NCHRP

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

# Guidebook for Managing Data from Emerging Technologies for Transportation

## EXECUTIVE SUMMARY

### NCHRP Research Report 952

*The National Academies of*
SCIENCES · ENGINEERING · MEDICINE
TRB
TRANSPORTATION RESEARCH BOARD

100 YEARS
MOVING IDEAS: ADVANCING SOCIETY

## AUTHORS

Kelley Klaver Pecheux
Benjamin B. Pecheux
Gene Ledbetter
Chris Lambert

*AEM Corporation*
*Herndon, Virginia*



## ACKNOWLEDGMENTS

J. D. Schneeberger and John Hicks of Noblis, Inc., Washington, D.C., and Brian Burkhard and Mara Campbell of Jacobs, San Francisco, California, and New Florence, Missouri, respectively, contributed to the report.





Research results for NCHRP Project 08-116, "Framework for Managing Data from Emerging Transportation Technologies to Support Decision-Making," have been published as *NCHRP Research Report 952: Guidebook for Managing Data from Emerging Technologies for Transportation*. The report is available for download from TRB's website at www.TRB.org by searching on *NCHRP Research Report 952*. The contractor's final report, *NCHRP Web-Only Document 282: Framework for Managing Data from Emerging Transportation Technologies to Support Decision-Making*, may also be downloaded from the website.

# INTRODUCTION, BACKGROUND, AND OVERVIEW OF THE GUIDEBOOK

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. Further developments in communications and technology have recently led to more advanced infrastructure and vehicle capabilities, mobile applications, and a host of mobility service offerings. These include 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.

*As such, the fundamental purpose of the guidebook is to help agencies begin to shift from their traditional data systems and management practices to more modern big data systems and management practices to make effective use of data from emerging technologies.* Table 1 contrasts, at a high level, 11 characteristics of traditional data management practices with their big data management counterparts. The table provides examples that demonstrate the stark contrast between the current state of the practice for most transportation agencies and the ideal state based on data industry best practices.

**Data from emerging technologies have tremendous potential to offer new insights for the transportation industry; however, 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. As such, modern big data methods to collect, transmit, store, integrate, analyze, apply, and share these data need to be accepted and adopted in order to reap the true benefits from these data.**

**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 pre-defined purpose; all requirements must be pre-determined before development and deployment. | *versus* | 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. | *versus* | 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. | *versus* | System features at the software level; hardware and software decoupled. |
| 4 | **Hardware Longevity** | As technology evolves, hardware becomes outdated quickly; system cannot keep pace. | *versus* | As technology evolves, hardware is disposable; system changes to keep pace. |
| 5 | **Database Schema** | Schema on write ("schema first"). | *versus* | Schema on read ("schema last"). |
| 6 | **Storage and Processing** | Data and analyses are centralized (servers). | *versus* | Data and analyses are distributed (cloud). |
| 7 | **Analytical Focus** | 80% of resources spent on data design and maintenance; 20% of resources spent on data analysis. | *versus* | 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). | *versus* | 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). | *versus* | Data governance is distributed between a central entity and business areas; data are open to many users. |
| 10 | **Data** | Uses a tight data model and strict access rules aimed at preserving the processed data and avoiding its corruption and deletion. | *versus* | 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." | *versus* | Many people can access the data; applies the concept of "many eyes" to allow insights and decision-making at all levels of an organization. |

*The 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:*

- New concepts and methodologies concerning modern data management and use are introduced, along with industry best practices and more than 100 associated recommendations for managing data in a modern, flexible, scalable, and sustainable way.
- An 8-step process and associated guidance is provided 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.
- More than 100 questions across 15 data management focus areas are included as part of a Data Management Capability Maturity Self-Assessment (DM CMSA), which will allow transportation agencies to gauge their data management practices as well as to identify specific areas for improvement.
- Examples and references are provided from transportation agencies that are currently exploring or are already 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.
- Big data governance recommendations, a description and framework for big data governance, and a tool for tracking the big data governance roles and responsibilities within an agency are included.
- A tool to help transportation agencies catalog existing and potential data sources is provided.
- Answers to frequently asked questions regarding big data implementation, management, governance, use, and security are provided.

Figure 1 shows the various components and supporting tools of the guidebook and illustrates how different agencies can implement them. In addition, *NCHRP Web-Only Document 282: Framework for Managing Data from Emerging Technologies to Support Transportation Decision-Making* is an online-only document provided as supplemental information for reference for agencies to support implementation of the guidebook. This report details the research conducted for the project, including findings from a literature review, surveys and interviews with state and local transportation agencies, emerging technology pilot project documentation reviews, an assessment of data management practices by transportation agencies, and a stakeholder workshop. The report provides background and supplemental details that the reader of the guidebook may find helpful, particularly during implementation of the roadmap and data management framework.

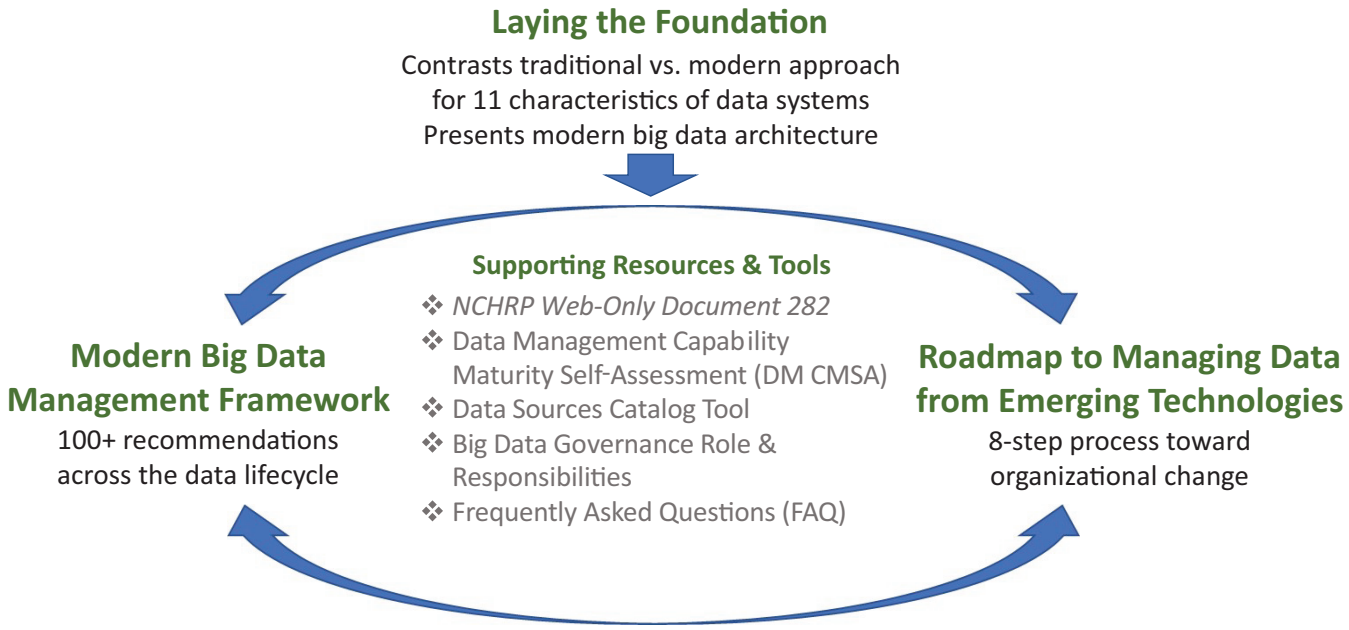



**Figure 1.** Components and application of the guidebook.

The guidebook is intended for engineering and information technology (IT) analysts and professionals working for state and local transportation agencies. The target audience is primarily those at the mid-level manager level, including positions such as operations manager, IT services program manager, traffic operations/traffic management center manager, IT manager, and database administrator, and it is the intent that these staff, in their positions, would play a primary role as champions in leading data initiatives for their agencies that support positive 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.

Whether an agency is starting from scratch with a new emerging technology data set, trying to make the business case for emerging technology data, has an issue or problem that might be solved with emerging technology data, is already working on a big data project, or is looking for a new enterprise data management solution, the steps and guidance are designed to walk them through the necessary data management policies, procedures, and practices to fully meet the needs of emerging technology data.

# ROADMAP TO MANAGING DATA FROM EMERGING TECHNOLOGIES FOR TRANSPORTATION

The steps and guidance outlined in this roadmap, in conjunction with the Modern Big Data Management Framework, can walk an agency through the process of developing the knowledge, projects, environment, and buy-in to grow iteratively and move incrementally from a traditional data management approach to establishing data management policies, procedures, and practices that 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. The roadmap 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. The roadmap includes eight steps. Figure 2 illustrates the steps.

**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.**


**Figure 2.** 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 in the roadmap 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 of the guidebook, and a variety of outside resources referenced at the end of Step 1. Step 1 includes guidance on the following:

- What is big data?
- Characteristics of big data.
- Concepts of big data.
- When to pursue big data.
- A common misconception regarding big data storage.
- A case study on the importance of understanding big data, per one transportation agency.

**The goal of Step 1 is to gain enough understanding to be able to lead the charge for a move toward a modern big data management approach within the organization.**

## STEP 2.

### Identify a Use Case and an Associated Pilot Project

Step 2 involves the champions and team identifying a use case and an associated pilot project for the data that will resonate with leadership. This use case is likely something that addresses the pain points of the group, division, or business unit and that cannot easily be addressed without the use of the data sets of interest. In some cases, a use case may be handed to the champions from the top down, with the charge of demonstrating value for a particular data set or project. Step 2 includes guidance on the following:

- Selecting a use case and pilot project that align with business unit, leadership, and organizational goals, including example drivers for change, example big data sources of interest, and associated example use cases and pilot projects.
- Engaging others in the cause, including those internal to the business unit, cross-business unit, junior and mid-level staff, and external partners.
- A case study on the Portland Urban Data Lake Pilot Project.

**The goal of Step 2 is to select and define a use case and pilot project that leadership will support.**

# STEP 3.

## Secure Buy-in from at Least One Person from Leadership for the Pilot Project

In Step 3, the champions and team work to communicate the value of the selected pilot project and to secure buy-in from at least one person from their leadership for the project. One champion from leadership can be key to ensuring success of the pilot and expansion to other groups, divisions, or business units within the agency. Step 3 includes guidance on the following:

- Establishing and communicating the value proposition for the pilot project, including example projects, value propositions, and questions to assist in developing the "pitch."
- Ways to create a sense of urgency and a fear of missing out.
- De-risking the decision by identifying and communicating risks and other potential barriers upfront.
- Knowing how and when to make the pitch.

**The goal of Step 3 is to obtain buy-in, support, and backing from leadership for the pilot project.**

# STEP 4.

## Establish an Embryotic Big Data Test Environment

Step 4 involves building an embryotic big data test environment or "playground," in which the pilot project can be developed and where there are little risks associated with the use of the data. This embryotic environment should be developed following as many of the big data best practices and recommendations identified in the Modern Big Data Management Framework section of the guidebook. Step 4 contains guidance on the following:

- Establishing buy-in from IT, including understanding potential challenges and barriers, as well as the pros and cons of on premise versus cloud storage.
- Establishing the playground, including both the data storage layer and the data processing layer.
- Taking ownership and responsibility for analytical projects.
- Common misconceptions on big data storage.
- A case study on storing data on premise versus in the cloud.

**The goal of Step 4 is to create a developmental and scalable platform that business units—through interaction and collaboration—can use to explore old and new data in a big data context using new data analysis tools and leveraging advanced analytical methods not in use by the business unit or the organization.**

# STEP 5.

## Develop the Pilot Project Within the Big Data Test Environment/Playground

In Step 5, the team develops the pilot project within the test environment with iterative feedback from the leadership champion. This development process requires the application of modern big data approaches and analytics and the development of data visualizations and products. Step 5 includes guidance on the following:

- Developing/ensuring the availability of the right expertise, including the pros and cons of various options (e.g., training/hiring in-house staff, trusted contractors and university partners, and big data experts/consultants).
- Developing the project by applying a data science perspective (e.g., collecting raw data, processing and cleaning the data, performing exploratory data analyses, and building data science pipelines).
- Iteratively developing and improving the project and the associated outputs/data products.
- Case studies on negotiating technical contracts for data services and building data knowledge.

**The goal of Step 5 is to develop the project to the point that it generates real value for the business unit.**

# STEP 6.

## Demonstrate the Value of the Data to Other Business Units


In Step 6, the team and the leadership champion begin to market the data visualizations and products developed in Step 5 to other business units across the organization. This horizontal organizational outreach will help to market the value of the data and the data products to identify other potential use cases and pilot projects that can be developed within the test environment. Step 6 provides guidance on the following:

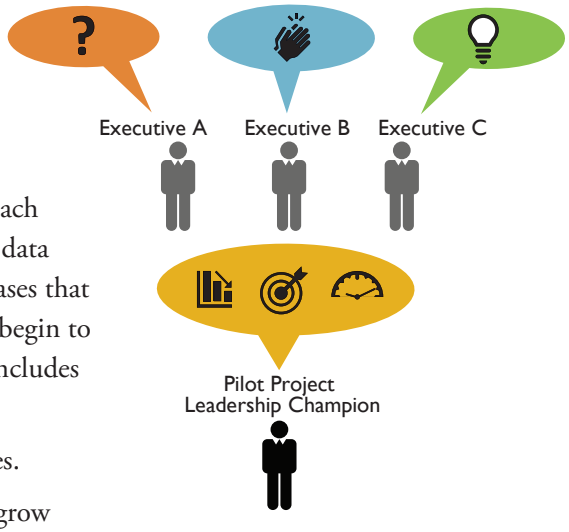
- Building support for the data and project across the agency, including other mid-level/branch managers that may have an interest in the data, project, and data products (or similar products) for their own business areas.

- Using the data to tell the story of success by crafting 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 unit.
- Getting others involved in sharing and using their data within the test environment, including iteratively expanding the use of the data to improved, enhanced, and new use cases.
- A case study on iterative success and growth of big data within a transportation agency.

**The goal of Step 6 is to generate interest in the data and the embryotic data environment that results in more data and additional use cases and pilot projects. 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 agency will be to shifting to a new culture of data awareness and use.**

# STEP 7.
## Demonstrate the Value of the Data to Executives

In Step 7, the team and the pilot project 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 and executives within the organization. This vertical organizational outreach will help to not only 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 organizational change. Step 7 includes guidance on the following:




- Presenting the success stories/business case to executives.
- Continuing to build support, foster data sharing, and grow iteratively and incrementally.
- Pushing for organizational change/adoption of a formal big data environment.
- A case study on buy-in from executive leadership.

**The goal of Step 7 is to develop enough use cases, projects, and data users throughout the organization to support the claim that the organization is not only ready for change but also that this change is vital to continue to develop and support data-driven decision-making organization-wide. Another goal of Step 7 is to gain enough recognition of the benefits of the data and the big data environment by leadership.**

# STEP 8.

## Establish a Formal Data Storage and Management Environment

After many iterations of Steps 2 through 7, as illustrated in Figure 3 (which could take several years), Step 8 establishes a formal, organization-wide data storage and management environment. Step 8 also institutionalizes policies, procedures, and practices associated with this formal data storage and management environment that represent an organizational shift from traditional management practices to modern data management practices. Step 8 includes guidance on the following:

- Establishing a clear vision and goals.
- Making data accessible yet secure.
- Integrating at the data level.
- Using data to make decisions.
- Merging existing projects into the same data infrastructure.
- Continuing to seek input from other stakeholders and to iterate evolving data governance plans and procedures.
- Seeking continuous improvement by periodically reviewing and revising data sets, technology, processes and procedures, documentation, security and privacy protection, metadata catalog, and so forth.
- A case study of continued room for growth within one transportation agency.

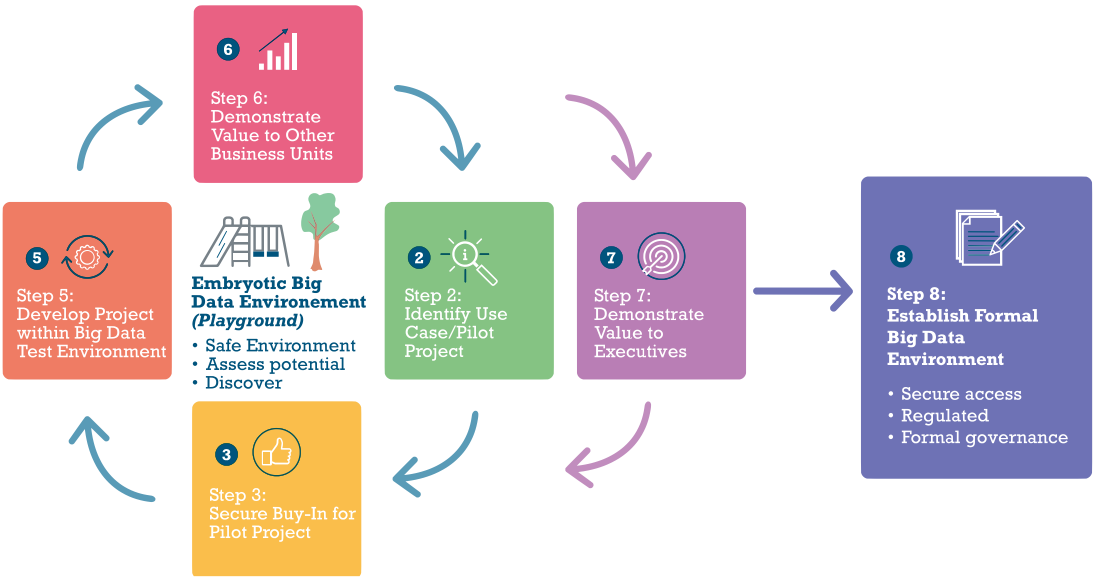



**Figure 3.** Iterative cycles of development toward organizational change.

The goal of Step 8 is to formally move the agency from a traditional data management system, practices, and policies to a modern big data management system, practices, and policies capable of fully managing new and evolving big data sources throughout their life cycle.

# MODERN BIG DATA MANAGEMENT LIFE CYCLE AND FRAMEWORK

The Modern Big Data Management Life Cycle (see Figure 3) defines the four major components of managing big data throughout their entire life cycle, including the *creation* of data, *storage* of data, *use* of data, and *sharing* of data (Figure 4). The Modern Big Data Management Framework builds on these data management components to include a discussion of *big data industry best practices* for each of the four data management components, as well as *more than 100 associated recommendations* for those looking to implement modern data management practices and systems. These big data industry best practices and associated recommendations should be referenced and implemented throughout each step in the roadmap, particularly in Steps 4 through 8.

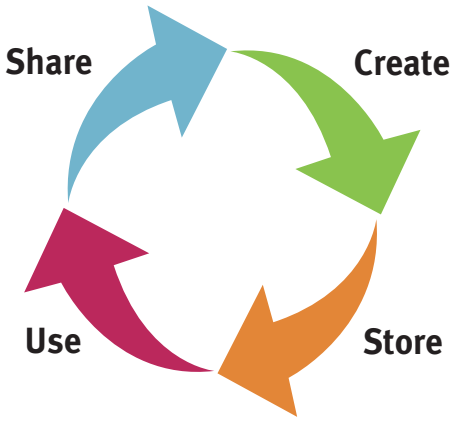



**Figure 4.** Big data life cycle.

# SUPPORTING TOOLS AND RESOURCES

To accompany the Roadmap to Managing Data from Emerging Technologies and the Modern Big Data Management Framework, there are a number of supporting resources and tools for agencies implementing the guidebook and the framework. Each of these tools is described in more detail as follows.

### *NCHRP Web-Only Document 282: Framework for Managing Data from Emerging Transportation Technologies to Support Decision-Making*

*NCHRP Web-Only Document 282* details the research activities, including results from a comprehensive state-of-the-practice review (50 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 online-only document defines data management and provides 65 modern big data management foundational principles organized by 15 data management "focus areas" to cover the full life cycle of big data. The document 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 the practice in data management within the transportation industry. The document ends with a list of common challenges reported by agencies during the research, as well as a list of associated needs.

The 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 online-only document and the 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 Data Management Capability Maturity Self-Assessment (DM CMSA) presents questions that will *allow transportation agencies to gauge their data management practices, as well as to identify specific areas for improvement.* The self-assessment consists of 104 questions divided across the following 15 focus areas:

- Data architecture
- Data modeling and design
- Data storage and operations
- Data security
- Data integration and interoperability
- Document and content management
- Reference and master data
- Data warehousing and business intelligence
- Metadata
- Data quality
- Data governance
- Data collection
- Data development
- Data analytics
- Data dissemination

The self-assessment was designed for ease of completion and to provide a high-level starting point for further inquiry. Due to the siloed nature of data within most transportation agencies, it is recommended that the self-assessment be taken by representative groups across an agency.

At the end of the self-assessment, a summary score sheet 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.

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

## 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 an organization. Traditionally, data governance dealt with the strict, authoritative control of data systems and users. And while traditional data governance operates on the fundamental premise that data itself cannot be governed—only what people do with the data—this approach is much more challenging to implement for modern data systems that incorporate agile development, big data, and cloud computing.

The Big Data Governance Roles and Responsibilities section of the guidebook includes big data governance recommendations, a description and framework for big data governance, and a tool for tracking the big data governance roles and responsibilities within an agency.

## Data Sources Catalog Tool

A data sources catalog tool is provided to help transportation agencies in cataloging existing and potential data sources. It is recommended that agencies periodically assess what data sources are in use, what data sources are available to be used, and what data sources might be obtained for use. Not only does this help prevent an agency from overlooking data sources that could be vital to current or future efforts but it also provides a better understanding of how data sets connect to support (a) prioritization and selection of data sources, (b) creation of a metadata catalog, (c) planning for data storage, (d) development of new data pipelines, and (e) 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 them.

The Data Sources Catalog Tool includes descriptions and examples of 11 characteristics of the data sources to be cataloged, as well as a simple table from which to build the catalog.

## Frequently Asked Questions

Frequently asked questions or FAQs include responses to 20 questions frequently asked regarding big data implementation, management, governance, use, and security.

## Examples, Case Studies, Quotes, and Common Misconceptions

Examples, case studies, quotes, and common misconceptions specific to transportation agencies are provided throughout the guidebook to support and supplement the information. These examples, case studies, quotes, and common misconceptions were gathered from information obtained from local and state transportation agencies that participated in telephone interviews and/or the stakeholder workshop conducted during the research project.

*The National Academies of*
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people's lives worldwide.

**www.nationalacademies.org**

ISBN-13: 978-0-309-67350-1
ISBN-10: 0-309-67350-X