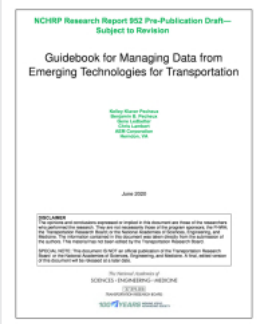


Guidebook for Managing Data from Emerging Technologies for Transportation (2020)

DETAILS

0 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-67939-8 | DOI 10.17226/25844

CONTRIBUTORS

Kelley Klaver Pecheux, Benjamin B. Pecheux, Gene Ledbetter, Chris Lambert, AEM Corporation; National Cooperative Highway Research Program; Transportation Research Board; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine 2020. *Guidebook for Managing Data from Emerging Technologies for Transportation*. Washington, DC: The National Academies Press. https://doi.org/10.17226/25844.

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Systematic, well-designed, and implementable research is the most effective way to solve many problems facing state departments of transportation (DOTs) administrators and engineers. Often, highway problems are of local or regional interest and can best be studied by state DOTs individually or in cooperation with their state universities and others. However, the accelerating growth of highway transportation results in increasingly complex problems of wide interest to highway authorities. These problems are best studied through a coordinated program of cooperative research.

Recognizing this need, the leadership of the American Association of State Highway and Transportation Officials (AASHTO) in 1962 initiated an objective national highway research program using modern scientific techniques—the National Cooperative Highway Research Program (NCHRP). NCHRP is supported on a continuing basis by funds from participating member states of AASHTO and receives the full cooperation and support of the Federal Highway Administration (FHWA), United States Department of Transportation, under Agreement No. 693JJ31950003.

COPYRIGHT INFORMATION

Authors herein are responsible for the authenticity of their materials and for obtaining written permissions from publishers or persons who own the copyright to any previously published or copyrighted material used herein.

Cooperative Research Programs (CRP) grants permission to reproduce material in this publication for classroom and not-for-profit purposes. Permission is given with the understanding that none of the material will be used to imply endorsement by TRB and any of its program sponsors of a particular product, method, or practice. It is expected that those reproducing the material in this document for educational and not-for-profit uses will give appropriate acknowledgment of the source of any reprinted or reproduced material. For other uses of the material, request permission from CRP.

DISCLAIMER

To facilitate more timely dissemination of research findings, this pre-publication document is taken directly from the submission of the research agency. The material has not been edited by TRB. The opinions and conclusions expressed or implied in this document are those of the researchers who performed the research. They are not necessarily those of the Transportation Research Board; the National Academies of Sciences, Engineering, and Medicine; the FHWA; or the program sponsors.

The Transportation Research Board, the National Academies, and the sponsors of the National Cooperative Highway Research Program do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of the report.

This pre-publication document IS NOT an official publication of the Cooperative Research Programs; the Transportation Research Board; or the National Academies of Sciences, Engineering, and Medicine.

Recommended citation: Pecheux, K. K., B. B. Pecheux, G. Ledbetter, and C. Lambert. 2020. *Guidebook for Managing Data from Emerging Technologies for Transportation.* Pre-publication draft of NCHRP Research Report 952. Transportation Research Board, Washington, D.C.

CONTENTS

Note: Numerous technical terms are used within this guidebook. While definitions are provided for some, others are used in a context without specifically defining them. As there are multiple sources that provide standard definitions to these and other terms commonly used in the field of information technology, rather than repeat these definitions herein, readers are encouraged to visit the following sites:

- Techopedia – https://www.techopedia.com/dictionary
- TechTerms – https://techterms.com/

List of Figures

List of Tables

INTRODUCTION

Transportation safety and mobility, which enhance American productivity, have advanced over the past three decades. This advancement is due in large part to various transformational intelligent transportation technologies including advanced traffic management systems, electronic toll collection, traffic signal coordination, transit signal priority, and traveler information systems to name a few. More recently, further developments in communications and technology have led to more advanced infrastructure and vehicle capabilities, mobile applications, and a host of mobility service offerings including connected vehicles, automated vehicles, on demand and shared mobility services, crowdsourcing, the Internet of things (IoT), and new mobility initiatives such as smart cities and communities, all of which are producing data at extraordinary volumes and speeds.

A wide range of institutions, both public and private, have initiated demonstration and pilot projects of these technologies, and many have invested in associated data sets. As these activities continue to expand, the amount of data is also expanding. Data from emerging technologies have tremendous potential to offer new insights and to identify unique solutions for delivering services, thereby improving outcomes. However, the volume and speed at which these data are generated, processed, stored, and sought for analysis are unprecedented and will fundamentally alter the transportation sector. With increased connectivity between vehicles, sensors, systems, shared-use transportation, and mobile devices, unexpected and unparalleled amounts of data are being added to the transportation domain at a rapid rate, and ***these data are too large, too varied in nature, and will change too quickly to be handled by the traditional database management systems of most transportation agencies***. Instead, ***modern, flexible, and scalable "big data" methods to manage these data need to be adopted by transportation agencies*** if the data are to be used to facilitate better decision-making. As many agencies are already forced to do more with less while meeting higher public expectations, continuing with traditional data management systems and practices will prove costly for agencies unable to shift.

This *Guidebook for Managing Data from Emerging Technologies for Transportation* provides guidance, tools, and a big data management framework, and it lays out a roadmap for transportation agencies on how they can begin to shift – technically, institutionally, and culturally – toward effectively managing data from emerging technologies. New concepts and methodologies concerning data management and use are introduced along with industry best practices for big data. Examples, references, and quotes are provided from transportation agencies that are currently navigating the implementation of big data to extend beyond traditional siloed use cases, including their challenges and successes. Common misconceptions within the transportation industry are discussed. Whether an agency is starting from scratch with a new technology data set, is trying to make the business case for emerging technology data, already working on a big data project, has an issue or problem that might be solved with emerging technology data, or is looking for a new enterprise data management solution, the steps and guidance outlined in this document are designed to walk them through the necessary data management policies, procedures, and practices to fully meet the needs of data from emerging technologies.

This guidebook provides guidance, tools, and a big data management framework, and it lays out a roadmap for transportation agencies on how they can begin to shift – technically, institutionally, and culturally – toward effectively managing data from emerging technologies.

This guidebook is intended for engineering and information technology (IT) analysts and professionals working for state, regional, and local transportation agencies. The target audience is primarily those at the mid-level manager level, including positions such as operations manager, ITS program manager, traffic operations/management center (TOC)/(TMC) manager, IT manager, and database administrator, and it is the intent that these staff, in their positions, would play a primary role as champion in leading data initiatives for their agencies that support positive organizational change. Successful implementation requires close cooperation between engineering and IT departments, which will provide mutual benefits and create a lasting and positive impact on the organization and its personnel. As such, as agencies progress through this guidebook, they are encouraged to revisit/assess existing processes and workflows that will need to be transitioned or replaced as a result of applying emerging technology data and modern data management processes within the organization.

Figure 1 shows the various components of this guidebook. The guidebook is comprised of three primary sections, which are described following the figure.

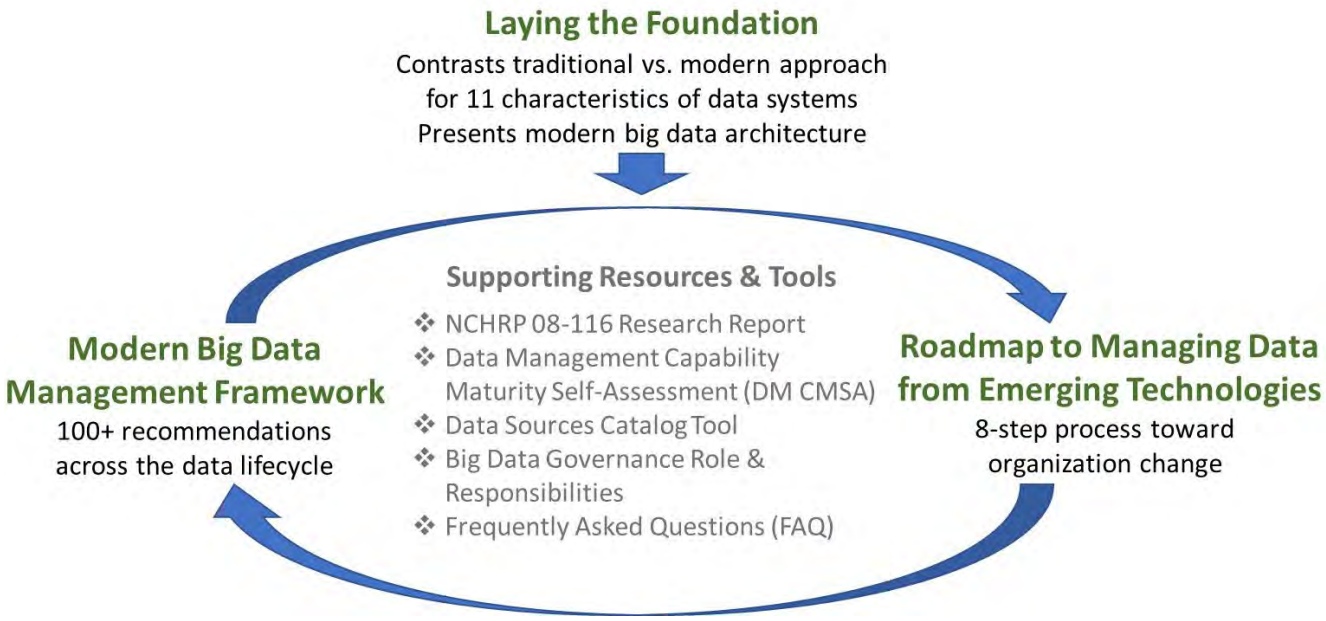


Figure 1. Application of Guidebook

- **Laying the Foundation** – for 11 characteristics of data systems and management, contrasts the traditional approach with the modern big data approach and provides an overview of a recommended big data architecture.
- **Modern Big Data Management Framework** – provides data industry best practices and more than 100 recommendations for modern data management across the full big data lifecycle, including the creation of data, storage of data, use of data, and sharing of data (Figure 2). This framework provides the details of the "how to" manage data from emerging technologies.

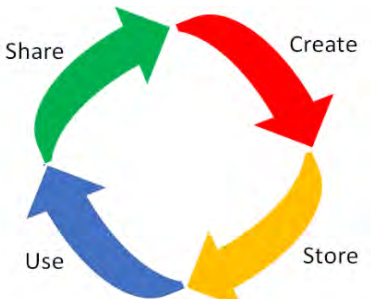


Figure 2. Big Data Lifecyle

- **Roadmap to Managing Data from Emerging Technologies** – presents an 8-step process and associated guidance for transportation agencies looking to begin or further efforts toward more effectively managing data from emerging technologies, with the goal of organization-wide change. For agencies just beginning, the roadmap provides a starting point. For agencies already on their way, the roadmap provides details on how to further those efforts and gain cross-organizational support.

These three primary sections will provide agencies with the foundation, recommendations, and steps to effectively implement a modern, flexible, scalable, and sustainable approach to managing data from emerging technologies.

In addition to these major sections of the guidebook, there are a number of supporting resources and tools for agencies implementing the guidebook. These resources/tools include:

- **NCHRP 08-116 Final Research Report, *Framework for Managing Data from Emerging Transportation Technologies to Support Decision-Making*** – details the research activities including results from a comprehensive state of practice review (49 sources cited) both external and internal to the transportation industry; an online survey of 25 organizations deploying emerging transportation technology projects; interviews with 11 city and state transportation agencies involved in managing data from emerging technologies; and a stakeholder workshop involving 17 representatives from 15 local, regional, and state agencies. The report defines data management and provides 65 modern big data management foundational principles organized by 15 data management "focus areas" to cover the full lifecycle of big data. The report presents a modern big data benchmark and assessment methodology, built from the foundational principles of big data management, which was applied to the information gathered from agencies participating in the research to further assess the state of practice in data management within the transportation industry. The report ends with a list of common challenges reported by agencies during the research, as well as a list of associated needs. This guidebook was developed specifically to address as many of these challenges and needs as possible. While there is some natural overlap in content between the final research report and this guidebook, an effort was made to preserve the brevity and applicability of the guidebook. It is recommended that agencies implementing the guidebook refer to this report for further background and details regarding the management of big data.
- **Data Management Capability Maturity Self-Assessment (DM CMSA)** – The DM CMSA is built from the modern big data benchmark and assessment methodology presented in the final research report. Questions within each of 15 data management focus areas will guide transportation agencies through a self-assessment to gauge their current data management practices, as well as to identify specific areas for improvement along their path toward shifting from traditional data management practices to more modern data management practices in order to handle data from emerging technologies.
- **Big Data Governance Roles and Responsibilities** – provides a list of recommendations to consider when developing a modern data governance approach, a description and frameworks for big data governance, and a tool for tracking the big data governance roles and responsibilities within an agency.

- **Data Sources Catalog Tool** – provided to assist transportation agencies in cataloging existing and potential data sources. This tool is useful for better understanding the data assets of an agency, prioritizing data sources, and informing the selection of those sources that might offer the most value before pursuing further development.
- **Frequently Asked Questions (FAQ)** – responses to frequently asked questions regarding big data implementation, management, governance, use, and security.

As indicated in Figure 1, it is recommended that readers start by reviewing the Laying the Foundation section. These seven pages concisely contrast the traditional data systems and management approach/practices of most transportation agencies with their modern data systems and management counterparts and presents a graphical representation and associated discussion of a recommended modern big data architecture. The concepts presented in this section should be understood by agency staff at all levels, including executive leadership.

From here, Figure 1 illustrates the cyclical and incremental nature in which this guidebook can be applied by agencies. The Framework and Roadmap are interrelated tools for agencies to achieve iterative and progressive transformation/change over time. How an agency proceeds will depend largely on the agency's needs, organization, level of data maturity, drivers for change, who's leading the charge, and the goals and anticipated outcomes of the effort. One agency may start by diving into the technical details and best practices presented in the Framework to further its understanding of modern data management. Another may start by reviewing the steps, requirements, and resource needs associated with implementing the Roadmap. Yet another may begin by convening an internal group of stakeholders to conduct the DM CMSA in order to baseline current practices and identify and prioritize areas for improvement. Once an agency begins its journey along the Roadmap, roadblocks faced can be overcome using recommendations in the Framework or by reviewing/applying the resources and tools provided (e.g., DM CMSA). For example, an agency might find that procurement does not support use of the cloud (a foundational recommendation and best practice of modern data management). To support agency champions in getting beyond this barrier, the Framework and Roadmap provide arguments they can use for why the cloud is needed.

There is natural overlap between the various components of the guidebook. For example, agencies will find discussions in Steps 4 and 5 of the Roadmap that expand on recommendations found in the Framework, as well as associated questions within the DM CMSA. The guidebook was not designed necessarily to be read from front to back, and agencies should refer to the various sections and supporting resources/tools as needed to support short-term incremental progress and long-term organizational change.

LAYING THE FOUNDATION

Data management is the practice of organizing and maintaining data and data processes to meet ongoing information lifecycle needs. It describes the processes used to plan, specify, enable, create, acquire, maintain, use, archive, secure, retrieve, control, share, and purge data. Data management is vital to every organization to ensure that data generated are properly managed, stored, and protected. Organizations are increasingly recognizing that the data they possess are an asset that must be managed properly to ensure success.

> *Data governance deals with quality, security and privacy, integrity, usability, integration, compliance, availability, roles and responsibilities, and overall management of the internal and external data flows within an organization (Roe, 2017).*

Central to data management is data governance. Data governance is a collection of practices and processes that help to ensure the formal management of data assets within an organization, including the planning, oversight, and control over management of data and the use of data and data-related resources. Data governance puts in place a framework to ensure that data are used consistently and consciously within the organization. It also deals with quality, security and privacy, integrity, usability, integration, compliance, availability, roles and responsibilities, and overall management of the internal and external data flows within an organization (Roe, 2017).

As data technologies have expanded, the purview of data management has also expanded. Increasing volumes of data and real-time processing of data have ushered in new big data frameworks. The variety of data has grown as well. Traditionally, relational database management systems have processed structured data. Data models described a set of relationships between different data elements formatted according to predefined data types in the database. Today, unstructured data such as emails, videos, audio files, web pages, and social media messages do not fit neatly into the traditional row and column structure of relational databases. As a result, enterprises are looking to a new generation of databases and analytical tools to address unstructured data. Collectively, these and other emerging data management technologies have come under the banner of “big data” (Rouse, 2013).

Traditional Data System and Management Approach vs. the Modern Big Data System and Management Approach

The fundamental purpose of this document is to help agencies shift from their traditional data systems and management practices to more modern big data systems and management practices to make more effective use of data from emerging technologies. It can be useful to define modern big data management approaches by contrasting them with traditional data management approaches with which an agency is already familiar. This section contrasts the two approaches (at a high level) and explains the value of the big data approach in today’s environment of emerging technologies.

Table 1 contrasts 11 characteristics of traditional data management practices with their modern big data management counterparts. The table is meant to provide examples that demonstrate the stark contrast between the current state of practice for most transportation agencies and the ideal state based on data industry best practices. Note that in no case can big data be effectively managed by simply adding more hardware or processing power to traditional methods; the nature of the data demands an updated approach.

Table 1. Traditional Data System/Management Approach Contrasted with Modern Big Data System/Management Approach

	Characteristics	Traditional Data System/Management		Modern Big Data System/Management
1	**System Design**	Systems are designed and built for a predefined purpose all requirements must be predetermined before development and deployment.	***VS***	Systems are designed and built for many and unexpected purposes; constant adjustments are made to the system following deployment.
2	**System Flexibility**	System designed as "set it and forget it;" designed once to be maintained as-is for many years. Systems are rigid and not easily modified.	***VS***	System is ephemeral and flexible; designed to expect and easily adapt to changes. Detects changes and adjusts automatically.
3	**Hardware/Software Features**	System features at the hardware level; hardware and software tightly coupled.	***VS***	System features at the software level; hardware and software decoupled.
4	**Hardware Longevity**	As technology evolves, hardware becomes outdated quickly; system can't keep pace.	***VS***	As technology evolves, hardware is disposable; system changes to keep pace.
5	**Database Schema**	Schema on write ("schema first").	***VS***	Schema on read ("schema last").
6	**Storage & Processing**	Data and analyses are centralized (servers).	***VS***	Data and analyses are distributed (cloud).
7	**Analytical Focus**	80% of resources spent on data design and maintenance; 20% or resources spent on data analysis.	***VS***	20% of resources spent on data design and maintenance; 80% of resources spent on data analysis.
8	**Resource Efficiency**	Majority of dollars are spent on hardware and software (requires a lot of maintenance).	***VS***	Majority of dollars are spent on data and analyses (requires less maintenance).
9	**Data Governance**	Data governance is centralized; IT strictly controls who sees / analyzes data (heavy in policy-setting).	***VS***	Data governance is distributed between a central entity and business areas; data are open to a lot of users.
10	**Data**	Uses a tight data model and strict access rules aimed at preserving the processed data and avoiding its corruption and deletion.	***VS***	Consider processed data as disposable and easy to recreate from the raw data. Focus instead is on preserving unaltered raw data.
11	**Data Access and Use**	Small number of people with access to data; limits use of data for insights and decision-making to a "chosen few."	***VS***	Many people can access the data; applies the concept of "many eyes" to allow insights and decision-making at all levels of an organization.

Each of the 11 characteristics listed in Table 1 is discussed in more detail herein:

1. **System Design** – Traditional systems are designed using the systems engineering approach. They are built to satisfy custom, predefined requirements, which are developed ahead of time and define in detail what the system will and will not do. With the rapid changes of software, hardware, data, and analytics, pre-defining such requirements is almost impossible, as they may be obsolete by the time the development of the system is complete. Modern data systems, therefore, are developed using different requirements. These requirements focus on the ability of the system to handle changes; support many different types of hardware, software, data, and analysis, including those that are unforeseen at the time of deployment; and the ability to remain stable as these changes occur.

> *In no case can big data be effectively managed by simply adding more hardware or processing power to traditional systems; the nature of the data demands an updated approach.*

2. **System Flexibility** – Traditional systems are designed to "set it and forget it." In other words, the system is designed to last for many years before any upgrades or a complete replacement is needed. This long-term approach to "sustainability" in traditional systems is achieved by imparting a significant rigidity by design, making it difficult to corrupt the system while also making it difficult to upgrade. This is no longer a valid approach to sustainability in an environment of rapidly changing hardware, software, data, and analytics. Modern systems instead achieve sustainability through more flexibility and their ability to rapidly correct issues, which allows them to be upgraded much more easily. Modern data systems do not "preset" anything ahead of time and instead rely on adjustable services and a "set at run time" approach, allowing changes to be made constantly. For example, if the data change suddenly (e.g., a new field is added), the system is able to detect these changes and adjust accordingly.

3. **Hardware/Software Features** – Having rigid predefined requirements, traditional system features are often developed by tightly integrating hardware and software to make the best of system performance, resources, and budgets. This tight integration unfortunately renders updates and upgrades difficult, complex, and costly. In addition, this approach limits the types of analyses that can be run using the data. Modern system features, on the other hand, are implemented using decoupled software and hardware, which allows for rapid updates and upgrades to be performed. There are often abstraction layers between the software and hardware, which allows for changes to happen on one layer without affecting the other. As such, in modern systems, analysts can select the right tool for the right analysis on top of the data directly and are not limited by system design.

4. **Hardware Longevity** – Traditional systems are often implemented on high performance, robust, and expensive hardware so they can last several years before the hardware becomes too obsolete and needs to be replaced. While traditional hardware can remain performant, given the current pace of hardware obsolescence, its depreciation often exceeds the cost of acquiring newer hardware. To avoid this depreciation, modern systems use hardware that is acquired at the lowest cost and that is "disposable." This approach ensures a constant refresh of the system hardware as technology evolves.

5. **Database Schema** – With traditional systems, data are organized based on a predefined database schema, and this organization occurs as the data are written to the system storage according to this schema. This is referred to as "schema on write" or the "schema first" approach. This approach requires extensive upfront data modeling to consider all the required datasets and data uses and to create optimal data organization (and as such requires compromises in order to satisfy these user/analytical needs). In today's environment, where datasets and data uses are added and updated constantly, efforts to create optimal data organization are more and more difficult, time consuming, and sometimes even impossible without significant compromise. To cope with the many possible ways of organizing data from many datasets, modern systems take a different approach to organizing the data where data are loaded into the system as-is (in their raw format) and are organized as they are pulled out of a stored location. This approach is known as the "schema on read" or "data last" approach; it requires no upfront modeling exercise, requires no specific data features or formats, and reduces the number of compromises necessary by allowing each data user to organize data according to their own analytical needs. The data are stored as they are, with no requirements, assumptions, formats, or schemas, and any formatting necessary is made at the time the data are used or "read."
6. **Storage & Processing** – In traditional data systems, requirements help to establish the maximum storage and computing capacity required by the system to support expected data uses. In other words, traditional systems are scalable only within certain limits – they are based on the maximum capacity of data and analyses that are expected at the time of design. Considering this need, and the available budget, a centralized, robust, high performance, hardware solution (i.e., servers) is selected and implemented. With big data, neither the storage nor the computing capacity can be easily (or accurately) defined upfront, and the needs will vary widely over time based on the users and their analytics needs. Therefore, modern systems need to be able to efficiently and cost-effectively scale to handle the surges in data storage and computing inherent in big data. As such, modern systems rely on what is known as shared and distributed data storage and processing (i.e., cloud). In distributed systems, multiple instances of the data are stored, and computing tasks are run in parallel across many servers. This distributed approach provides fast and reliable storage and computing capacity, allows for surges in data storage and processing (without designing a system to meet the maximum expected capacity at all times), and allows copies of data and server tasks to be restarted on another server in the case of server failure.
7. **Analytical Focus** – Traditional data systems are complex and leave little room for negligence or for innovation. As such, upwards of 80 percent of the time/resources is spent maintaining and preparing the data (e.g., schema editing, table index maintenance, database archiving) so that they can be analyzed, leaving only about 20 percent of the time/resources for actual analyses.[1] In modern data systems, automated cluster management, disposable and replaceable hardware and hardware/software abstraction, the infrastructure sharing model (i.e., cloud), and schema-less datastores automatically handle many of the traditional data maintenance tasks, such as schema updates, and many others such as archiving do not even need to be performed. As such

[1] Based on the Pareto Principle or the 80/20 Rule (Dam, 2019)

only 20% of the resources traditionally assigned to a system are needed for data design and maintenance, and the remaining 80% of the resources can then be spent on exploring and deriving value from the data.

8. **Resource Efficiency** – Resources assigned to traditional systems are largely spent on system operation and maintenance. In modern data systems, the operation and maintenance costs are reduced through resources sharing and automation, and these costs are paid as a small percentage of the per-use cost of data storage and analyses. This leaves the majority of the budget available to spend on data and analysis tasks rather than maintenance and operation tasks.
9. **Data Governance** – In traditional data systems, data governance is partly handled by the system design (some aspects of governance are performed passively), and the rest is addressed by a set of policies that enforce a strict and detailed approach to application development and data analysis. In modern data systems, data governance cannot be performed the traditional way, because it would be too restrictive. In order to extract the most value from vast and varied data, big data approaches require that many people across an organization (and at all levels) have access to data and analyses. The rigid policing and one-size-fits-all governance of traditional systems simply will not allow for the many possible uses of the data. To support this kind of flexibility, modern data governance is distributed across data applications, shifting responsibility and decision-making from a central authority to a shared model across IT and business areas. As such, data governance in a modern environment is more complex than traditionally, and it will require new tools to monitor and track people in real-time, as well as new processes to manage what data are collected, stored, created, and shared and what resources are consumed across the entire organization.
10. **Data** – Traditional data systems apply an extract, transform, and load (ETL) process, a rigid data model (database schema), and a schema on write approach to bringing data into the system. During this process, data are transformed, filtered, and deleted in order to fit the data into the tight structure. As such, the data that are stored are a processed version of the original data. In addition, strict data access rules (governance) are applied, which aim to preserve these processed data and to avoid potential corruption or deletion of the data. Therefore, the focus of the traditional approach is on maintaining the transformed and processed data as the ground source of truth. The focus of the modern big data approach, on the other hand, is on preserving the unaltered, unprocessed "raw" data as the ground source of truth. The modern approach considers processed data to be disposable and easy to recreate from the raw data. The schema on read approach allows raw data to come into the system, and the shared and distributed approach allows many users to create an unlimited number of processed data sets, analyses, and data products from the raw data. This is a fundamental concept of big data that needs to be understood.
11. **Data Access and Use** – In an effort to avoid overloading a traditional system and potentially corrupting or deleting the processed data, only a small number of people are allowed to freely access and perform operations on the data. This approach limits the number of people to a "chosen few" who can effectively use and gain insights from the data for decision-making. Due to the size and nature of big data, modern data systems require a different approach that relies on many people across an organization – the concept of "many eyes" – to access, explore, and use the data. This approach allows and encourages insights and decision-making at all levels of an organization.

The information presented in Table 1 and the expanded information on each of the characteristics summarizes the shift that needs to happen for transportation agencies to be able to manage data from emerging technologies, and it is the basis of the Modern Big Data Management Framework provided in this guidebook.

Modern Big Data Architecture

The traditional data warehouse architecture pattern has been used for many years. Traditional, schema first architecture assumes that data sources and business requirements need to be understood before storing the data – e.g., what is the source system structure, what kind of data does it hold, are there any anomalies in the data, what data are to be stored in the system, which data should be discarded, how should the data be modeled based on the business requirements. A limited and predefined number of structured data sources are studied extensively to develop the ETL process to organize the storage of the resulting data and to allow reporting and business intelligence to query the data according to the relationships defined in the model. This is a tedious and complex task that grows exponentially with the variety and velocity of data considered and can take months to years to complete.

While this data system architecture pattern is quite ubiquitous and has served transportation agencies well for many decades, it cannot scale in the era of big data, including data from emerging technologies. As data sources become more varied and change more and more rapidly, this approach cannot cope with the complexity and cannot be redesigned quickly enough to handle frequent data and business requirement changes.

Modern data system architecture has taken a radically new approach to circumvent the limitations of the traditional architecture by “flipping it on its head” and taking a schema last approach to data modeling and distributing it across end-users rather than centralizing it. By doing so, it dissociates data storage and management to better allow for rapid changes. This modern architecture is represented in Figure 3, where structured and unstructured data are loaded raw/untouched into a “data lake” where minimal centralized data management is applied. The raw data are then available to many end-users to create their own individual analytical pipelines, and these pipelines/analyses can be similar to traditional ETL and leverage relational database systems. In the modern architecture approach, however, these pipelines/analyses are not set in stone as they are in the traditional architecture; they are disposable and can be thrown away and rebuilt quickly and easily.

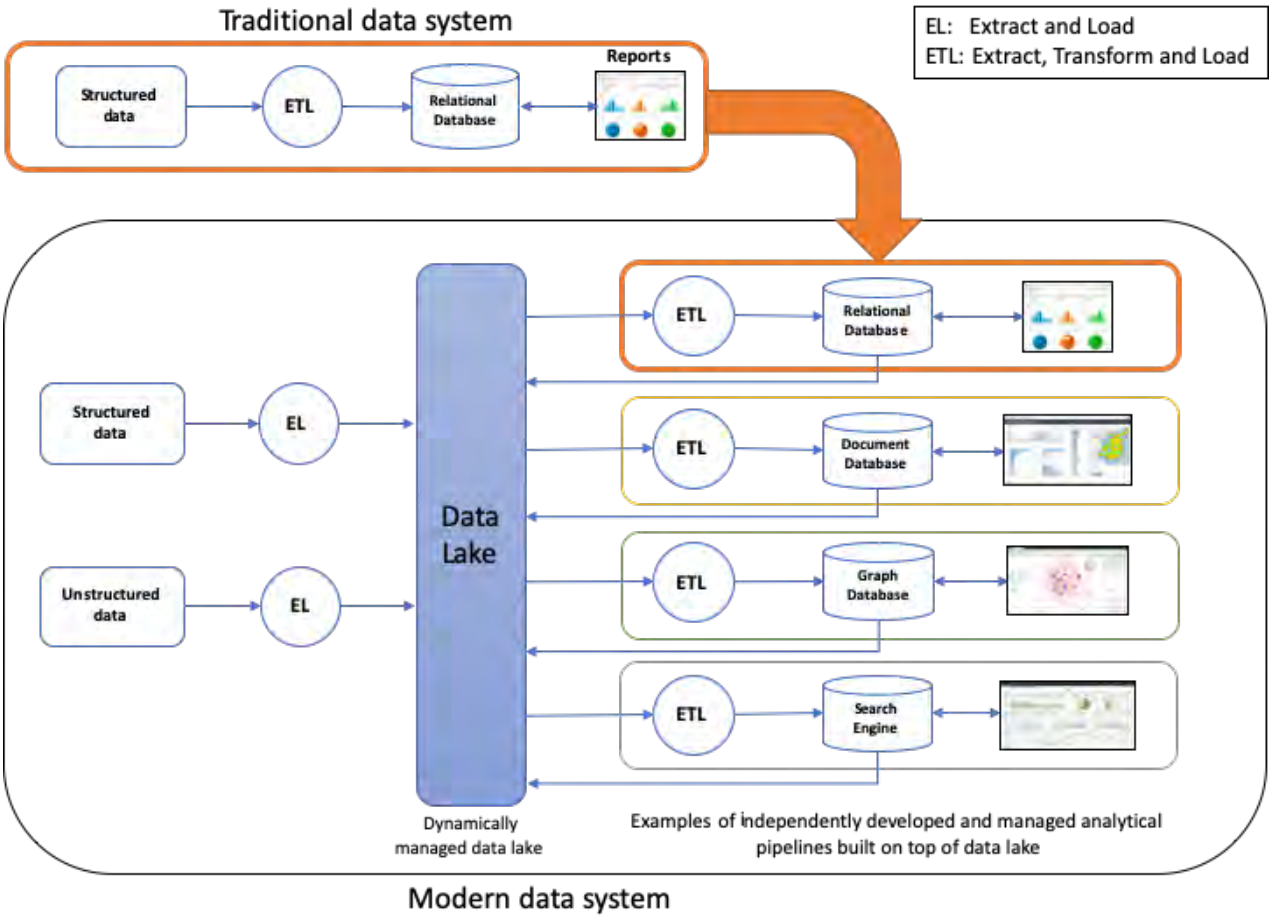


Figure 3: From Traditional to Modern Data System Architecture

ROADMAP TO MANAGING DATA FROM EMERGING TECHNOLOGIES FOR TRANSPORTATION

An agency may embark on this roadmap and guidance for various reasons. There may be a data source that cannot be effectively used or processed due to limitations in the agency's existing traditional data system (e.g., connected vehicle or crowdsourced data that are too big/unstructured to fit within the systems). Or maybe there is a specific problem that could be addressed using data (e.g., data from a third-party scooter provider to help manage scooter drop locations). Or perhaps the city is implementing a smart city project that involves building a system that will handle the requisite volume and variety of data from all partners. Whether an agency is starting from scratch with a new technology data set, has an issue or problem that might be solved with emerging technology data, is already working on a big data project, or is even looking for a new enterprise data management solution, this guidebook will be applicable.

This road map to big data represents an organic, bottom-up approach for transportation agencies that relies on an iterative process to grow big data use cases, pilot projects, and ultimately value for an organization.

The steps and guidance outlined in this roadmap – in conjunction with the best practices and recommendations in the Modern Big Data Management Framework – are designed to walk an agency through the process of developing the knowledge, projects, environment, and buy-in to move incrementally from a traditional data management approach to establishing data management policies, procedures, and practices that fully meet the needs of data from emerging technologies. This roadmap to big data represents an organic, bottom-up approach for transportation agencies that relies on an iterative process to grow big data use cases, pilot projects, and ultimately value for an organization. It allows an agency to start small – at the day-to-day operations level – and to expand and grow interest and use both horizontally and vertically across the organization over time, with the ultimate goal of effective organizational change.

Overview of the Roadmap

The roadmap includes eight steps and is illustrated in Figure 4. Each step is described briefly following the figure.

Figure 4. Big Data Roadmap for Transportation Agencies

- Step 1 – Develop an understanding of big data – With a new data source or a new big data project (or desire for either) at hand, the first step involves an agency champion or champions developing a general knowledge and understanding of big data using the information presented in this step; information in the Modern Big Data Management Framework section; and additional resources referenced herein. The goal at the end of Step 1 is to have enough understanding to promote a big data approach within the organization.
- Step 2 – Identify a use case and an associated pilot project – In this step, the champion(s) and team identify a use case and an associated pilot project that will resonate with their leadership. This use case is likely something that addresses the pain points of the group/division/business unit and that cannot easily be addressed without the use of the dataset(s) of interest. In some cases, a use case may be handed to the champion(s) from the top down, with the charge of demonstrating value for a particular dataset or project.
- Step 3 – Secure buy-in from at least one person from leadership for the pilot project – In this step, the champion(s) and team work to communicate the value of the pilot project and to secure buy-in for the project from at least one person from their leadership. This is a critical step in that, without this buy-in, the project is likely to fail. One champion from leadership can be key to ensuring success of the pilot and expansion to other groups/divisions/business units within the agency.
- Step 4 – Establish an embryotic big data test environment – As this is a big data initiative, it will require building an embryotic big data test environment or "playground" in which the pilot project can be developed. In this step, this embryotic environment is developed following as many of the big data best practices and recommendations identified in the Modern Big Data Management Framework section of this guidebook.

- Step 5 – Develop the big data project within the test environment/playground – In this step, the team develops the big data pilot project within the test environment with iterative feedback from the leadership champion. This development will require the application of modern big data approaches and analytics and the development of data visualizations and products. As such, expertise with these techniques is required, which can be acquired in various ways, the pros and cons of which are presented.
- Step 6 – Demonstrate the value of the data to other business units – In this step, the team and the leadership champion begin to market the data visualizations and products developed in Step 5 to other business units horizontally across the organization. This horizontal organizational outreach will help to market the value of the data and the data products to other mid-level managers and to identify other potential use cases and pilot projects that can be developed within the test environment.
- Step 7 – Demonstrate the value of the data to other leadership/executives – In this step, the team and the leadership champion begin to market the data visualizations and products developed in Step 5 (including any new use cases/pilot projects that have been developed by or for other business units) to other leadership/executives within the organization. This vertical organizational outreach will help to market the value of the data and the data products to executive management to identify other use cases that can be developed within the test environment, but also to begin to gain executive support for organization change.
- Step 8 – Establish a formal data storage and management environment – After many iterations of Steps 2-7 (which could take several years), this step will establish a formal data storage and management environment and will institutionalize policies, procedures, and practices associated with this environment that represent an organization-wide shift from traditional management practices to modern data management practices. Recognizing that changes happen rapidly, and that technology is now disposable, this step requires continuous improvement including enhancing the use of datasets, reevaluating data pipelines, and reviewing system architecture.

Implementing modern big data management principles within an organization requires a multidisciplinary team, and the skills and personnel required will evolve over time. As an agency makes its way through this guidebook, those resource needs will become clearer. For example, early on in the process (Steps 1-3 of the roadmap) staff with specific business knowledge and needs and others with influence are more involved. Developing the big data environment and projects (Steps 4-5 of the roadmap) will involve more technical people, including cloud architects, big data analysts, and counterparts from the information technology (IT) department. After the environment has been built and projects have been developed, demonstrating value for the approach and data products (Steps 6 and 7 of the roadmap) will involve those who can successfully show business value and promote and achieve wider organizational change.

Case Study – The Road to Big Data

The Kentucky Transportation Cabinet (KYTC) is likely the best example (at the time of writing of this document) of a state transportation agency's journey to implement real-time transportation management solutions using a modern big data approach. The journey was sparked by three events: the increasing costs of "snow and ice" operations, the Cabinet's new data sharing partnership with Waze, and interest in a new "database" technology called Hadoop.

In the winter of 2012-2013, KYTC experienced record costs for snow and ice operations. Costs for that winter were approximately $70 million, a significant hike from a historic average of approximately $50 million a year. And based on historical data, KYTC could expect to experience high costs for the following two winters as well. Therefore, decision-makers set in motion a plan to better leverage existing real-time automatic vehicle location (AVL) data from snowplows to help control those costs. The Director of Maintenance tasked the ITS development team with addressing this issue using the AVL data. By the summer of 2014, the ITS team had developed a rudimentary, real-time, proof-of-concept "snow and ice" system to show the value of tracking snowplow activities in conjunction with Doppler radar.

In September of 2014, KYTC signed an agreement to be part of the Waze Connected Citizen Program (CCP). Executive leadership had been briefed on the partnership, but the full benefits of the partnership and the associated data were completely unknown at that time. The goal was simply to help Kentucky motorists better navigate the roadways and to provide additional reporting options on the 511 system, GoKY. Handling the details of the partnership and data sharing agreement were again assigned to the ITS development team. When the ITS personnel responsible for the proof-of-concept snow and ice system discovered that road weather reports were also included in the Waze data, they decided to add the Waze data into the snow and ice system to see what would happen.

Two months later, Kentucky received the first snow of the season. At that point, the snow and ice system was processing real-time data from approximately 200 snow plows every 10 seconds, approximately 200-300 Waze reports every 2 minutes, and was pulling statewide Doppler radar images every 5 minutes. The system crumbled. It required constant reboots and didn't provide the stability or throughput to sustain the operations. Worse yet, ITS personnel understood that KYTC would need much more data in future iterations to build a true, real-time statewide snow and ice decision support system.

Fortunately, the developer assigned to the task knew someone in the IT department who was trying to find a good use case for a new big data database and processing architecture called Hadoop. A few days following a meeting of the minds, the team started pushing data from the real-time snow and ice processor into Hadoop using one of the low latency technologies contained within the system. This approach provided the stability and throughput needed for the real-time processor to continue working.

Over the years, the system has grown by steadily meeting the needs of additional use cases either by incorporating new data sources or by repurposing existing data to be used by a different group of specialists within the Cabinet (Figure 5). The system and much of the data has developed into an enterprise-ready solution and has far outlived the original snow and ice use case for which it was originally designed. As of fall 2019, the system is being incorporated into the long-term enterprise architecture plans of the organization where it will undoubtedly take on a much larger and more fundamental role within the agency. The system has matured through the phases of proof-of-concept to being enterprise-ready with a few production use cases to finally being recognized and adopted as an integral part of the enterprise for integrating, processing, storing, analyzing, reporting, and republishing data.

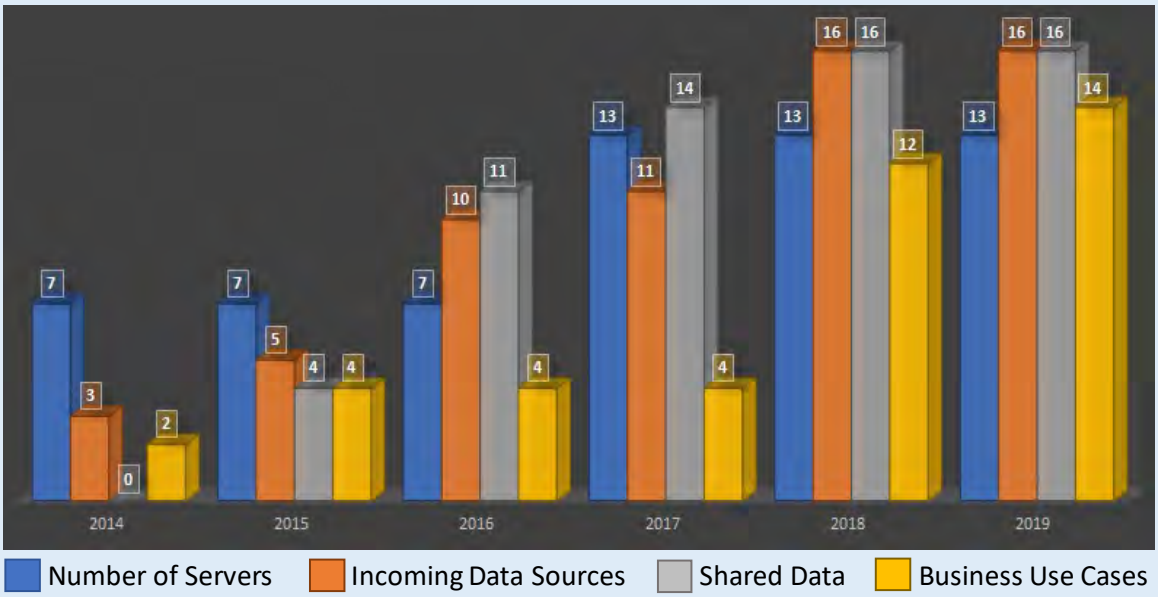


Figure 5. Kentucky Transportation Cabinet Big Data System Growth (KTYC, 2019)

Step 1. Develop an Understanding of Big Data

The first, most crucial challenge in building a modern data system is overcoming a lack of knowledge about big data. Indeed, accomplishing the steps along the big data roadmap will rely on informing and educating more and more stakeholders within the agency about the value of big data, how the big data approach differs from the existing traditional approach, why the agency needs to make fundamental changes in how it views and manages data, and what organizational outputs and outcomes can be anticipated from these changes.

One or more champions within the organization will need to gain a solid foundational knowledge of big data concepts. These data champions need not be experts, nor must an organization hire data scientists or a Chief Data Officer (CDO) to begin this process. However, every organization must know at a high level what big data is, how it must be handled, and why it matters to them. Furthermore, these champions must also be comfortable enough in that knowledge to be able to effectively communicate it with anyone, from top level executives to front-line data users.

While foundational information on the basics of big data will be covered in this section, the Modern Big Data Management Framework included in this guidebook and the NCHRP 08-116 final report Framework for Managing Data from Emerging Transportation Technologies to Support Decision-Making are other good sources to reference to as part of this step. Both documents contain industry best practices and the Framework contains associated recommendations for creating, storing, using, and sharing data. Developing a rudimentary understanding of proven modern big data management approaches early on will pay dividends throughout the process as costly errors and pitfalls are avoided.

What is Big Data?

Big Data is a popular term, but what does it really mean? Below are just a few of the many definitions of big data:

- Big data may refer to data sets, typically consisting of billions or trillions of records, that are so vast and complex that they require new and powerful computational resources to process (Big Data, 2019).
- Big data is an approach to generating knowledge in which a number of advanced techniques are applied to the capture, management, and analysis of very large and diverse volumes of data – data so large, so varied and analyzed at such speed that it exceeds the capabilities of traditional data management and analysis tools (Burt, Cuddy, & Razo, 2014).
- Big data may encompass all of the non-traditional strategies and technologies needed to gather, organize, process, and generate insights from large datasets (Ellingwood, 2016).
- Big data is extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions (Lexico Powered by Oxford, 2019).
- Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It is what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves (What is Big Data, 2019).
- Big data is a new attitude by businesses, non-profits, government agencies, and individuals that combining data from multiple sources could lead to better decisions (Press, 2014).

Within just these few definitions, a number of concepts are revealed that demonstrate that the term "big data" represents more than just the volume of data – it's an "approach," it's about analyzing the data to extract information to inform better decision-making, it's an "attitude." The last definition is an acknowledgment that storing data in "data silos" has been the key obstacle to getting the data to work in ways to improve businesses, work, and lives.

> *Storing data in "data silos" has been the key obstacle to getting the data to work in ways to improve businesses, work, and lives.*

Big Data Characteristics

Most people have encountered the five characteristics or "V's" of big data: *volume*, *variety*, *velocity*, *veracity*, and *value*. Without some context, however, these terms may seem nebulous. Following is a brief review of these terms as they apply to transportation agency data.

"Volume" characterizes the main aspect of a big dataset. Currently, big data are generally considered to be anything over a terabyte (TB); however, the size characterization of big data is continuously changing. As an example, Walmart processes 2,500 terabytes of data every hour (Marr, 2017). While most transportation agencies are still working with "small" data, the amount of data available to transportation agencies continues to grow. Examples of data volumes that are reaching big data levels for transportation agencies include the following:

- 0.222 TB per year – Waze alerts for one state
- 0.65 TB per year – data generated by 300 TMC field devices (Gettman, et al., 2017)
- 1.2 TB per year – data from 300 CCTV cameras if stored (Gettman, et al., 2017)
- 2.3 TB per year – statewide vehicle probe speed data for one state
- 4 TB of data every day – estimate for a single automated vehicle (Swaney, 2019)

"Variety" refers both to the "structured" and "unstructured" data present in big data workflows, as well as the ability to combine and use these various data types to gain insights that were difficult or impossible to obtain prior to big data analytics. "Structured" data can be thought of as data that are structured to be easy for machines to handle, especially when it comes to searching, sorting, or storing data in relational databases. Unstructured data are the opposite of that: video files, audio files, freeform text, and other data that do not conform to traditional data structures and are therefore difficult for machines to categorize. There are entire fields within the study of data science, such as natural language processing and computer vision, that are devoted to helping machines do more with unstructured data.

> *"We have so much data and so much technology giving us data that there is no human that can keep up with it."*
>
> – Delaware DOT

Just as traditional relational databases cannot store different varieties of data, traditional data analysis techniques cannot quantify various data types against each other. With the advent of modern big data techniques, however, data from two separate sources can be combined to generate new insights. One example of this is traffic incident management, where a structured table of incident report data can be enhanced with roadside camera data captured at the time and location of an incident. Using computer vision techniques, vehicles and license plate numbers could be identified and speeds could be calculated that could then be added to the incident report. Another example could be extracting slowdown or weather information from unstructured commuter tweets and using that to add context to speed data coming from global positioning systems (GPS) probes.

"Velocity" is typically defined as the speed at which data are generated. What is not often covered or explained is the variation of velocity that can take place for any given data source. Crowdsourcing and social media are good examples. The velocity of these data sources is typically much higher than traditional transportation data sources and can be highly responsive to newsworthy events. One agency utilizing social media and crowdsourced data asked users to report road closures due to water over the roadways during flooding. As a result, users responded to the request by sending in twice as many reports as usual, jumping from 1,500 to 3,000 in a single day. If an agency is unprepared for high, and highly variable, volumes of data, they will not be able to make effective use of the data.

"Veracity" refers to how accurate or truthful a dataset may be. In the context of big data, veracity is not just about the quality of the data itself, but how trustworthy the data source, type, and processing of the data are (Veracity: The Most Important "V of Big Data, 2019). For example, using sentiment analysis to extract information from traveler tweets is inherently more uncertain than analyzing road sensor data. Other common big data techniques, such as classification and predictive analytics, do not generate an exact result, rather a predicted value with an associated confidence score; this results in reported values or statistics that have confidence bands and levels of uncertainty, rather than absolute "truths."

Just as big data analytics deal with confidence levels, the management and preparation of big data for analysis deals with confidence levels as well. With traditional systems, the rigid structure of relational database management systems requires that data go through extensive preprocessing before they can be loaded into a database. This means that incomplete or aberrant data are often purged before ever being stored in the system, producing cleaner data at the cost of analytical flexibility. Because big data techniques make use of all data, even incomplete entries or outliers, the modern approach instead calls for scoring and flagging suspect data rather than removing them. This approach allows analysts and researchers more flexibility in the data they use but calls for more awareness of the relative veracity or trustworthiness of the data as they have not necessarily gone through a strict preprocessing step like traditional data must.

"Value" denotes how big datasets contribute to improving the status quo. Value involves determining a benefit and estimating the significance of that benefit across any conceivable circumstance. If a new dataset provides answers to important questions of interest, provides new business opportunities or leads to better decisions, then it can be deemed valuable to an organization. Because of this, value is perhaps the most important of the five V's (as evidenced by the previously listed big data definitions).

Big Data Concepts

Beyond the 5 Vs, there are a number of important concepts to understand about big data:

- **Data lake** – a system where a wide variety of data is communally stored in their raw, unprocessed format. This is the opposite of "siloed data," which are stored on disconnected systems that cannot easily communicate with each other. A data lake architecture is particularly beneficial when working with big data that rely heavily on large amounts of raw data and can effectively combine data from multiple types and sources to produce valuable insights.
- **Cloud** – Online accessible virtual infrastructure, software, or other IT services that are hosted on very large external server clusters rather than in-house. Cloud services are popular and almost ubiquitous when working with big data, because they offer – due to the large number of servers and the pay-as-you-go model – scalability, flexibility, reliability, availability, and cost effectiveness that cannot be obtained using on-premise infrastructure alternatives. Cloud storage services are typically the first cloud service organizations adopt, as they can greatly reduce the costs associated with storing, managing, archiving, sharing, and securing large amounts of data (as compared to on-premise).
- **Distributed computing** – A method of performing a single computing task more efficiently by dividing it across multiple servers. An analogy is assembling a team of horses to pull a carriage as opposed to using a single large horse. The concept of distributed computing is widely used in big data, as individual servers are often too small to handle big data processing tasks on their own.

It is implemented through distributed computing frameworks that run directly on a cluster of servers. Distributed computing framework allow for computing tasks to scale easily, as they only need the addition of new servers to their cluster to improve their performance rather than having to upgrade or replace them. Apache Hadoop is the most well known framework for creating clusters of distributed computing resources.

- **Distributed storage** – Similar to distributed computing, distributed storage is the technique of storing large amounts of data on a distributed network or "cluster" of drives/servers. This technique requires a distributed file system to manage the files and present the storage as a single file system, the most widely known of these being the Hadoop Distributed File System (HDFS). Cloud service providers typically handle the management of this process in a method that is transparent to the user.
- **Nonrelational databases** – Because big data involves a large variety of data changing at a rapid pace it can be difficult, at times even impossible, to fit these data neatly and efficiently into a single relational table structure. To remedy this, nonrelational databases have been developed for storing and processing big data. Nonrelational databases are also called "NoSQL" or "Not Only SQL" databases. NoSQL databases do not comply with the ACID (Atomic, Consistent, Isolated, and Durable) model, which guarantees safe data operations on relational databases; rather, they use the BASE (Basically Available, Soft-state, Eventually consistent) model, which is looser than ACID but allows them to adjust to changes in the data rapidly. While these databases often use their own query language, more recently, many have adopted SQL-like syntax for ease of use.
- **Common big data analytics techniques** – A few of the most common techniques enabled by big data include classification (where a classifier is trained on known data to be able to sort new data entries into a particular class), prediction (where existing data points are used to predict future data points), and natural language processing (where language as it is naturally written or spoken is converted into machine-readable data). Transportation applications of these techniques can vary from predicting the location and severity of traffic crashes to extracting road condition data from tweets and blog posts.

When to Pursue Big Data

While a single new large dataset (such as data from a connected vehicle pilot project or probe speed data purchased from a third-party) may drive a transportation agency to pursue big data, it is not necessarily the volume of data at hand that indicates an agency's readiness or need to pursue big data. It could be a combination of the "Vs" that drive the need. It could also be the need for a new enterprise data management system and the recognition of the need for more flexibility and scalability. And while transportation agencies certainly are dealing with larger datasets than they ever have, as well as a variety of data types, value might just be the single most important "V" driving the need for big data.

Most agencies are driven to pursue modern data management practices for one of two reasons: either they have encountered an exciting new data set or use case that requires big data management in order to use, or they see an opportunity to reduce costs, improve efficiencies, or gain some other benefit from modernizing their current data approaches. Most agencies have already encountered one or both of these two situations, and with the rapid advancement of new technologies and data management techniques, it is inevitable that every agency will be forced to deal with big data at some point soon.

Even in the increasingly rare situation where an agency does not yet see a need for big data, by preparing for big data and updating data management approaches now, agencies will be much better equipped to deal with the transition when it becomes an absolute necessity.

The copious amounts of unstructured data common among connected vehicles, automated vehicles, and other emerging technologies cannot feasibly be stored or analyzed using traditional data management techniques. These technologies, along with probe speed data, crowdsourced data, and data from IoT devices, are actively being used by public and private organizations today. Without some knowledge of big data management techniques, public agencies may find themselves too far behind the curve to realistically catch up with private sector capabilities, leading some to find themselves forced to rely on third-party contractors more heavily than would be their preference.

Even when not strictly necessary, adopting modern data management approaches, such as a cloud-based data lake environment, can result in less administrative overhead and more efficient workflows. This increased efficiency may help make a business case for developing modern data capabilities now in preparation for the big data applications that will surely present themselves in the near future. It takes time and effort to build big data management capabilities; starting to demonstrate the value of more modern data management – even on a set of pilot data – can get agencies moving in the right direction so that they are not in a position of needing to play catch up after it is too late.

Common Misconception: Transportation Agencies Only Need to Purchase More Servers to Store Big Data

Despite what the name may imply, “big data” is not simply a larger version of traditional data. Rather, big data are so radically different from traditional data that they cannot adequately be collected, stored, or analyzed using traditional techniques. As big data are not “traditional data only larger,” big data management cannot be “traditional data management only larger.”

For example, if an organization has a traditional relational database and wants to add real-time streaming data from a connected vehicle project, it will be impossible to do so effectively regardless of how performant the relational database system is. The challenge is not simply a matter of bandwidth or processing speed, it is that the traditional database is ill-suited to handle the kind of live geospatial data common among connected vehicle technologies. No number of additional servers will allow a data system to manage data structures for which it was not designed.

Case Study – The Importance of Understanding Big Data

In one transportation agency, the lack of education about big data architecture and methodologies among decision-makers resulted in several mistakes including hardware procurement, development, data maintenance, and reporting. Traditional thinking, and the lack of understanding concerning the benefits and pitfalls of horizontal versus vertical scaling, resulted in the purchase of servers with inappropriate specifications for the data being managed. Utilizing on-premise architecture requires a certain level of understanding about properly scaling central processing units (CPU), random access memory (RAM), and storage ratios. In addition to misunderstanding the concept of scaling, the agency was not fully aware of the breadth of software solutions available to perform similar functions. In this case, the agency decided to use a very complex data aggregation tool typically used for IoT data for simple once-per-day batch jobs. The agency also had trouble understanding the concept of a data lake, specifically how it contains raw data as opposed to processed data. The perception was that preserving and storing raw data, much of which were outside the scope of the original pilot project, served little to no value. Ultimately, the agency found that these raw data, once thought to be useless, enabled several new use cases that went above and beyond the original expectations of the project.

Additional Resources

Suggested additional resources for review include the following:

An Introduction to Big Data Concepts and Terminology – This article provides a definition of big data, addresses why big data systems are different, discusses the concepts and tools associated with the various steps of the big data lifecycle, and provides a big data glossary (Ellingwood, 2016).

Big Data – Concepts, Applications Challenges, and Future Scope – This paper introduces big data concepts and provides multiple case studies showing how big data are used in real-world applications today. It also includes a helpful section outlining common obstacles organizations face in implementing big data. In this section the authors offer insight into why these practical challenges persist and how they can be overcome (Mukherjee & Shaw, 2016)**.**

Big Data and Cloud Computing: Innovation Opportunities and Challenges – This journal article provides clear descriptions of current big data technologies along with a frank discussion of the challenges present when adopting big data technologies. Notable chapters include chapter 3, which provides insight into obstacles faced within specific aspects of big data management and chapter 5 where cloud computing benefits and approaches that overcome these obstacles are discussed (Yang, Huang, Li, Liu, & Hu, 2017).

New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe – This book provides in-depth information on big data storage, processing, and analysis that is generalizable to nearly any organization, including U.S. transportation agencies. Of particular interest is chapter 7 on data storage, which provides informative detail on modern databases, distributed platforms, and methods to secure sensitive data on each (Cavanillas, Curry, & Wahlster, 2016)**.**

Beyond the Hype: Big Data Concepts, Methods, and Analytics – This paper provides a clear and concise definition of big data. It also provides an overview of available big data applications including text mining, video/audio analytics, and predictive analytics (Gandomi & Haider, 2015).

NCHRP Research Report 865 Guide for Development and Management of Sustainable Enterprise Information Portals – This document sets forth guidance and recommendations for state transportation agencies to build sustainable systems to collect, store, analyze, and disseminate data. Sustainability refers to the ability of a system to handle changes and disruptions (e.g., sudden growth in data volume, sudden changes in technology, sudden changes in data quality, security breaches) without being taken down and rebuilt at a large cost. The concept of sustainability is foundational for the collection, aggregation, analysis, and dissemination of data from emerging technologies. The document includes technology recommendations for building sustainable platforms, data governance guidance, software deployment guidance, acquisition recommendations, and more (Pecheux, Shah, & Miller, 2019).

NCHRP Research Report 904 Leveraging Big Data to Improve Traffic Incident Management (TIM) – This document provides guidelines and recommendations for transportation agencies to leverage big data to improve TIM (although the guidelines extend well beyond the TIM use case). The document provides an in-depth explanation of big data, big data architecture, and big data analytics techniques. The document presents and applies two data maturity models to 30+ transportation-related data sources, including data from emerging technologies (Pecheux, Pecheux, & Carrick, 2019).

Big Data's Implications for Transportation Operations: An Exploration – The purpose of this white paper is to expand the understanding of big data for transportation operations, the value it could provide, and the implications for the future direction of the U.S. Department of Transportation (USDOT) Connected Vehicle Real-Time Data Capture and Management (DCM) Program. This paper also identifies two additional, broad areas where big data analytical approaches may be able to provide further value, including transportation system monitoring and management and traveler-centered transportation strategies (Burt, Cuddy, & Razo, 2014).

Big Data and Transport Understanding and Assessing Options – this report examines issues relating to the arrival of massive, often real-time, datasets whose exploitation and amalgamation can lead to new policy-relevant insights and operational improvements for transportation services and activity. It is comprised of three parts. The first section gives an overview of the issues examined. The second broadly characterizes big data and describes its production, sourcing, and key elements in big data analysis. The third section describes regulatory frameworks that govern data collection and use and focuses on issues related to data privacy for location data (OECD/ITF, 2015).

Step 2. Identify a Use Case and an Associated Pilot Project

By this point, the agency champion(s) should have developed a general understanding of big data concepts, benefits, applications, and best practices in Step 1. In Step 2, the agency champion(s) will identify a use case for the data of interest and an associated pilot project in which to demonstrate the value of the data. Note: if a use case / pilot project has already been identified by leadership, and the agency champion(s) is charged with carrying out the project, proceed to Step 4.

In most cases, agencies will be driven to this guidebook because of the need to work with new data that does not fit with their traditional database management systems and practices. To give the reader more context, Table 2 lists a few potential drivers for change and examples of associated use cases and pilot projects.

> *DelDOT conducted a 30- to 60-day pilot project with a few selected datasets as a proof of concept for a new cloud-based solution. This pilot project duplicated critical data from on-premise storage to the cloud for more efficient access and use.*

Select a Use Case and Pilot Project that Align with Business Unit, Leadership, and Organizational Goals

The primary goals of the pilot project are to demonstrate value for the data at hand (as well as the associated modern data management approach) and to create a success story that will drive additional use cases and pilot projects throughout the agency. To improve the chance that the pilot project chosen will resonate with leadership, select a use case and pilot project that meet one or more of the following:

- Addresses a clear and evident need of the business unit – The use case and pilot project will inevitably be linked to the business unit of the champion leading the initiative and should address a clear and evident need of that business unit. For example, a TMC manager (champion) knows that there are routes outside of the current safety service patrol (SSP) areas where incidents are increasing, and incident response and clearance times are longer than those inside the service area. He/she needs to use data that will help to make the business case for additional routes and vehicles. There are new data that can provide information on incident types, locations, and speeds that can be assessed to help make this business case.
- Addresses leadership pain points – While the pilot project might help the business unit with a problem, leadership may care more about solving other problems under their purview. Pay attention to leadership. Figure out what is bothering them and how the pilot project can help them. For example, instead of presenting the project as a way to make things better for staff, frame it as a way to make staff more productive.
- Helps leadership meet their goals – Understand where leadership wants to be (i.e., their goals) and how the selected project can help them get there. They will want to know how the project can help them too.

Table 2. Example Drivers for Change, Big Data Sources, and Use Cases/Pilot Projects

Drivers for Change	Example Big Data Source(s)	Example Use Cases/Pilot Projects
An agency faces an issue or problem that requires new data and new methods, as it cannot easily or efficiently be addressed with the current systems and data alone.	Crowdsourced data – Crowdsourced data generated through mobile apps such as Waze can help address transportation issues or problems more easily and efficiently. Crowdsourcing turns transportation system users into sensors, providing real-time data on traffic conditions, operations, and driver behaviors well beyond the boundaries of the fixed sensors and cameras currently available to transportation agencies.	• Demonstrate early incident detection to improve traffic incident response and clearance times statewide • Combine with automatic vehicle location (AVL) data to improve the treatment for snow and ice during winter storms, while reducing costs
An agency has acquired a new dataset, but it is not being used to its fullest potential due to limitations in infrastructure, tools, and skill sets. A business unit within an agency wants to purchase (or is currently testing) a new dataset, and they need to demonstrate the business case for purchasing it.	Vehicle probe speed data – Like actively crowdsourced data, mobile apps can produce passively crowdsource data such as vehicle probe speed data. These data – usually purchased from a third-party – can provide agencies with new insights into the status of their roadways beyond the boundaries of their fixed sensors and cameras.	• Better manage traffic through work zones and detours • Re-time traffic signals more frequently and without sending staff to conduct field studies • Inform transportation planning strategies and investments
An agency is conducting a connected vehicle pilot project, which is producing data that cannot be used directly by the agency due to its size and structure.	Connected vehicle data – Anonymous signals in connected vehicles are generating new data about how, when, and where vehicles travel, as well as vehicle status information such as vehicle position, heading, speed, and predicted path. This new data-rich environment is the beginning of new safety and mobility applications that will improve safety, help to keep traffic flowing, and make it easier for people to plan their travel.	• Assess the quality, reliability, and trustworthiness of the data being generated and potential uses and applications for the agency • Combine with agency data to better understand where, when, and why crashes occur • Exchange data with roadside infrastructure like traffic signals to improve mobility and safety
An agency needs access to data about new mobility solutions that are emerging within its city to develop more informed policies and plans to help maximize the benefits of these services, while reducing the potential drawbacks.	Mobility and shared mobility data – Data generated by private shared fleet operators deploying vehicles, including Uber, Lyft, scooters, and bikes. These services are amassing large amounts of data on when, where, and how people travel.	• Identify where to expand micro-mobility infrastructure (e.g., lanes and parking) to bring about needed change in how streets are allocated • Assess and incentivize expanding access to mobility in low-income and historically underserved areas
An agency needs a more efficient method for meeting data and reporting demands/requirements, and the current database management system isn't getting the job done.	Agencies may already be experiencing the limitations of their existing systems with respect to data storage and processing.	Demonstrate throughput and cost efficiencies associated with modern data management systems and practices

- Aligns with or links to organizational goals or objectives – Beyond the needs of the leadership of the business unit, select a pilot project that aligns with or links to the overall goals, objections, or mission of the organization. This will resonate with executive leadership and help the business unit leadership look good.
- Addresses policy makers' concerns or goals – Even beyond the executive leadership, how can the pilot project address the concerns or goals of policy makers?
- Can be expanded to include other datasets and use cases – Start small but think big. The pilot project is not a one-off; it can be expanded to include other data and use cases to demonstrate the benefit of the data and the approach to other business units.

> *Start small but think BIG.*

Engage Others in the Cause

In addition, consider engaging others to help identify a project that will resonate with leadership:

- Internal to business unit – Others within the business unit, particularly those on the front lines of the day-to-day operations, are likely to be the most interested in and supportive of use of the data, especially if they understand the potential uses and benefits. These staff understand the existing limitations and gaps and can help to advocate for the need for change.
- Internal, cross business unit – Other business units may also have an interest in or use for the data at hand and may be willing to support the pilot test or even offer up potential/future use cases for the data. Most of the emerging technology datasets offer potential across a wide range of use cases, especially when combined with other data sources, including agency siloed data.
- Junior or mid-level staff – Younger generations are often more tech savvy and may have an interest in applying new technology to problems both old and new. They are often interested in getting involved in using technology to address challenges and may be willing to put in extra hours for the charge.
- External partners – There is a wide potential for the involvement of external partners, depending on the use case. Local partners such as cities that manage traffic signals, metropolitan planning organizations that conduct regional planning activities, and partners such as law enforcement could all be good advocates for the pilot project. Having conversations with an extended group of potential champions to identify uses and associated benefits could give the project more clout.

Remember, there is strength and value in numbers. The more people interested, and the more diverse the interest, the better the chance of gaining support. Collaborating with other business units or organizations could also help to defray the costs of the project by sharing resources across the groups/organizations.

Case Study – Portland Urban Data Lake Pilot Project

The goals of the Portland Urban Data Lake (PUDL) pilot project are to collect and store data from a variety of sources; develop analytics that create new insights from the data; and explore technologies and architectures for providing standardized, documented access to data for public sector agencies and local innovators. The city of Portland recognized that it did not yet have good, centralized systems in place for managing, integrating, and analyzing the data it currently has, much less the large volumes of data coming from Smart Cities technologies like sensors, connected vehicle infrastructure, and private sector services. As an example, the city found itself ill-equipped to handle the streaming data from a Mobility on Demand (MOD) scooter project, which put a tremendous strain on its existing data systems. This project underscored the need for a more advanced way of managing these types of technology data. By building a unified data lake environment, Portland aims to "help City leadership and City staff to make and evaluate decisions, design and evaluate policies and programs, enhance community engagement, and allow us to better partner with the private sector, researchers, and non-profits to meet City goals around livability, affordability, safety, sustainability, resiliency and equity" (Portland Urban Data Lake (PUDL)).

PUDL itself handles various data including data from IoT devices, origin-destination data from MOD scooters and bikes, Waze traffic data, pedestrian counts and more. Having these data in one place allows analysts to merge datasets and perform deep analyses more efficiently than they would have otherwise been able to do. This data lake also supports the Portland Bureau of Transportation's goal of achieving a more nimble, agile, and efficient process of deploying smart cities projects.

Step 3. Secure Buy-In from at Least One Person from Leadership for the Pilot Project

In Step 3, the agency champion(s) and team will work to secure buy-in for the pilot project from at least one person from leadership (likely the next-in line or senior manager) within the business unit. The success or failure of the pilot project will largely depend on this buy-in and support. Securing buy-in for use of a new dataset, as well as taking a modern approach to data management, could be a challenge – the goals or benefits of the project may not be well understood, the project may be seen as unnecessary or frivolous, or the price of the data/project may be perceived to be too high. The champion(s) and team will need to translate their project needs and goals into business needs and goals, demonstrate the value of the project, and do so within a five-minute "elevator pitch."

In addition to the efforts made in Step 2 to select a pilot project that resonates with leadership, following are a few tips for increasing the likelihood of gaining buy-in:

- Establish and clearly communicate the value proposition for the pilot project
- Create a sense of urgency and a fear of missing out (FOMO)
- De-risk the decision by identifying and communicating risks and other potential barriers
- Know how to make the pitch

Establish and Clearly Communicate Value Proposition for the Pilot Project

The team has already selected a pilot project that meets a business need, addresses leadership pain points, and/or aligns with organizational goals. Now, the team needs to clearly communicate the value proposition for the project. The value proposition is a statement that communicates to leadership why they should support the project (e.g., how it will solve problems) and makes the benefits of the project and the resulting products crystal clear from the onset. Whenever possible, develop and communicate the anticipated or estimated return on investment or benefit-cost ratio, even if it relies on ballpark estimates. Having tangible numbers for leadership to mull over can help the cause. If this is not possible, consider using quantitative or qualitative/anecdotal benefits from other transportation agencies to support the argument. Table 3 lists various example projects, along with their value propositions and associated questions to assist in further developing the pitch. (Note: there is natural overlap between these examples and the corresponding questions.)

EXAMPLE VALUE PROPOSITIONS

The smartest way to get around – *Uber*

Rides in minutes – *Lyft*

All your tools in one place – *Slack*

Save money without thinking about it – *Digit*

Table 3. Example Projects, Value Propositions, and Questions to Assist in Developing the Pitch

Project	Value Proposition	Questions to Assist in Developing the Pitch
The project will use new data in an effort to improve performance in a particular area with less effort from staff.	*Improved performance, less effort*	• What is the issue/problem? • How is it currently being addressed? • Why is the current approach insufficient and where is it lacking? • Why is the new data/proposed approach superior? How can it improve over the current approach? • What is the nature of the data (size, structure)? • How accurate and reliable are the data? • What is the cost of the data versus the potential value its use brings to the agency? • Who are the potential users of the data/resulting data products? • What are their use cases? • How might the data be used or combined with other datasets to improve organizational efficiency or performance? • How will performance be improved with the new approach? • How will performance be measured? • What quantitative or qualitative/anecdotal benefits are available from other transportation agencies that could support the argument?
The project will address an issue or solve a problem that cannot be addressed or solved (either efficiently or at all) with existing data/approaches.	*Addressing a challenging issue at a low cost*	
The project will allow the agency to explore/better understand how a promising new data source can be managed and used to the agency's benefit.	*From exploration to improvements*	
The project will leverage new data in an effort to develop more informed policies/plans to maximize the benefits of emerging technology services, while reducing the potential drawbacks.	*Maximizing benefits, minimizing drawbacks*	
The project will make the business case for the procurement of data from a third-party.	*Data for informed and effective decision-making*	

The potential benefits will depend on the nature of the project. High-level benefits that are likely to resonate with leadership include the following:

- Reduces congestion/travel times
- Improves safety
- Increases efficiency
- Reduces costs
- Increases productivity
- Makes things faster
- Makes things easier
- Increases awareness
- Improves processes/procedures
- Develops new capabilities
- Supports new plans/policies
- Balances inequalities
- Reduces negative environmental impact

An example of the "improved performance, less effort" or the "addressing a challenging issue at a low cost" value proposition is the case of traffic operations and signal timing in Louisville, Kentucky. The city of Louisville now uses free crowdsourced data, low cost cloud storage, and free business intelligence software to optimize signal timing. To demonstrate the value of this approach, the city conducted a pilot on a corridor in a fast-growing part of the city that served 40,000 vehicles per day. The city had recently implemented a new traffic control plan for the corridor to account for a 15% increase in traffic and wanted a more efficient method of verifying the results. Using the crowdsource data, a small team developed a dashboard that verified the effectiveness of the newly implemented plan, observing a 30% overall drop in traffic jam reports and a 38% drop during peak hours. The traffic jam dashboard is now available to everyone and used by traffic engineers across Louisville to identify signal timing needs. The

project inspired the expansion of the concept and working with multiple data providers to enhance the signal timing of the other 1000+ signals across the city. The efforts associated with retiming isolated signals and area-wide corridors have been greatly reduced as staff have become more familiar with the data-driven decision-making.

Create a Sense of Urgency and a Fear of Missing Out

Many people and organizations can be professionally conservative and may not want to risk their reputations when taking on new organizational policies or procedures. Thus, they can be averse to change and may avoid making bold decisions that might risk their reputations. It is often easier and less risky to kill an idea than it is to risk failure. However, this position can be overcome by creating a sense of urgency and a fear of missing out (de Ternay, 2018).

Sense of Urgency

When champions create a sense of urgency, they alert the agency (in this case, their direct leadership) why change must occur, and they begin preparing the agency for the change process (via the pilot project). Urgency is important to change, because meaningful organizational change cannot occur without the cooperation of the affected stakeholders. This is why creating a sense of urgency for a needed change is the first step champions should take to gain the cooperation of leadership (Llewellyn , 2015). Champions can create a sense of urgency by:

> *Create a compelling narrative that tells leadership why it is not in their best interest for the organization to stay in its current state.*

- Selling the value of a future state – What is the future state and why should the business unit strive to achieve this state (what's in it for them)?
- Demonstrating that the status quo is a dangerous place for leadership to remain. Clarify the consequences of inaction.
- Communicating clearly, effectively, and with consistency to demonstrate confidence in the proposed approach.
- Being outcome-focused (instead of task-focused) – What are the anticipated outcomes of the pilot project as opposed to the steps in the process?
- Identifying causes of complacency and how to eliminate them.
- Getting to the point quickly – Leadership should understand the project's goals, benefits, and the consequences of inaction within a five-minute pitch.
- Securing internal and external partner/stakeholder input and buy-in (as discussed in Step 2).

The impending influx of data and the associated data management and use requirements are not yet seen as a problem by many transportation agencies. Below are a few statements heard from transportation agency representatives regarding big data from emerging technologies:

> *The impending influx of data and the associated data management and use requirements are not yet seen as an issue by most transportation agencies.*

- *The systems we have right now are meeting our needs, anything new is an additional cost.*
- *In aggregate, transportation agencies do not understand the need for a shift in how data are managed and used.*
- *Unless you are a local city testing connected vehicles, then you are not getting a flood of data. What is the data overload you are talking about? We are not seeing it.*
- *Even though we understand at a high level that big data are coming, we are also balancing the needs of many jurisdictions that are not facing a data problem. It isn't currently a priority for most jurisdictions.*

MISCONCEPTION

"Our systems and processes are good enough."

When making the pitch, a sense of urgency is needed to overcome these perceptions.

Fear of Missing Out

Mentioning that other agencies are already using the data and the benefits they are reaping can trigger a fear of missing out (FOMO). FOMO is a pervasive apprehension that others might be having rewarding experiences from which one is absent. A strong case can be made by establishing that some agencies have implemented a similar project/approach with success. Telling the stories of peer agencies can also help leadership understand how the project would work in their own agency, and buy-in is more likely to be achieved if these comparisons involve agencies that leadership respects (de Ternay, 2018).

De-risk the Decision by Identifying and Communicating Risks and Other Potential Barriers

Leadership is more likely to support a project that is low risk and high reward. The reward is how helpful the pilot project will be to them (e.g., the project helps them solve their problems, achieve their goals, or makes them look good inside and outside the organization). The risk involves costs, the likelihood of failure, and the consequences of failing (de Ternay, 2018).

TALKING POINT

Due to the size of the data and the need for flexibility and scalability, this effort will require a different approach to data storage/management.

While inevitably there will be risks and other barriers to success, the lower the risk, the more attractive the pilot project will be. Therefore, potential risks and barriers to success, along with plans to mitigate them, should be identified and communicated up front to leadership when seeking buy-in. Leadership will be impressed with this risk-mitigation strategy and feel more confident in the overall approach.

The primary risks associated with a new data project will likely involve resistance from the IT department, as well as the traditional procurement process. IT may push back on the establishment of an embryotic big data test environment in the cloud, because it doesn't follow recognized processes or make use of approved tools. There may be perceived security issues related to storing and analyzing data in the cloud. And the cost model of the cloud is likely to be misunderstood or even rejected.

The champion(s) and team should be prepared to defend the concept and needs for the embryotic big data environment – demonstrating that the environment will be separate from formal organizational systems and processes; it will foster the assessment, exploration, and analysis of a new dataset; and it will help to demonstrate the benefits of these data to the organization.

> *The primary risks associated with a new data project will likely involve resistance from the IT department regarding the establishment of an embryotic big data environment, as well as the traditional procurement process.*

Know How to Make the Pitch

There are different situations and ways to make the pitch to leadership. The pitch could be discussed informally over coffee, a meeting with other stakeholders could be arranged, or a one-pager could first be developed and shared to present the idea. The approach that will work best will depend largely on the parties involved and their relationship. A rule of thumb is to keep it low profile at the beginning; it makes things less scary and more human (de Ternay, 2018).

If efforts to recruit a champion from leadership are not successful, speak to others in leadership positions wherever possible. If no one from leadership is willing to support the project, it may be necessary to return to Step 2 and refine the project proposal or choose a new project altogether. It is not recommended to continue on to Step 4 without backing from leadership. Creating a successful project will be nearly impossible under such circumstances, and even if the project does succeed, there is no guarantee that it will lead to organizational change. It is far more efficient to spend extra time finding the right pilot, use case, and value proposition early on than it is to press forward on a project that goes nowhere.

If the efforts put forth thus far fail to convince leadership to support the project, return to Step 2 and refine or select a new project that better meets the needs of leadership. It may take several attempts before landing on a project that gains buy-in.

Step 4. Establish an Embryotic Big Data Test Environment

After gaining support from leadership for the pilot project, the next step is to establish an embryotic big data environment. This data environment is often referred to as a "playground." It is considered to be a test environment where there are little risks associated with the use of the data. Typically, the playground is a scalable and developmental platform used to explore an organization's datasets through interaction and collaboration. The playground is primarily for business units to explore new data in a big data context using new data analysis tools and leveraging advanced analytical methods not currently in use by the business unit or the agency. This playground should support the needs of the pilot project and should follow as many big data best practices as possible from the Modern Big Data Management Framework so that it will be easily scalable to allow for the addition of more datasets or analytics when needed.

The playground should provide the capability to work with both small and large datasets coming from both historical and streaming data feeds. It should allow users to perform data analyses, from simple analyses such as aggregation, to complex analyses requiring massive parallel processing, large amounts of memory, and high-capacity storage and input/output (I/O) capacity. It needs to be separated from production data warehouses to facilitate data experimentation.

> *When planning the big data project and associated environment, make sure to refer to the industry best practices and recommendations in the "Store" section of the Modern Big Data Management Framework in this guidebook. Also, refer back to Figure 3, as it visually demonstrates the differences between traditional data architecture and modern data architecture.*

The playground is typically created using cloud services, as they allow access to large storage and computing power on demand on a pay-as-you go model, which dramatically reduces the cost of running the test environment. Setting up the big data test environment within the agency will require collaboration with, support from, and approval of the agency's IT department. It needs to be understood, however, that ***the goal at this stage is not to propose or force an organizational change***. Rather, the goal is to establish a separate, independent test environment in which the benefits of the data can be evaluated on a small scale and the benefits of the platform can be demonstrated to others with different use cases to develop interest and drive adoption across the organization.

Step 4 includes the following activities:

- Establish buy-in from IT
- Establish test environment
- Take ownership and responsibility for analytical projects

Establish Buy-In from IT

It is important to emphasize that the environment developed in Task 4 is a test environment or playground that needs to comply with the modern data systems and management approach described in Table 1 and the big data best practices and recommendations presented in the Modern Big Data Management Framework (e.g., needs to be scalable, flexible, and allow access to a range of users). If IT demands that a traditional, more rigid and controlled approach be adopted instead, the test environment simply will not work or be successful.

There will almost certainly be a need to share the big data knowledge gained in Step 1, the information contained in this step (Step 4), and the recommendations in the Modern Big Data Management Framework with IT, especially if this is the organization's first foray into big data. A clear understanding of how hardware, software, and cloud pay-as-you-go models differ from traditional procurement is needed, as they are likely to be quite different from what the IT team normally encounters. For example, most cloud-based data storage and processing services charge monthly usage-based fees that can be difficult to compare with the costs of purchasing and maintaining local hardware. Explaining that the computing needs of the test environment are of an unpredictable, elastic nature, and that they can peak to levels that are higher than all the computational capabilities found in the organization for a few hours, will be needed. Also, preparing an estimate of data access needs ahead of time will make it easier

for the IT team to understand and compare the costs involved. It may also be helpful to recruit at least one IT professional to be integrated into the pilot project team to support close collaboration and clear communication.

If this approach fails, or if the cloud is not allowed, the test environment technically could be developed on-premise by either deploying appliances from typical vendors such as IBM or Oracle or deploying an on-premise cloud setup. It should be noted, however, that this approach is highly discouraged. Building a cloud-like environment on-premise is a large and challenging task. Such environments require advanced server clustering expertise that is often not found within transportation agencies and that is expensive to acquire. They are also much more expensive and time consuming to deploy, operate, and maintain than their cloud counterparts, which as a consequence lead them to be deployed under strict control policies that limits data experimentation. The deployment and management of an on-premise, cloud-like environment requires a significant amount of resources to support, including but not limited to:

- Purchase of a large quantity of commodity servers and network hardware to build a cluster.
- Constant replacement of the cluster hardware due to failure, obsolescence, and scaling.
- Deployment and constant maintenance of the cluster software stacks.
- Real-time monitoring of the cluster software stacks and hardware resources.
- Real-time monitoring of users and data to ensure openness and security at the same time.

These challenges are the reason why online cloud services were created to begin with – mainly to allow organizations small and large to share the burden of managing large computer clusters. The benefits offered by online cloud environments are numerous and make the adoption of an on-premise cloud reasonable only for very large systems such as global banking or video streaming. Below are some of the benefits that online cloud service providers offer over on-premise cloud:

- Quick adoption of new hardware technologies such as graphical processing units (GPU), field programmable gate arrays (FPGA), and solid-state drives (SSD)
- Inexpensive data storage
- Very large computing capability
- Built-in data and user management
- Large choice of software implementation from vendors to open source
- Shared management and security

While cloud architecture is highly recommended, there may be situations where a cloud environment is not feasible – e.g., where data use is reliant on software that cannot run in a cloud environment because of compatibility (e.g., aging software) or legal reasons. Such situations are increasingly rare, however, leaving most transportation agencies free to pursue the benefits of adopting modern cloud-based data architectures.

> *"IT management wants a clear definition of why you need a data source and what parts of that data source you'll be using. It took us a few years, but we finally convinced them that people need to see the data first; they can't just read a 70-page API document and expect to understand what they need."*
>
> – Kentucky Transportation Cabinet

The establishment of a big data playground will not come without barriers. Below is a list of barriers that business areas should expect to encounter when establishing a data playground:

- The pay-as-you-go cost model for cloud computing is often perceived as an issue (resulting from a shift from centralized IT procurement/billing to decentralized/distributed billing of groups across an organization that are using the cloud services).
- Data privacy concerns due to a lack of confidence or knowledge with regards to securing data in the cloud.
- Resource concerns originating from not knowing the amount of resources that will be needed to establish the environment or how much it will really cost in addition to maintaining the current infrastructure.
- Resistance to establishing a new data environment when there is already an infrastructure in place with people maintaining it.

Depending on the business model, IT teams may push back on big data development. This pushback not only stems from a lack of understanding, but also from a fear that big data adoption will reduce the team's size or importance to the organization. Finding the right funding structure to where big data initiatives can use modern technology without fear of eliminating IT payroll positions is a long-term problem that may require extensive retraining or reorganization. The adoption of a cloud-based data playground should not be dependent on the adoption of cloud at the organizational level (this will come much later). Leadership champions should work on obtaining IT department support in the short-term so that the pilot project development is not delayed while a long-term solution is sought.

The process of obtaining buy-in from IT will most likely be met with negation, and it is expected that it will take time and require the involvement and support of reputable and highly trusted individuals in the organization.

Common Misconception: Data Stored on the Cloud are Less Secure Than Data Stored Locally

Storing data on a cloud service does not make it any less secure than storing it locally. In fact, data are often more secure on cloud services, because modern cloud service providers employ large teams of cybersecurity experts who focus on securing data as a critical aspect of their business. In contrast, most transportation agencies are under budgetary constraints that limit how much cybersecurity expertise they can develop and retain in-house. Major cloud service providers therefore have stricter security procedures that employ more up-to-date algorithms than most transportation agencies.

Establish Test Environment

Establishing a big data test environment / data playground differs from the traditional IT system deployment that agencies currently follow. Allocating a server or a relational database on the cloud as a data playground will not achieve the intent and goals of the playground, as this approach will inherently limit the amount of data that will be able to be explored as well as how and how fast they can be explored. This approach will also require data preparation that will need to be done before moving data into the playground, which will dramatically reduce the value that can be derived from them.

Rather, the data playground needs to be established in a more flexible yet controlled fashion that is split into two independent layers: a data storage layer and a data processing layer.

Data Storage Layer

The data storage layer is the part of the data environment where the data to be explored are stored. This part of the data environment should be implemented on a cloud storage service so as to benefit from its ability to easily and inexpensively store, organize, and secure very large datasets and make them available to many different data processing software applications.

> *Cloud storage services have allowed DelDOT to more easily integrate their data sources (which was a challenge to do in-house), store all the data into a single repository so end-users can access them directly, and leverage cloud tools that they otherwise would not have access to.*

Once the storage service is acquired, data of interest need to be moved to the playground storage. This can sometimes be a problem, as some organizations do not trust that cloud services can store data securely, especially when data are, or are perceived to be, sensitive. It is then essential to assure management, with the support of IT, that the data can be stored securely on the cloud. A solution that is often implemented to help secure data and alleviate the fear of exposing the data in a public cloud environment is a virtual private cloud. Virtual private clouds are on demand, configurable pools of shared computing resources allocated within a public cloud environment and are designed to provide a certain level of isolation between the different stakeholders using the cloud resources. This solution should be strongly considered to reduce the risks and eliminate existing fears (perceived or real) of exposing sensitive data in the cloud.

While the data playground only uses a simple storage solution to store datasets, the way in which the data will be stored will need to satisfy several conditions in order for the data playground to provide the most benefits to its users and avoid becoming messy (i.e., a data swamp):

- The data need to be stored as-is or raw, that is unedited or transformed from the way they were provided to the agency (from sensors or third-party APIs). This is rather important, as early transformation of the data may inadvertently filter out data that may be perceived as useless but that in fact are essential to the exploration of the data and the establishment of their veracity and value.
- The data need to be stored under strict “read-only” privileges for all users expect individuals in charge of uploading and managing the data. Indeed, within the playground no users should have

the ability to alter or delete the uploaded raw data, as they represent the ground source of truth for each individual data exploration effort and need to remain unaltered.

- The data need to be organized logically so that they can be easily found and understood by analysts. To do so, a taxonomy created from keywords describing each dataset and documentation describing each dataset and its content should be developed. The taxonomy will then be used as a basis for a simple folder structure into which each dataset will be stored. The resulting folder structure and documentation can then be shared within the playground to help users understand what datasets are where and how they are structured. This goes a long way to providing a head start to most data analysis projects.
- Should some of the data be sensitive, access should be restricted to a few users only. While a virtual private cloud implementation will protect the data from other cloud users, some data will also need to be protected from cloud users within the organization. There are two options available to secure sensitive data in the data playground storage: folder access restriction and encryption. Depending on the data and the nature of the processing of the data, either or both solutions can be implemented. Folder access restriction should be considered in the taxonomy/folder structure so that limiting access to a folder where sensitive data are stored does not also limit access to non-sensitive data. Data encryption should be considered when there is a need to provide a clean version of a sensitive dataset so that analyses can still be performed on the data without risk. Encryption algorithms should be selected carefully as many have become compromised over the last few years.

Once the data are stored and organized into the data playground storage, they are now ready to be explored using cloud analytics services.

> *Understanding the concept and utilization of a cloud storage layer (i.e., data lake) is extremely important at this stage. Agencies that do not fully grasp the need for storing raw, unprocessed, data are far more prone to encounter mistakes or pitfalls as the architecture and use cases mature over time. One agency described the data lake as the "big undo button." This agency has been forced to deal with several "versions" of the same dataset due to the inability to implement a data lake much sooner in the process.*

Data Processing Layer

The data processing layer is the part of the big data test environment where the datasets stored in the data playground storage layer are processed to create new data that will then be stored back into the data storage layer.

Again, as with the playground storage layer, the processing layer should be able to support multiple data analysis tools as needed to explore the data in the storage layer. *Providing a relational database or traditional statistical software to users in the playground will not suffice, as they will not provide the analytical capabilities needed to process the large datasets in parallel or the unstructured and semi-structured datasets stored in the storage layer.* They will also not support a set of analyses varied enough to explore datasets, as traditional analytical solutions often can only apply a limited set of analyses to well curated datasets.

The data processing layer should not prescribe specific analytical tools; rather, it should allow users to pick and choose the tools they would like to use to analyze the data. This approach is a complete departure from traditional IT management and methods used to control data processing; but in order to explore the potential value of the raw datasets, multiple analysis tools will be needed to discover the characteristics and hidden patterns in the data and develop those into data analytics pipelines. Furthermore, the questions of interest and the skills of individual users will also affect the choice of solutions that should be used for each project.

The data processing layer should not prescribe specific analytical tools; rather, it should allow users to pick and choose the tools they would like to use to analyze the data.

The data processing layer should then make available to its users as many tools as possible, ranging from the tools provided by the cloud service provider to open source tools:

- Cloud provider services solutions are the easiest way to deploy analytical solutions on a cloud environment but are designed to lock the user in to their specific cloud.
- Open source solutions are ideal to implement cloud analytics solutions without risking a cloud vendor lock, but they require additional skills, such as container development and management, to be deployed.

The use of non-cloud provider, commercial data analysis solutions is not recommended in the data processing layer, as their costs are often non-negligible and make them better suited for a production system rather than a system meant for exploration. Indeed, cloud provider analytics solutions offer flexibility at a relatively low financial commitment, and open source technologies can also avoid the need to pay substantial sums for every new component. By following these recommendations, the cost of the data playground can remain reasonable without forcing the data processing layer into a commercial analytical solution.

Take Ownership and Responsibility for Analytical Projects

Managing and controlling the playground data storage and processing layers may also raise concerns with IT and management. Indeed, with many users developing multiple data analyses and using multiple analytical solutions, performing data management across the data playground can be seen as overly complex, overwhelming, or even impossible from a traditional point of view. And risk of the environment becoming out of control and overly costly could compromise its flexibility or even existence after only a few months.

To remain in control of the data playground, a different approach needs to be taken that combines the following:

- Complete ownership of the data project by the champion and team. They should be held responsible for development, maintenance, and expenditures, and only in exceptional cases should they place the burden of supporting the project on the IT department.
- Clear control of who has access to what data, including real-time tracking of who accesses and processes what data and the implementation of alerts and triggers to avoid abuse and

violations. This can be implemented by the data governance team or the IT department and should be done using the activity log analysis tools made available by the cloud provider.
- A clearly defined starter budget for each data analysis project in the playground and a defined process for additional funding to limit excessive spending.

By implementing such a big data test environment, agencies will be able to safely explore new data and achieve many benefits without massive outlay. By investing in the development of team knowledge, giving them access to the type of data environment and responsibilities typically under the authority of IT departments, they will also start to develop a data understanding from within the organization that can begin to foster a culture of data.

> *A major concern shared amongst almost every transportation agency is the liability or risk involved with handling sensitive data and PII. Many agencies protect themselves from PII by immediately aggregating away sensitive information and permanently deleting the original data. This approach effectively removes the risk of a data breach, as there is no sensitive data to be compromised. It also removes the risk of being subpoenaed, as personal information cannot possibly be identified from data that no longer exists. The downside of this approach, however, is that if anomalies are found in the data, it may be impossible to determine whether the values are legitimate outliers or data collection errors. If the source data are preserved in an encrypted format on a separate, restricted-access server, a similar level of protection can be achieved while avoiding the loss of data usability.*

Common Misconception: Transportation Agencies Must Regularly Delete Data to Keep Data Storage Affordable

Traditionally it has been good practice to regularly review data for archival or destruction in order to keep data storage costs low, even to the point that data management lifecycles typically included a step to destroy or purge data. In the world of big data, destroying old data no longer needs to be a focus. Most cloud storage providers will automatically transition less-used data to archival storage with no input required from the data owner. Furthermore, under the usage-based fee structure that is common among cloud providers, the less data are used the less they costs to store. These factors, along with an overall decrease in cloud storage costs and an overall increase in the value of data for modern analytical techniques, creating an environment whereas much data are retained as possible is much more feasible than it traditionally has been.

Case Study: On-Premise vs Cloud

The Kentucky Transportation Cabinet started the road to big data in 2014. Cloud computing during that era was still considered to be too new and too risky for government agencies. As such, for the first two years, the team developed and scaled the system using on-premise architecture and expertise developed in-house. After a staffing change on the development team, and the need for additional resources, the proof-of-concept pilot project graduated to an official project within the Office of Information Technology. Soon after hiring more staff in early 2017, the agency decided to continue using on-premise architecture. Even though cloud computing had matured by this point, it still wasn't an approved architecture by the centralized IT department. Instead of challenging that policy, the development team was told to continue using and scaling on-premise architecture. Executive leadership still did not fully understand the benefits and potential of big data or further benefits from a cloud architecture. During the process of scaling data inputs and processing, the on-premise method of scaling big data started to experience very serious pain points. As layers of complex tools and additional data were added, the system stopped operating as expected and the team was forced to dedicate entire sprints (3-week time periods) to system tuning and optimization. Issues such as small files and CPU/RAM resource management for processing became extremely problematic, taking valuable time away from use case development. An outside organization was eventually hired for the sole purpose of managing the network servers so the developers could spend more time working towards business use case functionalities. In 2019, the development team finally received the go-ahead to proceed with a proof-of-concept to move the real-time data pipeline to the cloud.

Step 5. Develop the Pilot Project Within the Big Data Test Environment/Playground

In Step 5, the champion(s) and team will work within the big data environment established in Step 4 to develop the pilot project. The development itself should be viewed as an iterative process with multiple feedback loops and revisions expected as the project evolves. The outputs of Step 5 will include data products (e.g., visualizations, dashboards, maps) that demonstrate the value of the data to the department. In Step 5:

- Develop/ensure availability of the right expertise
- Develop the project applying a data science perspective
- Iteratively develop/improve the project and associated outputs

Develop/Ensure Availability of the Right Expertise

To support the development of the big data environment and pilot project, an interdisciplinary team must be gathered. This team should include cloud architects, modern data management specialists, big data analysts, and business area specialists. This team can consist of in-house personnel, contractors, university personnel, or all three. When choosing which resources to engage in the development of the project, keep in mind that while the pilot project itself ought to provide benefit to the organization, that is not its only goal. This project will eventually serve as evidence when arguing for the adoption of modern big data management practices across the organization. Therefore, there is an inherent benefit to using in-house resources as much as possible; they will naturally develop into data champions as they work on these projects. This in-house knowledge, experience, and enthusiasm for big data often proves invaluable to generating the momentum needed for organizational change.

> *"Training in-house staff is a major issue, and we will always argue that in-house staff need to be trained, even if the contractor performs the work or builds the solution. This cannot be understated, even if your role is just to lead and advise. We put in the hours, learned the technology, and designed the system instead of just asking for things without understanding what we were asking for. As a result the agency has benefited greatly."*
>
> – Kentucky Transportation Cabinet

There is a variety of options for developing/ensuring the availability of the right expertise to develop the big data pilot project within the test environment. These options include:

- Developing the expertise with staff already in-house
- Acquiring/hiring new staff with big data expertise
- Contracting with trusted contractors or universities
- Contracting with big data experts/consultants

Table 4 provides a high-level overview of some of the pros and cons of each of these options for developing or working with different resources to develop the big data pilot project.

Table 4. Pros and Cons of Different Potential Support Resources

Resources	Pros	Cons
Training or Hiring In-House Personnel	The same resources that built the system will support the system Skills developed during the pilot will be retained for other projects Resources become data champions within the organization New staff hired with requisite skillsets can be immediately effective and productive	Difficult to attract big data professionals to transportation agencies (e.g., salaries not competitive with private sector) Can be costly to hire for big data skillsets Training for big data skillsets requires time and dedication Training in-house staff make them more marketable to other organizations (may lose them after they are trained)
Trusted contractors and university partners	Trusted/vetted resources May be able to get things done more quickly if in-house resources are limited and there are competing priorities Usually local or "pre-approved" Cost accountability	Do not often employ staff with the requisite big data skillsets Cannot verify quality of work without some in-house expertise High turnover rate University students involved in projects have little business experience
Big Data Experts/ Consultants	Already possess requisite big data skillsets Understand languages and tools	Cannot verify quality of work without some in-house expertise High turnover rate Not vetted Not usually local or "pre-approved" Comes at a price – expertise more widely available but still not common

How an agency decides to obtain the right expertise will likely need to be dealt with on an agency by agency basis. Some agencies may prefer to develop the expertise in-house, which will require a long-term commitment from the agency to not only develop the appropriate job descriptions and skill requirements – e.g., modern big data pipelines, modern tools (machine learning, natural language processing), Python, JSON, object oriented programming, applied statistics – with commensurate and competitive salaries but to ensure that these personnel receive ongoing training to keep their skills up to date given the fast-changing environment. As private companies can adjust more quickly to changes, other agencies may prefer to contract the necessary resources to big data experts/consultants; however, it is strongly recommended that in-house staff develop a core set of skills through education and/or experience to provide oversight, ask the right questions, and verify the quality of data and work. Agencies are cautioned regarding the use of their traditional, on-call contractors, as many have also not yet developed or hired staff with the requisite skillsets to develop big data systems and projects.

Develop the Project Applying a Data Science Perspective

Given the overall goals of the pilot project, it is important that the team approach the project from a data science perspective (as opposed to a more traditional approach) – i.e., extracting value from the data. Project development steps applying a data science perspective include:

- Identify the goal of the project
- Collect raw data
- Process and clean the data
- Perform exploratory data analyses
- Build data science pipeline(s)

> *Be sure to reference the Modern Data Management Framework section of this guidebook for more details.*

Each of these is discussed in more detail herein.

Identify the Goal of the Project

While the champion and team likely accomplished at least some of this in Step 2 of the roadmap, this first step of project development involves developing a very clear sense of what the team is trying to achieve through analyzing the dataset(s) in question. Questions answered at this stage include (Turner, 2019):

- What decisions need to be made from the data?
- What questions do the team wish to answer?
- For answers, what level of confidence would the team be happy with?
- Can the team formulate hypotheses relating to these questions? What are they?
- How much time does the team have for the exploration?
- What decisions would the team like to make from the data?
- What would the ideal result(s) look like?
- How is the team to export and present the final results?

The team should consider brainstorming, whiteboarding, and workshopping during this stage to answer these and other questions to develop a clear line of sight to the overall objectives of the project.

Collect Raw Data

The next step in developing the project is to identify and collect data that will help to provide insights needed to formulate a solution for the problem at hand. This part of the process involves thinking through what data the project team will need and how they can be obtained. The latter is often the least easy of the two. Data can be obtained in two ways: by obtaining historical datasets and by collecting data directly from real-time data feeds. Unlike some traditional data providers, most modern third-party data providers generate so much data that they do not provide historical datasets, just a live data stream. The project team will need to make sure to collect enough data from the data streams to ensure a good understanding of the data collected.

It is also useful to ensure that the team has a bigger picture understanding of what is there. Questions asked during this stage include (Turner, 2019):

- What is the size of the data?
- How many files are there?
- To what extent does the data originate from different sources?
- Automated exports or manual spreadsheets?
- Does the data have consistent formats (dates, locations etc.)?
- What is the overall data quality?
- What is the level of cleaning required?
- What do the various fields mean?

- Are there areas in which bias could be an issue?

Answering these and other questions about the data can aid the team in deciding how to go about the analyses and determining which aspects of them are most important for the analyses.

As a start, organizations should consider adding the following datasets to their data playground:

- Agency traffic/speed data
- Road weather information systems (RWIS) data
- Operations/traffic management center data (incident reports, etc.)
- Social media and/or crowdsourced data (e.g., Waze)
- Third-party vehicle probe speed data

Process and Clean the Data

Once the raw data have been collected and a copy has been stored, the next step is to transform and clean a *working copy* of the data before performing any in-depth analyses. In many cases, raw data can be quite messy (duplicate values, missing values, corrupted values, non-standard timestamp formats, time zone differences, unexpected data in columns or fields). As such, this processing and cleaning can take a long time and can be relatively tedious work, but the results are well worth the effort. This step includes (Turner, 2019):

- Combining all data into a single, indexed database
- Identifying and removing data that are of no relevance to the defined project goal
- Identifying and removing duplicates
- Ensuring that important data are consistent in terms of format (dates, times, locations)
- Dropping data that are clearly not in line with reality, these are outliers that are unlikely to be real data
- Fixing structural errors (typos, inconsistent capitalization)
- Handling missing data either by dropping or interpolation
- Labeling and organizing the data efficiently so that there is no confusion about what is contained within the data or what the data mean

Processing and cleaning can be a bit of a manual discovery process and is greatly facilitated by data analysis experience and domain knowledge expertise. To uncover errors, the project team will want to look through various aggregates and plots of the data and assess if the values make sense. Once uncovered, depending on the findings and problem to be solved, the project team will likely need to correct, rename, or remove data; however, not all errors need to be removed through this process. Indeed in some cases, the errors themselves need to be considered as trusted data to solve certain problems.

Additionally, the project team may want to enrich or augment the datasets by adding new values needed to solve the problem. This is often done by joining the data with another dataset containing the data to be added. An example of enrichment is adding historical weather data to incident data by joining the two datasets by location and time.

Perform Exploratory Data Analysis

Once the data have been processed and cleaned, the project team can begin inspecting, exploring, and modeling the data to find patterns and relationships that were previously unknown. This process is of a heuristic nature and requires a lot of poking and testing to uncover patterns from the data. This may include (Turner, 2019):

- If there are time-based data, explore whether there exist trends in certain fields over time — usually using a time-based visualization software
- If there are location-based data, explore the relationships of certain fields by area — usually using mapping software
- Explore correlations (r values) between different fields
- Classify text using natural language processing methods
- Implement various machine learning techniques to identify trends between variables/fields
- If there are many variables/fields, dimensionality reduction techniques can be used to reduce these to a smaller subset of variables that retain most of the information

Here again, experience in big data analysis and domain expertise can be of great help. The difficulty of this step is to come up with ideas and tests that can quickly lead to valuable patterns so as not to lose too much time and money exploring areas of the data that do not provide value. This can be difficult when the datasets are very large, very small, or unstructured and when the domain they cover is not well known. From the uncovered insights and patterns the project team can now use them to develop a more in-depth analytic pipeline to.

This step can sometimes reveal patterns in the data that may require the data to be transformed and cleaned in a different way that was done in the previous step, in which case the process and clean step should be repeated.

Productize a Data Science Pipeline

This step focuses on developing a prototype analytical pipeline based on the findings of the previous steps. A data science pipeline is a sequence of processing and analysis steps applied to data for a specific purpose. Figure 6 shows an example of a modern data pipeline, developed from data in a data lake. The development of a data science pipeline is composed of two main subprocesses – one subprocess that first applies a refine, transform, and clean process to the raw data, then a subsequent subprocess that applies descriptive, inferential, predictive, prescriptive, or causal analyses to the transformed and cleaned data to generate a resulting data product.

> *Once established, this data pipeline will automatically pull in the appropriate raw data, transform it as necessary, and apply pre-defined and pre-established data analytics techniques to develop the end data product.*

After the raw data are cleaned, better understood, and prepared for analyses and the exploratory analyses are conducted on the data, the team will likely be ready to develop a pipeline that will generate a specific end-product – such as a report or dashboard – that may run automatically to continuously inform a specific business unit function. Once established, this data pipeline will automatically pull in the appropriate raw data, transform it as necessary, and apply predefined and pre-established data analytics techniques to develop the end data product.

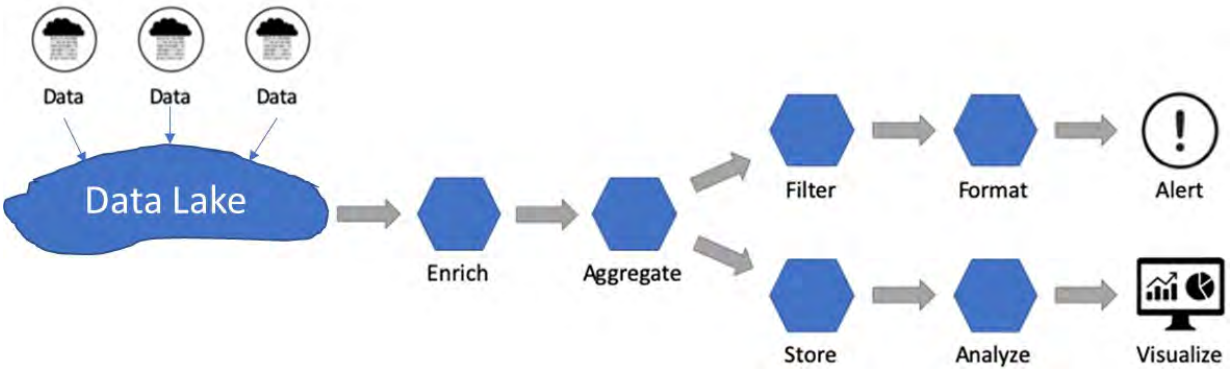


Figure 6. Example of a Modern Data Pipeline

When developing the analytical pipeline, the team should consider the use of:

- Open source data science friendly programming languages such as Java, Scala, Python, R, and Go as opposed to traditional programming languages such as C# or Visual Basic, as these support few data science libraries.
- Open source data science frameworks such as Pandas, Numpy, Matplotlib, Jupyter Notebooks, MapReduce, Scikit-learn, Tensorflow, and Keras as opposed to commercial relational database modules, as these are too costly, limited, and slow to evolve prototype pilot projects.
- Advanced data science algorithms such as stochastic gradient descent, random forest, Lasso, ElasticNet, NaiveBayes, and deep Learning as opposed to classic statistical techniques such as linear regression, logit, and least squares.

The project team should use whatever they can afford to build a pipeline that is of value. Once the pipeline is successful and capable of demonstrating a value product, the project team should prepare a "data story" combining qualitative insights and quantitative analyses designed to move people to action.

It should be noted that there is no need at this point to spend resources "above and beyond" those necessary in an attempt to perfect the data pipeline or the resulting end-products; the project is still at a prototype stage and these extra resources may not provide a good return on investment. Data science and cloud solutions are evolving constantly and at a rapid pace; as such, a prototype pipeline needs to be modified constantly to keep up with these advancements. Therefore a pipeline that is "good enough" is more resource efficient than a "perfect" data pipeline even at the production stage. Also, most modern data science pipelines are ephemeral – built, deployed, and run at start time and dismantled at stop time, leaving many opportunities to update them between runs without ever being perfected.

> *DelDOT found the cloud was more efficient both in processing speed and in the available pipeline tools. Even with experienced IT personnel on staff, developing and maintaining a large suite of data tools was less efficient than simply using the tool suite that the cloud vendor had already built, tested, secured, and optimized. One data process that used to take days for their on-premise hardware and software to process was completed in an afternoon on the cloud.*

Iteratively Develop/Improve the Project and Associated Outputs

The development of the project is likely to be more successful if it is treated as an iterative process with multiple feedback loops and revisions taking place throughout the development process. This development will also serve as a "test run" of a larger organization-wide evolution of data management practices, where challenges are encountered and overcome, and lessons are learned and documented by a small and dedicated team before they are later faced by the agency as a whole.

An iterative development process will serve as a "test run" of a larger organization-wide evolution of data management practices, where challenges are encountered and overcome, and lessons are learned and documented by a small and dedicated team before they are later faced by the agency as a whole.

While being able to obtain valuable results from the developed pipeline is important, the ability for these results to be useful to the rest of the organization is just as important and even essential to adoption of more modern data practices within the organization down the road.

As the project becomes more and more polished, it may be prudent to add additional analyses, both to make the project itself more successful and to experiment with different techniques. Many big data analyses require a relatively clean and complete dataset in order to be effective, so this is likely to be the first time they can be attempted. Applying such methods successfully will help the project display not only the benefit of modern data handling practices but also the benefit of modern data analyses that are only possible with big data. This is also a great time to experiment with additional data sources where applicable. Seek out additional data sources that can be used to augment the original data. If there are additional data sources that are unstructured, ill-formed, or otherwise difficult to work with, now is a good time to attempt to clean and process them. The most convincing arguments for adopting modern data management practices can be made by pointing to applications that are difficult or impossible to achieve without said practices. This stage of development is an ideal time to stretch the team's limits and attempt challenging data work, because the cost of failure is low and the rewards for success are high. The lessons learned and challenges overcome during this phase will improve the pilot project and future projects to come.

Also, putting analytic pipeline results back into the data playground storage layer and sharing them with others as a potential data source to solve new problems is crucial. This is how the organization will be able to quickly build an advanced data practice and adopt a more data-driven approach.

Case Study – Negotiating Technical Contracts for Data Services

The Los Angeles County Metropolitan Transportation Authority (LA Metro) initiated a Mobility on Demand (MOD) project to provide first mile/last mile transportation services. To develop this project, LA Metro partners with Via, a third-party, shared-ride transportation company. Under this partnership, Via not only provides shared rides to LA citizens, they also provide data services including collecting, storing, cleaning, preparing, and aggregating data associated with the service. Via also generates data visualizations and makes them available to LA Metro via an online dashboard. In order to make this project and partnership a success, LA Metro needed to overcome significant challenges with contract development and data sharing.

At the time the partnership agreement was negotiated, very few employees at LA Metro had any experience with highly technical IT contracts and had to work closely with legal counsel to fully understand and navigate all the details. Both LA Metro and Via wanted to maintain maximum control over the data while limiting liability. Via did not want to share raw data to avoid inadvertently disclosing trade secrets, while LA Metro did not want to be held liable for potential security breaches on systems over which they had no control. It took a considerable amount of time and effort, but LA Metro was able to successfully navigate this process and draft a mutually beneficial contract for both parties.

Data sharing was another issue that took time and attention to resolve. At the time the agreement was developed, LA Metro did not yet have any organization-wide internal data sharing policies. As such, they had to coordinate individual tasks with each data silo owner, many of whom had reservations about sharing their data with an outside company like Via. Talking to these stakeholders to individually address their concerns about data security, appropriate data use, and competent data analyses was time consuming but proved vital to the success of the project.

Case Study – Building Data Knowledge

In their transportation technology strategy document “Urban Mobility in a Digital Age” LADOT recognizes the need for modern approaches to use vast amounts of data, noting that “LADOT and other city departments must access and understand underlying data to make strategic decisions about prioritization. While the city generates large volumes of data, it lacks comprehensive, quality data to plan for all modes, evaluate existing programs, and understand how to adapt” (Hand, 2016). Since that document’s publication in August 2016, LADOT has worked to select projects that provide specific benefits while advancing their data management practices generally. One of these projects was a data inventory.

When working with the city council and mayor to fund broader transportation technology initiatives, LADOT found it useful to begin with a data inventory to assess how the city’s many departments and business units obtained, managed, and shared data. This inventory resulted in immediate benefits to the city by identifying areas where data efficiency could be improved, including business units that heretofore had been operating with no concept of internal data sharing. The inventory has also helped support the building and expansion of other data-related initiatives by the DOT.

One success story resulting from these projects has been the formalization of a Bureau of Transportation Technology. The growth of LADOT’s nascent Bureau of Transportation Technology is sustained by making efficient use of available resources and avoiding potential budgetary or administrative roadblocks. Six positions were moved from the IT department into the Bureau as a means of quickly onboarding skilled talent. When partnering with consultant services, the internal team prioritizes knowledge transfer to augment internal training and development.

Step 6. Demonstrate Value of Data to Other Business Units

Once the project has developed to a point where it generates real value for the business unit, it is time to share it with others outside of the business unit. The goal of Step 6 is to share the approach, outcomes, and value of the data, project, and resulting data product(s) to develop interest and buy-in on the project and approach from these groups. To begin to drive change within the organization, others need to know about the project. In Step 6 the champion(s) and team should:

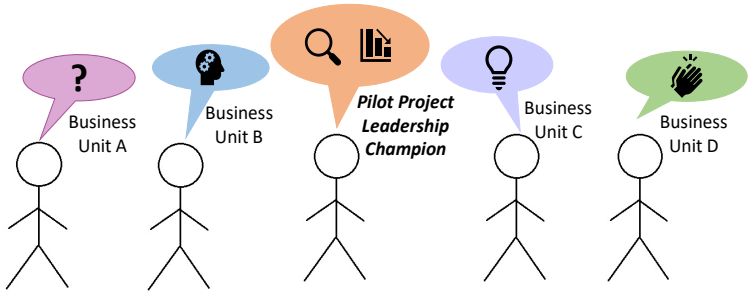
- Build support for the data and project horizontally within the organization.
- Use the data to tell the story of success.
- Get others involved in sharing and using their data within the test environment.

Build Support Horizontally

Depending on the organization, it is likely the leadership champion's role (with support from the project champion) to share and demonstrate the value of the pilot project to others within the organization. A good place to start is with other mid-level/branch managers that may have an interest in the data, project, and data products (or similar products) for their own business areas.

So that a larger potential group of stakeholders is reached in this step, volunteer to give presentations or demonstrations at various organizational events (e.g., coalition meetings, district meetings/conferences, statewide maintenance conference).

Be careful with sharing the project too soon with those at the executive level. Sharing the project vertically (Step 8) is likely to be more successful after multiple iterations of Steps 2-6 have generated more successful use cases.



Use the Data to Tell a Story

The data itself (and the resulting data products) are the best ammunition for selling the benefits of the data and the approach. Using the data and the resulting data products, the project team will need to craft a compelling story using understandable and persuasive visualizations that tie the insights uncovered in the data to the ability to address an issue or solve a problem of the business units.

When seeking buy-in from other business units, create a storyline for the project that includes the following:

- Introduction (characters and setting are introduced) – Explain who/what group within the organization is involved in the pilot project, their roles and functions, and their goals.
- Problem (main characters face a series of conflicts) – Detail the challenges and roadblocks currently keeping the characters from effectively/efficiently completing their functions and meeting their goals.
- Climax (the most exciting part of the story) – Describe the new/exciting data, the potential for the data to improve work functions and meet goals, how the data pipeline was established and what data were combined, how the data were assessed, what was discovered (about quality and potential uses of the data), new/exciting tools and analyses conducted, and challenges and how

they were overcome. Visualize the data to enhance the story. Make the visualizations beautiful and easy to understand.

- Resolution (events leading to the end of the story where the outcome is revealed) – Reveal the resulting data products and how they help to resolve the problem. State that big datasets, including messy and "dirty" data, and big data tools and analytics techniques can be used within the agency and that this approach adds value by providing insights that the agency simply would not have without the data and the approach.

> *DelDOT sees firsthand the value of the data they have been collecting and building. They realize now more than ever the importance of storing raw data and then integrating and analyzing these data to drive decisions rather than basing decisions on experience, conjecture, or gut feelings. They are now able to back-up decisions with real data and are helping others within the department, as well as partners outside the department, to better understand this approach.*

- Conclusion (the end of the story, judgement/decision reached) – Conclude with how these data and this approach have fundamentally changed the way the characters view data and how they will use data in the future. Deduce that the same approach could work for other business units (and the organization as a whole).

The results should be shared in two different ways. First is the compelling story that will be used to communicate to management and other business areas. Second is by sharing the results with others in the data playground environment so that they can be used as a data source for other analytics projects.

Get Others Involved in Sharing and Using their Data Within the Test Environment

A primary goal of this step is to generate interest in the data and the embryotic data environment that results in more data and additional use cases and pilot projects. This road map to big data represents an organic, bottom-up approach that relies on an iterative process to grow use cases, pilot projects, and value to the organization. With enough marketing of the project and data products, interest undoubtably will be generated in other groups in sharing their data and developing their own data products. The more people/groups that add data to the test environment and that interact with the data in this environment, the more use cases will be generated, and the closer the organization will be to shifting to a new culture of data awareness and use. Work with new users to gather feedback and identify areas where the data could be further cleaned or enhanced and use those enhancements to further extend the success of the project.

> *"We didn't force anyone to change as we implemented the project and explored the data within our test environment. We simply asked if we could copy others' data into our system. It was a very non-confrontational approach. Most people were flattered that we wanted their data and they're now impressed with the outcome."*
>
> – Kentucky Transportation Cabinet

It's important at this point to not try to force change. Let others see the value of the data and the environment and envision how they might develop a similar pilot project for their business unit.

Consider this an iterative process that seeks to build a stronger case for the data and the big data environment across the agency. Figure 7 illustrates this iterative process – after demonstrating the value of the pilot project to various business units and generating interest from one or more of these groups, loop back to Step 2 to identify new pilot projects (and associated data) specific to these groups. If necessary, secure buy-in from business unit leadership (note, however, that given the process, these new pilot projects might be more top-down initiatives). Once new use cases and pilot projects are identified, work with the other business units to develop their projects and data products within the data environment and subsequently share these results across the organization.

> *KYTC first implemented Elasticsearch – an open source tool for storing and discovering data – simply so users could browse the data. This helped with resource management and allowed users a safe, read-only, environment for accessing, interacting with, and reporting on data. The visualizations of the data also greatly helped to sell the idea of investing in big data to upper management.*

Figure 7. Iterative Process to Generating Interest and Buy-In Horizontally Across the Organization

Enough iterations will prepare the leadership champions to market the data, data products, and data environment to organizational executives in Step 7. Some agencies may find that they are ready to go to executive management after the first pilot project, while others may iterate through multiple projects before they are ready. This decision will be driven largely by the structure, culture, and relationships within each agency.

If the efforts put forth thus far fail to result in interest or support beyond the business unit, return to Step 2 and refine and expand on the current project and/or develop a new pilot project with data from other business units to better meets their needs and appeal to them. It may take several attempts before landing on a project that gains wider spread buy-in throughout the organization. Remember that ***the business unit can still derive value from the data and the data environment even if wider spread interest is not generated immediately or even after several iterations or projects***. With continued development of use cases and projects within the business unit, eventually, the team will have more to communicate, and others within the organization will catch on and want to learn more. Once it takes off, it will be impossible not to talk about it.

> *"Everyone who has interacted with our system has had something to say about it to someone."*
>
> – Kentucky Transportation Cabinet

Case Study – Iterative Success and Growth

For KYTC, the original proof-of-concept (pilot project) for big data was to develop a real-time snow and ice decision support system with the goal of making more efficient use of materials, trucks, and labor. The first iteration of the system was based on just three data sources – Doppler radar, snowplow AVL, and Waze. After demonstrating the ability to process a variety of data sources with different volumes and velocities, the project was considered a success and quickly gained the attention and support from leadership. Over time, as the platform evolved through the iterative process, additional data sources were added. Eventually, the system combined 11 additional data sources with the original three, all in real-time. These additional sources include data from: HERE Traffic, iCones, Twitter, KYMEsonet, CoCoRahs, TMC reports from two TMCs, RWIS, county activity reports from 120 counties, dynamic message signs, and truck parking.

The capabilities of the system and the data drew considerable attention from multiple divisions within KYTC including the Division of Environmental Analysis, Office of Highway Safety, Department of Motor Carriers, Division of Planning, Division of Traffic, Division of Maintenance, and Division of Construction. Each of these divisions now benefits from the data being collected and processed through the big data system in some manner:

- The Division of Environmental Analysis uses some of the data to help them report on environmental impacts to the roadway network as part of the MAP-21 requirements.
- The Intelligent Transportation Systems group uses the system for real-time crash detection across the entire roadway network, publishes information to Waze to help mitigate congestion due to road closures or other issues, and leverages big data to provide real-time monitoring of the roadway with 0.01-mile and 2-minute precision.
- The Office of Highway Safety can leverage the system to produce more precise after-action reviews of crashes.
- The Department of Motor Carriers now publishes real-time, post-processed data to mobile applications such as PrePass to inform the trucking community of roadway issues.
- The Division of Planning leverages the big data architecture to quickly run calculations on historic traffic data to measure the performance of the roadway network.

- The Division of Traffic uses the system to assist with signal optimization research in addition to providing data to an existing software vendor for dynamic signal timing.
- The Division of Maintenance fully leverages big data to provide a robust and comprehensive snow and ice decision support system.

KYTC was also able to replace the legacy, vendor-based 511 system with a new in-house traveler information system. By doing so, KYTC was able to eliminate a $750,000 per year contract. With the big data environment in place, this repurposing of the data and recreation of the 511 system required less than 200 hours of combined labor.

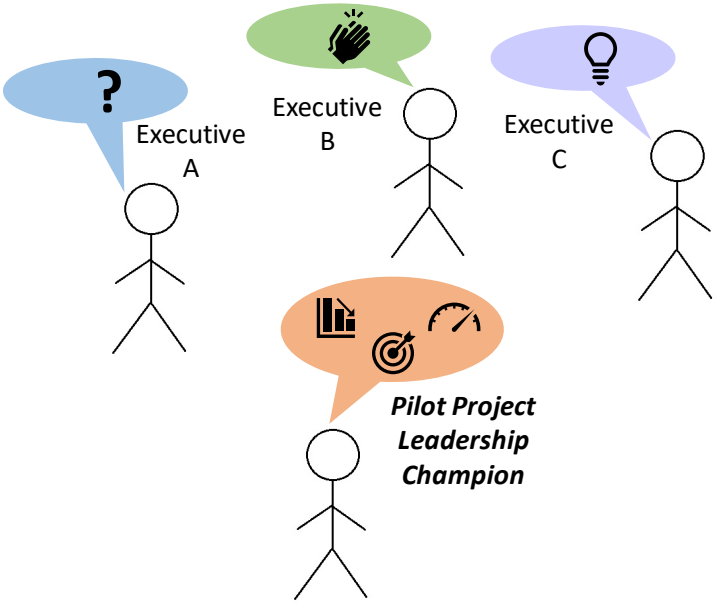
Step 7. Demonstrate Value of Data to Executive Leadership

After one or more iterations of Steps 2-6 and growing the number of use cases/pilot projects within the big data environment, it may be time to begin marketing the data, environment, results, outputs, and benefits to higher-level executives within the agency. While Step 7 is similar to Step 6 in many ways, the task of selling the data and approach vertically within the organization will likely come with a different set of challenges than selling them horizontally within the organization. In Step 7:

- Present the success stories/business case to executives
- Continue to build support, foster data sharing, and grow incrementally
- Push for organization change/adoption of a formal big data environment

Present the Success Stories / Business Case to Executives

The importance of communicating effectively with leadership in terms that they value cannot be overemphasized. Make certain that clear business needs are addressed within the first few minutes of the conversation or presentation. Avoid discussing specific technical details except where absolutely necessary; focus instead on presenting the project results in terms of measurable benefits, resources saved, or new capabilities/efficiencies gained. Create a logical link between the demonstrated benefits of the pilot(s) and the benefits of expanding them more widely across the organization.



Similar conversations/presentations should be sought with executives representing a variety of business areas. It should be anticipated that executives across the organization will have special needs or concerns when it comes to migrating their data from local data silos to a unified data lake architecture.

Taking the time to educate leaders about data access management and security within a data lake environment could prove to be invaluable in reducing hesitation and avoiding implementation delays.

Taking the time to educate leaders about data access management and security within a data lake environment could prove to be invaluable in reducing hesitation and avoiding implementation delays.

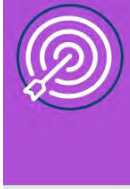
Each step of building the big data management capabilities to effectively handle data from emerging technologies has involved communicating concepts and educating colleagues. As the scope of effort expands from individual pilot projects to organizational change, many agencies will find it is no longer feasible to have a small team shoulder the burden of spreading knowledge. Unless the organization is very small, recruiting more data champions across as many teams within the organization as possible is strongly recommended. By delegating some of the education effort to these champions nearly every member of the organization can gain essential understanding without placing an unreasonable burden on any one team.

Common Misconception: Data Owners Have Less Control Over Their Data After Uploading it to a Data Lake

Business unit siloes, and associated data siloes, are and will continue to be a barrier to the ability of transportation agencies to leverage data from emerging technologies. Rigid processes around how data are managed and analyzed is a traditional approach to data management that was developed to facilitate fast query speeds and low error rates when working with relational database management systems. Today, this approach is outdated and cannot be applied to new, big data sources.

One common misconception that drives the traditional approach of "keeping the data close to the vest" is that the data owners will have less control over their data if they are moved into a common data lake environment. However, cloud-based storage follows the same concepts of user authentication and authorization as in traditional storage: one is no more "open" than the other. With cloud storage, the data owner can restrict the data to as many or as few users as they like. Unauthorized users can be prevented from accessing the data, or even seeing that the data exist.

In addition to the fear of losing control over data access, some data owners also believe that migrating data to a data lake will mean a loss of control over data quality. These data owners may have been the sole authoritative source of the data, or they may have business needs that require data be cleaned or manipulated in some particular way prior to being presented. Such data owners may mistakenly perceive that if their data are shared in a data lake, then other business units or analysts may interpret their data incorrectly or present conflicting results. Such a situation can be easily avoided by limiting access to data that may be prone to misinterpretation or misuse or only making the final version of that data visible to all users.



Continue to Build Support, Foster Data Sharing, and Grow Incrementally

As iterations of Steps 2-6 may be necessary to get to Step 7, further iterations involving Step 7 may also be required to gain buy-in from business unit or higher-level executives (as illustrated in Figure 8). A successful outcome of Step 7 (and any associated iterations) would be for an executive to declare that "This is the way we're going to go."

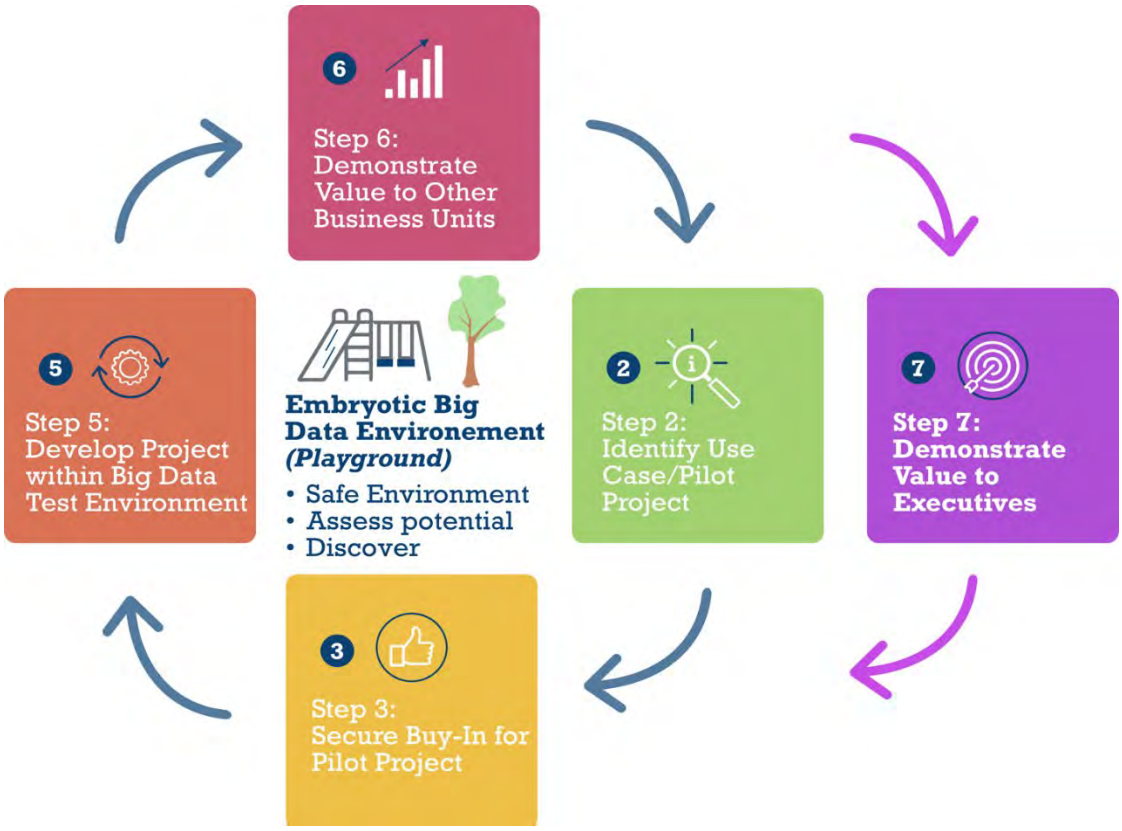


Figure 8. Iterative Process to Generating Interest and Buy-In Vertically within the Organization

Push for Organizational Change/Adoption of a Formal Big Data Environment

After what will likely be many iterations of Steps 2-7, there will come a point in which there are enough use cases, projects, and data users throughout multiple business units across the agency, and there will be enough recognition of the benefits of the data and the big data environment by leadership. This data sharing and use throughout the agency should support the claim that the organization is not only ready for change, but that this change is vital to continue to develop and support data-driven decision-making organization-wide. At this point, it is time to push for organizational change and adoption of a formal big data environment. This push ultimately will need to come from executive leadership as a top-down initiative within the agency. As necessary, arm these executives with as many success stories and business cases to support the argument that change is absolutely necessary for the long-term success of the agency.

Case Study – Buy-In from Executive Leadership

At KYTC, after the big data proof-of-concept system had proven itself useful for snow and ice operations and real-time monitoring for the Division of Maintenance and Intelligent Transportation Systems, the architecture and data gained additional attention and support from executive leadership.

With this additional executive support, the three-person big data team became the single point of contact within the agency for all things related to real-time data. The project was even officially renamed to portray a broader, agency-wide scope to increase acceptance from other business units. Executive support escalated to the point that all other development efforts were compared and contrasted against the big data architecture and development processes. In one specific case, a business unit within the agency had a project that included creating a real-time system for notifying a third-party application about roadway hazards. Once executive leadership learned of the specifics of the project, both the project team and the big data team were told to meet and determine where they could eliminate redundancies. Once the redundancy of the system was identified, the project team was told to use the existing big data architecture.

Another phenomenon that happened rather organically was that executive leadership started referring other divisions to consult with the big data team any time they perceived that the data or architecture may be of assistance to the different areas of the agency. Over time, the big data group became something of an internal consulting service to the other departments within the agency, thereby growing the influence and exposure of big data. One example of this was when the agency wanted to solve an issue of frequent crashes and congestion within a specific corridor. In addition to the traditional engineering team members, the big data team was tasked by leadership to develop an analysis of the corridor. In the end, the results obtained by the big data team greatly reinforced what the engineering staff already understood about the corridor, but the extent of the problems came at a surprise to the team.

Step 8. Establish a Formal Data Storage and Management Environment

Step 8 ends the focus on individual pilot projects and the test environment and begins the establishment of organization-wide big data management. An agency will likely arrive at Step 8 only after many iterations of Steps 2-7, as illustrated in Figure 9. The key focus of this step should be on scaling-up existing capabilities to serve a wider audience. There will be many possible new capabilities to develop or improvements to pursue, but it will be more effective to first expand the reach of the existing technology and proven data management approaches to avoid getting lost among the fine details. The more well-developed the initial data projects are, and the more the data playground reflects the planned data lake infrastructure, the easier this process will be.

The more well-developed the initial data projects are, and the more the data playground reflects the planned data lake infrastructure, the easier this process will be.

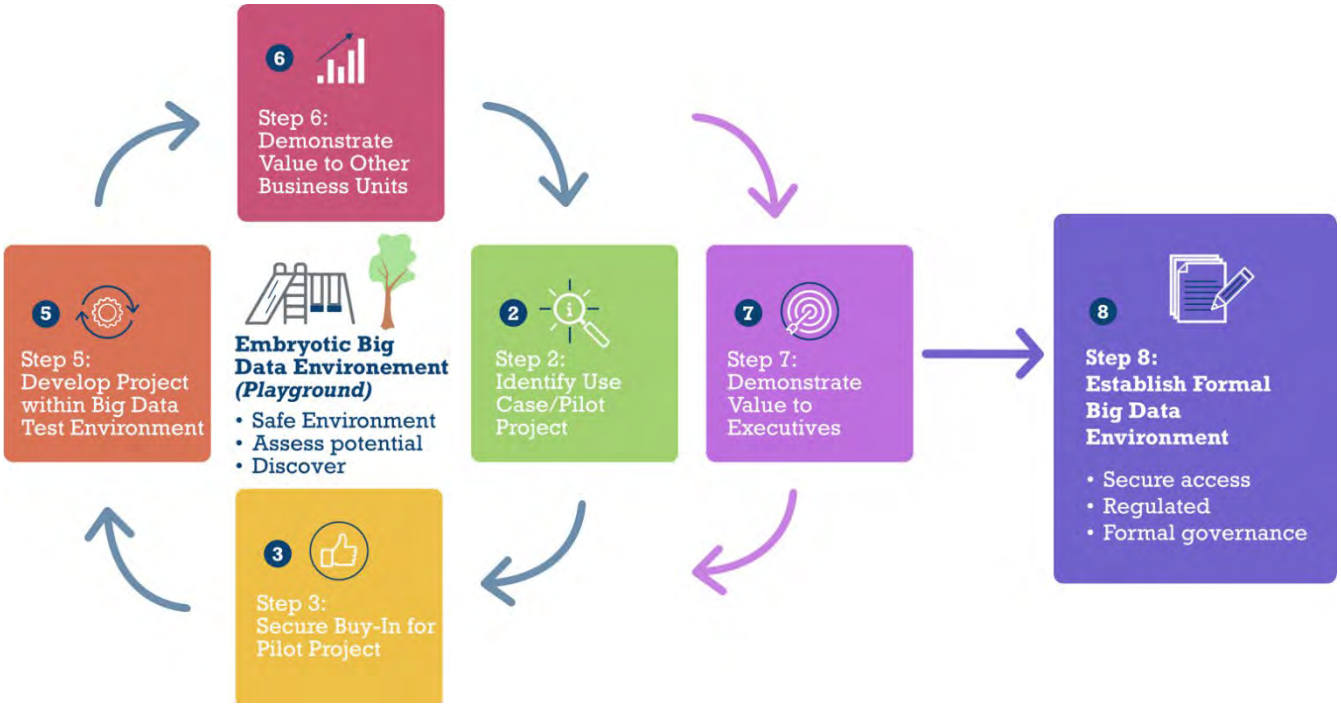


Figure 9. Iterative Process to Arrive at Step 8

To successfully complete Step 8, the agency will need to commit to a paradigm shift, a culture change, and a process capable of building on the data playground projects to shift the organization from opinion driven to data-driven decision-making. As was heard from a state DOT participant in this research, "Let's be honest, it's 'data-informed,' not 'data-driven.' We just use that term because people understand what it means. There's no way we'll ever let data make the decisions over a human because we can't always trust our data." However, as has been stated, the data are becoming too big for humans to process. Nevertheless, the shift will occur progressively as people within an organization begin to trust in the data, the process, and the novel data products being generated.

What is data-driven decision-making?

Progress in an activity is compelled by data not by intuition, personal experience, or political agenda. Data are now too large, fast, and change too quickly for the latter.

First, business areas that are the easiest to migrate and that boast the most supportive and enthusiastic teams should be shifted. Then, more difficult business areas should be progressively migrated. Slowly, as business areas are added incrementally, an organization-wide modern data environment supporting both pilots and production projects and associated modern data management policies should be developed.

This step is often where organization fall into a trap; they often go too fast and try to integrate the newly created data pipelines into their traditional IT infrastructure. This is very common and results in promising modern data projects being redesigned into more traditional ones in order to comply with the traditional IT policies. The result is that the progression towards modern data practices is completely stalled. Instead, organizations need to proceed with the transition slowly and progressively by growing the data playground, in parallel with the traditional data projects, into an organization-wide data environment that eventually will become the state of the practice in data management within an agency.

To achieve this, organizations need to focus on the following objectives:

1. **Establish clear vision and goals** – It is very important to develop a vision before introducing modern data practices agency-wide, and executives will need to develop and present this vision to the rest of the organization. In the vision statement, it will be vital to provide rational arguments for this change and shift in organizational culture, present the benefits and future plans, and resolve doubts.

2. **Make data accessible yet secure** – Data are at their most valuable when they are accurate, completely secure, and trusted; this has been achieved traditionally by applying data governance policies that limit access to the data. Yet a data-driven culture requires data openness, letting teams access data and consider new data-driven approaches. Executives will need to develop policies that allow employees across the organization to easily access and process data while at the same time keeping it under control and secure.

   > *A data-driven culture requires data openness, letting teams access data and consider new data-driven approaches.*

3. **Integration at the data level** – Traditionally, data integration has been accomplished at the IT infrastructure level based on predefined requirements. This approach is too rigid and costly to support the many rapidly changing data integration needs arising within a data-driven organization. Instead of adopting IT infrastructure level integration, executives need to support integration at the data level and develop governance policies that will allow data located in the shared modern data storage to be integrated by each business area as they see fit using the tools of their choice. By integrating at the data level, the organization will be able to adapt more quickly to changes in data sources, data tools, business goals, and objectives without requiring an expensive infrastructure expenditure each time.

   > *Executives need to support integration at the data level and develop governance policies that will allow data located in the shared modern data storage to be integrated by each business area as they see fit using the tools of their choice*

4. **Connect data to business goals** – Executive and middle management need to tie data to business goals by developing data based goals though the establishment of a set a valuable business key performance indicators (KPIs) that are satisfactory to the end-users and can be tracked and monitored to support the organization's internal processes.

5. **Use data to make decisions** – Finally, executives should lead by example and foster the use of data for decision-making within the organization. Most organizations are still hesitant to make data-driven decisions due to the lack of fidelity, granularity, and details found in traditional datasets and the risk associated with making the wrong decision. To create a data-centric culture in an agency, executives should design processes that support data-driven decisions.

Step 8 will encompass all four stages of the Modern Big Data Management Framework: *create*, *store*, *use*, and *share*. If not done already, it is strongly recommended to complete the accompanying Data Management Capability Maturity Self-Assessment (DM CMSA) before moving forward. Doing so will help an agency understand its strengths and weaknesses and will help to more effectively plan an expansion into a formal data storage and management environment. In addition to the DM CMSA, this guidebook

contains a description and framework for big data governance as well as a tool for tracking the big data governance roles and responsibilities within an agency(pages 106-110).

During the expansion, questions may be encountered that hadn't been previously considered, or some approaches that worked at a project level may be determined to be inadequate for data management at a larger scale. Do not be afraid to remain flexible with developing plans and processes even at this stage. Continue to seek input from other stakeholders and iterate on the evolving data governance plans and procedures.

If not already done, put a priority on merging existing projects into the same data infrastructure. This will help data users recognize this effort as a nascent organizational big data management effort and not simply a collection of pet projects. Delegating additional outreach efforts to identified data champions will also help fight this perception.

It is important to understand that even after the data playground has evolved into a production-ready data lake, and all associated projects have been adopted by relevant stakeholders across the organization, the efforts to evolve and improve these modern data management practices never truly end. New systems, datasets, and best practices will emerge that provide opportunities to improve and refine the approaches developed while following this roadmap. Below is a list of items that should be periodically reviewed and revised as part of continuous improvement efforts:

- **Currently used and potential datasets** – A catalog of available datasets that includes the contents of the dataset, the applications that use the data, and the potential opportunities for new users for the data will assist stakeholders in prioritizing their data development efforts and reduce costs. The accompanying Data Usability Assessment Tool can be used to assist with this process.
- **Currently used and potential technology** – A list of technology, including what the technology is used for and the costs involved, is intended to calculate the potential RoI of migrating from an aging piece of technology to a newer piece of technology. This could entail replacing a part of the data pipeline with a more efficient process or updating data analysis/visualization tools with more feature-rich options.
- **Processes and procedures** – Reviewing procedures includes both finding areas where more uniform processes would be helpful to reduce confusion and finding areas where over-regulation or red tape bureaucracy is hindering efficiency.
- **Documentation** – Identify what areas have missing or outdated documentation. One agency has adopted the use of wikis to handle this issue. Two different wikis are used for documentation: one for developers, where detailed technical documentation and emergency contact information exists and another for the entire organization, where users can find details about the different data sets and hints on how some of the data are being used already. This has greatly reduced the agency's overhead in maintaining that documentation by a few, select, people.
- **Security and privacy protection** – Identify new cybersecurity technology and techniques that have been developed, as well as what algorithms and methods are no longer secure. Ensure all security software is up to date. If working with a cloud provider who manages security, ensure they continue to meet contractual obligations.
- **Metadata catalog** – Find areas where additional metadata fields may be helpful. Review existing data to ensure that metadata enrichment is applied appropriately.

When reviewing these areas for continuous improvement, agencies may find it useful to complete the accompanying DM CMSA. This tool is designed to help identify areas in which an agency's data management approaches can be updated and improved. It can also be helpful to review the accompanying Modern Big Data Management Framework and compare current practices with best practices and recommendations listed in each stage of the data management lifecycle.

Common Misconception: Transportation Agencies Must Permanently Delete All Sensitive Data Immediately to Protect Themselves Against the Risks Involved with PII

Total avoidance of collecting or storing any personally identifiable information (PII) or other sensitive data is rarely necessary to protect an organization and doing so may cripple data analysis now and in the future. The modern approach is to instead preserve and secure the raw data first then anonymize or aggregate sensitive data as needed. These techniques are quite effective at removing sensitive data, and by storing the raw sensitive data on a separate server with restricted access, that data can be made very secure.

Raw data are the lifeblood of big data analytical techniques, so taking simple steps to preserve raw data is critical to modern data management. Even when using the most basic analytical techniques, it is important to have raw data to check data quality and investigate outliers. If a data analyst sees something unusual in the data, it is often impossible to confirm whether or not the anomalous data readings are legitimate if the raw data have been discarded. Due to both the value of raw data and the relative ease at which raw data can be preserved, it is almost always preferable to secure and use sensitive data rather than discarding or avoiding it.

Case Study – Continued Room for Growth

While many organizational units within KYTC have started to understand and experience the benefits of big data architecture, the system still has much room to grow before all the benefits can be fully realized. The democratization of data and end-user engagement remains a barrier to this day, and the reasons are complicated, as with any new technology adopted by any agency. One reason for slow adoption is simply the lack of time available to end-users to learn about new technology that may not be considered critical to that individual's day-to-day tasks. With public sector employees taking on additional responsibilities as staffing dwindles, it can be difficult to prioritize the need to learn something slightly outside of a worker's norm in the hopes that it may produce a gain. Another reason is simply the infrastructure, in terms of network bandwidth, available to users in remote offices and the tools they may have available to them. Using client-side business intelligence tools can be time consuming and frustrating to end-users connected to the network by slower than optimal network connections. But this also adds to the urgency of needing to adopt different architecture so agencies can move computing and reporting to lighter and faster cloud-based tools.

MODERN BIG DATA MANAGEMENT LIFECYCLE AND FRAMEWORK

This section of the guidebook presents a Modern Big Data Management Lifecycle and Framework. The lifecycle defines the four major components of managing data throughout their entire lifecycle. The framework builds from these data management components to include big data industry best practices as well as over 100 associated recommendations for those looking to implement modern data management practices and systems. A review of the framework will introduce the reader to modern data management principles, concepts, and specific recommendations for creating, storing, using, and sharing data from emerging technologies. These practices can be implemented following the Roadmap to Managing Data from Emerging Technologies, particularly in Steps 4-8.

Modern Data Management Lifecycle

According to the Data Management Association International (DAMA), data management is "the development and execution of architectures, policies, practices, and procedures that properly manage the *full data lifecycle* needs of an enterprise" (DAMA International, 2011; DAMA International, 2017).

The full big data lifecycle generally involves the following four phases (Figure 10):

- Create – The first time new data are observed, gathered, or created.
- Store – Writing collected data to secured, managed storage.
- Use – Performing analyses and developing data products.
- Share – Disseminating data to appropriate internal and external recipients.

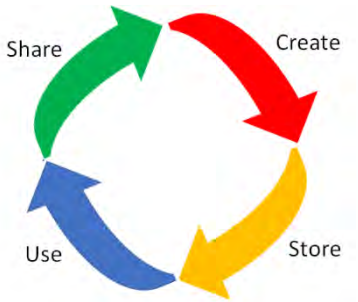


Figure 10. Big Data Lifecycle

Note that while some data management lifecycles include steps for certain data upkeep tasks, such as an "archive" step and/or a "purge/destroy" step, with big data, these steps have become more automated, less necessary, or otherwise require less focus from an organization. Therefore, while these tasks may still be performed under certain circumstances, they do not merit a full phase in the big data management lifecycle.

Modern Big Data Management Framework

The data management framework presented in this section provides recommendations that are based on big data industry best practices across the full data management lifecycle, including the creation of data, storage of data, use of data, and sharing of data.

Create

The creation of data could include new information generated from sensors, the discovery of a new internal dataset, access to a new external partner dataset, or the purchase of a new dataset from a

third-party provider. ***Correctly identifying the most appropriate data to acquire is one of the most vital first steps in building a practice to manage data from emerging technologies***, as these data form the foundation for all future projects, tools, and analyses. Figure 11 illustrates the most common data source types used by transportation agencies, which include:

- Raw data collected and controlled by the agency – includes both existing/traditional data (e.g., ITS devices, crash, asset) and data from emerging technologies such as connected vehicles and smart cities.
- Data obtained from third parties – includes data from vendors (e.g., AVL, advanced traffic management systems (ATMS)), partnership agreements (e.g., Waze Connected Citizen Program), crowdsourcing (e.g., HERE, INRIX), and social media platforms (e.g., Twitter, Facebook).
- Data processed at the edge – signal timing plans, signal traffic counts, or other raw DOT data that are processed at the edge instead of being sent to storage.

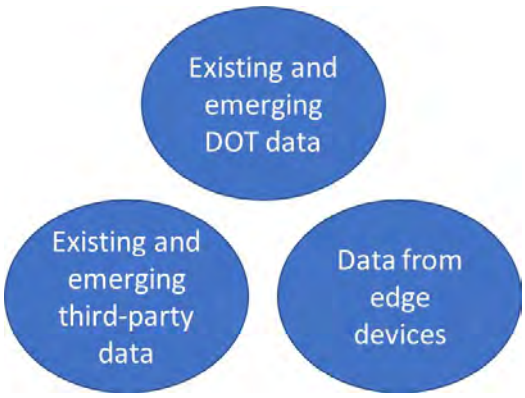


Figure 11. Most Common Data Source Types Used by Transportation Agencies

Following are recommendations associated with the collection of each of these types of data.

Recommendations for Managing Data within the "Create" Lifecycle Component

Recommendations are presented for the three most common data source types used by transportation agencies.

Data collected by the agency and originating from infrastructure owned and managed by the agency

➢ *Collect all data as they are generated, raw, and unaggregated. Do not discard data during collection.*

With the cost of data storage continuously getting lower, there is no reason not to keep all data, even outliers and erroneous data.[2] Following are a few reasons to do so in a context where data and data tools are varied and changing quickly:

- Data lineage is easily traced in any statistical analysis, as the data are in the same format they were when generated.
- Query design is no longer restricted to the specific data model that the data are being transformed and filtered to fit. Any type of data analysis is theoretically possible.
- Extensive time and resources spent planning the perfect ETL (extract-transform-load) process(es) and accounting for all possible source data variations and desired analysis is no longer needed and in fact is risky. Transformations can now be done at the time of query and can be corrected easily without affecting data collection.
- All results will be statistically significant, as no data will be omitted in the datasets available for analysis.

➢ *Don't limit the analytical potential of the data:*

[2] https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/

- Collect data in accessible, open (non-proprietary) file formats (easily interoperable with the most tools, e.g., CSV, JSON).
- Collect data even if they are sensitive – acquire knowledge on how to encrypt data soon after collection (e.g., National Institute of Standards and Technology AES 256 and SHA3).
- Learn how to anonymize the data in ways that protect the underlying usability of the data. Over-aggregating data, such as taking a rolling 3-minute average of connected vehicle data to obfuscate the movements of any one vehicle, will reduce the value of data and should be avoided where alternative methods of securing data are available.
- Collect "early" data from pilot/experiment/side projects to think about processes for later on.

➢ *Assess, tag, and monitor data as they are collected:*
- Assess data – determine format, size, costs, level of granularity, usability, openness, provenance, sensitivity and associated legal restrictions, etc.
- Identify data quality risks.
- Design data quality rules and metrics to flag identified risks.
- Deploy data quality rules and metrics to tag each incoming piece of data with its assessed level of quality and risk.
- Actively monitor data collection and quality in real-time using automated checks and alerts to identify issues rapidly.
- See the Data Usability Assessment Tool on page 112.

➢ *Ensure data collected are both technically and legally open. Avoid or resolve potential infrastructure software and hardware vendor lock restricting data usage:*
- Technically open data are data available in a machine-readable standard and open format, which means they can be retrieved and meaningfully processed by as many computer applications as possible.
- Legally open data, including publicly available data, are data explicitly licensed in a way that permits commercial and non-commercial use and reuse without restrictions.

➢ *Do not limit the collection of data to known or familiar data. Each business unit should be aware of what data are available outside of the unit. Investigate to understand if and how these data could support decision-making:*
- Create a regular process to review existing and potential data sets for business value – Assess to what level the data are open, ready, and exploitable.
- Consider data from other divisions within the agency.
- Consider data from other agencies and third parties.
- Establish standard procedures for creating new data pipelines.

➢ *Do not collect data with only selected users in mind:*
- Open and share data as a whole at no more than a reasonable reproduction cost to allow authorized users to reuse, redistribute, and intermix with other datasets
- Data should be available to any person, group, or field of endeavor with a genuine interest.

➢ *Maintain accurate data lineage for all pieces of collected data:*
- Create lineage metadata that uniquely describe where a datum originated, what happened to it, and where it has been as it was collected.
- Tag each incoming datum with the appropriate lineage metadata.

- Use lineage metadata in combination with data quality metrics to identify the source of data collection, issues, and subsequent corrective actions.

➢ *Do not segregate (i.e., silo) collected data. Apply the same collection approach to all incoming data using the same platform or system:*
- Use data lineage and quality tagging to distinguish immature data from production data.

➢ *When data are too sensitive to be collected and shared as-is, do not restrict them entirely:*
- Use anonymized, encrypted, or obfuscated versions of the original data to maintain most of their analytical value by enabling them to be still be shared with many albeit at a lower risk, and secure data at time of creation.

Data originating from third-party data providers where the transportation agency does not own or manage the infrastructure used to collect and process the data

Data collected from third parties are often not raw data, but proprietary data products generated for many different data consumers of which transportation agencies are often only one (and therefore they have little leverage). As there is often very limited insight into the lineage and quality of the data, it is necessary to assess, validate, and gain trust in it.

➢ *Do not rely solely on contractors/vendors to collect, aggregate, and provide data:*
- Strike a balance between maintaining control of data and the use of third-party data so as not to lose control over the data.
- Do not engage in agreements with partners, vendors, or service providers that severely limit access to actual data both internally and externally.
- Do not attempt to share ownership of the data.

➢ *Establish a clear understanding of the purpose, lineage, value, and limitations of the data product(s).*
- Develop a clear and concise understanding of what source data and processes are being used to generate the data products.
- Request raw data to validate the quality and value of the data products.

EXAMPLES

The Waze Connected Citizen Program is an example of a third-party data provider. While Waze is mainly focused on gathering data for advertising, it provides a custom data product to transportation agencies in exchange for information on roadway closures. Waze provides "reliability" and "confidence" information for reports, but calculation of these measures is not made clear.

CoCoRaHs is a provider of crowdsourced precipitation data collected by trained volunteers. The data are stored in a central repository and made available to the public through downloads and an API. While CoCoRaHs has established controls to ensure the quality of its data, it is still very dependent on the performance of volunteers.

Data providers such as INRIX, Here, and Cellint are extremely dependent on cellphone activity data from cell phone companies, and their accuracy is significantly dependent on the density of cellphones in an area and privacy law aggregation requirements.

➢ *Establish data quality rules and metrics for third-party data rather than relying on the quality metrics provided by the data provides (if provided at all):*

- Develop customized quality rules and metrics based on knowledge of how the data products are generated.
- Tag and monitor data product quality in real-time.
- Use data quality metrics to measure data providers' performance.
- Collect ancillary data to enrich data from other agencies (e.g., NOAA, law enforcement) or from external data providers (e.g., automotive manufacturers, supply chain managers).

➢ *Augment or customize third-party data products to allow better understanding of their quality and to establish contract clauses or communication channels with providers to fix potential issues:*
- Benchmark third-party data products against agency-collected sample data in selected areas.
- Compare third-party data products against data products from competing data providers.
- Design and seek agreement on lineage metadata to be added to the third-party data products.
- Develop and seek agreement on the quality metrics to be used to assess the data.
- Establish a formal process between the agency and the data provider to communicate, track, and correct data issues.

Data originating from infrastructure where, prior to being collected, they are processed at the edge by Internet of things (IoT) devices using machine learning algorithms

Collecting data from artificial intelligence (AI) based IoT edge devices requires attention and diligence. Machine learning algorithms have limitations – they are unalterable black boxes whose learning processes cannot be directly edited; the algorithms can only be retrained as a whole and completely replaced. Machine learning algorithms need to be retrained frequently as they deviate from normal behavior. One example is how a neural network learned to differentiate between dogs and wolves. It didn't learn the differences between dogs and wolves, but instead learned that wolves were more often shown on snow in pictures, while dogs were shown on grass. It learned to differentiate the two animals by looking at snow and grass. The network learned incorrectly. If the dog was on snow and the wolf was on grass, it would be wrong.[3]

➢ *Data coming from the edge devices are not the sole source of data for any particular purpose or application. Collect a sliding history of the last few minutes of raw data ingested by the edge device to help diagnose variations/abnormal behavior and improve edge device algorithms.*

➢ *Conduct edge device performance assessments using the collected raw data and edge devices, and audit edge device data regularly to measure the performance of the edge devices.*

➢ *Monitor the edge device data in real-time to detect slow drift or abnormal behavior rapidly.*

➢ *Adopt an edge device maintenance approach based on disposability to quickly replace devices as soon as they start to drift or act abnormally.*

Store

The "store" data lifecycle management component encompasses the management and use of data storage architecture to house existing and newly acquired datasets. Properly managed data are securely stored in an architecture built to support their individual formats and use cases while remaining

[3] https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982

scalable, resilient, and efficient. All data management and configuration that are performed on collected data to prepare it for future use falls under "store."

While many core concepts relating to the proper management of traditional data storage systems remain valid today, such storage systems and schemas are often insufficient to meet the needs of emerging data and modern machine learning applications. The capacity, scalability, structure, backup and recovery process, data quality management procedures, and oversight of data are all approached somewhat differently when managing big data. New architectural patterns need to be adopted to cope with the wide variety of fast-changing data that will need to be processed to guide decision-making. To fully capture the value of this extensive amount of data, transportation agencies need to develop a flexible and distributed data architecture capable of applying many analytical technologies to stored data of interest as opposed to the nearly impossible task of creating an all-inclusive data model capable of organizing each and every data element.

To fully capture the value of this extensive amount of data, transportation agencies need to develop a flexible and distributed data architecture capable of applying many analytical technologies to stored data of interest.

Few transportation agencies have developed to the point that they have metadata catalogs, database diagrams, or comprehensive data quality monitoring in place. Following are best practices within the big data industry for storing and managing big data.

Few transportation agencies have developed to the point that they have metadata catalogs, database diagrams, or comprehensive data quality monitoring in place.

- **A cloud-based, object, storage solution, also called "data lake," is used to store all data** – A cloud storage solution allows for data storage to be elastic and provisioned on demand. As such, modern data systems are flexible; they can adjust up or down based on a change in raw data (as opposed to rigid traditional storage systems that must be sized for a predetermined maximum storage capability).
- **All data are stored, both structured and unstructured data** – Data of any kind, be it text files, videos, documents, spreadsheets, audio files, etc., is stored. Cloud storage solutions are a kind of object storage. As such, they are meant to store large binary objects up to several terabytes each, not just numbers, characters, small text, or pictures typically allowed in a traditional database schema.
- **No filtering or transformation is imposed on the data prior to storing it; each user defines and performs their own filtering/transformations** – Applying a "schema last" or "schema on read" approach allows for each data user to define their own data model fitted to their own business needs on top of the raw data and update it at their own pace without having to compromise or compete with other data users.
- **Inexpensive cloud storage solutions are used for inactive data rather than performing traditional backups** – Rather than using a traditional approach of moving data out of data storage to dedicated archival storage, very low cost cloud storage, such as tape-based, cloud object storage, is used to store inactive data. This allows for the data to be recovered much

more quickly in case it is suddenly required by an end-user. Furthermore, most cloud storage solutions include an automated archiving system that automatically migrates unused data to low cost storage after an established period of time and allows end-users to recover it to faster storage when needed/requested.

- **Isolated cloud storage solutions are used if strong security requirements are needed** – Commercial cloud storage solutions, while fairly secure, are sometimes not used by agencies due to the fear of exposing sensitive data in a shared data center environment. Should such concerns exist, additional cloud storage solutions are available. These solutions have been spear-headed by the federal government, are built on strong authentication and encryption practices, and are hosted in dedicated data centers. Agencies can use the Federal Risk and Authorization Management Program (FedRAMP) certification to assess each cloud storage solution and the security level it provides.
- **Data are organized using the "regular file system" like structure offered by cloud-based object storage** – While modern data architecture forgoes the design of an all-inclusive data model to organize the data, it still relies on the basic folder like structure of cloud storage as a way to catalog the datasets. This folder like structure is essential to allow data users to find their way through the many stored datasets. Contrary to data models, this folder structure should not be designed to support potential data analysis but to support the discovery and understanding of the data and how it relates to business needs.
- **Raw data are augmented/enriched by adding metadata to each record to help end-users understand and use the data** – Modern data architecture does not filter out bad or incorrect data but still needs a way to help data users understand the value of the data they are using. To do so, modern data storage approaches add several columns or fields to raw data records to qualify them in terms of business areas (e.g., district code, responding unit number, data quality such as custom metric quality levels and data lineage or provenance such as location, sensor ID, sensor firmware version, etc.). While some of these record qualifiers, especially ones related to quality and provenance, can be added during collection, they are not sufficient to completely understand the data within each dataset; additional ones based on entire record populations and other relevant datasets need to be developed and added.
- **Folder structures, datasets, and access policies are managed to accommodate end-users' needs while maintaining the security and quality of the data** – While traditional data architecture approaches allow for data access and data use to be controlled at the record level in terms of reading and writing data, creating temporary tables and executing specific queries. Modern data architecture, except for a few specialized solutions, only controls data at the file and folder level, that is reading and writing files, creating folders, and deleting folders. The modern data architecture does not have the ability to control the use of data except by denying access to it altogether. Consequently, cloud storage folder structure and datasets are created to allow sensitive data to be denied from data users without blocking access to entire datasets by duplicating datasets without sensitive information in different folders or encrypting sensitive data inside the dataset. Access policies are established to strictly control what data users have

access to, what folder structure, and which users can create new folders and add new data within the structure.

- **Accessibility of the raw data is maximized by using open file formats and standards** – Following modern data architecture principles, files created in the cloud storage solutions do not restrict the potential use of the data and as such are as open and accessible as possible without imposing the use of a specific software solution to read the data, especially costly proprietary ones that some data users may not be able to afford. Open file formats relevant to the data being stored are used as much as possible over proprietary ones.
- **Data discoverability is maximized by maintaining a searchable metadata repository** – Without using a data model or schema to convey the structure of the stored data, modern data architecture approaches still need to maximize the ability for any data user to find, understand, and use any of the datasets they are allowed to access. To do so, the data stored in the cloud storage solution are described in detail in a searchable metadata repository accessible to all data users, allowing them to learn about each dataset and explore the possible ways they can relate to each other. This metadata repository is maintained dynamically and updated quickly as new datasets are created and deleted.
- **End-users' data access and use is monitored and controlled in real-time** – While traditional data architecture relied on a somewhat static data structure and established data access and use control, modern data architecture cannot do so as it deals with a much more lose and dynamic data structure where new folders and datasets can be created and destroyed rapidly. Therefore, data users' needs are maintained closely, access to datasets is monitored and checked against it in real-time, and unauthorized access is denied as quickly as possible.
- **Open file compression standards are used to limit storage space used** – When dealing with high frequency data, such as connected vehicle data, to be stored in cloud storage solutions, these data objects are compressed upon storage to save space and limit their impact on cloud costs. Several open file formats, such as Apache Parquet or Apache Avro, were designed for this purpose. They also offer reduced data scanning time, which greatly reduces the time it takes to scan entire datasets.

Recommendations for Managing Data within the "Store" Lifecycle Component

The following recommendations associated with the "store" data management lifecycle component are based on the big data industry best practices just reviewed.

Data architecture

- Use solutions that are built on common yet distributed architectures (i.e., cloud) and that possess good commercial support.
- Rely on cloud hosting services in combination with either cloud provider or open source data storage services to be able to quickly respond to fluctuations and changes in data storage needs without incurring excessive downtime and cost increase.
- Do not adopt commercial solutions that restrict the system's scalability and responsivity and its ability to keep data open.

- Follow a distributed architecture to allow for data processes to be developed, used, maintained, and discarded without affecting other processes on the system.

Data storage and operations

- Use a common (i.e., cloud) data storage environment (i.e., a big data environment); they are easily scalable and inexpensive.
- Don't store/use proprietary, on-premise storage systems; they no longer make economic sense when considering the scale and rapidity with which new data are being added and deleted.
- Store the data as-is (raw, unprocessed, uncleaned) and augment it into new and more usable datasets without altering the original data.
- Use common, open data file formats.
- Migrate datasets progressively, one-by-one, from agency silos to a data lake; transform data silos to make them interact with the data lake instead.
- Executive buy-in is a must for migrating silos to the cloud.
- Structure the data for analysis.
- Do not move the data out of storage; instead, move the data analysis software to where the data are stored.
- Use cryptographic hashes (i.e., an alphanumeric string generated by an algorithm from raw data that can be used to snapshot it upon storage in the common data store to ensure that the dataset has not been corrupted and/or manipulated).
- Do not modify or edit collected data; rather, create new datasets derived from the raw data to suit analytics needs.
- Manage and monitor in real-time users' data access and cloud resources usage across the entire system.
- Create logs, dashboards, and automated alerts to better track user activities across the entire system.
- Manage users' privileges using roles that can be assigned to each user and grant them privileges associated with each role.
- Setup thresholds and maxima for each user role to prevent abusive use of the data environment.

Data quality

- Do not delete or correct original/raw data of lower quality.
- Setup quality ratings methods and metrics for each dataset.
- Augment datasets by adding quality metrics for each record in a dataset, effectively rating each record, and allow for the same dataset to be used at multiple quality levels depending on the analysis performed.
- Leverage data crawling tools to continuously measure data quality trends across each stored dataset.
- Develop dashboards and alerts to better track and control overall data quality trends.
- Develop an environment where data quality is maintained not only by a governing entity but also by each and every data user allowing them to report or flag erroneous or defective data they encountered.

Data security

- Develop privacy protocols for the data considering the different data stakeholders (funding agencies, human subjects or entities, collaborators, etc.).

- Prior to distribution, remove or obfuscate/encrypt sensitive data that are not necessary for end-users.
- Learn how to encrypt/obfuscate (data masking), such as hashing techniques and encryption, to anonymize personal information, or hire third parties to perform and maintain encryptions and take responsibility over the security of the shared data.
- Create multiple versions of sensitive datasets with different levels of obfuscation/encryption to allow for analysis to still be performed on the data at different level of access.
- Outsource cybersecurity expertise and audit – less expensive to outsource cyber security experts (e.g., AWS).

Data integration and interoperability

- Use variable names within each usable dataset that are mapped to existing data standards (MMUCC, TMDD, etc.) to allow datasets to be joined together easily.
- Organize the data into logical folder structures following taxonomies and metadata relevant to the entire organization.

Data governance

- Focus on managing user access to data and user resources usage, not prescribing or enforcing what applications and language data analysts should use. Do not overregulate the use of data by imposing tools or platforms.
- Focus on exposing as much data as possible to as many users as possible by not using data access restrictions methods that arbitrarily limit or rigidify its possible exploitation.
- Maintain a flexible, appendable, and evolvable data governance that can adapt quickly to new data or unexpected data use.
- Allow users to use multiple tools so they can identify which ones will work best for their analyses based on their needs, resources, and knowledge.
- Focus on controlling, maintaining, and controlling data quality across every dataset.

Data modeling and design

- Augment datasets with metadata pertaining to data provenance and quality to maximize the data user perceived usability and understanding of each dataset.
- Adopt “continuous development” practices when managing/augmenting datasets to quickly modify or correct as usage changes rather than trying to optimize it from the time they are first generated.
- Make use of data masking techniques rather than removing data from datasets to solve data sensitivity issues.

Use

The “use” data management component includes the actual analyses performed on the data and the development of other data products such as tools, reports, dashboards, visualizations, and software. Proper management of this process includes educating end-users on how best to derive decisions from the data, using effective software development cycles to create new data products, and supporting

architecture that allows data to be effectively analyzed where it is stored without unnecessary computational overhead.

All interactions with the data by end-users, analysts, or software programs made to gain some insight or drive some business process fall under "use." Big data and data from emerging technologies can generally be used in support of traditional analyses such as spreadsheets and reports; however, the true value of such data is often found using new analytical techniques such as text analysis, clustering techniques, predictive models, and deep learning applications.

Traditional data analysis is heavily reliant on the traditional data system architecture and its approach of shaping stored data to fit predetermined analyses. Traditional data systems are optimized for a specific data model, which converts raw data to structured data, removing the fuzziness and outliers and rigidly organizing it using predetermined relationships between each data element. Traditional data analyses, whether a filtering and aggregation designed using SQL queries or a statistical analysis such as linear regression or probability distribution, can be performed either by a statistical software or by a relational database using structured data extracts exported from the relational database. This approach to data analysis has been the standard for more than 30 years supporting real-time analytics, also called OnLine Transactional Processing (OLTP), and historical analytics, also called OnLine Analytical Processing (OLAP).

There is value in both types of analyses. For OLTP (real-time analytics), the value of a single datum is very high immediately after it has been created – a quick analysis can lead to immediate corrective or augmentative action(s). As the datum ages, however, analyses supporting immediate actions is less valuable. Conversely, for OLAP (historical analytics), the value of data in aggregate is very low immediately after creation of the first datum. Small datasets composed of recent data are of less value; however, as more data are created over time, they accumulate into larger and more diverse datasets that may be analyzed effectively to reveal patterns and trends that can be acted on.

As more and more data became available and the need for more detailed analytics emerged, traditional data systems became more and more complex and costly to develop. As such, data warehouses, which combine and coordinate multiple traditional data systems, were created to cope with the increasing size and complexity of the data. But RDBMS and data warehouses were able to handle real-time analytics from the millisecond range and historical data analytics up to 3-5 years before they became too costly to operate and too rigid to maintain.

Moreover, such traditional systems were often proprietary, and data analysts were dependent on vendors upgrading their software or creating additional components supporting additional algorithms before they could perform new analytics.

In the 2000s, companies like Google and Yahoo were trying to index the entire Internet. When faced with the sheer volume of data from websites to be index and the overwhelming frequency with which each of these websites was updated, they quickly realized the limits of traditional data systems. This was the beginning of modern data system architecture capable of handling much larger and detailed datasets and encountering millions of data changes per second. The first data system created, called Hadoop, was designed to run on a large group of servers on which it distributed large-scale historical data analytics. In the following years, Hadoop was the base model for new data analytics tools capable of handling an ever-increasing amount of rapidly changing data more efficiently and at a lesser cost.

The data environment that agencies face when dealing with data from emerging technologies is no different than the one that Google and Yahoo faced in 2001. Agency data that will need to be searched and monitored will include (but not be limited to):

- Millions of connected and automated vehicles, which are predicted to produce as much as 25 gigabytes per hour at a frequency of no less than 50 milliseconds
- Crowdsourced data such as Waze and Twitter generating hundreds of millions to billions of records a year reaching many terabytes to petabytes in size
- Numerous and ever-increasing smart cities data sources ranging from traditional utility, transit and police data to new IoT technologies data.

None of the traditional data systems currently in use or being developed by transportation agencies will be able to handle these analyses; modern data analysis approaches will need to be adopted and become central to each level of agency decision-making.

None of the traditional data systems currently in use or being developed by transportation agencies will be able to handle these analyses; modern data analysis approaches will need to be adopted and become central to each level of agency decision-making.

Following are recommendations for analyzing and managing big data within the "use" data management lifecycle component based on best practices within the big data industry.

Recommendations for Managing Data within the "Use" Lifecycle Component

The following are recommendations for analyzing and managing data within the "use" lifecycle component. Each recommendation is described in more detail following this list:

- Adopt a distributed approach to data processing using cloud infrastructure to benefit from abundant and low cost computing power.
- Do not dictate or limit the deployment or use of data analytics tools; use many, varied tools to meet the needs of individual business areas.
- Move data tools to where the data reside, because data are now too large to be moved around to specialized data processing environments; process the data where they are stored.
- Make data users responsible for the development, deployment, maintenance, and retirement of their data pipelines.
- Make each business area responsible for developing their own custom ETL processes.
- Make data accuracy and quality of the analytics processes and products the responsibility of the business area that develops them.
- Delegate analytics control to analysts within each business unit and make them responsible as the owner of their data analyses and products.
- Closely monitor the evolution of data analysis pipeline products to ensure reliability and quality as they evolve.
- Understand the nature and limitations of modern data analysis algorithms.
- Use open source software or proprietary cloud-based software as a service (SaaS).
- Do not impose analytics solutions and resources limits on the analysts upon design.
- Control data analysis activities at the data level not at the software level.

- Make room within the cloud storage environment for experiments, trials, and pilots to test and discover the most appropriate solution.

- **Adopt a distributed approach to data processing using cloud infrastructure to benefit from abundant and low cost computing power** – Except for very performant relational databases and supercomputer environments, traditional data analyses are done with the range of memory available on a single server and within the computing power provided by its CPUs. Traditional data analyses are optimized to leverage these resources. Modern data analyses, on the other hand, are performed in a distributed fashion, processing data where they are stored across multiple servers using a method called "parallel computing," which is similar to the ones used in super computing. Therefore, traditional algorithms no longer work in modern data systems, and new distributed and parallel algorithms are used instead to perform tasks such as aggregation, filtering, linear regression, etc.

  > *Modern data analyses are performed in a distributed fashion, processing data where they are stored across multiple servers using a method called "parallel computing."*

  - Use distributed algorithms and data analysis tools when processing very large amounts of data.
  - Use distributed data stream analysis tools when processing very large amounts of streaming data in real-time.

- **Do not dictate or limit the deployment or use of data analytics tools; use many, varied tools to meet the needs of individual business areas** – Indeed, the sheer volume, variety, and velocity of data to be analyzed are too ample for data to be analyzed using one or only a few tools. To satisfy the needs of each business area of each agency, data from emerging technologies require many different kinds of analytics from classification of images, to the detection of patterns in video feeds, to the mining of topics and words in social media, to the discovery of outliers in traffic operations. These analyses vary widely in terms of time and resource requirements and are not limited by a predetermined consensus on what resources should be available for analysis but left to business areas to determine which analytics tools best satisfy their analytical needs with their means. As such, do not dictate what tools should be used by data users to build their data analysis pipelines; let each data user define what tools are best suited for their analyses based on their data, resources, and knowledge. In addition, allow software to be deployed, used, and modified at will across the many servers at no cost and with no restriction.

  > *Data from emerging technologies require many different kinds of analytics. As such, do not dictate what tools should be used by data users to build their data analysis pipelines.*

- **Move data tools to where the data reside, because data are now too large to be moved around to specialized data processing environments; process the data where they are stored** – In a traditional system, data extracts are created to be loaded on statistical software to perform analyses on many months or years of data. This is no longer possible using modern data systems; data are now so large that it is very costly to move the data to different data analysis environments to be processed. Instead, data analysis processes are brought to the data where they reside, and the

outputs are saved at the same location to avoid the additional cost of moving analysis results across servers.

- **Make each business area responsible for developing their own custom ETL processes** – As already mentioned, in a modern data system, data are stored raw, unaltered, and possibly augmented with quality and provenance metrics. It is therefore the responsibility of each business area to develop its own ETL process for every analysis developed. The IT department should not be in charge of preparing and maintaining epurated (i.e., purified) datasets for analyses; rather, the business units should learn how to develop and maintain their own.
- **Delegate analytics control to analysts within each business unit and make them responsible as the owner of their data analyses and products** – In a traditional data system, once developed and tested, data analyses and products are often no longer under the control and responsibility of the developer but under the authority of the IT division that oversees and maintains the servers on which they run. This authority often is the sole responsible party for the quality and accuracy of the data products and the sole authority in allowing improvements or additions to the existing data analysis. In modern data systems, this approach quickly becomes overwhelming and impossible to manage when dealing with the specificities and needs of each custom data analysis. Instead, an approach of making the creator of the data analysis responsible for the maintenance and operation in production, leaving to each business unit the responsibility of creating accurate data analysis and usable data products.

> *Given the very large amounts of data produced by emerging technologies and the many different types of analyses, tools, and data products that will be desired by users, each business unit should be made responsible and accountable for their own ETL processes, analyses, data products, and quality control of the analytics processes. These can no longer be managed by a central entity.*

- **Make data accuracy and quality of the analytics processes and products the responsibility of the business area that develops them** – As many different data analyses are developed using many different tools over a large amount of data, no central organization can reasonably be responsible for each and every one of them. Rather, each of the business areas in need of data analytics develops the analytics process, becomes the owner, and is responsible for the quality and accuracy of the process/analytics from development to retirement.
- **Closely monitor the evolution of data analysis pipeline products to ensure reliability and quality as they evolve** – Data analytics pipelines developed in cloud environments are not as rigid and static as their traditional counterparts. They can change frequently under the influence of cloud services updates, new data, newer and better software solutions, or simply a change request its end-users.
- **Understand the nature and limitations of modern data analysis algorithms** – Traditional data algorithms were designed to derive as much value as possible from a small amount of high quality data following precise mathematical steps. This is not the case for modern data analysis algorithms. These algorithms have been designed to take advantage of a large amounts of noisy, unfiltered data, and they often rely on a brute force approach enabled by inexpensive cloud computing power. Because of the nature of the algorithms, the results are often approximated. As such, their results

are susceptible to variations and disruptions not commonly seen in the traditional data analysis approach; therefore, results are carefully reviewed and monitored.

- **Use containerization and microservices to develop custom data analyses** – Traditional data analysis algorithms are intrinsically linked to the database or statistical software on which they are built. Often, they are programmed using a specific language or software development kit provided by the companies providing the software. This allows all traditional analyses to use the same underlying libraries and configurations and avoid potential conflicts between different analyses. In the context of modern data analysis, this is no longer doable, as each analysis developed on top of the data is often customized and does not necessarily use the same underlying libraries and configurations as the others. In a traditional data system, this would typically lead to a server configuration nightmare. In modern data systems, containerization and microservices – distributed virtual data applications that are instantiated (created as an instance as a process on a computer), configured, and run just for the time of the analysis – are used to avoid this issue by effectively creating a custom compute environment for each analysis on the fly. Design data processing pipelines using a serverless or containerized design approach so that analytical stacks configuration and deployment can be reduced to a simple script that can be modified, redeployed, and tested easily.
- **Adopt a more iterative and interactive approach to the development of analytical products, building off existing analytical processes and visualizations rather than creating them from scratch** – Do not reinvent the wheel. Collect the analytical results of data analyses using the process used for external data sources, storing them in the common data storage, and augmenting them with provenance and quality metadata. Leverage, adapt, and reuse the analytics already developed and shared by others and benefit from the support of a much larger community of experts that is within the agency or even beyond the transportation community.

Leverage, adapt, and reuse the analytics already developed and shared by others and benefit from the support of a much larger community of experts that is within the agency or even beyond the transportation community.

- **Understand the ephemeral nature of modern data analyses** – Adopt "continuous development" practices when managing analyses and visualizations that can be quickly modified/corrected (as opposed to slow deployment of perfect products). Traditional data analyses often attempt to perfect data analysis queries and algorithms to fit the data at their best. In the context of modern data analyses, data change fast. As such, attempting to perfect algorithms is often a waste of time, as the data are short lived. Rather, a "good enough" approach to the development of analytics is used. This approach does not attempt to develop analytical queries or algorithms too far, because the results are sufficient to support decision-making. Further, these queries/algorithms are updated often to follow the fast-changing nature of the underlying data. This process is called "continuous integration" and supposes that analytical processes are constantly monitored for quality and performance and are updated as soon as their performance declines.

- **Use open source software or proprietary cloud-based software as a service (SaaS)** – In traditional data systems, proprietary software solutions are often the guarantee of a robust hardware and software combination capable of performing analyses efficiently and reliably up to the limit of the system resources. In the context of modern data analyses and the use of cloud computing, proprietary solutions are rarely used and can in fact be disadvantageous. Instead, cloud provider services or open source solutions are the preferred choice. Indeed, modern data analyses often rely on a large quantity of servers for a limited amount of time to support historical data analyses or surges in streaming data created by special events.
  The proprietary license model often follows a per server license scheme, which can have serious cost consequences as the number of servers licensed must be able to support the maximum level needed for surges. Instead, use pay-as-you-go / pay-for-what-you-use open source or proprietary cloud-based SaaS to limit the cost incurred when large amounts of data need to be analyzed for just a short period of time.

  Use pay-as-you-go / pay-for-what-you-use open source or proprietary cloud-based software as a service (SaaS) to limit the cost incurred when large amounts of data need to be analyzed for just a short period of time.

- **Do not impose analytics solutions and resources limits on the analysts upon design** – In traditional data systems, stability and order are often maintained by tightly restricting the type of software or languages than can be used to develop data analyses and specifying or allocating a maximum amount of resources or priority with which the data analyses can be run. This is no longer needed in modern data systems, as there are enough resources on the cloud to support all data analyses, and each of them is separate, customized, and contained. Instead, data analysts are given ways to independently control and manage their use of cloud resources, and they are alerted or stopped when they are exceeding these resources.

- **Control data analysis activities at the data level not at the software level** – Traditional data systems attempt to control the quality and reliability of data analyses and products at the design and deployment level, focusing on how the analyses are performed and how much resources are allocated for them to be performed. In modern data systems, without such ways to control data analyses and products, data management adopts a different approach to controlling the outputs of the many data analysis processes running in the data environment. This new approach focuses on data analyses to identify and detect bad data products and inefficient data analysis processes. Two types of datasets are used to develop this approach: the data generated by each analysis and the cloud activity data generated by each analysis. This approach is typical of a system-of-systems and treats each data analysis process as a black-box that can be observed and monitored using its output and resources use. This is a radical change from the traditional approach and will require extensive development of real-time data analyses capable of identifying deviating outputs and abnormal resource usage. Fortunately, cloud services providers have already designed services capable of monitoring cloud processes in real-time, and they can be customized to fit the monitoring of specific data processes. Also, it is relatively common not to approach data product quality assessment from the sole point of view of data management but to engage and involve data users familiar with the

data or domain of the analysis and solicitate their feedback to more effectively design and detect deviation in data products quality.

- **Make room within the cloud storage environment for experiments, trials, and pilots to test and discover the most appropriate solution** – In traditional data systems, experiments and trials are often conducted on separate systems with smaller datasets so as not to compromise the stability of the production data system. Upon successful testing, the data/analyses are migrated and integrated to production. Modern data systems are a combination of independently run data analyses that do not affect each other. Therefore, once an agency has fully migrated to an agency-wide big data environment, experiments and trials should not be kept separate; they should be developed directly in the same cloud environment that supports production data analyses, and they use the same data by creating a dedicated space within the data store where their results can be stored.

> *Modern data systems are a combination of independently run data analyses that do not affect each other.*

Share

The "share" data lifecycle management component involves disseminating data, analytics, and data products to all appropriate internal and external users. This includes creating an open data policy where appropriate, maintaining updated documentation and other content support, and providing some means by which authorized users may easily access relevant data products. Efficient management of this component balances the desire to provide the most use out of the data as possible with concerns over safeguarding privacy, ensuring security, and limiting liability.

> *Sharing includes creating an open data policy where appropriate, maintaining updated documentation and other content support, and providing some means by which authorized users may easily access relevant data products.*

Sharing data from emerging technologies brings its own challenges and risks. The volume of data involved makes sharing datasets more difficult, and the results of big data analytics require some effort to understand and summarize compared to traditional data reports. Sophisticated machine learning techniques enable the extraction of identifiable information from large combined datasets in ways that were impossible previously, requiring additional caution and care when preparing datasets for public use. When these modern challenges are managed effectively, however, sharing data analyses and receiving validation of their conclusions from external sources could provide valuable benefits to transportation agencies and the public users they serve.

The World Business Council for Sustainable Development (WBCSD) recently published a report titled, Enabling Data Sharing: Emerging Principles for Transforming Urban Mobility (Chitkara, Deloison, Kelkar, Pandey, & Pankratz, 2020). Working in partnership with a range of mobility stakeholders including auto manufacturers, operators, and industry experts, WBCSD identified five principles, shown in Figure 12, for data sharing as best practice.

In the traditional data system approach, the sharing of data products is rather limited. The sharing of data analysis processes is reserved to an even smaller set of advanced data users. Data products are often shared through access to dashboards, delivery of reports through email or PDF attachments,

delivery of email or SMS alerts, or just plain data export from a database provided in an Excel or CSV format. The latter offers the recipient the ability to reuse the exported data by further processing it using a spreadsheet, database, or statistical software to create new data products. When sh aring is required on a regular basis, traditional systems employ a machine-readable web interface, also called an Application Programming Interface or API, to quickly share or export predefined data products from the main database each time a web request is submitted from a browser or other program. Typically, these APIs are development following the strict eXtensible Markup Language (XML), Simple Object Access Protocol (SOAP), and Web Service Description Language (WSDL) standards and protocols to tightly control what requests are made, how they should be formatted, what should be returned, and how they should be formatted, leaving very little room for atypical requests or a response format that may be preferred by the end-user.

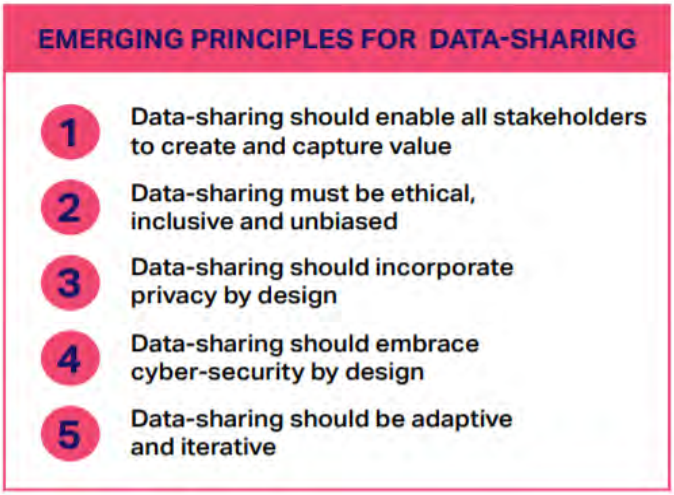


Figure 12. Emerging Principles for Data Sharing (Chitkara, Deloison, Kelkar, Pandey, & Pankratz, 2020)

These sharing methods are in line with the design philosophy of the traditional data system architecture. Data sharing needs are first studied and defined, then developed using the language and tools provided by the data system platform vendor and deployed in a way that is not too taxing on the limited amount of resources available in the system to preserve the stability and reliability of the system.

> *In an environment where many agencies are losing leverage in their data partnerships, the city of Portland is using its data to drive better behavior from partners. The city provides high quality location data to scooter companies that operate within the city. Scooter rental companies connect to Portland's data feed API to help their software navigate trips. In turn, Portland asks them to prevent rented scooter trips from beginning or ending in public parks. To support this request, the city has added scripting to its API connections so that if the scooter company software requests location data for trips starting or ending in a park, the city's data feed will not provide it. By providing a valuable data service that companies rely on for their business, the city of Portland has gained an effective tool in guiding partner company behavior.*

Traditional data systems can also be shared directly with end-users to allow them to run queries and analyses on top of their data models. This is often done using standardized protocols such as the Open DataBase Connectivity (ODBC) or the Java DataBase connectivity (JDBC), which allow end-users to connect various software clients or programs to traditional data system databases. This type of sharing practice is very limited, and it is often only granted to trusted data users and bounded with strict usage limits such as table access restrictions, read-only access, maximum query time, and created data size limits to avoid the risk of taking too many resources from the system and possibly corrupting the data or processing capabilities. While this is the highest level with which traditional data systems can share their data and

data analysis processes with external users, it is also non-trivial, as the end-users granted such access must understand two things prior to being able to efficiently process the available data using the system. First, the user needs an understanding of the database model used in the data system to make sense of the abbreviations, categories, units, and relationships it uses to inform the data. Second, the user needs to be familiar with the peculiarities of developing data analyses on the vendor platform on which the data system was built. Indeed, despite supporting standards such as Structured Query Language (SQL), most vendors mix non-standard modules and extensions into their products to provide their customers with easier or better performing services and to retain them as customers.

Modern data systems, on the other hand, are not developed with the same constraints. Storage and computing resources on cloud infrastructure are plentiful and inexpensive, and data are not modeled for predefined analyses and can be understood easily by looking at the metadata catalog. Yet, modern data systems offer a very similar way of sharing data products as traditional systems, with some caveats. Data products such as dashboards and reports created and maintained by individual business units can be used once access is granted to a user; however, access to these dashboards/reports is often not under the same usage restrictions as with traditional systems, which limit the number of users out of necessity. Because of the low cost of data compute and the parallel processing capabilities of the cloud, modern data systems can support many more users (e.g., tens of thousands); however, these users must be managed, which is not a trivial task. Data products can also be in the form of data tables or hierarchical data documents available as downloads directly from cloud storage or through an API running on the cloud infrastructure. As opposed to traditional data systems, modern data systems do not use the SOAP standards for the development of APIs, as these standards are too strict to handle the fast changes occurring in large data environments. Instead, APIs are designed using the Representative State Transfer (REST) protocol, which is simpler, more flexible, faster, and less expensive than SOAP web services. Modern data systems also shy away from the XML standard when exchanging data through an API, as in the context of large varied datasets, it is too strict, difficult to change, too verbose, and too slow to parse. Instead, data are exchanged using formats such as Comma Separated Format (CSV) or JavaScript Object Notation (JSON), which are much more flexible file formats that are easily read by humans.

> *Modern data systems use different standards for the development of APIs and the exchange of data through APIs. The REST protocol, which is simpler, more flexible, and less expensive than SOAP is preferred for API development. CSV and JSON, which are more flexible and more easily read by humans than XML, are preferred for data exchange.*

When considering the ability for external users to query or process data to create their own data products from data stored and managed on modern data systems, cloud infrastructure offers new possibilities. As data storage and compute are two distinct services in a cloud environment, external users can be given access to large datasets and process them at their own expense on the same cloud infrastructure using the tools they choose. Whether small or large, real-time or historical, the analyses desired by external users do not compete for resources within the production services. For example, researchers at a local university would be able to request access to a data use project or grant money to perform various analyses of the store data without ever having to extract the data or move them to a server inside the university. This is of particular importance when dealing with very large and fast-

changing datasets that cannot be easily moved across institutions. In this fashion, some datasets can even be made public without incurring any additional cost than storage itself.

Overall, modern data system architecture favors an open approach to data sharing with the understanding that a few sets of eyes will not suffice to extract value and intelligence from large and complex datasets. Rather, ***gathering the inputs and insights from many eyes from other agency divisions, universities, and even the public, can help agencies understand and successfully derive value from the data***.

> *A few sets of eyes will not suffice to extract value and intelligence from large and complex datasets.*

Recommendations for Managing Data within the "Share" Lifecycle Component

The following are recommendations within the "share" lifecycle component. Each recommendation is described in more detail following this list:

- Share data extracts, data APIs, or even entire large datasets with external users by giving them access to the data directly on the cloud environment where they can search and analyze data at their own cost.
- Create data documentation that can be easily viewed by internal and external users and as much as possible automate documentation update when data changes.
- Do not use proprietary file formats or API protocols when sharing data; use open web APIs to allow the shared data to be used by a large number of data tools rather than just the ones of a single vendor.
- Be transparent with those with whom you share data and provide information that helps them get the most out of the data.
- Do not be concerned with corrupted data but flag it when discovered.
- Encrypt, obfuscate, or remove sensitive data prior to sharing.
- For the best protection, frequently update encryption methods used to obfuscate sensitive data.
- Create different versions of datasets based on who they need to be shared with.
- Beyond sharing data, also share data analysis process code.
- Establish live data streams in addition to sharing historical data.
- Share the responsibility of providing data services and products to external users between the central governance and the business areas that own the products.
- Identify and track external users allowed to access the data.
- Implement a method to collect feedback from internal and external stakeholders using the shared data to continually improvement how the data are being shared and identify opportunities for new data sources and data products.

Each of these recommendations is discussed in more detail herein:

- **Share data extracts, data APIs, or even entire large datasets with external users by giving them access to the data directly on the cloud environment where they can search and analyze data at their own cost** – Traditional data system owners are typically wary of sharing their data with users that do not fall under the predetermined use cases the system was built to support. Indeed, traditional data systems are designed and sized for specific requirements and cannot easily accommodate more users without the risk of overloading the system, exposing data to unknown or non-trusted users, corrupting the data, etc. On the contrary, by leveraging cloud infrastructure storage and computing scalability, modern data systems were designed to establish an environment that enables large and complex datasets to be searched and analyzed in many different ways by many different users inside or outside an organization to its benefit.

  > *Modern data systems were designed to establish an environment that enables large and complex datasets to be searched and analyzed in many different ways by many different users inside or outside an organization.*

- **Do not use proprietary file formats or API protocols when sharing data; use web APIs to allow the shared data to be used by a large number of data tools rather than just the ones of a single vendor** – Traditional data system owners typically share data using the features available in their main data systems, which often implies that the system only offers the possibility to share data using vendor proprietary file formats and sometimes interfaces. This is done as a way to coax stakeholders that want to use the data to acquire systems from the same vendor. This can be observed today across many transportation agencies sharing geospatial data using a single vendor and its proprietary file format and interfaces. Instead agencies should focus on using non-proprietary file formats and APIs to share their data both internally and externally. The open file formats and APIs are often included in vendor solutions but not enabled by default.

- **Be transparent with those with whom you share data and provide information that helps them get the most out of the data***:*
  - Augment the data being shared with external users with provenance and quality metadata. Engage them to identify metadata updates or additions that may further improve the value of the data.
  - Mandate the creation and maintenance of web documentation for every dataset and analytical process currently existing on the system, and maintain an easy, flexible web documentation environment (rely on existing web documentation frameworks such as Readthedocs).
  - Leverage data crawl tools to automatically extract information from datasets and create/update web documentation.
  - Allow for the representation of the datasets and their relationships to be flexible and adapt quickly to the addition of new relationships or data to the environment.

- **Do not be concerned with corrupted data** – One of the most notable differences between traditional and modern data systems is the way in which they deal with corrupted data. Traditional data systems use tight data models and strict access rules aimed at preserving processed data to avoid corruption and deletion. On the other hand, modern data systems consider processed data as disposable and easy to recreate from the raw data. They focus instead on first preserving unaltered, raw data and making it available to all users as read-only and second on using containerization and microservices data processing methods that can be saved, rebuilt, and redeployed on the fly so that any lost or corrupted data can be recreated directly from the raw data. This approach greatly reduces the need to heavily control access to shared data.

  Modern data systems consider processed data as disposable and easy to recreate from the raw data.

- **Encrypt, obfuscate, or remove sensitive data when sharing** – Traditional data systems often rely on the vendor software security feature they are built on to limit access to sensitive data by either not providing access to a specific table or by moving all sensitive data to a different and more secure data system. In modern data systems, this approach is often not applicable; while data can be moved to a more secure cloud infrastructure and folder where access can be restricted, it is sometimes detrimental to the exploration of such large datasets. Also, in raw format, sensitive data are often combined with non-sensitive data that could still be used for analysis. Thus, the approach taken by modern data systems is two-fold. First, sensitive data can be removed from the dataset to be shared. This can imply a duplication of the entire dataset, which can lead to non-negligible additional storage costs for large datasets. Second, sensitive data can be obfuscated or encrypted to allow it to be seen by only those users able to decrypt it. The advantage of the latter is that it does not necessitate duplication of the datasets and still allows non-authorized data users to perform analyses involving sensitive data fields while never seeing them. For example, encrypting vehicle license plate or vehicle identification numbers would still allow an analyst to identify individual vehicles in records and to perform an aggregation or join while never knowing anything about the actual vehicles.

  Sensitive data can be obfuscated or encrypted to allow it to be seen by only those users able to decrypt it.

- **For the best protection, frequently update encryption methods used to obfuscate sensitive data** – The combination of large amounts of data with the massive parallel computing offered by cloud infrastructure has rendered some of the traditional encryption methods ineffective. Indeed, parallel processing enabled by cloud infrastructure has allowed decryption software to test solutions at a rate of tens of millions in a few seconds. Common encryption algorithms such as SHA1, developed by the National Security Agency (NSA) in 1995 are no longer safe on their own when facing such massively parallel decryption attempts. In 2012, 117 million usernames and passwords were stolen from the website LinkedIn and decrypted in a matter of days before being offered for sale online.[4] Given how quickly the effectiveness of encryption algorithms is changing, these algorithms need to be carefully chosen. Among the algorithms recommended by the National Institute of Standards and

[4] https://money.cnn.com/2016/05/19/technology/linkedin-hack/

Technology (NIST), some (e.g., 3DES) already had their key space exposed and searchable and are no longer suitable to encrypt data. Others, such as AES 256 and SHA3, are still effective.

- **Create different versions of datasets based on who they need to be shared with** – Traditional data systems are built around the principles of ensuring that the data are not duplicated and that they are controlled concurrency in reading, writing, deleting, and updating to better preserve the filtered and prepared data they manage. Using raw data and leveraging the inexpensive storage and computing costs of the cloud infrastructure in modern data systems allows for a much more flexible approach, where data can be duplicated in several datasets catering to the needs of specific user groups without over-expending resources or costs. This approach is used especially when large user groups request similar sets of data for analysis.
- **Beyond sharing data, also share data analysis process code** – In traditional data systems, sharing data and data only is fairly limited, but sharing data analysis processes or code is even more limited. Analysis code in traditional systems is platform-dependent and requires that the recipient of the shared code be knowledgeable about that vendor system and deploy a similar system with similar data in order to rerun the analysis using the same code. Modern data architecture is different in the sense that it eliminates this dependency when using cloud infrastructure and allows anyone with the adequate resources to quickly replicate an analysis on a cloud system. Often such data analysis process code does not only contain the code processing the data but also the code implementing the cloud services necessary to perform these actions. This is often referred to as Platform as a Service (PaaS) or Infrastructure as a Service (IaaS). Therefore, sharing data analytics processes in modern data systems is much easier and less costly than with traditional systems. This provides a significant opportunity that traditionally is only performed by sharing results on paper through publications – that is to share both data and code to allow other agencies, institutions, or universities to review, validate, and improve. Code for data analysis can be shared using dedicated cloud services such as GitHub or direct as part of the cloud storage.

  Modern data systems offer a significant opportunity that traditionally is only performed by sharing results on paper through publications – that is to share both data and code to allow other agencies, institutions, or universities to review, validate, and improve.
- **Establish live data streams in addition to sharing historical data** – In traditional data systems, live data are typically shared using SOAP web servicing, which returns a snapshot of ongoing events stored in the database in XML format upon request. Modern data systems possess enough computing power and storage to analyze and share massive amounts of data in real-time. Therefore, data shared from emerging technologies should not be restricted to historical data stored on the cloud but should also provide some live stream interfaces exposing the massive amounts of data collected in real-time to allow internal and external users to develop real-time analyses directly on these data streams.
- **Share the responsibility of providing data services and products to external users between the central governance and the business areas that own the products** – Traditional data systems maintain a centralized control over who is allowed to receive data and perform queries or analyses on the data system. In modern data systems, this responsibility is shared between the data management/governance team and the business areas responsible for the analyses or data. The data

management/governance is able to assert the possibility of exposing the data environment to external users and assigning the privileges with regards to the whole system. Business areas review the request for access to their system, assess the impact it may have on their system, and create and manage the new accounts once granted. Also, share data products using a subscription model to keep track of users' activities and potentially share their cost.

- **Identify and track external users allowed to access the data** – While traditional data systems intend to control access to the data upfront by tightly controlling it, such an approach is less likely to be successful across the large and complex datasets, distributed processing, and extensive sharing of modern data systems. Rather, modern data system management teams are more dynamic and responsive, keeping detailed records about data users and their activities and monitoring in real-time their data access, processing activities, and results publishing. By building various alerts and thresholds over this activity monitoring, they are able to quickly, even automatically, alter access and privileges to rogue users as they attempt unauthorized operations.
- **Identify and track external users allowed to access the data** – While traditional data systems allow for strict access rules to the data for a limited number of users, modern data systems can do the same thing over a much larger number of users. This can render it difficult to ensure that each user has correct access to the data, as there are so many users to oversee. Therefore, in addition to assigning data access privileges to each user and group of users and trusting that this will be enough to secure the data, organizations should also track in real-time what each user does with the data, which data are accessed and what is being done with the data. Organizations should also set up alerts and access denial triggers as soon as activities no longer match current user or group access privileges.
- **Implement a method to collect feedback from internal and external stakeholders using the shared data to continually improvement how the data are shared and identify opportunities for new data sources and data products** – While traditional data systems identify user needs from pre-established data sharing requirements developed during system design, modern data systems can no longer assume that pre-established data sharing requirements will be relevant long enough when using rapidly changing data sources. Therefore, modern data systems need to implement a method to continuously collect and review user data sharing needs, as well as a process to rapidly change the way data are shared on the system to reflect users' needs and behaviors.

Case Study – Data Sharing Platform

One example of good data sharing practices is the city of Columbus, Ohio. Fueled by a $40-million grant from the federal government, the "Smart Columbus" project goes beyond the typical open data policies by developing and hosting two major data sharing projects: the data and analytics hub, "Smart Columbus OS," and the more application-focused, "Integrated Data Exchange," or IDE.

The Smart Columbus OS system hosts over 3,200 publicly available datasets. These datasets are held to minimum quality standards before being allowed on the website. Once accepted into the system, they are enriched with metadata and tagged with keywords to enhance usability and searchability. Users can access the datasets via a searchable web interface or via an API. Advanced users can request access to

the data through a hosted online “Jupyterhub” interface, which is a popular development environment for data scientists and other analysts.

Where Smart Columbus OS provides resources for researchers and analysts, the IDE supports IoT devices and business applications. Smart Columbus developed the IDE with funding and collaboration from local businesses to provide a more unified data platform that can integrate with emerging technologies and the businesses that use them. The goal is to gather data from multiple IoT sources, ensure the privacy of that data, and govern access to the data to ensure usability.

It takes significant effort and resources to build a data sharing platform as robust as either the Smart Columbus OS or the IDE, but both are exemplary in how they fully embody the spirit of not just allowing but enabling and promoting open data sharing among business partners, researchers, and everyday users.

SUPPORTING TOOLS

This section contains a variety of tools in support of the roadmap. This section contains:

- **Data Management Capability Maturity Self-Assessment (DM CMSA)** – The DM CMSA is built from the modern big data benchmark and assessment methodology presented in the final research report. Questions within each of 15 data management focus areas will guide transportation agencies through a self-assessment to gauge their current data management practices, as well as to identify specific areas for improvement along their path toward shifting from traditional data management practices to more modern data management practices in order to handle data from emerging technologies.
- **Big Data Governance Roles and Responsibilities** – provides a list of recommendations to consider when developing a modern data governance approach, a description and frameworks for big data governance, and a tool for tracking the big data governance roles and responsibilities within an agency.
- **Data Sources Catalog Tool** – provided to assist transportation agencies in cataloging existing and potential data sources. This tool is useful for better understanding the data assets of an agency, prioritizing data sources, and informing the selection of those sources that might offer the most value before pursuing further development.
- **Frequently Asked Questions (FAQ)** – a list of questions and responses to frequently asked questions regarding big data implementation, management, governance, use, and security.

Data Management Capability Maturity Self-Assessment (DM CMSA)

This section contains the Data Management Capability Maturity Self-Assessment (DM CMSA) developed as part of the research. The DM CMSA will allow transportation agencies to gauge their data management practices, as well as to identify specific areas for improvement. The self-assessment was designed for ease of completion and to provide a high-level starting point for further inquiry.

The self-assessment consists of 104 questions divided across 15 "focus areas." These focus areas were derived from the 11 data management "knowledge areas" described in the DAMA DMBOK2 (shown in Figure 13[5]), as well as four additional focus areas (data collection, data development, data analytics, and data dissemination) that expand the scope of data management to the full data lifecycle. Brief descriptions of each of the 15 focus areas are provided below:

©DAMA International

Figure 13. DAMA DMBOK2 Knowledge Areas (DAMA International, 2017)

- Data collection – Acquiring new data, either directly or through partnerships, in such a way that the value, completeness, and usability of the data are maximized without compromising privacy or security.
- Data modeling and design – Analysis, design, building, testing, and maintenance of data.
- Data architecture – The overall structure of data and data-related resources as an integral part of the enterprise architecture.
- Data storage and operations – Structured physical data assets storage deployment and management.
- Data security – Ensuring privacy, confidentiality, and appropriate access to data.
- Data quality – Defining, monitoring, maintaining data integrity, and improving data quality.
- Data governance – Planning, oversight, and control over management of data and the use of data and data-related resources.
- Data integration and interoperability – Acquisition, extraction, transformation, movement, delivery, replication, federation, virtualization, and operational support.
- Data warehousing and business intelligence – Managing analytical data processing and enabling access to decision support data for reporting and analysis.
- Data analytics – Investigating processed data to drive actionable insights and answer questions of interest for an organization.
- Data development – Designing, developing, and creating new data products, as well as augmenting, customizing, and improving existing data products.

[5] Single use permission for the DAMA-DMBOK2 Guide Knowledge Area Wheel. No redistribution rights. Contact DAMA for use in other documents.

- Document and content management – Storing, protecting, indexing, and enabling access to data found in unstructured sources (electronic files and physical records) and making these data available for integration and interoperability with structured (database) data.
- Reference and master data – Managing shared data to reduce redundancy and to ensure better data quality through standardized definition and use of data values.
- Metadata – Collecting, categorizing, maintaining, integrating, controlling, managing, and delivering metadata.
- Data dissemination – Sharing data products and data analysis results effectively with appropriate internal and external audiences.

The DM CMSA contains 15 tables (Table 5 through Table 19), one table corresponding to each of the 15 data management focus areas. The questions in each of the tables were derived from the foundational principles of big data management and the corresponding benchmark methodology, which are presented in the NCHRP 08-116 final report: Framework for Managing Data Emerging Transportation Technologies to Support Decision-Making. Every question calls for a self-assessment score of "low," "moderate," or "high" and provides examples of what practices or procedures would merit each score. At the end of the self-assessment a summary scoresheet (Table 20) is provided where all recorded answers can be totaled across the 15 data management focus areas. The summary score sheet should provide an overall measure of an organization's data practices, particularly as they relate to managing data from emerging technologies for transportation. Given the rather siloed nature of data within most transportation agencies, it is recommended that the self-assessment be taken by representative groups across an agency.

After completing the self-assessment, it is recommended that the individual responses be reviewed to identify areas where the most improvement can be made. The descriptive examples included in each question will help identify changes that need to be made and goals to pursue that will advance the organization's data management processes and practices.

Table 5. Data Management Focus Area: Data Collection

Question	Scoring			Score
	Low	Moderate	High	
Considering the data collected by the agency, how relevant are they to the current agency needs?	Data collected are not relevant to current agency needs.	Data collected are somewhat relevant to current agency needs.	Data collected are highly relevant to current agency needs.	
Are the data collected relevant to the future agency needs?	Data collected are not relevant to future agency needs.	Data collected are somewhat relevant to future agency needs.	Data collected are highly relevant to future agency needs.	
How significant are the gaps in the current data for meeting agency needs?	Significant gaps in current data to meet agency needs.	Some gaps in current data to meet agency needs.	Little or no gaps in current data to meet agency needs.	
In what formats are the data collected by the agency?	Data collected are in outdated or proprietary formats.	Data collected are in a usable, but not ideal, formats.	Data collected are in modern, open source formats.	
To what extent are the source data collected by the agency preserved prior to any editing or other modification?	Source data are usually deleted or modified.	Source data are modified but not deleted.	Source data are never deleted or modified.	
How is personally identifiable information (PII) within the source data handled?	PII is collected and handled via an insecure process.	PII is secured or anonymized at some point after collection.	PII is collected securely.	
Are documented data collection procedures available and routinely updated?	There are no documented data collection procedures.	Documented data collection procedures exist but are infrequently reviewed/updated.	Documented data collection procedures are frequently reviewed and updated.	
			# of Low Scores	
			# of Moderate Scores	
			# of High Scores	

Table 6. Data Management Focus Area: Data Modeling & Design

Question	Scoring			Score
	Low	Moderate	High	
Has the agency referenced any existing data models or frameworks when designing their data architecture and processes?	No data models or frameworks referenced.	Some research performed prior to data workflow design.	Extensive knowledge of applicable data frameworks used.	
Has the agency performed any data usability assessments for the current data sources?	No data usability assessments have been performed.	Basic inventory of data sources performed with limited information.	Gull data usability assessment performed and regularly updated.	
How flexible is the data workflow model?	Data model does not allow for ad hoc data augmentation or other continuous development practices.	Data model allows for some ad hoc data augmentation or other continuous development practices.	Data model is designed to fully implement continuous development practices, including ad hoc data augmentation.	
Does the data model include augmenting datasets with metadata designed to enhance usability, such as data quality and provenance identifiers?	Data model does not account for adding any data usability focused metadata.	Data model has steps to add some usability focused metadata.	All appropriate metadata additions that enhance usability are fully accounted for by the data model.	
Does the data model allow for data masking techniques to be applied, where sensitive information is anonymized or obscured?	Data model does not include any data masking techniques. Source data are deleted rather than masked.	Data model includes insufficient masking techniques, or masking techniques are inconsistently applied throughout the model.	Data masking is fully accounted for by the data model in all areas where it may be useful.	
Has the agency designed new processes specific to handling data from emerging technologies?	No new processes for emerging technologies have been considered.	Some modifications have been made to existing data models and processes to handle emerging technology data.	New processes have been designed from the ground up to support emerging technology data.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 7. Focus Area: Data Architecture

Question	Scoring			Score
	Low	Moderate	High	
How well is the agency's data organized?	Data are organized haphazardly.	Data are organized adequately.	Data are organized optimally.	
Do the file and folder names follow a documented naming convention?	Folders and files follow no common standard or convention.	Folder and file names generally make sense, but do not follow documented conventions.	All folder and file follow a documented naming convention.	
How well do the data schemas meet the needs of your analysts?	Data schemas do not meet the needs of analysts.	Data schemas are generally functional.	Data schemas that best meets the analysts' needs are used.	
Are data tables generally "well-formed," with one subject per column and one piece of information per row?	Tables are not generally well-formed.	Tables are generally organized.	Tables are fully well-formed.	
Do data administrators, both internal and external, provide sufficient support to maintain the data architecture?	There is little or no ongoing support for maintaining the data architecture.	There is adequate support for most maintenance needs.	There is optimal support for all maintenance needs.	
If the agency is needed to quickly respond to a change in data storage needs, how difficult or costly would that change be?	Data architecture is outdated, rigid, or vendor locked. Rapid change is impossible.	Data architecture relies on closed source software/hardware or is not well documented. Rapid change is difficult and costly.	Data architecture is built on well understood open source services. Rapid change is possible with little difficulty.	
Is the data architecture distributed enough for some processes to be changed or discarded without affecting the whole?	All processes are fully dependent on each other such that any change in one process necessitates systemwide modifications.	Some processes have well-documented dependencies that require additional work to be done if they are to be updated or changed.	All processes are independent enough to be individually modified with minimal impact on the other processes.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 8. Focus Area: Data Storage & Operations

Question	Scoring			Score
	Low	Moderate	High	
How much of the data collected by the agency (and that could be useful/relevant to the organization's needs) are stored?	Only some data are stored.	Most data are stored.	All relevant data are stored.	
Are the agency's data consolidated in a central location for storage and analysis?	Data are stored in separated data silos.	Data are centrally stored but must be copied to a separate system for analysis.	Data are stored in a fully functional data lake architecture.	
How long are data preserved/stored for future use?	Some data are only stored for a short period of time due to size or legal reasons.	Some data are stored only long enough to perform needed analyses. Data that are not perceived to be useful are seldomly stored.	All data are stored as long as possible to support current and future analyses, even if those analyses are not actively in use today.	
In what type of format re the data stored?	Data are stored in an outdated or proprietary format.	Data are stored in a usable but obscure or difficult format.	Data are stored in a well known, modern open source format.	
How often are backups of the data created?	Backups are rarely performed.	Backups are performed every so often.	Backups are frequently performed and verified.	
Where are the backup data stored?	Backup data are stored onsite.	Backup data are stored both onsite and at a single offsite location, such as a separate on-premise storage facility or data center.	Backup data are stored at multiple offsite locations or by a reputable cloud service provider.	
How quickly can the agency recover data from backup storage after a disruption?	Unacceptable time to recovery.	Adequate time to recovery.	Excellent time to recovery.	
Are records of the data's history and origin maintained?	No history or origin of the data is maintained.	The history and origin of some data files are maintained.	The history and origin of all data files are maintained.	

Question	Scoring			Score
	Low	Moderate	High	
Does the agency maintain a documented disaster recovery plan?	No disaster recovery plan.	Some processes are in place for disaster recovery, but they are not frequently reviewed or updated.	Disaster recovery plan is frequently reviewed and updated.	
Does the data system architecture fully meet the needs of analysts?	Data system architecture is insufficient and error prone.	Data system architecture is adequate.	Data system architecture is optimal for analyst's needs.	
Does the organization rely on closed source/proprietary software to manage the data?	Software and environment are closed source and proprietary.	Some software used is open source.	All software used is well supported and open source.	
Does the architecture incorporate cloud-based systems where appropriate?	The data architecture is on-premise.	Some systems are cloud-based.	All systems are in a cloud-based environment.	
Are the agency's data processes designed to be independent of the underlying systems used?	Data processes are built directly into the system and are difficult to update.	Some data processes are kept independent of the system itself.	All data processes are independent and can be upgraded or replaced easily.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 9. Focus Area: Data Security

Question	Scoring			Score
	Low	Moderate	High	
How is sensitive information/PII within the data stored?	Sensitive information/PII is stored in plaintext.	Sensitive information/PII is stored in a somewhat secure manner.	All sensitive information/PII is fully secured from collection to data product.	
Are privacy filters applied to anonymize the data?	No privacy filters are applied.	Some privacy filters and/or encryption is employed for PII.	Privacy filters and other safeguards are applied at the time of collection.	
Are privacy filters granular enough to allow different analyses to be performed at different levels of access?	No granularity in the privacy filters. A record is either flagged as sensitive and filtered or it is not.	Records are labeled with some granular level of sensitivity, but no easy means exist to selectively filter records based on this label.	Records are given detailed labels that describe their sensitivity, and analytical processes are able to filter or obfuscate records at varying levels according to their access and needs.	
How are the network and other infrastructure secured?	No network encryption or endpoint protection.	Basic level of network and endpoint security.	All relevant security software and procedures are employed.	
Does the agency employ secure user authorization processes?	Insecure authentication processes fail to prevent unauthorized use of data.	Outdated or inadequate authentication processes fails to fully secure data.	Authorization processes are up to date and fully prevents all unauthorized use.	
Do the authentication processes hinder authorized access to the data?	Rigid authentication structures hinder authorized use of data.	Authorization structures somewhat hinder authorized use of the data.	Fluid and convenient authorization structures do not hinder authorized use of the data.	
How efficiently is the agency able to add or manage user access to the data?	A great amount of time and effort is required to grant access to a new user.	Some amount of time and effort is required to grant access to a new user.	It is easy to grant new users access when warranted.	
Are customized privacy protocols, tailored to the different data stakeholders, employed?	No changes in privacy protocols are made with respect to data stakeholder groups.	For some, but not all stakeholders, customized privacy protocols are applied.	All relevant data stakeholder groups are handled with their own customized privacy protocols.	
Is outsourced cybersecurity expertise utilized?	Outside expertise is never utilized on any cybersecurity matter.	Outside expertise is infrequently consulted, or the outputs from that consultation cannot be independently audited.	Outside expertise is heavily utilized, and in-house experts are qualified to independently audit and verify third-party findings and recommendations.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 10. Focus Area: Data Quality

Question	Scoring			Score
	Low	Moderate	High	
Is data quality monitored?	Data quality is unknown.	Data quality is somewhat known.	Data quality is fully known and actively monitored.	
Is a standardized system of ranking data quality employed?	No data quality rankings are performed.	Basic data quality rankings are performed.	Detailed data quality rankings are performed.	
Are dashboards used to visualize data quality statistics?	No data quality dashboards are available.	Some data quality dashboard are available.	Many data quality dashboards and supporting tools are available.	
How are suspect data flagged for review?	Few processes are in place to flag low quality data.	Manual processes are in place to flag low quality data.	Both automated and manual processes are used to flag low quality data.	
Are users and data processes able to select a level of data quality to use for their analysis?	No ability to select data based on data quality rankings or flags.	Users can filter data at a low granularity or with some difficulty.	Users are able to easily filter data based on data quality rankings at a high level of granularity.	
When working with data quality, are the original data ever corrected, modified, or deleted?	When data quality concerns arise, the source data are almost always deleted or heavily modified.	When data quality concerns arise the source data are sometimes modified, corrected, or deleted.	When data quality concerns arise, the source data are flagged and scored but never modified or deleted.	
Are any data crawling tools employed to continuously monitor data quality trends?	No data crawling is performed at any level.	Data crawling is performed infrequently or through a manually initiated process.	Fully automated data crawling is continuously performed, generating timely and detailed alerts on data quality trends.	
Are users able to report data quality issues?	End-users of the data are unable to report data quality issues.	End-users may report data quality issues, but those reports are infrequently reviewed using a fully manual process.	End-users may report data quality issues, and those reports are frequently and easily reviewed via a partially automated process.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 11. Focus Area: Data Governance

Question	Scoring			Score
	Low	Moderate	High	
Does the agency have full ownership of and unrestricted access to the data that they obtain from third parties?	In most cases, the third-party owns the data and severely restricts access and use.	In some cases, the agency owns the data from the third-party but must comply with rigid use restrictions.	In most cases, the agency owns the third-party data and may fully use it with few restrictions.	
Is the agency limited by high costs in accessing and using data relevant to their needs?	Very high cost to use data.	High cost to use data.	Reasonable cost to use data.	
Are users restricted to using only certain tools when analyzing data of interest?	Users are restricted to a small number of proprietary analysis tools.	Users can use a range of tools (with some restrictions), mostly proprietary but some open source, to analyze the data.	Users can use any number of tools to analyze the data and with few restrictions.	
Is access, use, or analysis of data limited by agency data management policies and practices?	Agency data management policies and practices severely limit access, use, or analysis of data.	Agency data management policies and practices somewhat limit access, use, or analysis of data.	Agency data management policies and practices do not limit access, use, or analysis of data.	
Is data management software used by the agency?	No data management software is used.	Data management software is used, but it is not optimal.	Optimal data management software is used.	
Does the agency follow a documented data management plan?	The agency has no documented data management plan.	The agency follows a loose, largely undocumented data management plan.	The agency follows a documented and frequently updated data management plan.	
Does the agency have in-house data management experts, or does the agency outsource data system management to one or more third parties?	The agency relies on outside parties for data management.	Some data systems are outsourced while others are managed in-house.	The agency conducts all data system management in-house.	
Does the agency actively monitor their data management systems?	There is no active system monitoring.	Some system activity dashboards are available.	The agency employs both reactive and proactive monitoring of their data management system.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 12. Focus Area: Data Integration & Interoperability

Question	Scoring			Score
	Low	Moderate	High	
Is data that the agency uses in a format that allows for easy integration into new systems?	Most data cannot be integrated to new systems without significant effort.	Some data must be converted into a new format before integrating into a new system.	All data can be integrated without conversion or modification.	
Do all systems that process data within the agency rely on a centralized data source?	Each system uses its own data type and siloed data source(s), making integration between separate systems difficult	Some systems connect to the same data source(s), while some retain their own siloed version of the data.	All systems referencing the same data connect to a common data source for that data.	
Is operational support provided for the agency's integrated data systems?	No operational support is provided for the agency's integrated data systems.	Some general support from IT is available for the agency's integrated data systems.	Full support from skilled resources with advanced system knowledge is available for the agency's integrated data systems.	
Are variable names in datasets mapped to existing data standards?	Variable names change from dataset to dataset, with no standardized nomenclature.	Some datasets have variable names mapped to some data standard.	All applicable datasets are mapped to the same standard wherever possible such that they can be easily joined.	
How are data organized across datasets?	No uniform organization plan; folder structures are unique to each dataset.	Some, but not all, datasets are organized using the same folder structure.	All datasets are organized using a single planned folder structure.	
How are data classified across datasets?	No uniform classification taxonomy.	Some datasets use similar classification taxonomies.	All datasets conform to a single, documented classification taxonomy.	
Are identification metadata consistently applied across datasets?	No uniform metadata enrichment is performed.	Some datasets have similar identifying metadata fields.	All datasets are enriched with a uniform set of identifying metadata.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 13. Data Management Focus Area: Data Warehousing & Business Intelligence

Question	Scoring			Score
	Low	Moderate	High	
Do stakeholders feel like the organization is getting its worth out of the data?	Few stakeholders recognize the value of the data; data are seldom used to meet real business needs	Some business needs are met, but data operations are not highly valued or prioritized.	Most stakeholders regularly derive real value from the data.	
Does the agency have access to sufficient Business Intelligence (BI) products?	Few or no useful BI products are available.	Some useful BI products are available.	All current needs are satisfactorily met by a suite of BI products.	
Does the agency have sufficient data visualizations to reference and understand their data?	Few or no useful data visualizations have been created.	Some useful data visualizations are infrequently used.	A variety of relevant and useful data visualizations are frequently referenced.	
Are data users able to develop their own BI products and visualizations?	It is not possible for data users to develop or customize their own BI products or visualization within current processes and procedures.	Data users can develop some limited customization of products and visualizations but only after a lot of administrative red tape.	Data users can develop new BI products and visualizations with minimal administrated red tape or oversight.	
Are successful BI products and visualizations shared with other users that could benefit from them?	BI products are "siloed" to where they are only used by the original stakeholders for the original use case.	Some BI products and visualizations are infrequently shared among stakeholders.	BI products and processes are regularly reviewed so that the most successful ones can be shared and emulated.	
Are stakeholders empowered to select the BI tools that are most useful for them?	Technical limitations or data format incompatibilities limit what BI tools can be used.	Some variety of BI tools are technically possible but limited by policy or organizational red tape.	Stakeholders are able to choose their own BI tools without undue technical or administrative limitations.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 14. Data Management Focus Area: Data Analytics

Question	Benchmarks			Score
	Low	Moderate	High	
Do the data need to be moved to a separate system for analysis?	Full migration to a separate system is necessary to perform any analysis.	Some data must be migrated to a separate system to perform some analysis.	All analysis can be performed without copying or moving data.	
Are the data analyses run and the results of the analyses saved to the same location where the data are stored?	All analysis results are saved on a separate system(s) from the data being analyzed.	Some analysis results are written to the same location as the data.	All analysis results are written to the same location as the data.	
Does the agency employ data analysis techniques designed for big data?	Only traditional analytical techniques and processes are used.	Some traditional analytical processes have been modified for infrequent use with big data.	Relevant big data analytical techniques are actively used.	
Does the agency leverage analyses that have been designed by other agencies or the online community?	No outside analyses have been referenced, copied, or built upon.	Outside analyses are infrequently referenced or rarely used for production data.	The agency frequently reviews, learns from, and uses relevant analyses from multiple outside sources.	
Does the agency have the means to perform analyses on live streaming data?	No capabilities to analyze streaming data.	Some capabilities to analyze streaming data exist, but they are limited or infrequently used.	Fully capable and actively deriving value from streaming data analyses.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 15. Data Management Focus Area: Data Development

Question	Scoring			Score
	Low	Moderate	High	
Does the agency perform or oversee the development of customized data products?	No customized data products are developed; out of the box solutions are used exclusively.	Some data product development is outsourced with little input from the agency.	New data products are frequently developed and effectively used.	
Is a review process in place to identify effective new data enrichment possibilities?	No review process is performed regarding new data enrichment.	New data enrichment is infrequently considered via an undocumented process.	Data enrichment opportunities are regularly reviewed via a well-documented process.	
Is a review process in place to identify effective new data products?	No review process performed regarding new data products.	New data products are infrequently considered via an undocumented process.	New data products are frequently considered via a well-documented review process.	
Can new analytical products be built on existing tools or must they be developed from scratch?	All new data products must be built from scratch.	Offshoot data products can be built with some difficulty.	Current tools support easy development of additional products and visualizations.	
Can new analytical products be swiftly developed and iterated on?	New products are developed slowly and "perfected" before being put into use.	New products take considerable development work before they can be put into use.	New products can be developed and put into use swiftly.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 16. Focus Area: Document & Content Management

Question	Scoring			Score
	Low	Moderate	High	
Does the agency maintain documentation for all data products and processes?	No documentation of data products or processes is maintained.	Some documentation of data products or processes in maintained in an offline format.	Detailed documentation of data products or processes is available in an online, web-based format.	
Is the documentation regularly reviewed and revised?	No reviews or revisions of the documentation since creation.	Some documentation is sporadically reviewed.	All documentation is regularly reviewed, revised, and updated.	
How often is documentation used by stakeholders?	Documentation rarely read or followed.	Some stakeholders are aware of documentation but seldom make use of it.	All relevant parties uniformly follow procedures as documented.	
Is documentation available in an easy to access web documentation framework?	Documentation is only unavailable online.	Documentation is available as a pdf download link only.	Documentation is available in a live, searchable, online documentation framework.	
Are groups or stakeholders held accountable for the accuracy and availability of their documentation?	No clear ownership of documentation responsibilities,	Documentation ownership is clear, but there is no incentive for owners to keep documentation updated.	All documentation is regularly reviewed, and owners are encouraged to update regularly.	
Are any automated processes employed to update data documentation?	No automated processes are employed.	Automated status checks are made, but there is no automated document editing available.	Automated processes regularly update web documentation with information extracted from live datasets.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 17. Focus Area: Reference & Master Data[6]

Question	Scoring			Score
	Low	Moderate	High	
Is reference documentation maintained for all databases and storage?	No documentation of databases and storage is maintained.	Some documentation of databases and storage is maintained in an offline format.	Detailed documentation of databases and storage is available in an online, web-based format that is updated regularly.	
Are reference data uniform across all business units?	There are mismatches in reference data across groups.	Reference data are siloed or duplicated but uniform.	Reference data exist in one location as a single source of truth for all users.	
Are master data values and identifiers consistently used across all systems?	Master data values are inconsistent.	Master data values exist in multiple siloed locations but are generally consistent.	Master data values are stored and managed in one accessible location.	
Does a visual representation exist that shows how each dataset relates to each other and/or how they can be combined?	No visual representation of dataset relations exists.	A visual representation exists, but it is outdated or otherwise inaccurate.	A visual representation exists in a regularly updated and highly legible format.	
Are data users able to easily access this visual representation of dataset relations?	Regular data users are unable to access the visual representation.	Data users can only access the visual representation with some difficulty.	Data users can readily access the visual representation.	
Can this visual representation of dataset relations be quickly and easily updated as more datasets are created?	Visual representation is stored in a format that is difficult to update (pdf).	Visual representation is infrequently updated via a manual process.	Visual representation is regularly updated through a largely automated process.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

[6] Reference data are data that define a set of permissible values to be used by other fields. Master data represent objects and all associated information about those objects that are relevant to the organization. In both cases, reference and master data management involve ensuring that these data remain consistent across all datasets in the organization (Entry on Reference Data, n.d.).

Table 18. Focus Area: Metadata[7]

Question	Scoring			Score
	Low	Moderate	High	
Does the agency keep and maintain a metadata catalog?	No metadata catalog is maintained.	A metadata catalog is maintained, but it only applies to some data.	A metadata catalog is maintained for all applicable data.	
Does the agency enrich the data with additional metadata fields?	No enrichment / additional metadata fields created.	Some enrichment / additional metadata fields created.	Optimal enrichment / additional metadata fields created.	
you are metadata practices regularly revised and updated?	Metadata practices are seldom reviewed or revised.	Metadata practices are infrequently reviewed or are ad hoc.	Metadata practices are regularly reviewed and updated following a documented process.	
Are metadata transparent and available to those with access to the data?	Metadata are never made available to data users.	Some users may be able to access metadata fields for some datasets.	All metadata for all datasets, along with associated documentation, are made available wherever appropriate.	
Is there a means of collecting feedback from data users regarding the available metadata?	No means of collecting or implementing feedback from data users.	Feedback from data users is not solicited or regularly reviewed but may sometimes be implemented if received.	Feedback from data users is openly solicited and regularly reviewed.	
Are all metadata fields that apply to multiple datasets applied uniformly across the datasets?	All metadata are dataset dependent.	Some groups of similar datasets are augmented with similar metadata fields.	All datasets are augmented with the same well-documented metadata fields wherever possible.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

[7] Metadata are "data about the data" and typically are found in a metadata catalog where users or programs can find information about the data such as how large a file is, what format that file is in, when the file was last modified, what data types are stored within each column of a table, or whether a numeric value represents hours or minutes.

Table 19. Focus Area: Data Dissemination

Question	Scoring			Score
	Low	Moderate	High	
Are open are the datasets within the agency?	Data are unavailable to all but a few users (e.g., IT)	Data are available to selected users who are expected to have use for them (some use within business units).	Data are available to whomever may have a potential use for the data, with the exception of sensitive data	
Has the agency implemented an open data policy?	No thought has been given to implementing open data policies	Some open data policies are in use.	Open data policies are applied wherever possible.	
Are there any technical barriers that prevent users from reaching the agency's open data?	Data can only be accessed internally.	Some technical barriers exist, or data are only available via simple download.	Data are easily reachable via APIs and/or hosted analytics platforms with no technical barriers.	
Are users with access to the data able to access it directly where they are stored?	All users must copy or download the entire dataset first before any analysis can be performed.	Some users are able to analyze some datasets directly through a process where the organization shoulders all costs involved.	All authorized users are able to access data directly where it is stored and analyze it at their own cost.	
Are any developed data products available to users via an open sharing portal?	No data products are shared.	Some data products are shared via an unmonitored process.	All relevant data products are shared with authorized users whose usage is monitored and who may bear some of the costs involved.	
Are users able to easily use their own tools and code with your open data API?	Proprietary file formats or closed access prevents the use of nearly all data tools.	Open file formats are used but outdated/incorrect documentation hinders use of non-standard data tools.	Open file formats and common protocols are used for maximum compatibility with a wide range of current and future data tools.	
			# of Low Scores	
			# of Medium Scores	
			# of High Scores	

Table 20. Self-Assessment Summary

Data Lifecycle Management Component	Focus Area	# Low	# Moderate	# High
CREATE	Data Collection			
CREATE	Data Modeling & Design			
	CREATE Subtotals			
STORE	Data Architecture			
STORE	Data Storage & Operations			
STORE	Data Security			
STORE	Data Quality			
STORE	Data Governance			
STORE	Data Integration & Interoperability			
	STORE Subtotals			
USE	Data Warehousing & Business Intelligence			
USE	Data Development			
USE	Data Analytics			
	USE Subtotals			
SHARE	Document & Content Management			
SHARE	Reference & Master Data			
SHARE	Metadata			
SHARE	Data Dissemination			
	SHARE Subtotals			
GRAND TOTALS				

Big Data Governance Roles and Responsibilities

Data governance is a collection of practices and processes that help to ensure the formal management of data assets within an organization, including the planning, oversight, and control over management of data and the use of data and data-related resources. Data governance puts in place a framework to ensure that data are used consistently and consciously within the organization. It also deals with quality, security and privacy, integrity, usability, integration, compliance, availability, roles and responsibilities, and overall management of the internal and external data flows within an organization (Roe, 2017).

Traditionally, data governance dealt with the strict, authoritative control of data systems and users. According to Wells, traditional data governance operates on the fundamental premise that data cannot be governed; only what people do with the data can be governed. While this may have been a feasible approach for traditional data systems, modern data systems, which incorporate agile development, big data, and cloud computing, have rendered this approach much more challenging to implement.

Below is a list of recommendations to consider when developing a modern data governance approach, based off of the work of "The Next Generation of Data Governance" by Dave Wells. Each recommendation has been divided into one of several aspects of data governance to consider during this development (Wells, 2017).

- **Agile data governance** – Governance that adapts quickly to changes in data or analysis.
  - Focus on value produced, not methodology and processes. This includes value to the project and enterprise value produced by meeting governance goals.
  - Govern proactively. Introduce constraints as requirements at the beginning of a project instead of seeking remedial action at the end.
  - Strive for policy adoption over policy enforcement. Make it easy to comply with policies, communicate the reasons for and value that is created by the policies.
  - Write brief, concise, clear, and understandable policies. Use simple language that is not ambiguous or subject to interpretation.
  - Include data governors and stewards on project teams. They bring valuable knowledge and are generally great collaborators.
  - Think "governance as a service" instead of "authority and control."
- **Big data governance** – Governance well-suited to handling very large amounts of data.
  - Do not attempt to govern all data. Writing policies that govern all data in a big data environment, not to mention enforcing such policies, is an enormous task.
  - Focus on policies for privacy-intensive, security-sensitive, compliance-sensitive data. This will direct governance efforts to where they will have the most impact.
  - Use automated methods to classify data. This can help identify what data are most important to govern, a process for which manual approaches often prove to be unfeasible.
  - Consider a govern-at-access approach. This approach determines the permissions of any given user at the time they attempt to access it, allowing for more flexibility, reactivity, and scalability than manually establishing such access beforehand.
  - Automatically detect and flag suspect access patterns. Attempts to access sensitive data on an unrecognized device or during unusual hours should be treated as suspicious.

- **Cloud data governance** – Governance of data on centralized cloud-based storage architectures.
  - Do not rely on physical server separations to enforce data governance. In a cloud-based data lake environment all data resides in a central location with unified data governance. Therefore, all data policies must consider all data users since there are no physical separations or data silos segregating user access.
  - Review national and local regulations. Some regulations have a direct impact on how cloud storage can be used.
  - Know the physical location of where cloud data are stored. This may have an impact on what data regulations apply.
  - Understand how governance is enforced by cloud partners. If a partner's implementation of data governance is insufficient a new service provider can be sought out.
- **Next generation data governance** – Horizontal governance rather than hierarchical governance.
  - Build a governance community.
  - Focus on proactive prevention and real-time intervention. Ideally, enforcing data governance rules after the fact should be a last resort to use only if prevention and intervention efforts have failed.
  - Embrace minimalist policymaking. A small number of important policies is more scalable and interpretable than a large, complex collection of minor policies.

Frameworks for big data governance have been developed to guide transition of organizations from traditional data governance to more modern data governance by decomposing and structuring the new data governance goals and objectives. Figure 14 presents one of these frameworks. The stated goals of this big data governance framework are to protect personal information, preserve the level of data quality, and define data responsibility (Kim & Cho, 2018).

The IBM Information Governance Council Maturity Model, represented in Figure 15, establishes a multi-level process for organizations to migrate from traditional data governance to next generation data governance (Soares, 2018). This maturity model includes setting goals associated with clear business outcomes that can be communicated to executive leadership; ensuring "enablers" including having the right organization structure and awareness to support data stewardship, risk management, and policy; establishing core disciplines including data quality management, information lifecycle management, and information security and privacy; and finally establishing the supporting disciplines of data architecture, classification and metadata, and audit information, logging, and reporting. As an organization becomes more able and develops capabilities within the core and supporting disciplines, the further they progress towards more modern data governance.

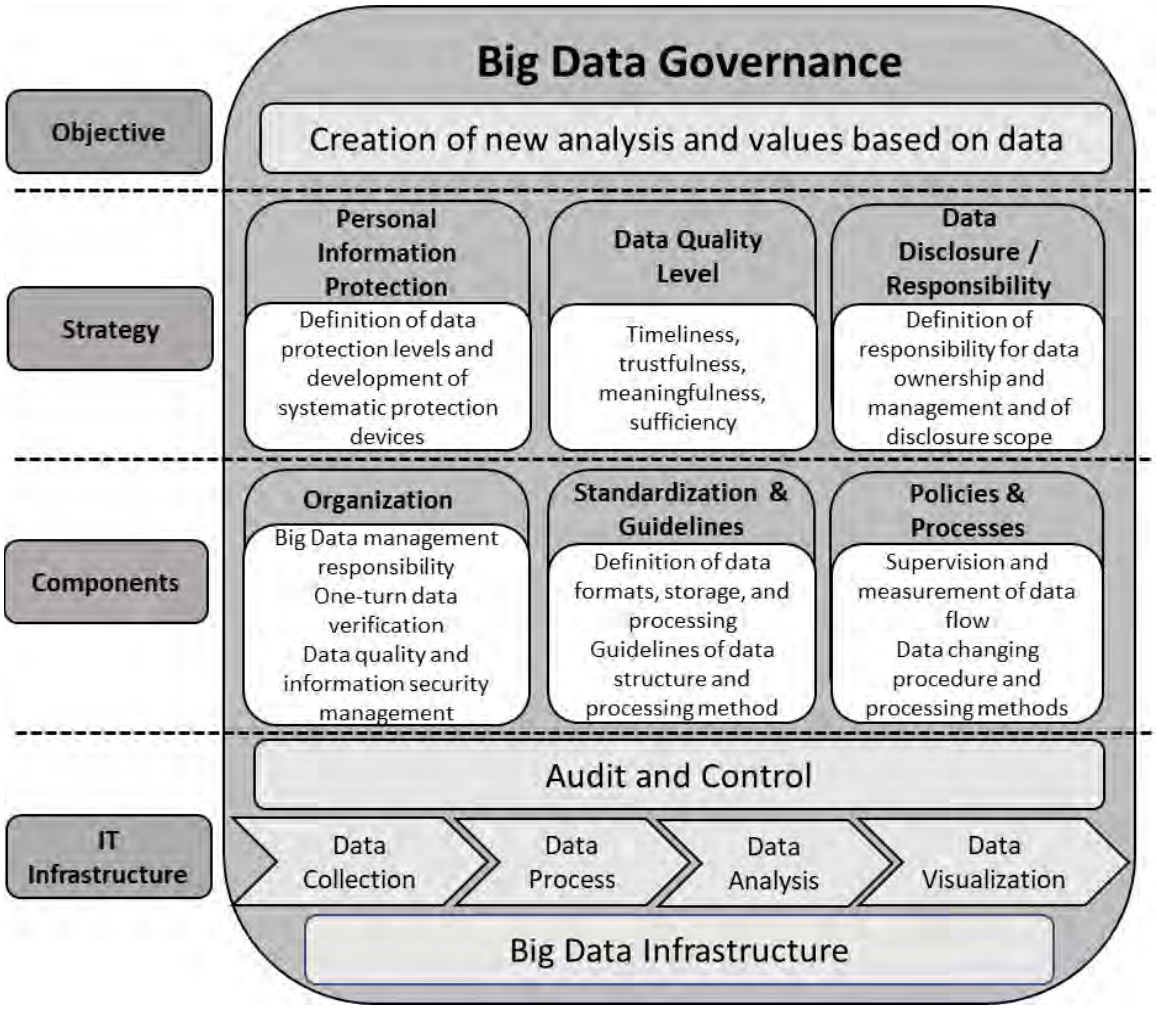


Figure 14. Big Data Governance Framework (Kim & Cho, 2018)[8]

[8] Available for use under the Creative Commons License: https://creativecommons.org/licenses/by/4.0/. Image was recreated for readability.

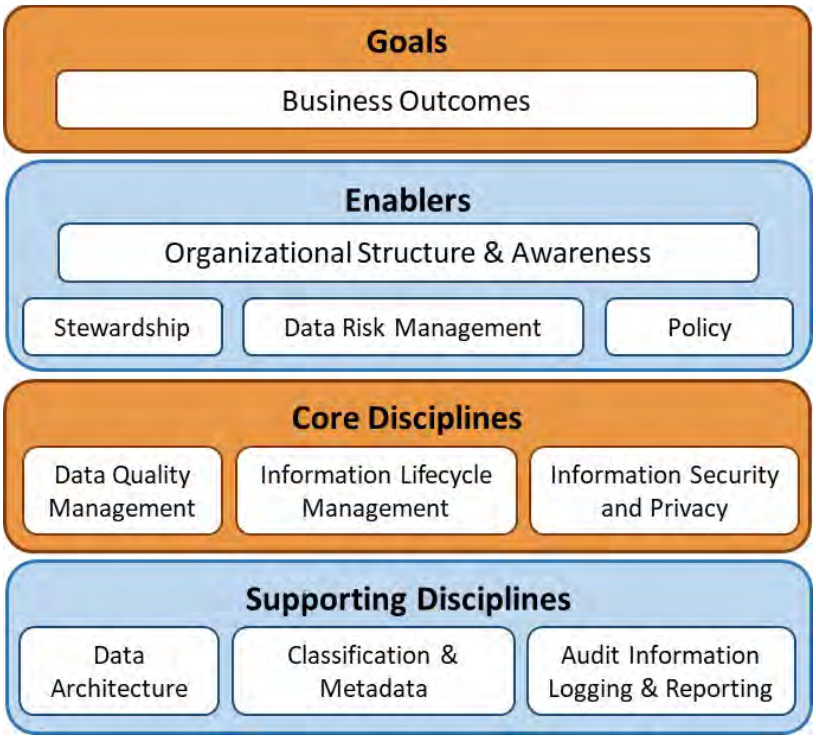


Figure 15. IBM Information Governance Council Maturity Model (Soares, 2018)

Roles

The commoditization of data and data analysis tools has fostered the adoption of self-service data preparation and analysis, where data tasks that were traditionally handled by an expert statistician or data analyst are now performed directly by a variety of end-users using visual and code-less tools requiring less technical expertise. To accommodate this move towards a distributed use of data, a distributed form of data governance has been adopted by many organizations. This approach builds on the concept of data governance roles, adding new roles that best support an expanded community of data users.

Following is a brief list of data governance roles that are commonly found in a distributed data governance model:

- Traditional roles:
  - Data Owner –Responsible for data access and administrative controls
  - Data Steward – Responsible for data quality and meaning
  - Data Custodian – Responsible for IT tasks and technical controls
- Additional roles for distributed data:
  - Data Curator – responsible for cataloging and describing datasets
  - Data Coach – Responsible for training and assisting data users

In the literature there are some who advocate for tracking additional data governance roles, such as data sponsors, data users, and data stakeholders (Wells, The Path to Modern Data Governance, 2019). Some agencies may find tracking additional roles, or even inventing new ones, to be useful. The key is to maintain a clear record of specific responsibilities without adding so many roles as to create unnecessary confusion or overhead. For most organizations, especially those organizations that are

adopting a distributed data governance approach for the first time, it is recommended to begin by focusing only on roles that are well known and well defined in the literature.

Data Governance Tracking Tool

To assist with assigning and tracking data governance roles two template forms are included herein. The first "information gathering" form (Table 21) is best used when determining roles for a given dataset. This form begins with basic identifying information including the name of the dataset, its logical storage address, a description of the data, and what potentially sensitive information the dataset contains. It then provides a space that lists each data role, provides a description of the role including what personnel typically take on that role, and a space to put the name of the organization member in that role.

Table 21. Information Gathering Form

Data Name	**Live Traffic Feed**
Data Location	Z:/DataLake/LiveFeeds/Traffic_XML/
Data Description	XML data pulled from roadside sensors every 10 seconds
Data Sensitivity	No sensitive information or PII

Data Governance Roles

Name of Role	Description of Role	Personnel Filling Role
Data Owner	Exercises administrative control over the data. Concerned with risk management and determining appropriate access to data. This role is typically filled by the most senior executive within the division that controls, created, or most often uses the data.	
Data Steward	Ensures the quality and fitness of the data. Concerned with the meaning and correct use of data. This role is typically filled by a division SME with domain knowledge relevant to the data or a member of the data team.	
Data Custodian	Exercises technical control over the data. Concerned with implementing safeguards, managing access, and logging information. This role is typically filled by IT personnel, such as system or database administrators.	
Data Curator	Manages the inventory of datasets. This includes cataloging the data, maintaining descriptions for the data, and recording the data utility. This role is typically filled by senior IT personnel or a member of the data team.	
Data Coach	Collaborates with business data users to improve skills and promote data utility. This role is typically filled by a member of the data team, or by a data SME within a division.	

The second "information cataloging" form (Table 22) collects the data roles for each dataset and condenses them into a single spreadsheet. This format allows executives or data team members to see at a glance all of their datasets and the relevant personnel associated with each.

Table 22. Information Cataloging Form

Data Name	Location	Description	Sensitivity	Owner	Steward	Custodian	Curator	Coach
Live Traffic Feed	Z:/DataLake/LiveFeeds/Traffic_XML/	XML data pulled from roadside sensors every 10 seconds	No sensitive information or PII					
Traffic Incident Performance Measures	Z:/DataLake/Historical/TIMPMs/	Traffic Incident data reported by responders	PII: License Plates					

Data Sources Catalog Tool

It is strongly recommended that transportation agencies periodically assess what data sources are in use and what data sources are available to be used. Not only does this help prevent an agency from overlooking data sources that could be vital to current or future projects, it also provides a better understanding of how datasets are connected to support the creation of a metadata catalog, planning for storage, development of new data pipelines, and better organization of an agency data lake structure. Maintaining a detailed catalog of data sources is one of the first and best ways to understand the nature of an agency's data and guide the development of the data analytics processes that can be built on it.

The provided herein is an example of how to structure a data sources catalog to summarize the specificities of each data source into a single table (see example in Table 23). To best review and assess the needs of available data sources, each data source is represented by its own row with columns briefly describing the various facets of the data source. Should more detailed information be needed, it is recommended to add this information to an appendix to preserve readability of the catalog. Below is a list of facets (table columns) that could be used to describe each data source, along with example entries for each:

- **Data Source** – The name of the data source.
  - Examples – "Local PD incident data," "Traffic light sensor data," etc.
- **Description** – Additional distinguishing details about the data source.
  - Examples – "Traffic incident performance measures from 2015 onward," "signal data from all intersections in the downtown area," etc.
- **Ownership** – Who owns or ultimately controls the data. This could be recorded as simply internal versus external, or the actual owners can be listed by name.
  - Examples – "internal," "external," "Vendor A," "FHWA," etc.
- **File format** – The format that the data are provided in. This is typically an open or closed file format but can also be an API or online dashboard when working with third-party vendors.
  - Examples – "csv," "xml," "json," "pdf," "API," "web-based report," "proprietary data format," etc.
- **Size** – How much capacity is required to store the data. This can be represented in terms of total storage used and/or how much additional storage is required per month depending on the nature of how the data source.
  - Examples – "10TB total," "5GB per month," "300MB daily," "600MB + 50M per month," etc.
- **Cost** – How much does it cost to use the data. For external data sources this is simply the amount charged by the vendor. For internal data sources this number represents various upkeep costs to process and manage the data.
  - Examples – "$500 per month," "$15 on average per day," "$5000 upfront and $250 per month for 4 years," etc.
- **Security level** – The level of security called for by the data. The exact classifications may vary between organizations but at a minimum this should include four basic levels: "PII" to identify the presence of PII that must be anonymized, "PII-possible" to identify data that is not

identifiable by itself but may become identifiable if combined with other data, "sensitive" to identify data that otherwise requires special attention or care, or "standard" for data that only calls for the standard level of security and encryption.
    - Examples – "PII," "PII-possible," "sensitive," "standard," "top secret," "secret," "confidential," etc.
- **Granularity** – How granular or specific the data are. Typically, the lowest level of granularity is having each individual item or event be represented by one row or record in the data. As data are aggregated each row/record may represent a group of many individual items/events, which affects how the data can be used and combined with other datasets.
    - Examples – "1 row per incident," "1 row per city block," "10 families per record," "incidents aggregated within 2-mile road segments each hour," etc.
- **Restrictions** – What restrictions are in place in how the data can be shared or used. Most commonly these restrictions are found with external data sources whose contracts limit how the data can be used. However, these can also apply to internal data sources that use proprietary or restrictive file formats.
    - Examples – "cannot distribute," "no access to raw data," "proprietary data format," "only usable with software from Vendor A," "limited to authorized users only," etc.
- **Update Frequency** – How often the data are updated. Most streaming data are updated in near real-time, while non-streaming data may or may not have a set update schedule in place.
    - Examples – "true real-time," "near real-time," "monthly," "weekly," "daily," "hourly," "upon request," "no longer updated," etc.
- **Projects** – A list of current or potential projects for which this data source could be useful. Very large organizations with many projects may find it helpful to separate this into two different columns for better visibility: one column for projects that currently use the data source and another for projects that could potentially use the data.
    - Examples – "In use by project A," "potential use for project B," "evaluation in progress for projects C and D," "vital component of system A," "necessary for monthly newsletter," etc.
- **Last Reviewed** – The date when this data source was last reviewed. This field is useful both for timing regular reviews as well as identifying at a glance whether new projects were created before or after the most recent data source review. A new project being created that could potentially use a data source is a good reason to perform a new review of that data source.
    - Examples – "2019-10-01," "Q3 2019," "05/10/2019," etc.

Table 23. Data Source Assessment Example

Data Source	Description	Ownership	Format	Size	Cost	Security Level	Granularity	Restrictions	Update Frequency	Projects	Last Reviewed
Waze Incidents	Traffic speeds based on GPS probe data	Internal	XML	2.1 TB total	$70,000 /year	Proprietary	Predefined roadway segments	Cannot share without permission	1 minute	Work Zones, Signal Timing	03/12/2019
Snowplow AVL	Probe data from snowplows	Internal	REST API	4 TB total	$4 /truck	No PII	0.01 mile point	None	1 minute	DOTPJ, Work Zones	01/15/2019
CoCoRahs	Certified crowdsourced weather reports	CoCoRahs Network	XML	380 MB total	Free	No PII	Interpolated from number of reports	None	24 hours	SNIC, Possibly DOTPJ	04/03/2019
Incident Reports	Individual incident reports collected from participating local agencies	Internal	CSV	500MB total	$15 /month	Sensitive	1 row = 1 incident	None	Monthly batch upload	A-110, possible use in A-123	02/22/2019

Frequently Asked Questions (FAQ)

Q. What exactly is big data?

A. "Big data" is more than a catch phrase. At its core, big data is a set of concepts and methodologies that allow for the storage, processing, management, and analysis of extremely large, diverse, and fast-changing datasets. As these datasets differ greatly from traditional datasets in terms of their volume, variety, and velocity, they require new and powerful ways of dealing with the data. Multiple definitions for big data are provided on page 17 of the guidebook. Table 1 on page 6 of the guidebook contrasts the fundamental differences between the traditional data systems and management approach of most transportation agencies and the modern big data approach that will be needed to effectively management data from emerging technologies.

Q. Why do we need big data?

A. Data from emerging technologies have tremendous potential to offer new insights and to identify unique solutions for delivering services, thereby improving outcomes. However, the volume and speed at which these data are generated, processed, stored, and sought for analysis is unprecedented and will fundamentally alter the transportation sector. With increased connectivity among vehicles, sensors, systems, shared-use transportation, and mobile devices, unexpected and unprecedented amounts of data are being added to the transportation domain, and these data are too large, too varied in nature, and will change too quickly to be handled by traditional database management systems. ***As such, modern big data methods to collect, transmit/transport, store, aggregate, analyze, apply, and share these data at a reasonable cost need to be accepted and adopted by transportation agencies if they are to be used to facilitate better decision-making***.

Q. Do other local or state agencies use big data?

A. Yes! And you can learn from their experiences. This guidebook references several agencies that have successfully transitioned or applied big data architectures and methodologies in different capacities with differing levels of success.

Q. How will this guidebook help my agency?

A. This guidebook provides guidance, tools, and a big data management framework, including over 100 recommendations, and it lays out a roadmap for transportation agencies on how they can begin to shift – technically, institutionally, and culturally – toward effectively managing data from emerging technologies. The guidebook will help transportation agencies identify a jumping-off point for managing big data, as well as a step-by-step process for gradually/incrementally building towards organizational change. Figure 1 and the associated discussion on page 2 of this guidebook can help an agency understand where they might begin to apply the guidance and tools provided in this guidebook.

Q. What is a data lake and how does it relate to big data?

A. Simply put, a data lake is a location where raw, unprocessed data are stored in their native form and organized to be subsequently accessed and used by various entities within an organization. Data lakes are simple and similar to a very large folder structure where data files are collected. They are meant to store data as long as possible and at a low cost allowing for the collection of all generated

data and the creation of very large data archives. Data stored in data lakes are available to data users in a read-only format to help guarantee that the original data will never be altered or modified from their original state. Data lakes also allow for the data to be used by many users at once even if they use very different analytical tools. This is a contrast from traditional data workflows where requirements define what data should be collected and how they should be modified and stored in order to support predefined analysis tasks. When using data lakes, agencies are now able to capture and store raw and unfiltered data and then explore the data and develop multiple use cases to support different areas of the organization. Indeed, within a raw dataset stored in a data lake, cleaned data from the dataset may be of interest to one business unit, outliers from the same dataset may be of interest to another, and only a few fields from that dataset may be of interest to another. The data lake allows for each business unit to use the same raw data independently from each other and shape them to the specific needs of their applications, business intelligence tools, and/or static reports.

Q. What skills are required for big data?

A. Typically skills required for big data are knowledge in programming (Python, Java, Scala, Go); modern data warehousing (data lake management); big data computation frameworks such as Map Reduce, Hadoop or Apache Spark; knowledge of statistics and linear algebra (summary statistics, probability distribution, hypothesis testing, etc.); and last but not least business domain knowledge to have a good understanding of what hides behind the data. The skill levels can vary greatly depending on the complexity of the big data analysis to be undertaken. Additional skills may also be needed depending on an agency's approach to big data. If the big data solution is implemented on-premise, it will require a much higher level of technical expertise than a cloud solution implementation. On-premise implementations require expertise in the development and management of the hardware and software of very large server clusters, and this expertise is not easily found or affordable. As such, on-premise big data implementations are known to be difficult, tedious, time consuming, and costly to implement. Cloud implementations do not require the acquisition of such skills, as they fall under the responsibility of the cloud provider, leaving agencies with only the need to acquire big data expertise.

Q. Can I do this on-premise?

A. Yes, but this approach is not advised. Not only does an on-premise implementation require a great deal of expertise, it also generally requires multiple people to implement and maintain. System administrators will need to know how to administer a very large cluster of commodity servers and deal with constant failure and optimization. Not only will developers need to know languages such as Java and Scala and a variety of distributed computing frameworks such as Kafka, Spark, and Hadoop, they will also need to know how to tune them so the performance of processing jobs they developed remain acceptable. Instead, using cloud services and architecture is a recommended best practice for big data.

Q. Why would I move to the cloud?

A. Cloud solutions were developed to allow organizations to benefit from large computing resources without having to bear the cost on their own. Indeed, big data projects could far exceed the organization's entire annual budget if developed and managed on-premise. This is due to the fact that big data projects require large bursts of computing power for short periods of time, which when implemented on-premise lead to the design of very large clusters that are seldomly used to their full potential. Cloud solutions solve this problem by adopting a shared server cluster model to maximize

its use. To allow the use of this shared cluster of servers, cloud providers have done the heavy lifting to make the accessibility of data storage and processing available through automation or easy to use APIs that can be leveraged with more common scripting languages such as Python. This greatly reduces the amount of time and resources spent on maintaining technology and allows for more time and resources to focus on deriving a better understanding of the organization and its operations from the data.

Q. When would I need to use machine learning algorithms?

A. Machine learning algorithms are a subset of data science techniques that use a multi-step machine learning process to perform advanced analyses. Because these applications rely on a huge amount of unstructured data to be effective, they are not something that most transportation agencies will need to concern themselves with until they have built a very mature set of big data management approaches. It will be more effective for agencies to focus first on using the guidance in this document to collect large amounts of data that are properly cleaned, stored, enriched, analyzed, and visualized before diving into deep learning.

That being said, once a strong data management foundation is in place and appropriate data have been acquired, deep learning algorithms can be used to support computer vision applications, classify existing data, and predict the attributes of future data. Computer vision, where a machine can be trained to distinguish and classify objects in image and video files, can be useful in turning roadside camera recordings into traffic observations or passenger records.

Q. What is data governance?

A. The DAMA Dictionary of Data Management defines governance as "the exercise of authority, control, and shared decision-making (e.g., planning, monitoring, and enforcement) over the management of data assets" (DAMA International, 2011). Data governance is a collection of practices and processes that help to ensure the formal management of data assets within an organization, including the planning, oversight, and control over management of data and the use of data and data-related resources. Data governance puts in place a framework to ensure that data are used consistently and consciously within the organization. It also deals with quality, security and privacy, integrity, usability, integration, compliance, availability, roles and responsibilities, and overall management of the internal and external data flows within an organization (Roe, 2017).

Q. What is big data architecture?

A. Big data architecture is the overarching system definition that an organization uses to build its big data environment and steer its data analytics work. Big data architecture is the foundation for big a data environment and consists of four logical layers:

- Big data sources layer
- Data massaging and storage layer
- Analysis layer
- Consumption layer

In addition to the logical layers, four major processes operate cross-layer in a big data environment (Taylor, 2017):

- Data source connection
- Governance (privacy and security)

- Systems management (large-scale distributed clusters)
- Quality control

Q. When do you know it is time to begin working with big data?

A. For many agencies the best time to start working with big data was probably a decade ago; the second-best time to start is today. There are two reasons to begin adopting modern data management practices: when new approaches will reduce costs or improve efficiencies, or when a new use case or application is identified that requires them. Often, migrating from siloed data storage to a cloud-based data lake alone will result in enough workflow improvements and cost reductions to make the pursuit of big data management worthwhile.

Any agency that is unsure if it is the right time to modernize their data management approaches may be well served by having a small team review current practices to identify potential cost savings from adopting new approaches. This same team may also review available data sets or recent big data-enabled achievements from their closest peers to see if they could benefit from pursuing new data products. The accompanying Data Management Capability Maturity Self-Assessment (DM CMSA) tool can be useful in identifying areas of improvement in data management, while the Data Sources Catalog Tool can be useful to identify the potential of new and existing data sets.

Q. How do we ensure that the third parties we work with are keeping the data secure?

A. When working with third-party data providers the best way to ensure strong data security practices is to incorporate clear requirements into the negotiated contract. These requirements should be flexible enough to accommodate for updated technology while including specific requirements that leave no room for ambiguity or loopholes. The agency must also have some means of monitoring for non-compliance so that these requirements can be effectively enforced.

When working with cloud service providers an agency may not have the leverage or opportunity to include data security enforcement clauses in the contract. If this is the case, then the next best option is to fully research and understand the standard security measures being used by the provider. Fortunately, due to the economies of scale and the vital importance of trusted data to their business model, nearly all major cloud service providers maintain standard levels of security that exceed those found at most transportation agencies.

Q. Some data management frameworks include a step where data are destroyed, where old or seldomly accessed data are deleted to preserve space in the system. Why is that step not included in the data management framework in this guidance?

A. Traditionally, there was a focus placed on managing free space on a data storage system in order to avoid unnecessary costs. One of the benefits of modern data management approaches is that managing free space on a server is handled automatically. For example, most modern cloud storage providers will monitor how frequently data are accessed and automatically migrate data that are seldomly used to archival storage in a process that is transparent to the end-user. This obviates the need to hire data maintenance workers to manually monitor and move data from active storage to archival storage.

Furthermore, the costs and benefits of storing unused data have changed. In traditional data storage models the server costs are the same whether the data are accessed or not. In modern use-based fee models that are common among cloud providers any data that are seldomly accessed are less

expensive to store, making long-term storage of unused data more feasible. With the advent of big data analytical techniques that can turn large amounts of seemingly uncorrelated data into actionable insights, the potential value of collected data is generally higher as well.

Because data are generally more useful to retain, less expensive to store, and easier to archive, there is no longer the need to expend as much energy on purging data. There are exceptions to this, where data are sufficiently large or expensive that long-term storage becomes unfeasible. These exceptions are rare enough that it is generally advised to preserve data as much as possible and any thought given to destroying data is not sufficient to merit the inclusion of a "destroy" step in the framework.

Q. What does it mean to be data-driven in a big data environment?

A. To be data-driven means that progress in an activity is compelled by the data itself, not by intuition, personal experience, or political agenda. While transportation agencies have been using data to make "informed" decisions for many years, big data are too large, fast, and change too quickly to be processed and understood by humans for informed decision-making. Through the processing of integrated and complex datasets, big data methodologies can provide decision-makers with more detailed, intricate, and timely outputs from which to base their decisions, which simply cannot be offered with the siloed nature of transportation agency data today.

Q. Is the flood of big data really coming?

A. It is already here! Private industry has been using and generating data at an increasingly rapid rate for several years now. Some transportation agencies may be waiting for the flood of data to hit them before they pursue modern data management practices. More commonly, it is the case that agencies do not recognize or pursue a new data opportunity and thus are never forced to act; they simply lose the chance to exercise control over the data as private industry fills the gap.

One example of this is online 511 systems. Private sector offerings like Waze and Google Maps have exceeded the capabilities of most, if not all, public transportation agencies. Had these agencies developed high quality online 511 systems in the first place, they could have exerted beneficial control over them, such as not posting the location of law enforcement vehicles which can impact the safety of officers. Now that the market share has become dominated by private sector offerings that window of opportunity may be difficult for any agency to re-open.

If agencies do not invest in their own data systems, they will be forced to pay third-party vendors if they ever want to use emerging technology data. These vendors may feel free to charge whatever rates they choose if they perceive they are working with an agency that has no other options available to them.

Q. Who determines the quality of the data?

A. When obtaining data from third-party sources, it is recommended to include minimum expected levels of data quality in the negotiated contract, along with clear repercussions for failing to meet these expected data quality levels. The contracting agency should then employ some base level of in-house data expertise to where they can independently verify the quality of the data coming in. With such an arrangement the third-party data provider is responsible for providing high quality data while the contracting agency takes on responsibility for validating the quality of the data coming in.

When dealing with internal data the dynamics are the same; the data creator or data pipeline owner is responsible for the quality of the data, which is independently verified. Effective data quality

verification processes employ both an automated validation process and periodic manual reviews. Because different applications require different levels of data quality it is recommended that low quality data be flagged with a data quality score rather than discarded entirely. This allows data analysts to make an informed decision as to what data sets are of sufficiently high quality to be included in any given analysis.

Q. What can be done if an agency is unable to obtain any big data?

A. Even if an agency does not have the inclination or resources to pursue large datasets immediately there are still benefits to modernizing data management practices. Advice contained in this guidebook relating to eliminating data silos, implementing data quality management procedures, and creating effective data product development practices can be useful even when only working with smaller, more traditional data sources. Following such guidance wherever it is relevant may not only improve the handling of traditional data sources but will also help prepare an agency for managing big data if they decide to pursue such datasets in the future.

Q. What is the value proposition in sharing data with others?

A. Most of the impetus behind open data policies and sharing data with external stakeholders is to foster innovation and promote development of data products that may or may not directly benefit the agency sharing the data. That said, there are situations where cost sharing may be appropriate. If two or more agencies collaborate on a single project, they may agree to equally contribute to the project's development.

There may also be situations where a transportation agency that is providing data to a partner may gain compensation or control from the arrangement. For example, one transportation agency shares data with private Mobility on Demand companies through a live API; however, that API will not respond to location requests sent inside of public parks. By sharing data in this way, the agency was able to gain a measure of control over private industry behavior that benefitted the public they serve.

Q. What formats and standards are used when sharing data?

A. Using open source data formats is strongly recommended whenever data are shared across internal applications or shared with external users to ensure that the data can be applied to multiple use cases without transformation. For example, the XLSX format is specific to the Microsoft Excel application, so spreadsheet data shared in this format will only be accessible to users who have licensed Microsoft Excel. If that spreadsheet data are instead shared in the open source CSV format then it can be accessed by far more applications, and therefore reach a wider audience. Other examples of commonly used open source data formats include JSON, GeoJSON and KML. Several successful open data platforms share data in multiple data formats, allowing the users to select the format that works best for them.

Consideration should also be made for the audience of the data and the intended use. Data aimed at a non-technical audience for manual review may be best delivered as an interactive web-based visualization, where data provided to partner companies for real-time application queries are better served through an API. Regardless of the delivery system that is used, the associated metadata and retrieval processes ought to be documented as well as possible to minimize the number of questions that must be fielded by department personnel over the lifetime of the data sharing system.

WORKS CITED

Big Data. (2019). Retrieved November 2019, from Lexico Powered by Oxford: https://www.lexico.com/en/definition/big_data

Big Data. (2019). Retrieved from Dictionary.com: https://www.dictionary.com/browse/big-data

Big Data Wikipedia Entry. (2019, 11). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Big_data

Burt, M., Cuddy, M., & Razo, M. (2014). *Big Data's Implications for Transportation Operations: An Exploration.* Washington, DC: US Department of Transportation. Retrieved January 2019, from https://rosap.ntl.bts.gov/view/dot/3542

Cavanillas, J. M., Curry, E., & Wahlster, W. (2016). *New horizons for a data-driven economy: a roadmap for usage and exploitation of big data in Europe.* Springer.

Dam, R. F. (2019). *The Pareto Principle and How to Be More Effective.* Retrieved March 24, 2020, from Interaction Design Foundation: https://www.interaction-design.org/literature/article/the-pareto-principle-and-how-to-be-more-effective

DAMA International. (2011). *The DAMA Dictionary of Data Management, 2nd Edition: Over 2,000 Terms Defined for IT and Business Professionals.* Bradley Beach, New Jersey: Technics Publications, LLC.

DAMA International. (2017, March 6). *DAMA Data Management Body of Knowledge Framework 2nd Edition (DAMA-DMBOK2).* Basking Ridge, NJ: Technics Publications. Retrieved from DAMA International: https://dama.org/sites/default/files/download/DAMA-DMBOK2-Framework-V2-20140317-FINAL.pdf

de Ternay, G. (2018, December 30). *Convince Your Boss: 11 Tips to Make Them Say "Yes!".* Retrieved November 2019, from Guerric: https://guerric.co.uk/convince-your-boss/

Demchenko, Y., Canh, N., de Laat, C., Membrey, P., & Gordijenko, D. (2013, August 30). *Big Data Security for Big Data: Addressing Security Challenges for Big Data Infrastructure.* Retrieved September 5, 2018, from Semantic Scholar: https://pdfs.semanticscholar.org/184b/b798f1f298e158fdcfe753559bbad65184ab.pdf

Ellingwood, J. (2016, 9 28). *An Introduction to Big Data Concepts and Terminology.* Retrieved November 2019, from DigitalOcean: https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology

Entry on Reference Data. (n.d.). Retrieved 12 2019, from Wikipedia: https://en.m.wikipedia.org/wiki/Reference_data

Gandomi, A., & Haider, M. (2015, April). Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management, 35*(2), 137-144. Retrieved from https://www.sciencedirect.com/science/article/pii/S0268401214001066

Gettman, D., Toppen, A., Hales, K., Voss, A., Engel, S., & El Azhari, D. (2017). *SubtitleIntegrating Emerging Data Sources into Operational Practice—Opportunities for Integration of Emerging Data for Traffic Management and TMCs.* Washington, DC: U.S. Department of Transportation.

Hand, A. (2016). *Urban Mobility in a Digital Age.*

Kim, H. Y., & Cho, J.-S. (2018). Data Governance Framework for Big Data Implementation with NPS Case Analysis in Korea. *Journal of Business and Retail Management Research*. Retrieved April 2020, from https://jbrmr.com/cdn/article_file/content_24232_18-04-20-02-28-48.pdf

Llewellyn , R. (2015, September 15). *20 Ways to Create a Sense of Urgency.* Retrieved November 2019, from The Enterprisers Project: https://enterprisersproject.com/article/2014/8/20-ways-create-sense-urgency?page=1

Marr, B. (2017, January 23). *Really Big Data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud.* Retrieved Novemer 2019, from Forbes: https://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/#1b22eb76c105

Mukherjee, S., & Shaw, R. (2016). Big data - concepts, applications, challenges, and future scope. *International Journal of Advanced Research in Computer and Communication Engineering*, 66-74.

OECD/ITF. (2015). *Big Data and Transport Understanding and Assessing Options.* Paris: OECD Publishing. Retrieved January 2019, from https://www.itf-oecd.org/sites/default/files/docs/15cpb_bigdata_0.pdf

Pecheux, B., Shah, V., & Miller, S. (2019). *NCHRP Research Report 865: Guide for Development and Management of Sustainable Enterprise Information Portals.* Transportation Research Board, Washington, DC.

Pecheux, K., Pecheux, B., & Carrick, G. (2019). *NCHRP Research Report 904: Leveraging Big Data to Improve Traffic Incident Management.* Transportayion Research Board, Washington, DC.

Portland Urban Data Lake (PUDL). (n.d.). Retrieved 11 2019, from Portland Bureau of Transportation: https://www.portlandoregon.gov/transportation/article/681572

Press, G. (2014, 9 3). *10 Big Data Definitions - What's Yours?* Retrieved November 2019, from Forbes: https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/#63db4d4413ae

Roe, C. (2017, December 18). *What is Data Governance?* Retrieved from Dataversity: http://www.dataversity.net/what-is-data-governance/

Rouse, M. (2013, October). *Big Data Management*. Retrieved from Search Data Management: https://searchdatamanagement.techtarget.com/definition/big-data-management

Soares, S. (2018, 8 15). *Big Data Governance: A Framework to Assess Maturity.* Retrieved from IBM Corporation: https://www.ibmbigdatahub.com/blog/big-data-governance-framework-assess-maturity

Swaney, R. (2019, 8 22). *Evolution of Data Management: The Role of Streaming Data and IoT Data Architecture.* Retrieved from Cloudera: https://blog.cloudera.com/evolution-of-data-management-the-role-of-streaming-data-and-iot-data-architecture/

Taylor, C. (2017, June 8). *Big Data Architecture.* Retrieved January 22, 2019, from Datamation: https://www.datamation.com/big-data/big-data-architecture.html

Turner, P. (2019, Feb 18). *A Data Scientific Method.* Retrieved Nov 2019, from Towards Data Science: https://towardsdatascience.com/a-data-scientific-method-80caa190dbd4

Veracity: The Most Important "V of Big Data. (2019, August 29). Retrieved November 2019, from Gut Check: https://www.gutcheckit.com/blog/veracity-big-data-v/

WBCSD. (2019). *Enabling Data Sharing: Emerging Principles for Transforming Urban Mobility.* Geneva: World Business Council for Sustainable Development. Retrieved March 2020, from https://docs.wbcsd.org/2020/01/WBCSD_Enabling_data_sharing_Emerging_principles_for_transforming_urban_mobility.pdf

Wells, D. (2017, January 17). *The Next Generation of Data Governance.* Retrieved January 22, 2019, from Eckerson Group: https://www.eckerson.com/articles/the-next-generation-of-data-governance

Wells, D. (2019, August 14). *The Path to Modern Data Governance.* Retrieved from Eckerson Group: https://www.eckerson.com/articles/modern-data-governance-problems

What is Big Data. (2019). Retrieved November 2019, from SAS: https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Yang, C., Huang, Q., Li, Z., Liu, K., & Hu, F. (2017). Big Data and cloud computing: innovation opportunities and challenges. *International Journal of Digital Earth*, 13-53.

Note: additional references reviewed and cited as part of the research can be found in the associated final research report: NCHRP 08-116 Framework for Managing Data from Emerging Transportation Technologies to Support Decision-Making. These resources may be of interest to the reader.