-in-one" Server – used to launch a single server that contains a web server (Apache), as well as your application (PHP) and database (MySQL).
2. In a Single Cloud Site Architecture, known as a standard three-tier website architecture, at least one dedicated server in each tier of the system architecture (load balancing server, application server, and database server).
3. Hybrid Cloud Site Architectures – leverages multiple public/private cloud infrastructures or dedicated hosted servers [72].

Table 5 shows AWS’s top five BD architectures.

Table 5. AWS's top five BD architectures [73]

Architecture Type	**Cost**	**Suitability**	**Caveats**	**Popular offering**
Streaming	$$$$$ (typically RAM intensive)	Mission-critical data, manic spikes in load, real-time response. You will want to build real-time dashboards of Key Performance Indicators (KPI)s.	Standalone streaming solutions can be expensive to build and maintain. Scaling can be challenging, especially if you are building on the Amazon Elastic Compute Cloud (EC2).	Kinesis (managed service), Kafka (EC2-based), Spark Streaming (both as a managed service and EC2-based), Storm. Because Kinesis is a managed service, it costs more than EC2-based Kafka or Spark Streaming.
General (or specific) purpose 'batch' cluster	$ - $$$$ (highly dependent on RAM needs)	Lowest cost, greatest flexibility. Good choice if you desire one cluster to do everything and are moving from Hadoop or Spark on-premise. Highly suitable for machine learning.	A system that can "do everything" but rarely "does everything well."	EMR (managed service – runs Spark as well), Cloudera (EC2-based), Hortonworks (both as a managed service via EMR, and EC2-based).

Architecture Type	**Cost**	**Suitability**	**Caveats**	**Popular offering**
NoSQL engines	$$ - $$$ (typically RAM intensive)	Three V's" issues. Simple and/or fast-changing data models	Give up transactions and rich, diverse SQL. Since it doesn't use SQL, data cannot be queried directly with visualization tools like Tableau and Microstrategy. Scaling, especially adding new nodes, can be difficult and affect user latency and system availability	DynamoDB (managed service), Neptune (managed service – still in beta), Cassandra (EC2-based), CouchDB (EC2-based), and HBase (both as a managed service via EMR, and EC2-based)
Enterprise Data Warehouse (EDW)	$$ - $$$$$ (typically need lots of nodes to store and process the mountain of data)	Analyze data specifically for business value or build real-time dashboards of KPIs.	Must understand the difference between OLAP and OLTP and ensure that they are using each in the correct way.	Redshift – there is really no other valid option regarding cost, performance, and flexibility.

Architecture Type	Cost	Suitability	Caveats	Popular offering
In-place analytics	$ - $$	Very low cost. No management whatsoever. Can act as a low-cost, moderately performant EDW. Does not require replicating data to a second system.	Not the lowest latency. To achieve decent performance, the stored data will need to be reformatted using a serialization format Parquet, compressing, re-partitioning, etc.	AWS Athena (managed service used to query S3 data), EMR (managed service – can install Presto automatically), Self-managed Presto (EC2 based – you would never want to do this in AWS).

3.6 Platforms

A BD platform is an integrated IT solution for BD management. It combines several software systems, software tools and hardware to provide easy to use tools to enterprises. A data platform combines all of the data together from various data sets and acts as a centralized hub where it can be stored, managed, and mined. A data platform harmonizes information and makes it easier to analyze.

The most important features of any good big data analytics platform include:

- Big data platform should be able to accommodate new platforms and tools based on business requirements, because business needs can change due to new technologies or due to changes in business processes.
- It should support linear scale-out.
- It should have capability for rapid deployment.
- It should support a variety of data formats.
- The platform should provide data analysis and reporting tools.
- It should provide real-time data analysis software.
- It should have tools for searching through large data sets [74].

A platform's ability to adapt to increased data processing demands plays a critical role in deciding if it is appropriate to build the analytics based solutions on a particular platform. To perform any kind of analysis on such voluminous and complex data, scaling up the hardware platforms becomes important. The scaling up of the hardware platform is the critical factor in choosing the right hardware/software technologies.

Scaling is a system's ability to add data processing resources as work increases to improve the capacity and performance of a system. The two categories of scaling are as follows:

1. Horizontal Scaling or scale out – distributing the workload across many servers, which may be even commodity machines. Unlimited nodes can be added. Prominent horizontal scale out platforms include those described below.
    - Peer-to-peer networks – millions of machines connected in a decentralized and distributed network architecture where the nodes in the networks (known as peers) serve as well as consume resources. It is the one of the oldest distributed computing platforms in existence. Communication between different nodes is difficult.

- Apache Hadoop – open source framework for storing and processing large datasets using clusters of commodity hardware. Hadoop can scale up to hundreds and even thousands of nodes and is highly fault tolerant. The Hadoop platform contains the following two important components.
  - HDFS is a distributed file system that can store data across a cluster of commodity machines while providing high availability and fault tolerance.
  - Hadoop Yet Another Resource Negotiator (YARN) is a resource management layer and schedules the jobs across the cluster.

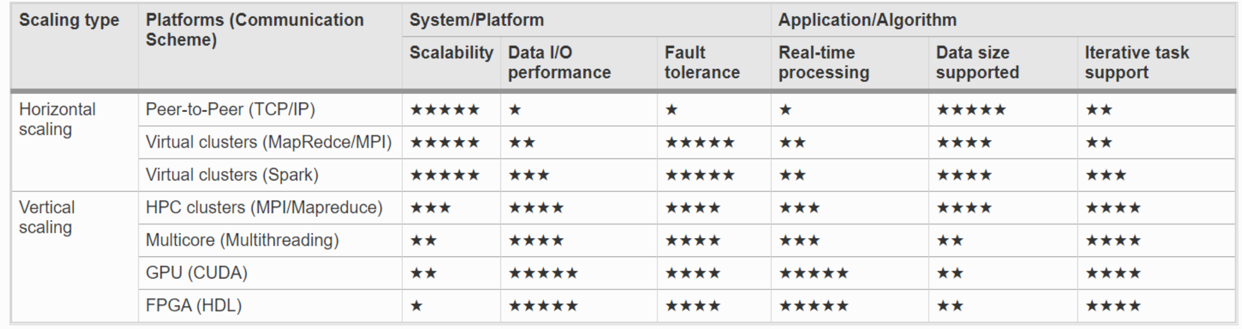
2. Vertical Scaling or scale up: installing more processors, more memory and faster hardware, typically, within a single server. This usually involves a single instance of an operating system.

Before selecting a BD platform, organizations should evaluate its architecture, infrastructure, requirements, budget, and business culture.

Before deciding on a particular platform for a certain application, an organization should consider the following factors:

1. Data size – If the data can fit into the system memory, then clusters are not required since data can be processed on a single machine. Platforms such as a graphics processing unit (GPU) or Multicore central processing unit (CPU) can be used to speed up the data processing. If the data does not fit the system memory, consider using cluster options such as Hadoop or Spark since they can handle a large amount of data. Hadoop has well developed tools and frameworks, although it is slower for iterative tasks.
2. Speed or throughput optimization – Speed refers to platforms that can process data in real-time. Throughput is a system that can handle and process data simultaneously. Consider Hadoop, Peer-to-Peer networks, etc. These platforms can handle large-scale data but it usually takes more time to deliver the results.
3. How long does it take to build a model? Does model building require several iterations or a single iteration? A model is typically applied in an online environment where the user expects the results within a short period of time (almost instantaneously). This creates a strong need for investigating different platforms for training and applying a model depending on the end-user application.

Table 6. Comparison of different platforms (along with their communication mechanisms) based on various characteristics [75]



Scaling type	Platforms (Communication Scheme)	System/Platform			Application/Algorithm		
		Scalability	Data I/O performance	Fault tolerance	Real-time processing	Data size supported	Iterative task support
Horizontal scaling	Peer-to-Peer (TCP/IP)	★★★★★	★	★	★	★★★★★	★★
	Virtual clusters (MapRedce/MPI)	★★★★★	★★	★★★★★	★★	★★★★	★★
	Virtual clusters (Spark)	★★★★★	★★★	★★★★★	★★	★★★★	★★★
Vertical scaling	HPC clusters (MPI/Mapreduce)	★★★	★★★★	★★★★	★★★	★★★★	★★★★
	Multicore (Multithreading)	★★	★★★★	★★★★	★★★	★★	★★★★
	GPU (CUDA)	★★	★★★★★	★★★★	★★★★★	★★	★★★★
	FPGA (HDL)	★	★★★★★	★★★★	★★★★★	★★	★★★★

The headings of table 6 are defined below:

- Scalability – A system's ability to add data processing resources as work increases to improve the capacity and performance of a system.
- Data I/O performance – Rate that data are transferred to/from a peripheral device, the rate that data are read and written to the memory (or disk), or the data transfer rate between the nodes in a cluster.
- Fault tolerance – Determines if a system can continue operating properly if one or more components fail.
- Real-time processing – System's ability to process the data and produce the results strictly within certain time constrains.
- Data size support – Size of the dataset that a system can process and handle efficiently.
- Support for iterative tasks – System's ability to do repetitive tasks efficiently.

3.6.1 Western Digital Case Study

In building a platform, Western Digital was concerned about futureproofing—how to build a platform that not only satisfies the business needs today, but also anticipates the needs of the future? Along with this, they also examined how to optimize the costs associated with building a platform. Western Digital took the following factors into consideration:

1. Third-Party vs. In-House Expertise – Where to host this platform—on the cloud, on-premise, or a hybrid? Because Western Digital knows their data and the way the engineers needed to use it, they decided to bring it into their own virtual private cloud and rebuild those platforms — the pipeline's, the data pipelines.

2. Data ownership – When implementing data into a cloud environment, it is expensive to store all the data in that environment and to get data out of that environment. Western Digital has the ability to ask for more data if it is needed. This saved Western Digital about $600,000 a year just by having their data and all the data that they want to have in their environment, while only sending the data they need to have into the cloud. When you have “that golden copy” – all the data in one environment – then you're not really beholden to any service provider. You could go to a different cloud if necessary or appropriate in the futureproofing of the platform.
3. Data quality – Bringing the data directly from the factories through the network, through buffer servers to the cloud, whenever there was a hiccup in the network led to corrupted data. It was expensive, but Western Digital constantly had to repair that data to provide it to users at a level of quality they expected. To solve this problem, instead of bringing that data through the buffer servers they brought it direct.
4. New Technologies – Used engineering team to study how to optimize cloud features and software investments.
5. Software Agreements – With an enterprise platform, it is possible to negotiate enterprise software agreements.
6. Scaling Internal Knowledge – Building up expertise and transitioning from external consultants into really building out the expertise within Western Digital’s team was definitely cost effective. [76].

3.6.2 Popular platforms

Cloudera, Hortonworks, and MapR Technologies (under Hadoop) emerged a decade ago to commercialize products and services in the open-source ecosystem around Hadoop, a popular software framework for processing huge amounts of data. Cloudera and MapR were designed for on-premises deployment and will lose support because of BD’s rapid transition to the public cloud. Vendors who do not aggressively embrace the public cloud and technologies like Spark will not be competitive.

Technologies like Spark and Kafka are rising to support modern data applications that use artificial intelligence and machine learning. However, Hadoop will not disappear and not every data workload will go to the cloud [77].

The U.S. Department of Transportation (USDOT) Secure Data Commons (SDC) is a collaborative transportation research and analytics access-controlled, cloud-based platform that enables traffic engineers, researchers, and data scientists to access transportation-related datasets [78]. It stores sensitive transportation data (made available by participating data providers) and grants access to approved researchers to work with these datasets. The SDC also provides access to open-source tools and allows researchers to collaborate, upload data sets to a data lake, and share code with other system users. With SDC, data can be uploaded in near real time throughout a project.

The USDOT created the SDC to provide a secure platform for sharing and collaborating on research, tools, algorithms, and analysis involving moderate sensitivity level (PII & confidential business information [CBI]) datasets using commercially available tools, without needing to install tools or software locally. SDC has a common platform for innovative data analysis and sharing of results that cuts across Department's data silos.

Key SDC features include the following:

- Built-in infrastructure for data analysis
- Effective project and team collaboration
- Multiple data formats
- Multiple data transfer frequencies
- Robust, cloud-based infrastructure for storage and management of data
- Strong user management controls

3.7 Enterprise Information Management

Enterprise Information Management (EIM) is a cloud-based platform that allows the FAA to manage data and information. The future vision for EIM is guided by agency-wide business user perspectives on how to improve data access and use to advance the quality of decisions and shorten decision-making cycles. EIM's purpose is to help employees use the data they get from many sources to make better decisions. EIM allows employees to access real-time information in the right format when it is needed; Hadoop is the heart of EIM.

EIM is a work in progress. EIM ingested data from many sources in the FAA and created the FAA Data Governance Center (DGC). The DGC is a data governance and catalog technology platform that enables the FAA to effectively organize and manage the agency's data, and comply with the OMB Order M-19-18 Federal Data Strategy – A Framework for Consistency.

The FAA EIM Data Platform (DP) is a cloud-based system that provides a platform of reusable data and information management services and BD processing capabilities for broad cross-agency use. This single repository helps FAA employees find and understand data that is available for their use, while back-end processes ensure the data's quality and accessibility. The EIM-DP currently ingests System Wide Information Management (SWIM) data as it continues to grow to include the data sets needed to provide FAA users, departments, and programs with an effective and efficient environment to perform post-operational data analysis and to provide information management support functions using FAA NAS, mission support, administrative, and other data.

The DGC contains a data catalog of ingested data sources as well as points of contact, should an employee need access to the data source. When data is ingested into EIM, it is tagged as either restricted, public, or private. Users are able to access the data based on user permissions set up in the system. To maintain the integrity of the original data, employees will have their own copy of the data so they can manipulate it according to their needs.

EIM Platform Benefits

- Programs and systems can focus on how to use data in a meaningful way and not waste time on infrastructure or finding and getting access to data (50-90% of their time) from other systems.
- Reuse the data for unique applications.
- Access all data in one place.
- Share tools and capabilities across the FAA enterprise.
- Employees can access data that is not private from their desks with their Personal Identity Verification (PIV) cards.
- Open architecture to support unique system processes and functions.

Figure 3 shows the EIM platform architecture. Because EIM is a work in progress, the data has been ingested and curated. However, analytical work has not yet been completed. EIM management is still in the process of deciding which vendor to use for the analytical tools. Palantir Technologies and Databricks seem to be the top contenders [79].

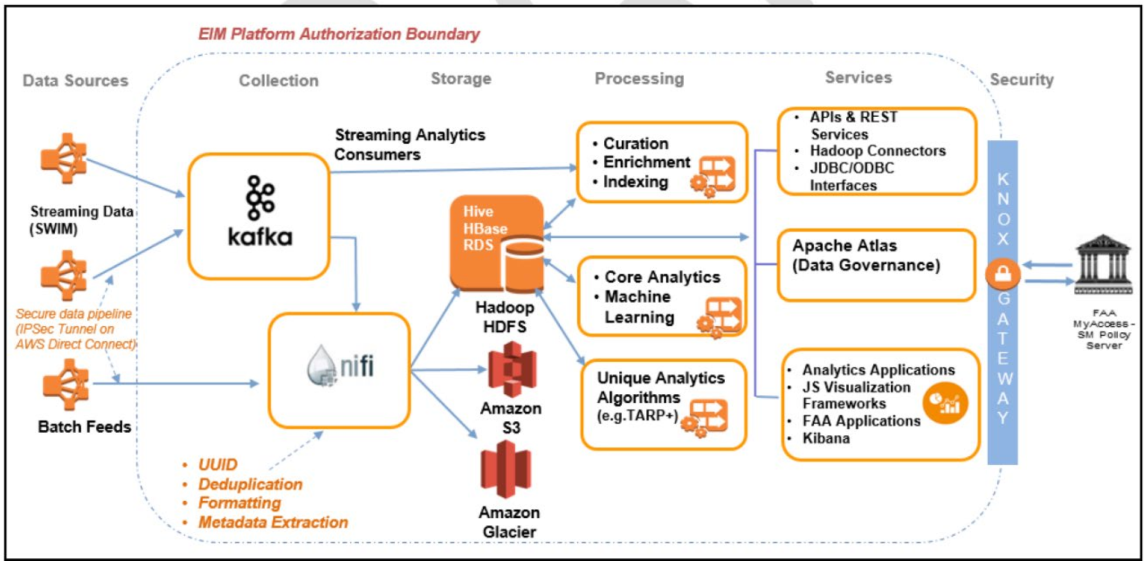


Figure 3. EIM platform architecture

3.8 Data governance

Data governance allows users to find data quickly, understand what it means and where it came from, and trust the integrity of the data/analysis. The FAA DGC is powered by Collibra. Collibra is a data governance tool, much like Amazon, but for data [80]. Data governance provides a data catalog that gives people what they need, when they need it, so they do not need to spend hours searching. Collibra has searching, drill-down, and simple provisioning capabilities.

The main benefits of Collibra include the following:

1. Automated data retrieval – Helps retrieve information regarding operations and processes.
2. Transparency of data – Aims to answer the five W's—who, what, when, where, why—for every data point that passes through its system.
3. Data intelligence – Interpreting data to help companies make better business decisions.
4. Federal Risk and Authorization Management Program (FedRAMP) Authorized.

3.8.1 Picking the right data governance tool

When choosing a data governance tool, be clear on how it will be used and the capabilities you require. Evaluation and selection depends on features, functionality, and on how you will add business value with these tools. If an organization cannot demonstrate visible value in a short period of time, getting continued funding and executive support for its governance initiative will be difficult.

When looking for a data governance tool, the following are important features to consider:

1. Identifies and tracks common create, read, update and delete activities for data elements.
2. Possesses data discovery capabilities that enable you to scan and identify data elements, plus data and metadata values.
3. Manages relationships between data elements through hierarchies or taxonomies.
4. Allows one to classify data based on its use or relevance.

When selecting a data governance tool, one should ask the following questions:

- What kind of reporting and dashboard features do products offer?
- Does the organization prefer tools that are deployed on premises or cloud-based tools?
- Are mobile capabilities needed?
- What kinds of connectors to existing software in your enterprise are needed? [81]

Table 7 compares a number of data governance tools in terms of features/functions, platforms, and price.

Table 7. Comparison table for data governance tools [82]

Data Governance Tools	Features & Functions	Platform	Price
OvalEdge	Data Governance, Data Catalog, Automated Data Lineage, Data Discovery, Self-Service Analytics	Windows, Unix, Cloud, On-Premise, Web, SaaS.	Starts at $50/user/month Contact company for more details
Truedat	Data Governance, Business Glossary, Data Catalog, Data Lineage, Data Quality	Cloud, On-Premise	Open source Professional Services Fee Contact company for more details
Collibra	Collaboration Features, Data Help Desk, Automation of Data Governance & Management.	Windows, Mac, iOS, Cloud, Web, SaaS	Contact company
IBM	Data Governance, Data Cataloging, Obtaining Information for Big Data Projects	Windows, Cloud, Web, SaaS	Contact company

Data Governance Tools	Features & Functions	Platform	Price
Talend	Data Governance, Cloud Integration, Data Integration, API, Application Integration	Windows, Mac, Cloud, Web, SaaS	Talend Open Source: Free Stich Data Loader: $100-$1000 Talend Cloud Data Integration: $1170/user/ month
Informatica	Manage GDPR Data Risks, Detect & Protect Sensitive Customer Data, Verify Contact Data	N/A	Starts at $2000/month
Alteryx	Discover, Prepare, & Analyze the Data, Deploy & Share Analytics, Collaboration Features	Windows, Mac, Cloud, Web, SaaS	Alteryx Designer: Starts at $5195/user /year Alteryx Server: Starts at $58,500

3.9 Services

3.9.1 FAA Cloud Services

The FAA Cloud Services (FCS) is simply a contract with a vendor to obtain cloud services from Microsoft Azure and AWS.

The FCS program delivers on-demand, pay-per-use computing, and data storage over the Internet. The FAA is moving the agency away from FAA-owned data centers. The FAA now purchases IT as-a-service rather than purchasing and maintaining expensive facilities and hardware. With cloud computing, the FAA will lease what it needs from the vendor instead of buying it, shifting IT from a capital expense to an operating expense. Cloud computing results in increased efficiencies, reduced operating and maintenance costs, greater flexibility, and improved information sharing across the FAA [83].

FAA customers have three service models, as follows:

1. Infrastructure as a Service (IaaS) provides a secure computing infrastructure—storage, virtual machines, and networking capabilities—as a pay-per-use service over the Internet. As hardware requirements change, IaaS will be leveraged to quickly and cost-efficiently scale up and down as needed. The FCS offers AWS and Microsoft Azure.
2. Platform as a Service (PaaS) provides virtualized computing resources with everything required to support the complete lifecycle of building and delivering web-based (cloud) applications, without the cost and complexity of buying and managing the underlying hardware, software, and hosting services. The FCS offers AWS and Microsoft Azure.
3. Software as a Service (SaaS) allows FAA customers to purchase third-party provider hosts applications and services that are owned and operated by others within FCS data centers, and that are accessed through the Internet. A good example is the FAA replacing Lotus Notes with Outlook [84].

3.9.2 FedRAMP

When the FAA wants to use a CSP, their service must meet the FedRAMP security requirements and implement FedRAMP baseline security controls through the Security Assessment Framework (SAF).

On December 8, 2011, OMB issued the “Security Authorization of Information Systems in Cloud Computing Environments" - also known also as the FedRAMP Policy Memo - mandating that for all agency use of cloud services, the agencies use FedRAMP for their risk assessments, security authorizations, and granting of ATOs, and ensure applicable contracts require CSPs to comply with FedRAMP requirements."[85]

FedRAMP is a government-wide program in the Federal Cloud computing program at the General Services Administration (GSA) that standardizes the approach to security assessment, authorization, and continuous monitoring for cloud systems by all agencies. FedRAMP intends to make the government's migration to cloud computing more cost effective and to ensure the safety, security, and reliability of the government's data.

The Federal Risk and Management Program Dashboard contains a "searchable, sortable database of all cloud services that are FedRAMP authorized, FedRAMP Ready, or In Process for an authorization." "Within the Marketplace, you can search for a particular service, provider, agency, or Third Party Assessment Organization (3PAO) to find exactly what you're looking for. You can sort the data by status (FedRAMP Ready, Authorized, or In Process), service model (IaaS, PaaS, or SaaS), deployment model (Public, Private, Government Community, or Hybrid Cloud), or impact level (Low, Moderate, or High)." [86]

FedRAMP provides a standardized framework for security assessment, authorization, and continuous monitoring for cloud products and services. This framework uses a "do once, use many times" approach that saves an estimated 30-40% of government authorization costs, by reducing both time and staff required to conduct Agency security assessments. [87]

The FedRAMP assessment process begins with CSPs working with the Joint Authorization Board (JAB) or an Agency to obtain a FedRAMP Authority to Operate (ATO). FedRAMP uses the Risk management Framework and FedRAMP baseline requirements that are Federal Information Security Management Act (FISMA) compliant and based on NIST 800-53 rev4.

There are two ways that the FAA can use a cloud, as follows:

1. Leveraging the Cloud Service Offering (CSO) that already has either an Agency or JAB ATO.
2. The CSO does not yet have an ATO, so the FAA may sponsor the FedRAMP cloud ATO activities. The FAA will review the CSP's FedRAMP required security authorization submittal, including assessment results from the 3PAO, and if the risk is acceptable grant the cloud service offering an Agency ATO.

Federal IT systems, including cloud systems (IaaS, PaaS, SaaS), must be authorized (be granted an ATO) before they can enter an operational environment. This holds true for cloud systems as well; a cloud system without a FedRAMP ATO cannot be used. The ATO status of clouds is identified on the FedRAMP web site (https://marketplace.fedramp.gov/#/products?sort=productName), which identifies the cloud authorization status in the FedRAMP Authorization process. The FedRAMP online secure repository contains security authorization documentation, guidelines, and templates that describe the FedRAMP ATO process. The FAA can leverage any of the CSOs for their use. Section 5 of the FY20 Security Authorization Process describes the process for leveraging and sponsoring a CSO.

There are two potential cloud adoption efforts, as follows:

1. Leveraging an existing FedRAMP ATO – CSO has an existing ATO granted by the JAB or another agency. The JAB is responsible for maintaining the CSP's continuous monitoring submittals and other FedRAMP requirements for JAB-P ATO. The original sponsoring Agency is responsible for CSPs that the Agency sponsors.
2. Sponsoring a cloud that does not have a FedRAMP ATO – the CSP does not yet have a FedRAMP ATO and a Federal Agency (i.e. the FAA) must sponsor the CSP in an effort to receive its FedRAMP ATO. The Agency works with the CSP to review their authorization documents, most importantly the SAR results, and Plan of Action and Milestones (POA&M), and actually grant the CSP Offering their FedRAMP Agency ATO. The Agency would then be responsible for the CSP maintaining their monthly continuous monitoring submittals and other FedRAMP requirements [88].

NOTE: In both efforts, a CSP is required to use all of the FedRAMP templates, and use a 3PAO to conduct the assessment of the target CSO.

A cloud system is deemed FedRAMP compliant, if it meets the following requirements:

- The system security authorization package was created using the mandatory FedRAMP templates (System Security Plan (SSP), Security Assessment Plan (SAP), Security Assessment Report (SAR) [89]
- The system meets the FedRAMP security control requirements
- The system has been assessed by an independent 3PAO
- The security package for the cloud system authorization needs to be submitted to the FedRAMP Program Management Office (PMO) to be listed in the repository so other agencies can leverage it
- An authorization letter for the system is on file with the FedRAMP PMO

Hosting a system in a cloud does not relieve the purchasing system owner from security requirements. Controls can and should be inherited from the CSP as appropriate but there will still be controls that remain the responsibility of the System Owner. According to the OMB FedRAMP policy memo, the purchasing entity is responsible to ensure applicable contracts appropriately require CSPs to comply with FedRAMP security authorization requirements.

If the system/application that does not fall in one of these categories, it is not considered a cloud and does not require FedRAMP authorization.

Before deciding to use a FedRAMP CSP system, agencies should review each candidate's package to make sure that it meets agency and FedRAMP compliance requirements. Only Federal Agencies may review FedRAMP security packages.

The Security Assessment Branch (AIS-230) is the FAA organization responsible for helping organizations within the FAA meet all agency and FedRAMP compliance requirements (refer to the FY20 FAA Security Authorization Handbook for details). The problem is that there are organizational issues with system owners not doing what is required and lack of security support.

Luckily, AWS is already FedRAMP compliant.

4 Big data laws and policies

Over the past ten years, multiple laws and policies have been enacted regarding BD. This section provides a timeline of these laws and policies.

4.1 Executive Order 12906

On April 11, 1994, President Clinton issued Executive Order 12906: Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure. The order describes "…the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data" (White House, 1994). It asks the Federal Geographic Data Committee (FGDC) "…to establish a National Geospatial Data Clearinghouse, orders federal agencies to make their geospatial data products available to the public through the Clearinghouse, and requires them to document data in a standard format that facilitates Internet search. Agencies were required to produce and distribute data in compliance with standards established by the FGDC." The FGDC supports the creation, management, and maintenance of the metadata required to fuel data discovery and access.

4.2 Federal Information Security Management Act (FISMA) of 2002

The Federal Information Security Management Act (FISMA) of 2002 established the standard used for federal agencies who are seeking an authority to operate (ATO) by government agencies

[90]. The act requires each federal agency to implement effective information security controls for Federal information systems, to provide oversight, and to develop minimum controls for securing Federal information systems. FISMA gave the National Institute of Standards and Technology (NIST) the authority to develop the standards and guidelines that are used for implementing and maintaining information security programs for risk management.

4.3 Data.gov

On March 5, 2009, the first Federal Chief Information Officer (CIO), Vivek Kundra announced the creation of Data.gov. The site introduced the philosophy of digital open data to the U.S. federal government. [91]

In May 2009, data.gov began to publicize a broad range of agency datasets, ranging from consumer complaint data to information about 911 emergency service areas [92]. This is good, but "…the datasets are not available in open formats or easily machine-readable (e.g., PDF versus XML) or are not updated more often than once per year, limiting the usefulness of the data provided on the site" [93]. In May 2009, www.data.gov began with 47 datasets and had over 200,000 datasets as of May 2019. FAA has data on www.data.gov to provide transparency and public information.

4.4 Open Government Directive

On December 8, 2009, President Obama issued the Open Government Directive or OMB Memorandum M-10-06 on Transparency and Open Government, which requires a senior agency official to be accountable for the quality of information on public websites. Federal agencies were asked to create an Open Government Plan and publish agency information online in "open and machine readable formats." OMB renewed agencies' focus to improve the effectiveness of Government by encouraging partnerships and cooperation within the Federal Government, across levels of government, and between the Government and private institutions. They issued a Transparency and Open Government memorandum based on three principles: transparency, participation, and collaboration. It sought to accomplish four main tasks:

1. Publish government information online.
2. Improve the quality of government information – identify and correct data quality problems, emphasizing immediate action on the quality of federal spending data.
3. Create and institutionalize a culture of open government – establish the key deliverables to encourage genuine and consistent progress on open government issues. First, the agencies must produce a detailed Open Government Plan within 120 days that will be

used to measure progress. Plans should be updated every two years. The directive provides details on what is to go into each agency's plan with regard to transparency, participation, and collaboration.

4. Create an enabling policy framework for open government – update current policies governing information management. The Office of Information and Regulatory Affairs must review existing policies, "such as Paperwork Reduction Act guidance and privacy guidance," to identify problems and issue revisions to allow openness to move forward.

The impact was that people began to discuss the potential value of open government and open data to citizens [94].

4.5 Big Data Interagency Working Group

The Big Data Interagency Working Group (BD IWG), formerly The Big Data Senior Steering Group (BD SSG), was formed in 2011 to facilitate the goals of the BD R&D Initiative. They "…coordinate Federal R&D to enable effective analysis, decision-making, and discovery based on large, diverse, real-time data." [95] The BD IWG is an interagency group under the National Science and Technology Council's Networking and Information Technology Research and Development (NITRD) Program. It was developed with the help of 17 participating agencies. The Federal Big Data Research and Development Strategic Plan is an important milestone in the Initiative to harness benefits from the rich sources of BD. The Plan is built around the following seven strategies:

1. Create next-generation capabilities by developing BD foundations, techniques, and technologies.
2. Support R&D to explore the trustworthiness of data, to make better decisions and enable breakthrough discoveries.
3. Build and enhance research cyberinfrastructure that enables BD innovation.
4. Increase the value of data through policies that promote sharing and management of data.
5. Understand the privacy, security, and ethical dimensions of BD collection, sharing, and use.
6. Improve the national landscape for BD education and training to fulfill increasing demand for analytical talent and capacity for the broader workforce.
7. Support a vibrant BD innovation ecosystem with collaboration between government agencies, universities, companies, and non-profit organizations. [95]

4.6 Federal Data Center Consolidation Initiative

On February 26, 2010, the Federal Data Center Consolidation Initiative (FDCCI) was launched. The goal was "…to promote the use of green IT by reducing the overall energy and real estate footprint of government data centers; reduce the cost of data center hardware, software, and operations; increase the overall IT security posture of the Federal Government; and shift IT investments to more efficient computing platforms and technologies." [96]

Through the FDCCI, agencies have formulated detailed consolidation plans and technical roadmaps to eliminate a minimum of 800 data centers by 2015 [97]. Since 2016, according to https://itdashboard.gov/, 210 tiered data centers and 3,005 non-tiered data centers have closed, resulting in a cost savings of nearly $2 million.

4.7 Cloud First

On December 9, 2010, Vivek Kundra, U.S. CIO, published the 25 Point Implementation Plan To Reform Federal Information Technology Management, known as "Cloud First," to deliver more value to the American taxpayer and to increase the operational efficiency of federal technology assets. These actions were implemented over the next 18 months with the help of OMB and agency operational center. Some highlights of the plan include:

- Turnaround or terminate at least one-third of underperforming projects in IT portfolio within the next 18 months.
- Shift to "Cloud First" policy. Each agency will identify three "must move" services within three months, and move one of those services to the cloud within 12 months and the remaining two within 18 months.
- Reduce the number of Federal data centers by at least 800 by 2015 [98].

4.8 Federal Cloud Computing Strategy

On February 8, 2011: Federal Cloud Computing Strategy, known as Cloud-First written by Vivek Kundra. This policy (a follow-up to the 25 Point Plan) is intended to accelerate the pace at which the government realizes the value of cloud computing by requiring federal agencies to evaluate safe, secure cloud computing options before making any new investments. Kundra made the point that "cloud computing can accelerate data center consolidation efforts by reducing the number of applications hosted within government-owned data centers."

This Federal Cloud Computing Strategy designed to achieve the following:

- Articulate the benefits, considerations, and trade-offs of cloud computing.

- Provide a decision framework and case examples to support agencies in migrating towards cloud computing.
- Highlight cloud computing implementation resources.
- Identify Federal Government activities and roles and responsibilities for catalyzing cloud adoption.

4.9 Big Data Research and Development Initiative

On March 12, 2012, the Obama Administration began to generate excitement with big data when it announced the Big Data Research and Development Initiative. The goal was to "…develop BD technologies, demonstrate applications of BD, and train the next generation of data scientists."[99] Six federal government departments and agencies will invest over $200 million to "…improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data."

4.10 Executive Order 13642

On May 9, 2013: President Obama signed Executive Order 13642: Making Open and Machine Readable the New Default for Government Information based on the simple principle that "openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth." [100]

The OMB issued M-13-13 Memorandum for the Heads of Executive Departments and Agencies. Under this order, agencies are directed to accomplish the following:

- Use machine-readable and open formats and data standards
- Ensure information stewardship through the use of open licenses
- Use common core and extensive metadata
- Build information systems to support interoperability and information accessibility
- Create and maintain an Enterprise Data Inventory
- Create and maintain a public data listing
- Create a process to engage with customers to help facilitate and prioritize data release.

Under the M-13-13 Memorandum, in order to ensure that agency data assets are managed and maintained throughout their life cycle, agencies must adopt effective data asset portfolio management approaches. Within six months of the date of the Memorandum, agencies and inter-agency groups must review and, where appropriate, revise existing policies and procedures to strengthen data management and release practices to ensure consistency with the requirements in the Memorandum, and take the following actions:

- Collect or create information in a way that supports downstream information processing and dissemination activities. This includes using common core and extensible metadata.
- Build information systems to support interoperability and information accessibility
- Strengthen data management and release practices. [101]

4.11 Official DOT Public Access Plan

The Official DOT Public Access Plan v1.1 of 2013 requires authors and/or Operating Administrations to submit all publications, to the DOT National Transportation Library (NTL) digital repository within a year of publication unless there are issues with privacy, confidentiality, or National/Homeland security. The plan was developed in response to the February 22, 2013 Office of Science and Technology Policy (OSTP) Memorandum Increasing Access to the Results of Federally Funded Scientific Research, while protecting the intellectual property of researchers. According to the plan, publications should be freely available to the Public within 12 months [102].

The dataset's metadata will be included in the DOT Enterprise Data Inventory. Through these mechanisms, datasets will be discoverable through data.gov, NTL, internet search engines and other tools leveraging open formats and standards.

The DOT Public Access plan is "…a framework for enhancing the tracking of the complete research lifecycle at the project level." It aims to "... maximize the potential for creative reuse to enhance value to all stakeholders" and to make it publically accessible for search, retrieval, and analysis.

The USDOT Research Hub (Research Hub) is a publicly available web-based database of USDOT-sponsored research, development, and technology project records. It is a central repository containing information on active and recently completed projects from USDOT's Operating Administrations, providing a comprehensive account of the Department's research portfolio at the project level. The Research Hub fulfills the following objectives:

> "(a) to aid the OST-R (Office of the Assistant Secretary for Research and Technology) in its research coordination, facilitation, and strategic planning efforts, (b) to leverage opportunities for cross-modal collaboration between USDOT agencies and between USDOT and external organizations, (c) to provide USDOT staff with a resource to quickly and accurately respond to requests for information from external stakeholders, (d) to provide a full account of USDOT's research portfolio to the transportation research community in the U.S. and abroad, and (e) to comply with the Fixing America's Surface

> Transportation (FAST) Act Consolidated Research Database mandates (49 U.S.C. 6502) and the Open Government and Public Access initiatives." [103]

The FAA can submit metadata to the USDOT Research Hub via https://ntl.bts.gov/submitting-content.

The question is can the FAA submit its research papers to the USDOT Research Hub. The answer is notably, "it depends" on a number of factors. Ultimately, the decision is up to the sponsor on whether the report can be made public.

In November 2014, The National Science Foundation (NSF) established a national network of BD Regional Innovation Hubs to foster cross-sector collaborations and partnerships around BD. In April 2015, stakeholders from academia, industry, government, and non-profit sectors discussed how to best structure and govern an Innovation Hub to address the challenges and leverage the opportunities around BD.

4.12 Federal Information Security Management Act of 2014

On December 18, 2014, President Obama signed the Federal Information Security Management Act of 2014 (FISMA). The new law amends the 2002 FISMA law that provides a leadership role for the Department of Homeland Security (DHS), including security incident reporting requirements, and other key changes, as follows:

- Authorizes the Secretary of DHS to assist the OMB Director in administering the implementation of agency information and security practices for federal information systems.
- Changes the agency reporting requirements, modifying the scope of reportable information from primarily policies and financial information to specific information about threats, security incidents, and compliance with security requirements. Directs the OMB Director to provide guidance on what constitutes a "major incident" as applies to agency reporting requirements.
- Addresses cyber breach notification requirements.
- Requires the OMB Director to revise Budget Circular A-130 to eliminate inefficient or wasteful reporting. This would allow federal agency information security personnel to allocate more resources to the protection of government systems. [104]

4.13 Establishment of BD hubs

In November 2015, four BD Hubs were established, one each in the Midwest, Northeast, South, and West regions of the country. The BD Hubs serve a convening and coordinating role—helping to bring together a wide range of BD stakeholders to connect solution seekers with solution providers. Each hub was given $4 million over four years.

To establish this national network, the NSF announced a funding opportunity: the Big Data Regional Innovation Hubs (BD Hubs) program.

Each BD Hub is a consortium of members from academia, industry, and/or government across distinct geographic regions of the United States and focuses on key BD challenges and opportunities for its region. The BD Hubs aim to support the breadth of interested local stakeholders within their respective regions, while members of a BD Hub should strive to achieve common BD goals that would not be possible for the independent members to achieve alone.

The network of four Hubs established a "BD brain trust" geared toward conceiving, planning, and supporting BD partnerships and activities to address regional challenges. Table 8 shows the four BD Hubs and the focus of each hub.

Table 8. BD Hubs

HUB	COORDINATED BY	REGIONAL PRIORITY AREAS
South	Georgia Institute of Technology and University of North Carolina.	Healthcare and health-related disparities, coastal hazards, industrial BD, materials and manufacturing, and habitat planning.
Northeast	Columbia University	Energy, finance, data science for education, climate, and the environment.
Midwest	University of Illinois at Urbana-Champaign	Agriculture; the food, energy, and water nexus; and smart cities.
West	University of California, San Diego, University of California, Berkeley, and University of Washington	BD technologies and data-intensive discovery, managing natural resources, and hazards and precision medicine.

Spokes of the BD Hub focus on a specific BD Hub priority area and address one or more of three key issues: improving access to data, automating the data lifecycle, and applying data science techniques to solve domain science problems or demonstrate societal impact. [105]

4.14 Executive Order 13719

On February 9, 2016, President Obama issued Executive Order (EO) 13719: Establishment of the Federal Privacy Council. The interagency council of senior officials from each of 24 federal departments and agencies will work to develop recommendations on federal government privacy policies and requirements, share ideas and best practices for protecting privacy, and advise on the hiring and training of professional privacy personnel for the federal government. [106]

OMB Memorandum M-16-24, Role and Designation of the Senior Agency Officials for Privacy (SAOP).

- Requires the head of each agency to assess the management, structure, and operation of the agency's privacy program, and, if necessary, designate or re-designate an official to serve as the SAOP
- Makes clear that the SAOP must serve in a central leadership position and have the necessary authority and expertise to lead the agency's privacy program and carry out all privacy-related functions.
- Requires the SAOP to take a central role at the agency in policy development and evaluation, privacy compliance, and privacy risk management.

4.15 Circular A-130

On July 27, 2016, OMB issued Circular No. A-130, Managing Information as a Strategic Resource. Circular A-130 provides general policy for the planning, budgeting, governance, acquisition, and management of Federal information resources. It also includes appendices outlining agency responsibilities for managing information, supporting use of electronic transactions, and protecting Federal information resources. [107]

4.16 Federal Information Technology Acquisition Reform Act

In November 2017, President Trump signed H.R. 3243, the Federal Information Technology Acquisition Reform Act (FITARA) Enhancement Act of 2017. It was the most significant federal IT reform in almost 20 years.

The act repeals the original expiration date in 2018 and extends the FITARA date to October 1, 2020. It seeks to eliminate duplication and waste in information technology acquisition for the federal government. The goals of the legislation were to reduce duplicative systems, examine software licensing options, make the business case for acquisition, and consolidate data centers.

It not only seeks to consolidate and close existing data centers but prevent new ones from being built. This intended move is to get Federal agencies to adopt the Cloud Smart strategy.

The main requirements of the act are as follows:

- Requires the Federal Chief Information Officer (FCIO) to develop and implement the DCOI to optimize the usage and efficiency of federal data centers.
- Sets forth permitted methods for agencies to consolidate data centers and achieve maximum server utilization and energy efficiency.
- Requires agencies to track costs resulting from implementation of the Initiative within the agency and submit an annual report on such costs to the FCIO.
- Permits CIOs to establish cloud service working capital funds.
- Requires federal computer standards to include guidelines necessary to enable effective adoption of open source software. [108]

4.17 Executive Order 13800

On May 11, 2017, President Trump unveiled Executive Order (EO) 13800 Strengthening the Cybersecurity of Federal Networks and Critical Infrastructure. It focuses on the Cybersecurity of Federal Networks, Critical Infrastructure, and the Nation. In May 2018, the Federal Cybersecurity Risk Determination Report and Action Plan was published.

According to this report, "…OMB and DHS also found that Federal agencies are not equipped to determine how threat actors seek to gain access to their information. The risk assessments show that the lack of threat information results in ineffective allocations of agencies' limited cyber resources. This situation creates enterprise-wide gaps in network visibility, IT tool and capability standardization, and common operating procedures, all of which negatively impact Federal cybersecurity."

The Federal Cybersecurity Risk Determination Report and Action Plan recommended the following four actions:

- Increase cybersecurity threat awareness among Federal agencies by implementing the Cyber Threat Framework to prioritize efforts and manage cybersecurity risks.
- Standardize IT and cybersecurity capabilities to control costs and improve asset management.
- Consolidate agency security operations centers (SOCs) to improve incident detection and response capabilities.

- Drive accountability across agencies through improved governance processes, recurring risk assessments, and OMB's engagements with agency leadership. [109]

4.18 President's Management Agenda

On March 20, 2018, President Trump unveiled The President's Management Agenda (PMA) that layed out a long-term vision for effective government on behalf of the American people. It "...focuses on actions that can have cross-agency impact and break down systems and information silos." PMA focuses on three areas for better government: IT modernization, workforce for the 21st century, and data, transparency, and accountability that will be met by the Cross-Agency Priority (CAP) Goals [110]. The CAP Goals fall into four categories: key drivers of transformation, cross-cutting priority areas, functional priority areas, and mission priority areas. The PMA seeks to improve the effectiveness and efficiency by which Federal agencies serve their constituents and carry out their mission.

In the area of IT modernization, OMB seeks to use the data stored within federal systems more effectively, securely and transparently.

4.19 Executive Order 13833

In May 2018, President Trump signed Executive Order (EO) 13833, Enhancing the Effectiveness of Agency Chief Information Officers to implement successful IT management practices from the private sector, enabling Agencies to reduce costs, execute IT programs more efficiently, reduce cybersecurity risks, better protect sensitive data, and improve services offered to the public [111].

The role of the agency CIO will include:

- Empowering agency CIOs to ensure that agency IT systems are secure, efficient, accessible, and effective, and that such systems enable agencies to accomplish their missions.
- Modernizing IT infrastructure within the executive branch and meaningfully improving the delivery of digital services.
- Improving the management, acquisition, and oversight of Federal IT.

4.20 Modernizing Government Technology Act

In December 2018, President Trump signed The Modernizing Government Technology (MGT) Act. The MGT Act is a powerful incentive for federal CIOs to update their legacy IT systems. OMB issued Memorandum M-18-12, Implementation of the Modernizing Government

Technology Act. The memo "…addresses how agencies can start to apply for funding from the centralized Technology Modernization Fund (TMF), who will staff the board that will oversee the TMF and some guidance on how CFO Act agencies can begin to set up their own IT Working Capital Funds." "The MGT Act set aside "up to $250 million appropriations for the TMF for each of fiscal years 2018 and 2019," bringing the total purse to only $500 million."[112]

In 2018, the Intelligent Transportation Systems (ITS) DataHub became the USDOT's primary storage and access system for ITS data. The system uses shared services such as the NTL and data.transportation.gov (DTG) to provide access to timely, discoverable, well-curated research data for public access.

4.21 Cloud Smart policy

On September 24, 2018, OMB drafted its Cloud Smart policy. The policy is an update to President Obama's "Cloud First" policy, introduced in 2010. It is said to be "a "practical implementation guidance" that aims to help government agencies "fully actualize the promise and potential of cloud-based technologies while ensuring thoughtful execution that incorporates practical realities." The strategy will help agencies achieve additional savings, security, and deliver faster services. It is broken down into three "inter-related" components — security, workforce, and procurement. The CIO Council developed 22 concrete "action items" to be accomplished over 18 months; to date, 13 actions have already been completed. [113]

4.22 Title II: OPEN Government Data Act

In 2017, the U.S. Commission on Evidence-Based Policymaking (CEP) was a 15-member agency in the federal government charged to examine how government could better use its existing data to provide evidence for future government decisions. They had five key recommendations: Strengthen Privacy Protections, Maintain Strong Confidentiality Protections for Sensitive Data, Institute Processes to Assess Data Risks, Enhance Public Trust in Data, and Establish Consistent Leadership on Key Data Issues.

These recommendations are the foundation of H.R. 4174 - Foundations for Evidence-Based Policymaking Act of 2018 ("Evidence Act"), including Title II, the Open, Public, Electronic, and Necessary (OPEN) Government Data Act that President Trump signed on January 14, 2019. Title II: OPEN Government Data Act builds on previous federal open data laws by adding an expectation that the federal government's data will be open and accessible, by default, unless there are restrictions or limits such as for protecting confidentiality and national security. It also makes key aspects of President Obama's May 2013 Open Data Policy permanent.

Key components of Title II: OPEN Government Data Act include the following:

- Key data management requirements, including data inventories, strong metadata standards, and a comprehensive data inventory includes all data assets created by, collected by, under the control or direction of, or maintained by the agency.
- Requires those data assets to be supported by strong metadata, which should help ensure that agency data assets are functionally useful to interested parties.
- Encourages agencies to engage with their data stakeholders, work to increase public and agency use of government data, and make government data more discoverable.
- Directs all federal agencies to publish their information online using standardized, machine-readable data, using searchable, open formats. It requires every agency to maintain a centralized Enterprise Data Inventory that lists all data sets, and mandates a centralized inventory for the whole government on data.gov.
- Requires the GSA work with OMB and the Office of Government Information Services (OGIS) to establish an "online repository of tools, best practices, and schema standards to facilitate the adoption of open data practices across the Federal Government." This new repository, resources.data.gov, will contain a series of tools that agencies can use to support a range of data-related activities. The repository is also available on Data.gov. (See the FDS section)
- Non-sensitive government data should be open by default, while ensuring privacy protections and other potential risks are adequately managed. [114]

The law gives researchers, evaluators, and statisticians inside and outside government the ability to securely use the data the government collects to drive better policy decisions. It requires agency data to be machine-readable (searchable) and requires agencies to plan to develop statistical evidence to support policymaking. It also specifies how people are supposed to access confidential data and information.

Tajha Chappellet-Lanier of FedScoop summarized it well:

- Agencies will be called upon to maintain comprehensive data catalogs and designate a nonpolitical chief data officer (CDO).
- The White House OMB will also create a CDO Council, comprised of CDOs, to "establish government-wide best practices for the use, protection, dissemination, and generation of data."

- “Within three years of the legislation being signed into law, the Government Accountability Office (GAO) is to conduct a study to assess whether agencies have complied with the law, and the value derived from the newly-public data.” [115]

Figure 4 shows how the OMB will implement the Evidence Act in four phases.

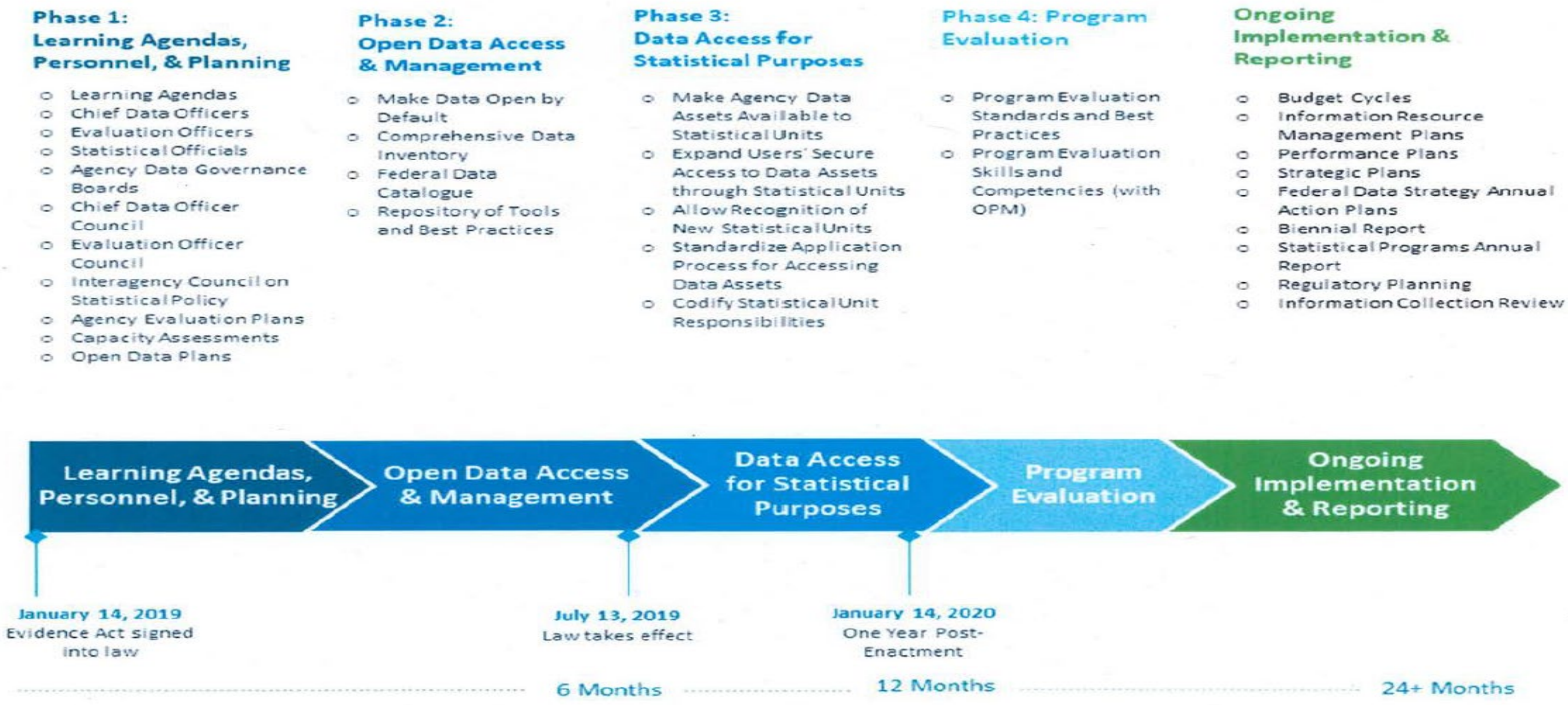
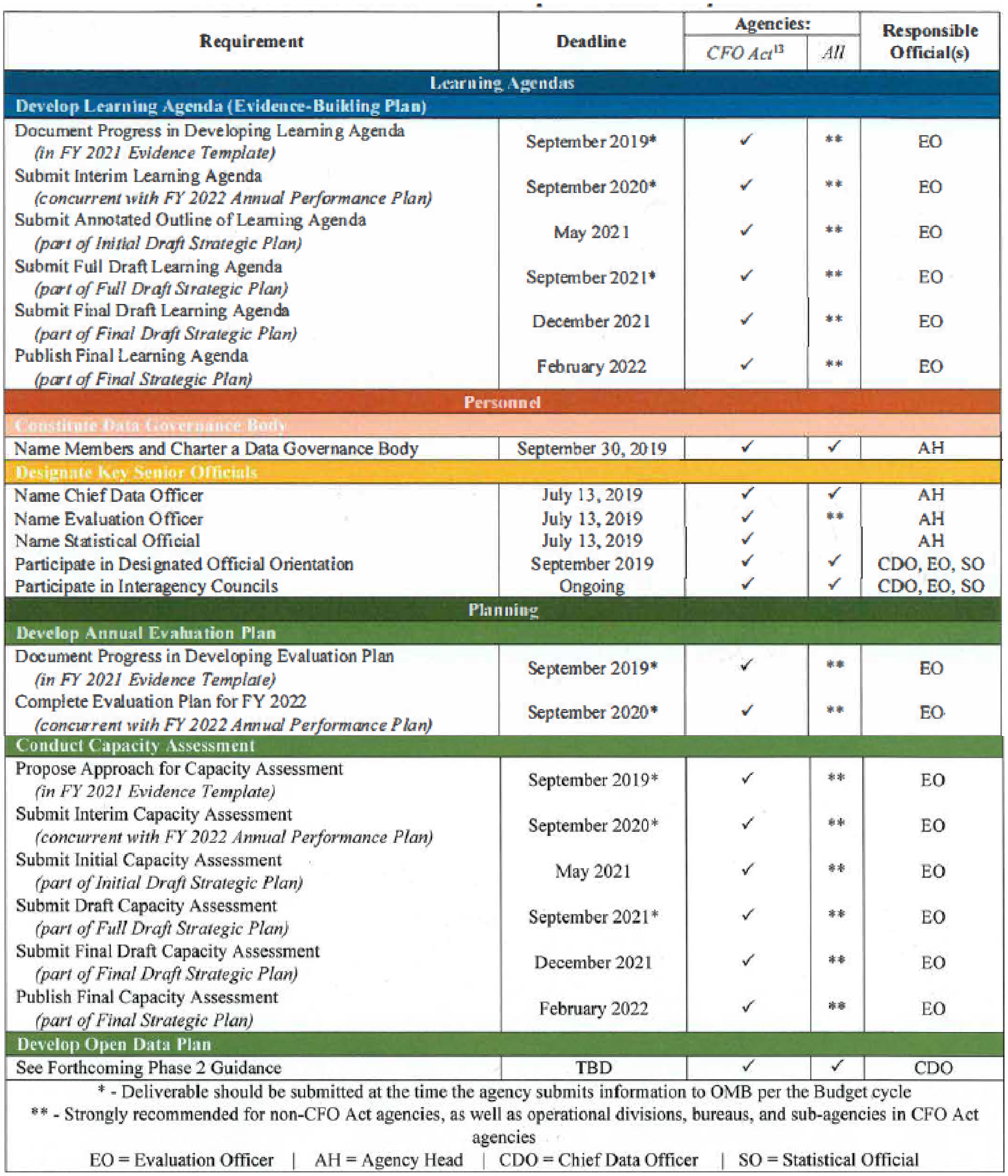

Phase 1: Learning Agendas, Personnel, & Planning
Learning Agendas
Chief Data Officers
Evaluation Officers
Statistical Officials
Agency Data Governance Boards
Chief Data Officer Council
Evaluation Officer Council
Interagency Council on Statistical Policy
Agency Evaluation Plans
Capacity Assessments
Open Data Plans
Phase 2: Open Data Access & Management
Make Data Open by Default
Comprehensive Data Inventory
Federal Data Catalogue
Repository of Tools and Best Practices
Phase 3: Data Access for Statistical Purposes
Make Agency Data Assets Available to Statistical Units
Expand Users' Secure Access to Data Assets through Statistical Units
Allow Recognition of New Statistical Units
Standardize Application Process for Accessing Data Assets
Codify Statistical Unit Responsibilities
Phase 4: Program Evaluation
Program Evaluation Standards and Best Practices
Program Evaluation Skills and Competencies (with OPM)
Ongoing Implementation & Reporting
Budget Cycles
Information Resource Management Plans
Performance Plans
Strategic Plans
Federal Data Strategy Annual Action Plans
Biennial Report
Statistical Programs Annual Report
Regulatory Planning
Information Collection Review
Learning Agendas, Personnel, & Planning
Open Data Access & Management
Data Access for Statistical Purposes
Program Evaluation
Ongoing Implementation & Reporting
January 14, 2019
Evidence Act signed into law
July 13, 2019
Law takes effect
January 14, 2020
One Year Post-Enactment
6 Months
12 Months
24+ Months

Figure 4. Evidence Act – implementation phases

On July 10, 2019, OMB wrote guidance document M-19-23 titled "Phase 1 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Learning Agendas, Personnel, and Planning Guidance." OMB has 16 actions to be implemented during 2019-2020, but the FAA is only responsible for implementing six of these actions. Table 9 shows the initial schedule of implementing the requirements. OMB will provide additional guidance on the other three phases as they approach.

Table 9. Evidence Act – initial implementation requirements



Requirement	Deadline	Agencies:		Responsible Official(s)
		CFO Act[13]	All	
Learning Agendas				
Develop Learning Agenda (Evidence-Building Plan)				
Document Progress in Developing Learning Agenda *(in FY 2021 Evidence Template)*	September 2019*	✓	**	EO
Submit Interim Learning Agenda *(concurrent with FY 2022 Annual Performance Plan)*	September 2020*	✓	**	EO
Submit Annotated Outline of Learning Agenda *(part of Initial Draft Strategic Plan)*	May 2021	✓	**	EO
Submit Full Draft Learning Agenda *(part of Full Draft Strategic Plan)*	September 2021*	✓	**	EO
Submit Final Draft Learning Agenda *(part of Final Draft Strategic Plan)*	December 2021	✓	**	EO
Publish Final Learning Agenda *(part of Final Strategic Plan)*	February 2022	✓	**	EO
Personnel				
Constitute Data Governance Body				
Name Members and Charter a Data Governance Body	September 30, 2019	✓	✓	AH
Designate Key Senior Officials				
Name Chief Data Officer	July 13, 2019	✓	✓	AH
Name Evaluation Officer	July 13, 2019	✓	**	AH
Name Statistical Official	July 13, 2019	✓		AH
Participate in Designated Official Orientation	September 2019	✓	✓	CDO, EO, SO
Participate in Interagency Councils	Ongoing	✓	✓	CDO, EO, SO
Planning				
Develop Annual Evaluation Plan				
Document Progress in Developing Evaluation Plan *(in FY 2021 Evidence Template)*	September 2019*	✓	**	EO
Complete Evaluation Plan for FY 2022 *(concurrent with FY 2022 Annual Performance Plan)*	September 2020*	✓	**	EO
Conduct Capacity Assessment				
Propose Approach for Capacity Assessment *(in FY 2021 Evidence Template)*	September 2019*	✓	**	EO
Submit Interim Capacity Assessment *(concurrent with FY 2022 Annual Performance Plan)*	September 2020*	✓	**	EO
Submit Initial Capacity Assessment *(part of Initial Draft Strategic Plan)*	May 2021	✓	**	EO
Submit Draft Capacity Assessment *(part of Full Draft Strategic Plan)*	September 2021*	✓	**	EO
Submit Final Draft Capacity Assessment *(part of Final Draft Strategic Plan)*	December 2021	✓	**	EO
Publish Final Capacity Assessment *(part of Final Strategic Plan)*	February 2022	✓	**	EO
Develop Open Data Plan				
See Forthcoming Phase 2 Guidance	TBD	✓	✓	CDO

* - Deliverable should be submitted at the time the agency submits information to OMB per the Budget cycle

** - Strongly recommended for non-CFO Act agencies, as well as operational divisions, bureaus, and sub-agencies in CFO Act agencies

EO = Evaluation Officer | AH = Agency Head | CDO = Chief Data Officer | SO = Statistical Official

DOT is complying with this law in a new way. It updated the NTL to become the ITS DataHub for public information. The mission of the ITS Joint Program Office (JPO) is to support "the overall advancement of ITS through investments in major research initiatives, exploratory studies and a deployment support program including technology transfer and training." To further this mission, the ITS JPO Data Center of Excellence is committed to providing timely access to public research data to support third-party research, evaluation, and application development in order to maximize the ITS JPO's investment in ITS research initiatives. [116]

The ITS Public Data Hub uses shared services such as the NTL and data.transportation.gov to provide a single point of entry to discover USDOT's publicly available ITS research data, including connected vehicle data. By providing access to these data, the USDOT aims to enable third-party research into the effectiveness of emerging ITS technologies, preliminary development of third-party applications, and harmonization of data across similar collections.

4.23 Federal Data Strategy (FDS)

FDS is the way that the data portion of the PMA will be implemented. The FDS is an ongoing effort co-led by the OMB to "…develop a 10-year, long-term strategy for better using government data to improve how agencies fulfill their missions, serve the American public, and use their resources efficiently." The mission of the FDS is to leverage the value of federal data for mission, service, and the public good by guiding the Federal Government in practicing ethical governance, conscious design, and a learning culture.

The FDS is intended to help the government accelerate the use of data to drive and deliver mission objectives [117]. The goal of the FDS is to leverage data as a strategic asset to grow the economy, increase the effectiveness of the Federal Government, facilitate oversight, and promote transparency. The common problem: federal agencies' do not share data across their agencies. OMB wants to break down these silos so that agencies owning data can share useful data with other agencies.

One of the Cross-Agency Priority (CAP) goals is Leveraging Data as a Strategic Asset that recognizes the value of modern and accessible federal data, and calls for developing a long-term enterprise-wide FDS. It seeks to define principles, practices, and a one-year action plan for delivering a more consistent approach to federal data stewardship, use, and access. [118]

The FDS consists of:

- Principles – The ten principles that serve as timeless guidance to agencies in the areas of Ethical Governance, Conscious Design, and Learning Culture. These principles were useful in developing practices, the 2020 Action Plan, and the subsequent action steps for the Strategy.
- Practices – The 40 practices (5-10 year goals) guide agencies that are organized into three categories: Building a Culture that Values Data and Promotes Public Use (Practices 1-10); Governing, Managing, and Protecting Data (Practices 11-26); and Promoting Efficient and Appropriate Data Use (Practices 27-40). The practices informed the development of the 2020 Action Plan and will inform the development of subsequent action steps for the Strategy.
- Annual Action Plan – Specifies measurable activities to implement the practices and that are the priority for a given year, providing timeframes for implementation and identification of responsible parties.

Under the FDS, federal agencies are required to undertake 20 actions in 2019-2020. These actions are divided into three sections: shared actions, community actions, and agency-specific actions that must be completed within a specific timeframe. Figure 5 shows the progress that federal agencies have made towards completing these 20 actions that were last updated on January 29, 2021.

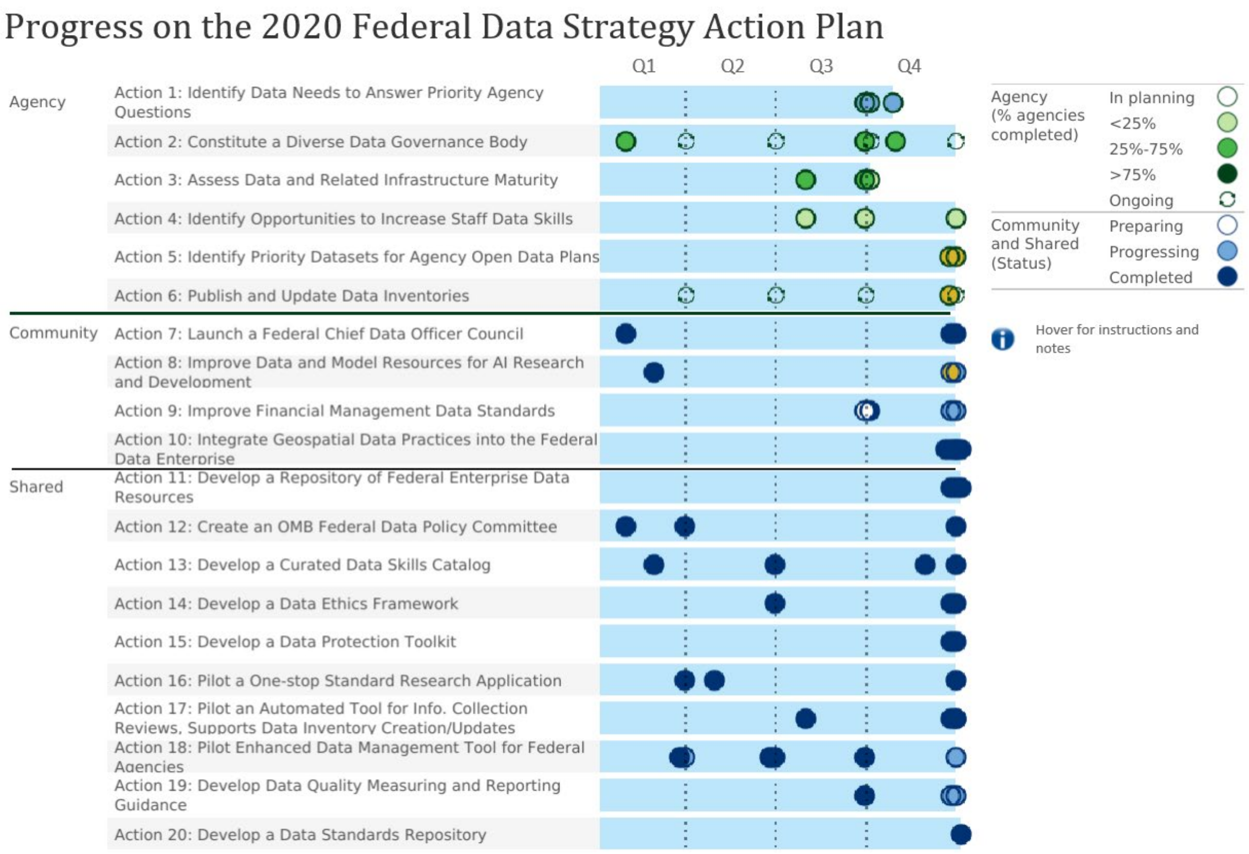


Figure 5. Progress on the 2020 Federal Data Action Plan

Note: To see more detailed info on Figure 5, please visit https://strategy.data.gov/progress/

Resources.data.gov is an effort by the OMB, the Office of Government Information Services of the National Archives, and the GSA to provide an online repository of guidance and tools to implement the OPEN Government Data Act. The content on the site is maintained by the Data.gov Program Management Office in GSA Technology Transformation Services (TTS), the Office of Government Information & Oversight, and the OMB.

Resources.data.gov is as an online repository for Federal Enterprise Data Resources. It contains tools, best practices, case studies, playbooks, and guidance on how to manage and use Federal data. The repository will be frequently updated with additional tools and resources. It will be developed in an open, collaborative manner and welcomes suggestions from the public. It has replaced the standard Project Open Data metadata schema.

This sounds good but this effort is new and there are not a lot of details about what is on resources.data.gov.

4.23.1 Process to upload data to data.gov

Under the OPEN Government Data Act and the Open Data Policy, federal agencies are required to publish metadata, provided as a data.json file, using the standard Project Open Data metadata schema. Data.gov does not host data but aggregates metadata. The machine-readable listing, as a standalone JSON file on the agency's website at `agency.gov/data.json.` The `data.json` file is what gets harvested to the Data.gov catalog.

A JSON file stores simple data structures and objects in a standard data interchange format known as the JavaScript Object Notation (JSON) format. Each record in the data.json files must meet metadata schema and have the minimum required fields. They are used to transmit data between a web application and a server. JSON files are lightweight, text-based, human-readable, and can be edited using a text editor. [119]

When an agency is ready for Data.gov to harvest its data.json for the first time, the agency should notify Data.gov via email. The Data.gov team will create a new Data.gov harvest source for the data.json. The team will assist agencies in generating the data.json file and provide tools that may help agencies prepare their data listings [120]. They can synchronize the source's metadata as often as every 24 hours.

4.24 Executive Order 13859

On February 11, 2019, President Trump issued Executive Order 13859: Maintaining American Leadership in Artificial Intelligence (AI). It is "[d]ubbed the 'American AI Initiative,' it aims to boost AI research and regulation efforts across the United States' federal, academic, and private sectors."[121]

Since AI will affect the missions of nearly all executive departments and agencies, they shall pursue the following strategic objectives:

- Promote sustained investment in AI R&D in collaboration with industry, academia, international partners and allies, and other non-Federal entities to generate technological breakthroughs in AI and related technologies and to transition those breakthroughs rapidly into capabilities that contribute to our economic and national security.
- Enhance access to high-quality and fully traceable Federal data, models, and computing resources to increase the value of such resources for AI R&D, while maintaining safety, security, privacy, and confidentiality protections consistent with applicable laws and policies.

- Reduce barriers to use AI technologies to promote their innovative application while protecting American technology, economic and national security, civil liberties, privacy, and values.
- Ensure that technical standards minimize vulnerability to attacks from malicious actors and reflect Federal priorities for innovation, public trust, and public confidence in systems that use AI technologies; and develop international standards to promote and protect those priorities.
- Train the next generation of American AI researchers and users through apprenticeships; skills programs; and education in science, technology, engineering, and mathematics (STEM), with an emphasis on computer science, to ensure that American workers, including Federal workers, are capable of taking full advantage of the opportunities of AI.
- Develop and implement an action plan, in accordance with the National Security Presidential Memorandum of February 11, 2019 [122]

This EO may sound good; however, there are serious issues with it. If Congress does not allocate funding to increase AI R&D efforts and to create incentives to ensure that agencies are prioritizing the initiatives they are asked to prioritize, nothing will probably happen with it. While the EO discusses the need to leverage federal data to improve AI systems while simultaneously protecting the privacy of Americans, the White House needs to ensure that the Justice Department has a role in protecting Americans' privacy. Because the EO does not define AI, it will be difficult for federal agencies to determine how to execute their responsibilities under the EO. The government and the public need more guidance as to what technology the order covers. [123]

Regardless, GSA and the Federal CIO are launching a government-wide AI Community of Practice (CoP) to harness those advancements and accelerate the thoughtful adoption of AI across the federal government.

4.25 OMB Memorandum M-19-19

On June 25, 2019, OMB issued Memorandum M-19-19, Update to Data Center Optimization Initiative (DCOI). M-19-19 rescinds the August 1, 2016 M-16-19 - Data Center Optimization Initiative. DCOI helps federal agencies meet OMB requirements to consolidate and modernize IT infrastructure.

"OMB is looking to retool security to provide flexibility for cloud access, improve the skills of the workforce when it comes to working with cloud and refine procurement methodology to accommodate the pay-as-you-go nature of commercial cloud computing." [124]

The FAA is affected by this initiative and it had until October 1, 2020 to meet the FITARA enhancements. The DCOI PMO helps agencies comply with federal policies impacting data center, cloud, and IT infrastructure optimization by acquiring technologies, tools, and evidence-based best practices to meet requirements in OMB Memo and the Federal Cloud Computing Strategy, better known as Cloud Smart. [125]

As part of the DCOI, CIOs must "…submit the following to OMB annually:

- A complete inventory of the data centers owned, operated, or maintained by (or on behalf of) the agency.
- A multi-year strategy to consolidate and optimize these data centers. Each agency (under its CIO's direction) must submit quarterly updates on their progress towards activity completion, consolidation and optimization metrics, and cost savings realized through the implementation of their strategy." [126]

Cloud Smart also ask agencies to reduce application portfolio by 1) assessing the need for and usage of applications and 2) discarding obsolete, redundant, or overly resource-intensive applications. Decreased application management responsibilities will free agencies to focus on improving service delivery and optimize their remaining applications.

4.26 GREAT Act of 2019

On December 30, 2019, President Trump signed the HR 150 Grant Reporting Efficiency and Agreements Transparency (GREAT) Act of 2019 into law. The GREAT Act gives agencies three years to update their data collection systems, as follows:

- Requires OMB to designate a single data standard-setting agency (HHS, the government's largest grantmaking agency) to work with it, and in consultation with other stakeholders, to develop a set of data set standards by December 30, 2021.
- OMB mandates that a set of unique identifiers for federal awards and grant recipients be developed that are consistently applied across the government.
- The data standards require that information collected by the federal government from grantees be fully searchable and machine-readable, be nonproprietary, and incorporate any standards already created under the DATA Act.

The GREAT Act will reduce costs, improve efficiency, enhance management, and encourage new technologies. The GREAT Act will improve oversight of federal funding, provide greater transparency about how funds are used, and enhance capabilities to compare grantees with interoperable information. [127]

5 Data analytics

This paper has discussed preparing data for analysis. Now that the grunt work is done, it is time for the magic: using the data. BD analytics is the complex process of examining large and varied data sets to uncover information (i.e. hidden patterns, unknown correlations, market trends, and customer preferences) that can help organizations make informed business decisions.

Big data analytics is the science of analyzing raw data so conclusions can be made about that information. It allows organizations in the public and private sectors to examine large data sets to uncover hidden patterns, correlations, and other insights. It helps organizations harness their data and use it to identify new opportunities [128]. For example, big data analytics helps the government in building smart cities by providing faster, reliable services to its citizens.

There are four types of analytics (ranging from simplest and of least value to most complex and of most value):

1. Descriptive analytics answers the question "what happened?" It seeks to find out the reasons behind precious success or failure based on the past. They are based on standard aggregate functions in databases.
2. Diagnostic analytics answers the question "why did is it happen?" It gives in-depth insights into a particular problem.
3. Predictive analytics answers the question "what is likely to happen in the future?" based on historical data.
4. Prescriptive analytics answers what is the best course of action based on previous trends and patterns. It highlights problems and helps businesses understand why those problems occurred. Businesses use the data-backed and data-found factors to discover solutions to business problems. Prescriptive analytics answers the question "what action needs to be taken to eliminate a future problem or take full advantage of a promising trend."

Prescriptive analytics equals data (internal and external) plus various business rules (preferences, best practices, boundaries, and other constraints). It uses the findings of descriptive and diagnostic analytics to detect clusters and exceptions, and to predict future trends, which makes it a valuable tool for forecasting [129]. It consists of statistical techniques from data mining, predictive modelling, artificial intelligence (AI), deep learning, neural networks, text analysis, and machine learning (ML).

Predictive analytics uses various statistical and ML algorithms to predict the likelihood of a future outcome. The accuracy of predictions is not 100%, because it is based on probabilities. It consists of capturing data, predicting outcomes, and acting on insights.

"Sentiment analysis is the most common kind of predictive analytics. The learning model takes input in the form of plain text and the output of the model is a sentiment score that helps determine whether the sentiment is positive, negative or neutral." [130]

Predictive analytics is just a term used to describe the application of data mining. It uses technology to learn and predict the future. It is useful for analyzing huge data automatically with multiple variables; it is inclusive of decision trees, clustering, neural nets, market basket analysis, regression modeling, hypothesis testing, decision analytics, genetic algorithms, text mining, etc. It contains different view approaches, like integrated reasoning and pattern recognition, along with predictive modeling. Many researchers use it to identify future events and measures.

A predictive model identifies patterns observed in historical and transactional data so that potential risks and opportunities can be determined. It is used to give a probability, not certainty. "Predictive analytics can be applied to a range of business strategies and has been a key player in search advertising and recommendation engines. These techniques can provide managers and executives with decision-making tools to influence upselling, sales and revenue forecasting, manufacturing optimization, and even new product development." [131]

Three fundamental strategies for predictive analytics include:

1. Data profiling and transformations – the functions that modify the row and column attributes, combine the fields, evaluate the dependencies, aggregate the records and data formats, and build rows and columns.
2. Sequential pattern analysis – determines that the relationships exist between the rows in a database. Sequential pattern analysis is involved with the identification of the sequentially occurring items that are frequently seen across the ordered transactions over time.
3. Time series tracking – a sequence that is ordered with values at different time intervals spaced with the equal distance. Time series analysis provides the conception of data points that are plotted over time. [132]

Data analytics may start with a specific hypothesis. That is, its purpose is testing that hypothesis. It advises on possible outcomes and results in actions that are likely to maximize key business metrics. Prescriptive analytics are comparatively complex in nature. Many companies are not using them in day-to-day business activities yet, because they are difficult to manage. Prescriptive analytics if implemented properly can have a major impact on business growth.

5.1 Data mining

Data mining (or knowledge mining) is "…[t]he analysis of large and complex data sets."[133] It is a process of extracting useful information and discovering hidden patterns or trends from large data sets.

Dimensionality reduction is a technique used to reduce noise and find important variables or combination of variables that are the most informative. Various algorithms, such as those used in machine learning, are used to find unknown patterns or relationships. Data mining uses computational strategies from statistics, ML, and pattern recognition as tools in its search for new knowledge without any preconceived notions.

Data mining involves the following three stages:

- Stage 1: Exploration – preparing data (clean and transform into another form); important variables and the nature of the data based on the problem are determined.
- Stage 2: Build a model and validate it – identify and choose the model that makes the best prediction. This is not always easy since there are so many from which to choose.

  There are seven popular data mining techniques:

  1. Tracking patterns is a basic technique that recognizes patterns in data sets. It seeks to recognize some aberration in the data happening at regular intervals, or an ebb and flow of a certain variable over time. [134]
  2. Classification constructs a model that labels a group of data objects into a specific category. In the classification model, the classes with their own labels are discrete in nature. It uses a decision tree or neural network-based classification algorithms. For instance, the same classification model can categorize people into groups of trustworthy and untrustworthy users of an online banking system. Some types of classification models include: classification by decision tree induction, Bayesian Classification, Neural Networks, Support Vector Machines (SVM), and Classification Based on Associations.

3. Prediction builds a model that produces continuous or ordered values that form a trend. For instance, a prediction model can provide estimated mean time to failure (MTTF) values for a computer.
4. Regression is used to identify and analyze the relationship between variables to identify the likelihood of a certain variable, given the presence of other variables.
5. Outlier detection helps to identify anomalies, or outliers in the data.
6. Association rules help find the association between two or more items and to discover hidden patterns in data sets.
7. Clustering is a process of grouping similar data objects into a class. Unlike classification, there are no predefined class labels. It is used to reveal features that distinguish one class of data objects from the other. This can lead to new discoveries on a dataset. It can be an effective way to distinguish groups or classes of objects. Clustering analysis range from pattern recognition and image processing to market research. Some clustering methods include Partitioning methods, Hierarchical Agglomerative (divisive) methods, Density-based methods, Grid-based methods, and Model-based methods.

- Stage 3: Deployment – patterns are implemented.

To ensure that data mining is successful, it is vital to follow the five commandments in data mining:

1. A copy of the raw data should always be maintained to ensure that one could access the history in case there is ever a question.
2. The data must not be archived or deleted, except under rare circumstances.
3. The data must be time-stamped.
4. The semantics of the data must be consistent and accurate.
5. The data must not be overwritten. [135]

Some challenges of data mining include:

- Skilled experts need to formulate the data mining queries.
- Overfitting: Due to small size training database, a model may not fit future states.
- Data mining needs large databases, which sometimes are difficult to manage.
- Business practices may need to be modified to determine how to use the information uncovered.
- If the data set is not diverse, data mining results may not be accurate.
- Integration information needed from heterogeneous databases and global information systems could be complex. [136]

While data mining can produce information that can be used to increase revenue and cut costs, data privacy is a growing concern. Work is focused on privacy-preserving data mining, proposing novel techniques to extract knowledge and protect the users' privacy.

Five of the Best Open-Source Data Mining Tools are as follows:

1. Orange – A component-based data mining and ML software suite that features a friendly yet powerful, fast and versatile visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting.
2. RapidMiner – Environment for ML and data mining experiments that is utilized for both research and real-world data mining tasks. It enables experiments to be made up of a huge number of arbitrarily nestable operators, which are detailed in XML files and are made with the graphical user interface of RapidMiner.
3. Weka – (Waikato Environment for Knowledge Analysis) is a well-known suite of ML software that supports several typical data mining tasks, particularly data preprocessing, clustering, classification, regression, visualization, and feature selection. It is written in Java.
4. jHepWork – Designed for scientists, engineers and students, and created as an attempt to make a data-analysis environment using open-source packages with a comprehensible user interface that is competitive to commercial programs.
5. KNIME – (Konstanz Information Miner) is a user-friendly, intelligible, and comprehensive open-source data integration, processing, analysis, and exploration platform. It allows users to visually create data flows or pipelines, selectively execute some or all analysis steps, and later study the results, models, and interactive views. KNIME is written in Java. [137]

RHadoop by Revolution Analytics is a collection of five R packages that allow users to manage and analyze data with Hadoop.

5.2 Machine learning (ML)

ML is based on self-learning or self-improving algorithms. It learns by itself through training data. It is a form of AI and requires little to no human interaction. It works based on a model, which continues to improve itself through trial and error. It can then provide meaningful insight in the form of classification, prediction, or clustering.

There are three types of ML:

1. Supervised learning – algorithms that learn the relationship between a given input and output from use training data and feedback from humans. An example of supervised learning is object recognition.
2. Unsupervised learning – there is no training data. The ML algorithm solely depends on clustering and continues to enhance its algorithm from the data itself.
3. Reinforcement (RL) – training ML models through trial and error to make decisions sequentially. It is considered the hottest branch of AI that tech-driven companies are adopting quickly. Unlike supervised learning, RL does not use a training dataset. An example of RL is the game of chess. [138]
   The top RL tools include OpenAI Gym, TensorFlow, Keras, DeepMind Lab, MATLAB, and Pytorch. [139]

Matthew Mayo [140] from KDnuggets outlines seven steps to ML:

1. Data Collection
   - The quantity and quality of the data dictate how accurate the model is.
   - The outcome of this step is generally a representation of data (Guo simplifies to specifying a table) which will be used for training.
   - Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step.
2. Data Preparation
   - Wrangle data and prepare it for training.
   - Cleaning (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.) may be required.
   - Randomize data, which erases the effects of the particular order in which it was collected and/or otherwise prepared.
   - Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analyses.
   - Split into training and evaluation sets.
3. Choose a Model
   - Different algorithms are for different tasks; choose the right one, which can be a difficult task.
4. Train the Model
   - The goal of training is to answer a question or make a prediction correctly as often as possible.

   - Linear regression example: For a linear regression model in the form of y = mx + b, where y is the output and x is the input, the model would need to learn the values of m and b in order to minimize the error between predicted outputs and actual outputs.
   - Each iteration of process is a training step.

5. Evaluate the Model
   - Uses some metric or combination of metrics to "measure" objective performance of the model.
   - Test the model against previously unseen data.
   - This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not).
   - Establish a good training/evaluation data split, such as 80/20, 70/30, or similar, depending on domain, data availability, dataset particulars, etc.
6. Parameter Tuning
   - This step refers to hyperparameter tuning, which is an "artform" as opposed to a science.
   - Tune model parameters for improved performance.
   - Simple model hyperparameters may include the number of training steps, learning rate, initialization values and distribution, etc.
7. Make Predictions
   - Using further (test set) data, which have until this point been withheld from the model (and for which class labels are known), to test the model: a better approximation of how the model will perform in the real world.

5.3 Data Visualization

Data visualization is the process of transforming raw data into tables, graphs, images, or video so people can gain insights and understand patterns easily in a meaningful way. It allows decision makers to see connections between multi-dimensional data sets. It also provides new ways to interpret data, see patterns and relations that occur between operations and business performance, identify emerging trends faster and in real time, track connections between operations and overall business performance, and interact with data directly. [141]

As the previous paragraph demonstrates, data visualization is a vital part of business intelligence, especially because it has interactive functionality. While data visualization is a powerful asset, the business world ignores it and does not see the value in it.

When working with massive amounts of data, it can be challenging to display summary data in a way that is not overwhelming. Displaying data in the best way possible requires collapsing and condensing data results in an intuitive fashion, while still displaying graphs and charts that decision makers are accustomed to seeing. To ensure that you use the best visuals to display your data:

- Understand the data you are trying to visualize, including its size and cardinality, and consider (honestly) the data preparation effort that will be required.
- Determine what content you want to visualize and what kind of information you want to communicate. This may require drilling down or secondary knowledge that is related to the dataset (such as what is not present, or private business models).
- Know your audience and understand how it processes visual information.
- Use a visual information in the best and simplest form for your audience.
- Consider speed, size, and diversity of data. [142]

To create meaningful visuals, you should consider data size, data type, and column composition.

The four benefits of data visualization are as follows:

1. Explores new patterns and reveals hidden ones.
2. Identifies relationships between data points and variables.
3. Identifies areas for improvement.
4. Makes it easier to explore data and simply complex quantitative data.

With so many data visualization tools (including Tableau, Microsoft Power BI, and SAP Crystal Reports) on the market today, it is hard to know which type to use when. The following list gives some guidance on the best type to use when.

1. Number Chart – gives an immediate overview of a specific value.
2. Line Chart – shows trends and change of data over a period of time.
3. Maps – visualizes data by geographical location.
4. Waterfall Chart – demonstrates the static composition of data.
5. Bar Graphs – used to compare data of many items.
6. Pie Chart – indicates the proportional composition of a variable.
7. Gauge Chart – used to display a single value within a quantitative context.
8. Scatter Plot – applied to express relations and distribution of large sets of data.
9. Spider Chart – comparative charts that are great for rankings, reviews, and appraisals.
10. Tables – shows a large number of precise dimensions and measures.
11. Area Chart – portrays a part-to-whole relationship over time.

12. Bubble Plots – visualizes two or more variables with multiple dimensions.

When deciding which data visualization tool to use, Dr. Kristen Sosulski, who develops innovative practices for higher education as the Director of the Learning Science Lab, suggests asking the following questions:

- Can others view and edit your visualization and analysis?
- Can you publish visualizations to the web, create PDF documents, and embed them into other applications?
- How easily can you connect to other data sources?
- What types of visualizations do you intend on building?
- Do you need a tool to explore and present your data visually, or to present a data visualization? Are you looking to create charts and graphs quickly?
- Do you think that you will have to go back and revise the visualizations you create? [143]

There are four broad categories of big data analytics challenges, as described below.

1. Data storage technologies do not possess the required performance for processing big data. The challenge is that algorithms do not always respond in an adequate time when dealing with these high dimensional data. It is essential to design ML algorithms to analyze data while improving efficiency and scalability.

   Automating this process and developing ML algorithms to ensure consistency is a major challenge in recent years. The solution is to transform the semi structured or unstructured data into structured data, and then apply data mining algorithms to extract knowledge.

2. Knowledge discovery and computational complexities

   While big data keeps growing exponentially, current tools may not be efficient enough to process these data and obtain meaningful information. Currently, systematic modeling is used when computational complexity issues arise. Consequently, there is a push to develop techniques and technologies that can effectively handle computational complexities, uncertainty, and inconsistencies.

3. Scalability and Visualization of Data

   Because data size is growing exponentially compared to CPUs, processor technology is now being embedded with an increasing number of cores, which has led to the development of parallel computing. A CPU lacks processing power to perform deep learning or ML because it has a finite number of cores and does single processes, which is called serial processing.

On the other hand, a graphic processing unit (GPU) is ideal for ML because it has an infinite number of cores and uses parallel processing. ML is also ideal for performing complex computations. Although GPUs are faster, the time taken to transfer huge amounts of data from a CPU to a GPU can lead to higher overhead time depending on the architecture of the processors.

4. Information Security

   Preserving sensitive information is a major issue in big data analysis. However, it can be enhanced by using the techniques of authentication, authorization, and encryption. To ensure privacy, a robust, multi-level security policy model and prevention system needs to be implemented. [144]

It should be noted that ML should be done on the cloud and not on-premises. If ML is done on the cloud, you get high performance and do not have to worry about updating GPU technology, which changes about once a year. Another reason to use the cloud is it is cheaper, you only pay for what you use, and you do not have to invest in infrastructure. [145]

The FAA plans to use ML on EIM. While it is certainly not up and running, there is a KSN that showcases some of the ways that the FAA is using ML. Employees are encouraged to add their own stories to the KSN.

6 Recommendations

Based on the information above, the best practices outlined below should be kept in mind when working with the big data pipeline.

Data ingestion

- Ensure that the data comes in clean and free of errors and in the right format. If data is not clean coming into the data lake, it will cost time, money, and poor data quality. Furthermore, this can lead to inaccurate or faulty results when running reports and data analytics.
- Find an automation tool that performs data ingestion, ETL, and metadata tasks.

Storage

- Object storage is ideal because it is less complex, requires fewer administrative resources, and is cheaper to use. This technology is ideal for R&D unstructured data because it clumps data, metadata tags, and a unique ID together to make one object.

- Before deciding which storage technology to use, organizations must consider the following factors: security capabilities to protect sensitive data, data storage locations, service-level agreements, technical support, how much space is needed, how frequently analytics will be performed, budget constraints, what features are available on mobile applications, and what types of data are to be processed.[44] [45]
- To develop a sound data storage management strategy, experts suggested that organizations need to know the data, understand compliance issues, have a data retention policy, look for a solution that fits the data, not be intimidated by upfront costs, and choose a storage provider carefully, based on an organization's needs. [46]

Security

- Use encryption whenever possible to secure sensitive data and do so before sending it to a cloud service provider (CSP).
- Use a trusted CSP and ensure that they cannot access data.
- To determine if any malicious operations are being performed or any malicious user is manipulating the data in the nodes, organizations should keep a log and do periodic audits of who changes data in a file.

Architecture

A big data architecture is the overarching system used to ingest and process data so it can be analyzed. The architectural design must support BD functional requirements: ability to capture, organize, integrate, and analyze data as well as to act on the results of the analysis.

A BD architecture varies on an organization's infrastructure and needs, but it usually contains the following four logical vertical layers:

1. Information integration – Responsible for connecting to various data sources. The layer is used by components to store information in big data stores and to retrieve information from big data stores for processing.
2. Big data governance – Guidelines that help enterprises make the right decisions about the data.
3. Quality of service – Layer responsible for defining data quality, policies around privacy and security, frequency of data, size per fetch, and data filters.
4. Systems management – Monitors the health of the overall big data ecosystem.[68]

The five architectural principles for building BD systems include:

1. Building decoupled systems – Decoupled systems in big data are important because they allow one to alter one aspect of the system without affecting the other.
2. In choosing the right tool, one should consider- data structure, latency, throughput, and access patterns.
3. Leverage managed services – Businesses should use CPS managed services because they are scalable/elastic, available, reliable, secure, and low maintenance. The CPS should perform management tasks so customers can focus on their core jobs.
4. Use log-centric design patterns – Since storage is cheap, most organizations do not need to delete any of their data. Organizations should build their big data system in a log-centric fashion. Immutable log files, which are protected from tampering, are copies of the original data in case anything happens to the system.
5. Be cost conscious – If the bill becomes too high, the company may need to consider some different products. As a rule of thumb, the lower cost tools are often the most used. [69]

Data Analytics

Data analytics is generally more focused than big data because instead of gathering huge piles of unstructured data, data analysts have a specific goal in mind and sort through relevant data to look for ways to gain support. On the other hand, big data is a collection of a huge volume of data that requires a lot of filtering out to derive useful insights from it. [146]

Predictive analytics is the most beneficial for R&D data because it answers the question "*w*hat is likely to happen in the future?" It consists of statistical techniques from data mining, predictive modelling, artificial intelligence (AI), deep learning, neural networks, text analysis, and machine learning (ML).

Platforms

- Before selecting a big data platform, organizations should evaluate its architecture, infrastructure, requirements, budget, and business culture.
- Before deciding on a particular platform for a certain application, an organization should consider a number of factors i.e., data size, speed, throughput optimization, and how long it takes to build a model.

Laws and policies

Table 10 gives a summary of all relevant Executive Orders (EOs), laws, and Office of Management and Budget (OMB) memorandums written to address what federal agencies under its jurisdiction must do to comply with various aspects of BD.

Table 10. Big data laws affecting the FAA

Law/Year/Topic	**OMB Requirements**
Open Government Directive 2009, Open Data	• Publish government information online • Improve the quality of government information -identify and correct data quality problems, emphasis on immediate action on the quality of federal spending data. • Create and institutionalize a culture of open government • Create an enabling policy framework for open government- update current policies governing information management. The Office of Information and Regulatory Affairs must review existing policies "such as Paperwork Reduction Act guidance and privacy guidance" to identify problems and issue revisions to allow openness to move forward.
Federal Data Center Consolidation Initiative (FDCCI) 2010, Storage	• Promote the use of green IT by reducing the overall energy and real estate footprint of government data centers • Reduce the cost of data center hardware, software, and operations • Increase the overall IT security posture of the Federal Government • Shift IT investments to more efficient computing platforms and technologies.
Executive Order 13642: Making Open and Machine Readable the New Default for Government Information 2013, Open Data	• Use machine-readable and open formats and data standards • Ensure information stewardship through the use of open licenses • Use common core and extensive metadata • Build information systems to support interoperability and information accessibility • Create and maintain an Enterprise Data Inventory • Create and maintain a public data listing • Create a process to engage with customers to help facilitate and prioritize data release.

Law/Year/Topic	OMB Requirements
Federal Information Security Management Act of 2014 (FISMA). 2014, Security	• Authorizes the Secretary of Department of Homeland Security (DHS) to assist the OMB Director in administering the implementation of agency information and security practices for federal information systems. • Changes the agency reporting requirements, modifying the scope of reportable information from primarily policies and financial information to specific information about threats, security incidents, and compliance with security requirements. Directs the OMB Director to provide guidance on what constitutes a "major incident" as it applies to agency reporting requirements. • Addresses cyber breach notification requirements. • The OMB Director is required to revise Budget Circular A-130 to eliminate inefficient or wasteful reporting. This would allow federal agency information security personnel to allocate more resources to the protection of government systems.
Executive Order 13719: Establishment of the Federal Privacy Council 2016, Privacy	• Requires the head of each agency to assess the management, structure, and operation of the agency's privacy program, and, if necessary, designate or re-designate an official to serve as the Senior Agency Officials for Privacy (SAOP). • Makes clear that the SAOP must serve in a central leadership position and have the necessary authority and expertise to lead the agency's privacy program and carry out all privacy-related functions. • Requires the SAOP to take a central role at the agency in policy development and evaluation, privacy compliance, and privacy risk management.
Federal Information Technology Acquisition Reform Act (FITARA) Enhancement Act of 2017 Storage	• Require the Federal Chief Information Officer (FCIO) to develop and implement an initiative to be known as the Federal Data Center Optimization Initiative to optimize the usage and efficiency of federal data centers. • Set forth permitted methods for agencies to consolidate data centers and achieve maximum server utilization and energy efficiency. • Require agencies to track costs resulting from implementation of the Initiative within the agency and submit an annual report on such costs to the FCIO. • Permit CIOs to establish cloud service working capital funds. • Require federal computer standards to include guidelines necessary to enable effective adoption of open source software.

Law/Year/Topic	OMB Requirements
Executive Order 13800: Strengthening the Cybersecurity of Federal Networks and Critical Infrastructure, 2017 Cybersecurity	• Increase cybersecurity threat awareness among Federal agencies by implementing the Cyber Threat Framework to prioritize efforts and manage cybersecurity risks. • Standardize IT and cybersecurity capabilities to control costs and improve asset management. • Consolidate agency Security Operations Centers (SOCs) to improve incident detection and response capabilities. • Drive accountability across agencies through improved governance processes, recurring risk assessments, and OMB's engagements with agency leadership.
Foundations for Evidence-Based Policymaking Act of 2018 ("Evidence Act"), including Title II, the Open, Public, Electronic, and Necessary (OPEN) Government Data Act 2018 Open Data	• Key data management requirements, including data inventories, strong metadata standards, and a comprehensive data inventory includes all data assets created by, collected by, under the control or direction of, or maintained by the agency. • Requires those data assets to be supported by strong metadata, which should help ensure that agency data assets are functionally useful to interested parties. • Encourages agencies to engage with their data stakeholders, work to increase public and agency use of government data, and make government data more discoverable. • Directs all federal agencies to publish their information online using standardized, machine-readable data, using searchable, open formats. It requires every agency to maintain a centralized Enterprise Data Inventory that lists all data sets, and mandates a centralized inventory for the whole government on data.gov. • Requires that the GSA work with OMB and the Office of Government Information Services (OGIS) to establish an "online repository of tools, best practices, and schema standards to facilitate the adoption of open data practices across the Federal Government." This new repository, found at resources.data.gov, will contain a series of tools that agencies can use to support a range of data-related activities. It will also be available on data.gov. • Non-sensitive government data should be open by default, while ensuring privacy protections and other potential risks are adequately managed.
The Modernizing Government Technology (MGT) Act. 2018	• Addresses how agencies can start to apply for funding from the centralized Technology Modernization Fund (TMF), who will staff the board that will oversee the TMF and some guidance on how CFO Act agencies can begin to set up their own IT Working Capital Funds.

Law/Year/Topic	OMB Requirements
Federal Data Strategy 2020 Action Plan	6 agency goals to be met during 2020: • Identify Data Needs to Answer Priority Agency Questions • Constitute a Diverse Data Governance Body • Assess Data and Related Infrastructure Maturity • Identify Opportunities to Increase Staff Data Skills • Identify Priority Data Assets for Agency Open Data Plans • Publish and Update Data Inventories
Executive Order 13859: Maintaining American Leadership in Artificial Intelligence 2019	• Promote sustained investment in AI R&D in collaboration with industry, academia, international partners and allies, and other non-Federal entities to generate technological breakthroughs in AI and related technologies and to rapidly transition those breakthroughs into capabilities that contribute to our economic and national security. • Enhance access to high-quality and fully traceable Federal data, models, and computing resources to increase the value of such resources for AI R&D, while maintaining safety, security, privacy, and confidentiality protections consistent with applicable laws and policies. • Reduce barriers to the use of AI technologies to promote their innovative application while protecting American technology, economic and national security, civil liberties, privacy, and values. • Ensure that technical standards minimize vulnerability to attacks from malicious actors and reflect Federal priorities for innovation, public trust, and public confidence in systems that use AI technologies; and develop international standards to promote and protect those priorities. • Train the next generation of American AI researchers and users through apprenticeships; skills programs; and education in science, technology, engineering, and mathematics (STEM), with an emphasis on computer science, to ensure that American workers, including Federal workers, are capable of taking full advantage of the opportunities of AI. • Develop and implement an action plan, in accordance with the National Security Presidential Memorandum of February 11, 2019.

Law/Year/Topic	**OMB Requirements**
Memorandum M-19-19, Update to Data Center Optimization Initiative (DCOI). 2019 Storage	• Agencies will have until October 1, 2020 to comply with federal policies impacting data center, cloud, and IT infrastructure optimization by acquiring technologies, tools, and evidence-based best practices CIOs must submit the following to OMB annually: • A complete inventory of the data centers owned, operated, or maintained by (or on behalf of) the agency. • A multi-year strategy to consolidate and optimize these data centers. • Each agency (under its CIO's direction) must submit quarterly updates on their progress towards activity completion, consolidation, optimization metrics, and cost savings realized through the implementation of their strategy.
HR 150 Grant Reporting Efficiency and Agreements Transparency (GREAT) Act of 2019	• Requires OMB to designate a single data standard-setting agency (HHS, the government's largest grantmaking agency) to work with it, and in consultation with other stakeholders, to develop a set of data standards by December 30, 2021. • OMB mandates that a set of unique identifiers for federal awards and grant recipients be developed that are consistently applied across the government. • The data standards require that information collected by the federal government from grantees be fully searchable and machine-readable, be nonproprietary, and incorporate any standards already created under the DATA Act.

7 References

1. “Untangling Big Data.” *U.S. Department of Transportation: My Access: Sign In*, https://my.faa.gov/focus/articles/2018/06/Untangling_Big_Data.html.
2. Rouse, Margaret, et al. “What Is Data Life Cycle ? - Definition from WhatIs.com.” *WhatIs.com*, https://whatis.techtarget.com/definition/data-life-cycle.
3. “Semi-Structured Data.” *What Is*, https://www.datamation.com/big-data/semi-structured-data.html.
4. “26 Best Data Integration Tools, Platforms and Vendors in 2019.” *Software Testing Help*, 29 Nov. 2019, https://www.softwaretestinghelp.com/tools/26-best-data-integration-tools/.
5. “Top 18 Data Ingestion Tools - Compare Reviews, Features, Pricing in 2019.” *PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices*, 28 Oct. 2019, https://www.predictiveanalyticstoday.com/data-ingestion-tools/.
6. Kranc, Moshe. “Data Ingestion Best Practices.” *CMSWire.com*, CMSWire.com, 12 Jan. 2018, https://www.cmswire.com/information-management/data-ingestion-best-practices/.
7. Dutta, Partha Pratim. Data Chaos. 6 June 2018, medium.com/@parthapratimdutta/data-chaos-882d7b9c864c.
8. "White Paper Taming Data Drift The Silent Killer of Data Integrity." 2016/07/Taming-Data-Drift-White-Paper
9. Pancha, Girish. “Big Data's Hidden Scourge: Data Drift.” *CMSWire.com*, CMSWire.com, 8 Apr. 2016, https://www.cmswire.com/big-data/big-datas-hidden-scourge-data-drift/.
10. “Unriddling Big Data File Formats.” *ThoughtWorks*, 29 Aug. 2018, https://www.thoughtworks.com/insights/blog/small-big-decision-file-formats.
11. DeBois, Pierre. “Data Drift: What It Is and How to Avoid It.” *CMSWire.com*, CMSWire.com, 10 Apr. 2017, https://www.cmswire.com/digital-experience/data-drift-what-it-is-and-how-to-avoid-it/.
12. “Data Ingestion: the First Step to a Sound Data Strategy: Stitch Resource.” *Stitch*, https://www.stitchdata.com/resources/data-ingestion/.
13. “The Best Data Ingestion Tools for Migrating to a Hadoop Data Lake.” *RCG*, 18 Oct. 2019, https://rcgglobalservices.com/the-best-data-ingestion-tools-for-migrating-to-a-hadoop-data-lake/.
14. “The Challenges, Parameters & Best Practices Of Big Data Ingestion.” *Datapine*, 20 Aug. 2019, https://www.datapine.com/blog/big-data-ingestion-parameters-challenges-and-best-practices/.

15. Kidd, Chrissy. “4 Reasons to Automate the Ingestion of Data.” *BMC Blogs*, 23 July 2018, www.bmc.com/blogs/automate-data-ingestion/.
16. Kassner, Michael. “Is Metadata Collected by the Government a Threat to Your Privacy?” *TechRepublic*, TechRepublic, 5 Mar. 2015, https://www.techrepublic.com/blog/it-security/is-metadata-collected-by-the-government-a-threat-to-your-privacy/.
17. Rizvi, Sanam Shahla. “A Comparitive Study of ETL Tools.” *Academia.edu*, https://www.academia.edu/354387/A_Comparitive_Study_of_ETL_Tools.
18. Maayan, Gilad David. “What Is the Relevance of ETL in 2018?” *RTInsights*, 28 Mar. 2018, https://www.rtinsights.com/what-is-the-relevance-of-etl-in-2018/.
19. Stedman, Craig. “What Are Key Features for Choosing the Best ETL Tools for Your Needs?” S*earchDataManagement*, TechTarget, 21 Dec. 2016, https://searchdatamanagement.techtarget.com/answer/ETL-tools-What-you-do-and-dont-want.
20. “ETL Ecosystem & Tools: In-depth Guide [2021 Update].” 1 Jan 2021. https://blog.aimultiple.com/etl-tools/
21. “Move Away from Batch ETL with next-Gen Change Data Capture.” *StreamAnalytix*, https://www.streamanalytix.com/blogs/move-away-from-batch-etl-with-next-gen-change-data-capture/.
22. Stevens, John P. “Why You Need Metadata for Big Data Success.” *Data Science Central*, https://www.datasciencecentral.com/profiles/blogs/why-you-need-metadata-for-big-data-success.
23. Russom, Philip. “Data Lake Management Innovations.” *Transforming Data with Intelligence*, https://tdwi.org/articles/2017/01/23/data-lake-management-innovations.aspx.
24. “The OPEN Government Data Act: A Sweeping Open Data Mandate for All Federal Information.” *Data Coalition*, 3 Dec. 2019, www.datacoalition.org/the-open-government-data-act-a-sweeping-open-data-mandate-for-all-federal-information/.
25. Zolly, L., Henkel, H.S., Hutchison, V.B., Langseth, M.L., Thibodeaux, C.J., 2015, USGS Data management training modules—metadata for research data: U.S. Geological Survey, http://dx.doi.org/10.5066/F7RJ4GGJ.
26. “Common Mistakes Companies Make with Metadata Management.” *MerlinOne*, 5 Sept. 2019, https://merlinone.com/common-mistakes-organizations-make-with-metadata-management/.
27. “10 Top Metadata Management Tools.” *Datamation*, https://www.datamation.com/big-data/top-metadata-management-tools.html.

28. “Geospatial Metadata Fact sheet.” Geospatial Metadata - Federal Geographic Data Committee, https://www.fgdc.gov/metadata.
29. “Metadata for Interoperability.” *Home Page*, https://www.naa.gov.au/information-management/building-interoperability/interoperability-development-phases/data-governance-and-management/metadata-interoperability.
30. Smith, Anne Marie. “Metadata Strategy Overview.” *EWSolutions*, https://www.ewsolutions.com/metadata-strategy-overview/
31. “Developing & Implementing a Metadata Strategy.” Unissant Corporate Website. (n.d.). https://unissant.us/portfolio_page/developing-implementing-a-meta-data-strategy/
32. AArete. “A Look at AWS Glue - Simplify Your ETL & Data Transfers on the Cloud.” 20 June 2020. https://www.aarete.com/insights/a-look-at-aws-glue-simplify-your-etl-data-transfers-on-the-cloud/.
33. Choudinard, Don. “Metadata Storage Solutions.” 3 May 2016. https://www.caringo.com/blog/metadata-storage-solutions.
34. Subramanyam, Shanti, and Shanti Subramanyam. “Big Data in the Enterprise Data Warehouse.” *Orzota*, 23 Jan. 2019, orzota.com/2014/04/28/big-data-enterprise-data-warehouse/.
35. Simplilearn. (2017 Aug 23) *NoSQL Tutorial For Beginners | RDBMS Vs NoSQL | NoSQL Database Tutorial | Simplilearn* [Video]. YouTube, https://www.youtube.com/watch?v=XFE0EgT5oQE Accessed May 3 2021
36. Traversy Media. (2017 May 24) *An Introduction To NoSQL Databases* [Video]. YouTube. https://www.youtube.com/watch?v=uD3p_rZPBUQ Accessed May 3 2021
37. *Difference between SQL and NoSQL*. GeeksforGeeks 23 Dec. 2020, https://www.geeksforgeeks.org/difference-between-sql-and-nosql/
38. Oracle Developers. YouTube (2015 June 8) *NoSQL, WTF! Let's Talk NewSQL* [Video]. YouTube. https://www.youtube.com/watch?v=Blrd9BVNZbs Accessed May 3 2021
39. Nazrul, Sadat. “CAP Theorem and Distributed Database Management Systems.” *Medium*, Towards Data Science, 5 Nov. 2018, https://towardsdatascience.com/cap-theorem-and-distributed-database-management-systems-5c2be977950e.
40. Atinder. “Different Cloud Storage Types.” *Best Cloud Storage | Cloud Reviews | Online Backup*, www.cloudstoragebest.com/cloud-storage-types/.
41. Kovacs, Gali. “Block Storage vs. Object Storage in the Cloud.” *Block Storage vs. Object Storage in the Cloud*, Netapp, 22 Feb. 2017, cloud.netapp.com/blog/block-storage-vs-object-storage-cloud.

42. Webster, John, and Evaluator Group. “Comparing Scale-out and Object-Based Storage Systems.” *SearchStorage*, https://searchstorage.techtarget.com/answer/Ask-the-expert-Comparing-scale-out-and-object-based-storage-systems.
43. PSSC Labs. “Major Benefits of Using Object Storage.” 8 Mar 2016 https://pssclabs.com/article/4-major-benefits-of-using-object-storage/
44. Ohlhorst, John. “Best Practices for Selecting Storage Services for Big Data.” CIO. 24 Apr. 2012 https://www.cio.com/article/2396760/best-practices-for-selecting-storage-services-for-big-data.html.
45. Delgado, Rick. “5 Things to Consider When Choosing the Right Cloud Storage.” *SmartData Collective*, 20 Oct. 2017, www.smartdatacollective.com/5-things-consider-when-choosing-right-cloud-storage/.
46. Schiff, Jennifer Lonoff. “14 Things You Need to Know About Data Storage Management.” *CIO*, CIO, 11 Sept. 2013, www.cio.com/article/2382585/14-things-you-need-to-know-about-data-storage-management.html.
47. Chai, Wesley. “Amazon Simple Service (Amazon S3).” https://searchaws.techtarget.com/definition/Amazon-Simple-Storage-Service-Amazon-S3?_ga=2.170225572.548285474.1606438043-1245633525.1606438043
48. HITInfrastructure. “Benefits of Object Storage to Health IT Infrastructure.” *HITInfrastructure*, 27 Apr. 2017, https://hitinfrastructure.com/news/benefits-of-object-storage-to-health-it-infrastructure.
49. Lowe, Scott. "The Leading Object Storage Vendors Offer Broad Range of Options." 28 Aug. 2017, https://searchstorage.techtarget.com/feature/The-leading-object-storage-vendors-offer-broad-range-of-options.
50. “Big Data Security - Issues, Challenges, Tech & Concerns.” *RDA*, 4 Sept. 2019, https://www.rd-alliance.org/group/big-data-ig-data-security-and-trust-wg/wiki/big-data-security-issues-challenges-tech-concerns.
51. Bhandari, Renu, et al. “Big Data Security – Challenges and Recommendations.” *ResearchGate*, March 2016, https://www.researchgate.net/publication/295907307_Big_Data_Security_-_Challenges_and_Recommendations.
52. Industry Perspectives “Nine Main Challenges in Big Data Security.” *Data Center Knowledge*, 19 Jan. 2016. https://www.datacenterknowledge.com/archives/2016/01/19/nine-main-challenges-big-data-security.

53. Rohloff, Kurt. “Why Encryption Holds the Secret to Data Security.” *Transforming Data with Intelligence*, 29 Mar. 2019. https://tdwi.org/articles/2019/03/29/dwt-all-why-encryption-holds-the-secret-to-data-security.aspx.

54. Ponemon, Larry, et al. “What's New in the 2019 Cost of a Data Breach Report.” *Security Intelligence*, 23 July 2019, https://securityintelligence.com/posts/whats-new-in-the-2019-cost-of-a-data-breach-report/.

55. “Big Data Security.” *Top Big Data Security Concerns*, https://www.datamation.com/big-data/big-data-security.html.

56. Y Z An, et el. “Reviews on Security Issues and Challenges in Cloud Computing.” 2016 *IOP Conf. Ser.: Mater. Sci. Eng.* 160 012106

57. “An Expert Guide to Securing Sensitive Data: 34 Experts Reveal the Biggest Mistakes Companies Make with Data Security.” *Digital Guardian*, 13 Nov. 2018, https://digitalguardian.com/blog/expert-guide-securing-sensitive-data-34-experts-reveal-biggest-mistakes-companies-make-data.

58. Sayegh, Emil. “Protecting Sensitive Data: When ‘Putting It in the Cloud’ Doesn't Cut It - Ntirety: Managed Cloud Solutions: Guaranteed.” *Ntirety*, 8 Oct. 2019, https://www.ntirety.com/protecting-sensitive-data-when-putting-it-in-the-cloud-doesn-t-cut-it/.

59. Suryateja, P.S. “Threats and Vulnerabilities of Cloud Computing: A Review.” *ResearchGate*, March 2018, https://www.researchgate.net/publication/324562008_Threats_and_Vulnerabilities_of_Cloud_Computing_A_Review.

60. Zeidler, Reto, et al. “When It Comes to Incident Response, Failing to Plan Means Planning to Fail.” *Security Intelligence*, 26 Mar. 2019, https://securityintelligence.com/when-it-comes-to-incident-response-failing-to-plan-means-planning-to-fail/.

61. *Usa.kaspersky.com*, https://usa.kaspersky.com/resource-center/definitions/data-breach.

62. “5 Best Practices for Big Data Security.” *KDnuggets*, https://www.kdnuggets.com/2016/06/5-best-practices-big-data-security.html.

63. Inukollu, Venkata, et al. “Security Issues Associated with Big Data in Cloud Computing.” *International Journal of Network Security & Its Applications (IJNSA)*, Vol.6, No.3, May 2014, https://www.researchgate.net/publication/276199542_Security_Issues_Associated_with_Big_Data_in_Cloud_Computing.

64. “Business Home.” *McAfee*, https://www.mcafee.com/enterprise/en-us/security-awareness/cloud/security-issues-in-cloud-computing.html.

65. (2012, Apr. 10) How To Get Cloud Architecture and Design Right the First Time 2012 [Video]. YouTube. www.youtube.com/watch?v=S-xbHC3xxKM. Accessed May 3 2021
66. “Exploring the Big Data Stack.” *Datamation*, https://www.datamation.com/data-center/exploring-the-big-data-stack.html.
67. Alley, Garrett. “What Is Big Data Architecture? - DZone Big Data.” *Dzone.com*, 17 June 2019, https://dzone.com/articles/what-is-big-data-architecture.
68. “Understanding the Architectural Layers of a Big Data Solution.” *IBM Developer*, https://developer.ibm.com/articles/bd-archpatterns3/.
69. Forrest, Conner. “5 Architectural Principles for Building Big Data Systems on AWS.” *TechRepublic*, TechRepublic, 7 Dec. 2016, https://www.techrepublic.com/article/5-architectural-principles-for-building-big-data-systems-on-aws/.
70. “Cloud Computing Architecture.” *Wikipedia*, Wikimedia Foundation, 10 Dec. 2019, en.wikipedia.org/wiki/Cloud_computing_architecture.
71. “7 Best Practices to Follow While Designing a Cloud Architecture.” *World-Class Cloud from India | High Performance Cloud Infrastructure | E2E Cloud*, 20 May 2019, https://www.e2enetworks.com/best-practices-designing-cloud-architecture.
72. *Cloud Computing System Architecture Diagrams*, https://docs.rightscale.com/cm/designers_guide/cm-cloud-computing-system-architecture-diagrams.html.
73. “Top Five Big Data Architectures.” *Worldwide IT Training*, https://www.globalknowledge.com/us-en/resources/resource-library/articles/top-five-big-data-architectures/.
74. *What Is Big Data Platform?*, https://www.roseindia.net/bigdata/what-is-big-data-platform.shtml.
75. Reddy, Chandan K. “A Survey on Platforms for Big Data Analytics.” Journal of Big Data, Springer Open, 9 Oct. 2014, https://journalofbigdata.springeropen.com/articles/10.1186/s40537-014-0008-6.
76. Team, Editorial. “[VIDEO] 6 Ways to Optimize the Cost of Your Big Data Platform.” *Western Digital Corporate Blog*, 3 Sept. 2019, https://blog.westerndigital.com/6-ways-optimize-cost-big-data-platform/.
77. “The Future of Open Source Big Data Platforms.” *InsideBIGDATA*, 15 June 2019, https://insidebigdata.com/2019/06/14/the-future-of-open-source-big-data-platforms/.
78. What is SDC? Secure Data Commons- About SDC (n.d.) https://its.dot.gov/data/secure/about.html
79. “Sign In.” U.S. Department of Transportation: My Access: Sign In, https://employees.faa.gov/tv/?mediaId=1850

80. O'Reilly Media (2017, Oct 25) *It's Like Amazon. But for Data - Dan Sholler (Collibra)* [Video]. YouTube. https://www.youtube.com/watch?v=ncLqaBYa0NE. Accessed May 3 2021
81. Ladley, John, and First San Francisco Partners. "How to Narrow down Your Choices for Buying a Data Governance Tool." *SearchDataManagement*, https://searchdatamanagement.techtarget.com/feature/How-to-narrow-down-your-choices-for-buying-a-data-governance-tool.
82. "10 Best Data Governance Tools To Fulfill Your Data Needs In 2019." *Software Testing Help*, 10 Nov. 2019, https://www.softwaretestinghelp.com/data-governance-tools/.
83. "Sign In." *U.S. Department of Transportation: My Access: Sign In*, https://my.faa.gov/focus/articles/2015/08/FAA_Awards_Cloud_Com.html.
84. "Sign In." *U.S. Department of Transportation: My Access: Sign In*, https://my.faa.gov/focus/articles/2016/07/AIT_Download_6__Prog.html.
85. VanRoekel, Steven (2011). Security Authorization of Information Systems in Cloud Computing Environments. *FedRAMP Policy Memorandum for Chief Information Officers*, https://media.bizj.us/view/archive/washington/fedbiz_daily/FedRAMPMemo.pdf
86. "Make the Most of the FedRAMP Marketplace." *Make the Most of the FedRAMP Marketplace | FedRAMP.gov*, https://www.fedramp.gov/make-the-most-of-the-fedramp-marketplace/.
87. "Federal Agencies." *FedRAMP.gov*, https://www.fedramp.gov/federal-agencies/.
88. FY20 FEDRAMP GUIDANCE, Section 5: CLOUD SECURITY AND FEDRAMP GUIDANCE issued by AIS-230
89. FedRAMP Concept of Operations (CONOPS) Version 1.2 Section 10 July 27, 2012
90. Geiger, Gene. "FedRAMP vs. FISMA: Choosing the Right Standard for Your Federal Clients." 19 Dec. 2016, https://a-lign.com/fedramp-vs-fisma
91. "Data.gov." *Wikipedia*, Wikimedia Foundation, 3 Feb. 2020, en.wikipedia.org/wiki/Data.gov
92. Data.gov. "Impact". https://www.data.gov/impact/
93. Data.gov has 36,529 (or 20%) of the total of 186,467 datasets are in PDF format. Source: Data.gov. "Data Catalog". http://catalog.data.gov/dataset#sec-res_format
94. M-10-06. Open Government Directive. 12/8/2009. https://www.whitehouse.gov/sites/whitehouse.gov/files/omb/memoranda/2010/m10-06.pdf
95. "The Federal Big Data Research And Development Strategic Plan." May 2016. https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf

96. "M-16-19 - Data Center Optimization Initiative." *M-16-19 - Data Center Optimization Initiative*, https://policy.cio.gov/dcoi/
97. "Federal Data Center Consolidation Initiative." *The IT Law Wiki*, https://itlaw.wikia.org/wiki/Federal_Data_Center_Consolidation_Initiative
98. Kash, Wyatt. "25 Point Implementation Plan to Reform Federal Information Technology Management." *Breaking Government*, https://breakinggov.com/documents/25-point-implementation-plan-to-reform-federal-information-techn/
99. Marzullo, Keith. "Administration Issues Strategic Plan for Big Data Research and Development." 23 May 2016, https://obamawhitehouse.archives.gov/blog/2016/05/23/administration-issues-strategic-plan-big-data-research-and-development
100. Executive Order 13642: Making Open and Machine Readable the New Default for Government Information. (2013, May 9), https://www.presidency.ucsb.edu/documents/executive-order-13642-making-open-and-machine-readable-the-new-default-for-government
101. M-13-13 Memorandum for the Heads of Executive Departments and Agencies. Project-Open-Data. (2016, January 20). *project-open-data/project-open-data.github.io*. GitHub. https://github.com/project-open-data/project-open-data.github.io/blob/master/policy-memo.md
102. *Official DOT Public Access Plan v1.1*. U.S. Department of Transportation. (n.d.). https://www.transportation.gov/mission/open/official-dot-public-access-plan-v11
103. "About." *Welcome to Research Hub*, https://researchhub.bts.gov/about
104. "FISMA Updated and Modernized." *Inside Government Contracts*, 2 Mar. 2016, www.insidegovernmentcontracts.com/2014/12/fisma-updated-and-modernized/
105. National Archives and Records Administration (n.d.) *Big Announcements in Big Data*. National Archives and Records Administration https://obamawhitehouse.archives.gov/blog/2015/11/04/big-announcements-big-data
106. "Federal Privacy Council." *Federal Privacy Council*, https://www.fpc.gov/
107. "Circular A-130 Managing Information as a Strategic Resource." *Managing Information as a Strategic Resource*, a130.cio.gov/
108. "Federal Information Technology Acquisition Reform Act." *Wikipedia*, Wikimedia Foundation, 15 Dec. 2019, en.wikipedia.org/wiki/Federal_Information_Technology_Acquisition_Reform_Act
109. "Federal Cybersecurity Risk Determination Report and Action Plan," May 2018. https://www.hsdl.org/?view&did=811093

110. Bur, Jessie. “Trump Management Agenda to Focus on Multiagency Goals.” *Federal Times*, 20 Mar 2018, https://www.federaltimes.com/management/2018/03/20/trump-management-agenda-to-focus-on-multi-agency-goals/

111. “The Federal Information Technology Acquisition Reform Act: Office of the Chief Information Officer.” *The Federal Information Technology Acquisition Reform Act | Office of the Chief Information Officer*, https://www.ocio.usda.gov/federal-information-technology-acquisition-reform-act

112. Chappellet-Lanier, Tajha. “OMB Releases Initial Agency Guidance for MGT Act Implementation.” *FedScoop*, 17 July 2018, https://www.fedscoop.com/mgt-act-implementation-guidance-omb/

113. Chappellet-Lanier, Tajha. “OMB Releases Finalized Cloud Smart Policy.” *FedScoop*, 24 June 2019, www.fedscoop.com/final-cloud-smart-policy/

114. “Future of Open Data: Maximizing the Impact of the OPEN Government Data Act.” *Data Foundation*, www.datafoundation.org/future-of-open-data-maximizing-the-impact-of-the-open-government-data-act

115. Chappellet-Lanier, Tajha. “The OPEN Government Data Act is Now Law.” *FedScoop*, 15 Jan 2019, https://www.fedscoop.com/open-government-data-act-law/

116. “ITS DataHub: Home.” *ITS DataHub | Home*, its.dot.gov/data/

117. “Analysis: Federal Data Strategy Action Plan.” *DLT Blog*, 6 June 2019, https://www.dlt.com/blog/2019/06/06/analysis-federal-data-strategy-action-plan/

118. Chen, April. “How the Federal Data Strategy is Transforming Data into a Strategic Asset.” *Federal Times*, 23 July 2019, https://www.federaltimes.com/opinions/2019/07/23/how-the-federal-data-strategy-is-transforming-data-into-a-strategic-asset/

119. “JavaScript Object Notation File.” *JSON File Extension - What Is a .Json File and How Do I Open It?*, fileinfo.com/extension/json

120. “How to Get Your Open Data on Data.gov.” *How to Get Your Open Data on Data.gov | Federal Enterprise Data Resources*, resources.data.gov/tools/how-to-get-your-open-data-on-datagov/

121. “What Is Trump's American AI Initiative?” Tech.co 2019. *Tech.co*, 14 Feb. 2019, https://tech.co/news/trump-american-ai-initiative-2019-02

122. “Executive Order 13859.” Executive Order 13859 - Wikisource, the Free Online Library, en.wikisource.org/wiki/Executive_Order_13859

123. Baker, Jim. “President Trump’s Executive Order on Artificial Intelligence.” *Lawfare*, 28 Feb. 2019, https://www.lawfareblog.com/president-trumps-executive-order-artificial-intelligence

124. Johnson, Derek B. "OMB Finalizes." *FCW*, https://fcw.com/articles/2019/06/25/cloud-smart-johnson.aspx

125. "Data Center Optimization Initiative (DCOI)." *GSA*, 30 Apr. 2018, https://www.gsa.gov/technology/government-it-initiatives/data-center-optimization-initiative-dcoi

126. "Reporting Requirements." GSA, 21 Aug. 2019, www.gsa.gov/technology/government-it-initiatives/dcoi/dcoi-data-center-resources/reporting-requirements

127. Kamensky, John. "The GREAT Act: Scaling the Tower of Babel." *Government Executive*, Government Executive, 23 Jan. 2020, www.govexec.com/management/2020/01/great-act-scaling-tower-babel/162596/

128. "Big Data Analytics - What It Is and Why It Matters." *SAS*, www.sas.com/en_us/insights/analytics/big-data-analytics.html

129. Bekker, Alex. "4 Types of Data Analytics to Improve Decision-Making." *Software Development Company - ScienceSoft*, ScienceSoft, 23 Jan. 2020, www.scnsoft.com/blog/4-types-of-data-analytics

130. "Types of Analytics: Descriptive, Predictive, Prescriptive Analytics." *DeZyre*, www.dezyre.com/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209.

131. "Pros and Cons of Predictive Analysis." *Georgetown University Online*, 28 Sept. 2018, scsonline.georgetown.edu/programs/masters-technology-management/resources/pros-and-cons-predictive-analysis

132. Pushpalatha, M., and S. Poornima. "A Survey of Predictive Analytics Using Big Data with Data Mining." *International Journal of Bioinformatics Research and Applications*, vol. 14, no. 3, 2018, p. 269., doi:10.1504/ijbra.2018.10009573

133. Finlay, Steven. *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*, Palgrave Macmillan, 2014.

134. Alton, Larry. "The 7 Most Important Data Mining Techniques." Data Science Central, 22 Dec. 2017, https://www.datasciencecentral.com/profiles/blogs/the-7-most-important-data-mining-techniques

135. Wheaton, Jim. "Best Practices in Data Mining: The Second Five Commandments." *Multichannel Merchant*, Multichannel Merchant, 29 Mar. 2013, multichannelmerchant.com/marketing/best-practices-in-data-mining-the-second-five-commandments/

136. Ramageri, Bharati M. "Data Mining Techniques and Applications." *Indian Journal of Computer Science and Engineering.* vol. 1, no. 4, pp. 301-305

137. Auza, J. 2010. 5 of the Best Free and Open Source Data Mining Software. [Accessed Online March 2013] http://www.junauza.com/2010/11/free-data-mining-software.html

138. BajajCheck, Prateek, and Prateek Bajaj. "Reinforcement Learning." *GeeksforGeeks*, 11 Dec. 2019, www.geeksforgeeks.org/what-is-reinforcement-learning/

139. Dhandre, Pravin, et al. "Top 5 Tools for Reinforcement Learning." Packt Hub, 27 June 2018, hub.packtpub.com/tools-for-reinforcement-learning/

140. "Frameworks for Approaching the Machine Learning Process." *KDnuggets*, www.kdnuggets.com/2018/05/general-approaches-machine-learning-process.html

141. Sreenivasan, Sreeram. "4 Business Benefits of Data Visualization." *4 Business Benefits of Data Visualization*, www.analyticbridge.datasciencecentral.com/profiles/blogs/4-business-benefits-of-data-visualization

142. "Data Visualization Techniques." *Analytics, Business Intelligence and Data Management*, www.sas.com/en_us/whitepapers/data-visualization-techniques-106006.html

143. 3 Sosulski. *Kristen Sosulski*, 24 Nov. 2018, www.kristensosulski.com/2018/11/criteria-for-evaluating-data-visualization-tools/

144. Shah, Shachi. "Do We Really Need GPU for Deep Learning? - CPU vs GPU." *Medium*, Medium, 7 Dec. 2018, medium.com/@shachishah.ce/do-we-really-need-gpu-for-deep-learning-47042c02efe2

145. IBM Cloud (2019 March 20) *GPUs: Explained* [Video]. YouTube. www.youtube.com/watch?v=LfdK-v0SbGI&t=319s Accessed May 3 2021

146. BySakshi, Posted. "What Is the Difference between Big Data (Hadoop) & Data Analytics." *Talentedge*, talentedge.com/blog/difference-between-big-data-and-data-analytics/

Appendix A - Acronyms

Acronym	**Definition**
AI	Artificial Intelligence
API	Application programming interface
APT	Advanced Persistent Threat
ATO	Authority to Operate
AWS	Amazon Web Services
BD	Big Data
BD IWG	Big Data Interagency Working Group
BD SSG	Big Data Senior Steering Group
BYOD	Bring Your Own Device
CAP	Consistency, Availability, and Partition Tolerance
CBI	Confidential Business Information
CCPA	California Consumer Privacy Act
CDO	Chief data officer
CIO	Chief Information Officer
CPU	Central processing unit
CSDGM	Content Standard for Digital Geospatial Metadata
CSO	Cloud Service Offering
CSP	Cloud service provider
DCOI	Data Center Optimization Initiative
DFS	Distributed file storage
DOI	Digital Object Identifier
DoS	Denial of Service
DOT	Department of Transportation
DP	Data Platform
DGC	Data Governance Center
EC2	Amazon Elastic Compute Cloud
EDW	Enterprise Data Warehouse
EIM	Enterprise Information Management
EO	Executive Order
ETL	Extract, Transform, and Load
FAA	Federal Aviation Administration

Acronym	**Definition**
FAST	Fixing America's Surface Transportation
FCS	FAA Cloud Services
FDCCI	Federal Data Center Consolidation Initiative
FedRAMP	Federal Risk and Authorization Management Program
FGDC	Federal Geographic Data Committee
FISMA	Federal Information Security Management Act
FITARA	Federal Information Technology Acquisition Reform Act
FLIR	Forward-Looking Infrared
GAO	Government Accountability Office
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
GUI	Graphical user interface
HDFS	Hadoop Distributed File System
HIDS	Host-based intrusion detection system
IaaS	Infrastructure as a Service
IDS	Intrusion detection systems
IoT	Internet of Things
IPS	Intrusion prevention systems
ISO	International Organization for Standardization
IT	Information technology
ITS	Intelligent Transportation Systems
JAB	Joint Authorization Board
JPO	Joint Program Office
KPI	Key Performance Indicator
KSN	Knowledge sharing network
MGT	Modernizing Government Technology
ML	Machine Learning
MTTC	Mean time to contain
NAS	National Airspace System
NIST	National Institute of Standards and Technology
NITRD	Networking and Information Technology Research and Development
NoSQL	Not only SQL
NSF	National Science Foundation
NTL	National Transportation Library

Acronym	**Definition**
OMB	Office of Management and Budget
OPEN	Open, Public, Electronic, and Necessary
OSTP	Office of Science and Technology Policy
PaaS	Platform as a Service
PII	Personal identifiable information
PIV	Particle Image Velocimetry
PMA	President’s Management Agenda
PMO	Program Management Office
POA&M	Plan of Action and Milestones
RAID	Redundant Array of Inexpensive Disks
R&D	Research and Development
RDBMS	Relational database management systems
RE&D	Research, engineering, and development
RL	Reinforcement learning
S3	Simple Storage Service
SaaS	Software as a Service
SAF	Security Assessment Framework
SAP	Security Assessment Plan
SAOP	Senior Agency Officials for Privacy
SAR	Security Assessment Report
SDC	Secure Data Commons
SSP	System Security Plan
STEM	Science, Technology, Engineering, and Mathematics
SWIM	System Wide Information Management
TMF	Technology Modernization Fund
3PAO	Third Party Assessment Organization
UML	Unified Modeling Language
USB	Universal serial bus
USDOT	U.S. Department of Transportation
YARN	Yet Another Resource Negotiator