Integrating Emerging Data Sources into Operational Practice

Opportunities for Integration of Emerging Data for Traffic Management and TMCs

www.its.dot.gov/index.htm
Final Report – November 2017
FHWA-JPO-18-625

U.S. Department of Transportation

Produced by Kimley-Horn and Associates, Inc. and Deloitte Consulting LLP
U.S. Department of Transportation
Office of the Assistant Secretary for Research and Technology

Notice

(Cover Image by iStockPhoto.)

Technical Report Documentation Page

1. Report No. FHWA-JPO-18-625	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Integrating Emerging Data Sources into Operational Practice— Opportunities for Integration of Emerging Data for Traffic Management and TMCs		**5. Report Date** November 2017 **6. Performing Organization Code**
7. Author(s) Douglas Gettman, Alan Toppen, Kelsey Hales, Alison Voss, Shane Engel, and Dayana El Azhari		**8. Performing Organization Report No.**
9. Performing Organization Name And Address Kimley Horn and Associates 7740 N. 16th St. Suite 300 Phoenix, AZ 85020 and Deloitte Consulting LLP 1919 North Lynn Street Arlington, VA 22209 Under Contract to: Cambridge Systematics, Inc. 3 Metro Center, Suite 1200 Bethesda, MD 20814		**10. Work Unit No. (TRAIS)** **11. Contract or Grant No.** DTFH61-12-D-00042
12. Sponsoring Agency Name and Address Federal Highway Administration Office of Operations 1200 New Jersey Avenue, SE Washington, DC 20590		**13. Type of Report and Period Covered** Final Report **14. Sponsoring Agency Code**

15. Supplementary Notes

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

16. Abstract

With the emergence of data generated from connected vehicles, connected travelers, and connected infrastructure, the capabilities of traffic management systems or centers (TMCs) will need to be improved to allow agencies to compile and benefit from using this information. New capabilities will be needed for data acquisition, communications bandwidth from the roadside to the TMC, new computing hardware, software, data storage and management systems, decision support subsystems, and data sharing and dissemination systems. The magnitude of what capabilities will be needed by individual TMCs, agencies and service providers will vary across regions. Regardless of the size of the traffic management system and TMC capabilities, the big data tools and technologies and systems are likely quite similar. The purpose of this report is to:

- Identify how big data tools and technologies can be used in traffic management systems or TMCs;
- Develop potential use cases for integrating big data technology and tools into traffic management systems or TMCs;
- Assess how connected vehicle and traveler related data could be used to enhance the operation of traffic management systems or TMCs;
- Analyze how the sharing of data with other TMCs, systems, connected vehicles and travelers; and agency business processes or systems could impact the performance of a traffic management system or TMC; and
- Identify the challenges and options to consider to compile, use and share this data.

17. Key Words Transportation System Management and Operations (TSMO), emerging data sources, connected travelers, connected vehicles, connected infrastructure, big data tools and technologies, Hadoop, real-time streaming, Hadoop Distributed File System (HDFS)	**18. Distribution Statement** No restrictions.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 86	**22. Price** N/A

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

The proliferation of data collected and stored from people and devices connected to the Internet is an important trend for businesses, individuals, and governments. Emerging data from travelers, vehicles, infrastructure, and other sources is expected to transform how agencies manage their transportation systems. The purpose of this report is to provide agencies responsible for traffic management with an introduction to big data tools and technologies that can be used to aggregate, store, and analyze new forms of traveler-related data. In addition, this report identifies ways these tools and technologies can be integrated into traffic management systems and traffic management centers (TMCs). While other functions of transportation system management and operations (TSMO) are affected by connected vehicles and related information, the primary focus of this report is on the TMCs of these systems that compile, aggregate, store, and disseminate this data for real-time and near-real-time traffic management. This data can then be shared with other systems and agencies for other uses such as improving transit system performance, evaluating traveler behavior patterns, and transportation planning.

The reader is encouraged to consider the four Federal Highway Administration (FHWA) reports as a complete set. The first report, *State of the Practice Review* (FHWA-JPO-16-424) provided a review of the state of the practice in big data tools and technologies and characterized the nature of emerging data sources. This report, *Opportunities for Integration of Emerging Data for Traffic Management and TMCs* is the second report in the series, which identifies specific use cases in common traffic management and TMC functions in light of the availability of new data sources. The third forthcoming report (*Capabilities and limitations of devices to collect, compile, save, and share connected vehicle data*) provides several proposed aggregation and edge-processing schemes to reduce the burden of the Department of Transportation's (DOT) Information Technology (IT) systems to consume and store "raw" data, while retaining the maximum amount of information from the new sources. The fourth forthcoming report (*plans and processes to enhance current transportation management systems*) then provides some recommendations on how these emerging sources in conjunction with acquisition, processing, and analytics techniques can be integrated into future next generation transportation management systems.

The intended audience of this report is individuals involved in, or responsible for, the planning, design, management, or maintenance of traffic management systems or TMCs and users of the data these systems may save or disseminate. These users have the opportunity to consider using big data tools and technologies covered in this report in the continued evolution of traffic management systems. The reader is assumed to have a general awareness of IT technologies, or wish to gain better appreciation of big data tools and technologies.

As more travelers and vehicles become connected and new sources of information emerge, new ways of acquiring, processing, and storing data will be required if the data is to be transformed into information and used to improve the day-to-day management and operation of the surface transportation system. After reading this report, the reader should be able to assess how certain new capabilities and technologies will be needed to allow traffic management systems or TMCs the opportunity to collect and use this data. The report also highlights methods available for agencies to consider as they collect, aggregate, preprocess data before it is stored and shared.

1.1 Overview of this Report

This report is designed to show how big data tools and technologies can be used to collect, process, store, and disseminate data from emerging sources to enhance the capabilities of existing systems, decision-making processes, and the day-to-day management and operation of traffic. The purpose of this report is to:

- Identify how big data tools and technologies can be used in traffic management systems or TMCs.
- Identify potential use cases for integrating big data technology and tools into traffic management systems or TMCs.
- Assess possible enhancements to enable traffic management systems or TMCs to collect and use connected vehicle or traveler related data.
- Analyze how the sharing of data with other TMCs, systems, connected vehicles and travelers; and agency business processes or systems could be accomplished with big data technologies.
- Identify the challenges and options to consider to compile, use and share emerging data sources.

The report is divided into four technical chapters. Chapter 1 provides an overview of emerging data sources and big data tools and technologies and discusses issues to consider with integrating these data sources and using these tools with traffic management systems. Chapter 2 reviews how existing and new traffic management functions could be enhanced with emerging data sources. Chapter 2 also identifies how the rate at which data may need to be processed and accessed may determine the need for certain types of big data tools and technologies to be used by traffic management systems.

> After reading this report, the reader should have an appreciation of the range of big data tools and technologies that could be used to assist traffic management systems and TMCs with collecting, saving, using, and sharing different sources of emerging data.

Chapter 3 discusses how big data tools and technologies may be provisioned to integrate new data sources with existing traffic management system functions. Chapter 3 then identifies how big data tools and technologies can be used for data sharing among inter-agency groups, inter-regional partnerships, and agency partners. Chapter 4 then identifies issues to consider in selecting, developing, or using different data aggregation techniques that can reduce the data velocity and volume which makes the acquisition of emerging sources more cost-efficient and technologically feasible.

This report provides an introduction to the subject of big data tools and technologies and how they could support transportation operations [1, 2, 3, 4, and 5]. In addition, it introduces the opportunities and challenges of using emerging data from connected sources (e.g. crowdsourcing, social media) and travelers for TSMO [6 and 7]. The reader is encouraged to review the references which are listed in each chapter for additional background and information.

1.2 State of the Practice of Big Data Tools and Technologies and Emerging Data Sources for Traffic Management and Traffic Management Centers

The purpose of the *Emerging Data Sources and Big Data Tools and Technologies: State of the Practice Review* (FHWA-JPO-16-424) was to provide an understanding of what emerging data sources relevant to traffic management systems could be available in the next 10 years. The report also provided information to

assist agencies assessing the implications if they decided to acquire, store and use data from these different sources. The second objective of the report was to raise awareness of the capabilities and potential benefits with using big data tools and technologies that have been developed and used in other industries.

Currently, the state of the practice in use of emerging data from connected travelers and connected vehicles for traffic management and TMCs is the use of aggregated data services such as INRIX, HERE, and Waze. Over the past five years, use of this data has become quite common. Since this these companies process the raw data from vehicles and perform validation and aggregation, it is not challenging for traffic management agencies and TMCs to use, process, and store it with IT technologies commonly used in the agency.

Traffic management agencies and TMCs have much deeper experience with connected infrastructure. Since the beginning of the intelligent transportation systems movement more than 25 years ago, traffic management agencies have been using data from connected infrastructure such as traffic signals, vehicle detectors, ramp meters, weigh-in-motion stations, and video cameras. While these traditional sources of transportation data for traffic management will remain, emerging data sources—largely those from Connected Travelers, Connected Vehicles, and new types of Connected Infrastructure—will represent a significant opportunity for DOTs and public agencies to improve how they manage travel on the surface transportation system. Since the data is expected to be of higher level of detail than what is currently available, new methods and tools will be needed to collect, process, save, store, manage, and use it. Connected vehicles are expected to generate status messages (i.e., “Basic safety messages” or BSMs) every 1/10th of a second. This data will be valuable to the traffic management agency for a variety of functions, as discussed further in this report.

If all data available to an agency through the emerging data sources was collected and stored, the volume of data consumed by a typical agency today could more than double by 2021 and increase to more than fivefold by 2026. This will require new hardware and software technologies, but also consideration by the agencies of what new data to collect, store, process, and use.

If all of this emerging data related to traffic operations is stored, the cumulative storage of a typical traffic management agency could be in the many thousands of Terabytes by 2026. Agencies will need to apply and use big data tools and technologies to transform the data into information that can be used by traffic management systems in the TMC This includes technologies and systems for processing, storage, analysis, and interfacing to existing software systems. It will also be important to determine where and how the data is stored and processed; at the roadside, in individual TMCs, and/or aggregated in central systems for an agency, a regional coalition, an entire State, or a multi-State group.

Currently, most traffic management agencies and TMCs have little to no experience or use of big data technologies and tools. Government agencies in general also have less experience with big data systems than do private companies. Private companies use big data tools and technologies for a variety of functions. Big data tools and technologies are the cornerstone of internet communications, social media data storage, and commerce. Over the next ten years, traffic management agencies will need to gain experience and improve the state of the practice with these tools and systems to take advantage of the higher levels of fidelity information available from emerging data sources.

1.2.1 Emerging Data Sources

Several emerging data sources are expected to improve the ability of traffic management agencies and TMCs to provide high-quality services to the public. These sources include connected travelers, connected vehicles, connected infrastructure, and other emerging sources. The following sub-sections define each category of emerging data and identify some of the key characteristics of each data source and the value for traffic management and TMCs. Status information from connected vehicles and travelers provide high-

resolution anonymous trajectories of user experience on the transportation network. In general, such detailed information has never been available to traffic management agencies and TMCs before. This level of detail on the transportation system user will unlock new functions and enhance a wide variety of existing functions at TMCs.

Connected Travelers

A connected traveler is one that is using a mobile device that generates and transmits status data which could be collected, saved and used by Intelligent Transportation System (ITS) devices and the corresponding traffic management system, other connected mobile devices, and connected vehicles. The information could be collected via Dedicated Short Range Communications (DSRC), Wi-Fi, Bluetooth, or cellular. Messages generated and distributed by Connected travelers could include data representing the traveler's location, trip characteristics (e.g., speed) mode and status (e.g., riding in a car, riding on transit, walking, biking, etc.). This information could be collected or compiled by a TMC or traffic management system. With appropriate privacy protections in place, connected traveler data may enable agencies to incorporate this information into how they manage and control traffic, share information back with the traveler, and share information with other systems (e.g., transit vehicles and systems). The archived data may also provide information on traveler behavior that is now only available via expensive and time-consuming travel surveys.

Many States and most metropolitan areas have a 511 system which provides travel condition information. Many of these systems have their own branded 511 application (apps) (or suite of apps for relaying traffic condition information and transit schedules). Some agencies also have citizen reporting apps, which allow the public to report incidents or travel conditions (e.g., congestion, location of potholes). Examples of agency-branded apps include the Utah Department of Transportation Citizen Data program and the Los Angeles Metropolitan Transportation Authority (MTA) 511 app. These systems and apps typically only provide general travel condition information to users and they do not collect, store or use information specific to the trips of travelers using these devices.

Connected traveler data is also collected by private companies (e.g., Google Maps, Waze, etc.) using the apps they have developed. Travelers allow companies to collect and use information about their trips, while the app provides the user information (e.g., directions). Similar to the trend in public agencies procuring link-speed data from private providers, it is likely that connected traveler data will be increasingly available from private sources. Sharing data from mobile devices on traveler activities and status can provide additional information that agencies can use to improve traffic management in the TMC.

Connected Vehicles

"Connected Vehicles" are vehicles that can communicate status information directly to other vehicles, other road users, and roadside systems so that every vehicle on the road is aware of where other nearby vehicles and travelers are located. Connected vehicles can identify threats, hazards, and delays on the roadway back to the TMC and provide drivers with alerts, warnings, and real-time information [1]. Key connected vehicle elements include wireless communications, onboard computer processing, advanced vehicle sensors, Global Positioning System (GPS) data, and smart infrastructure.—There are three basic types of Connected Vehicle functions, (1) Vehicle-to-Vehicle (V2V) principally for active safety purposes, (2) Vehicle-to-Other Objects (V2X) encompassing vehicle interactions with vulnerable road users such as pedestrians and cyclists, and (3) Vehicle-to-Infrastructure (V2I) providing data between vehicles and traffic management devices and management centers. V2I is also commonly used as a blanket term to refer to data exchanges in any combination of transmission and reception: (1) from the infrastructure to the vehicle (I2V), (2) from the vehicle to the infrastructure only (V2I), and (3) bi-directional data exchange

(V2I-I2V) Connected vehicles may use a variety of communication media for data exchange include DSRC, cellular, and Wi-Fi.

Commercial connected vehicles have cellular connections to a private cloud from the vehicle's infotainment system or third-party in-car systems for vehicle tracking and data collection. Currently, commercial in-vehicle systems primarily are used for providing Internet access for passengers' nomadic devices and infotainment systems, though additional functions include navigation, driver assistance, health monitoring, remote diagnostics, insurance premium evaluation, road user charging, and automated driving. Several private companies share the information related to their commercial fleets with DOTs through aggregation of individual vehicle travel data into summaries of vehicle speeds on roadway links. Real-time link-speed data is currently invaluable to traffic management agencies since it is becoming more and more challenging for agencies to maintain their spot-speed detection equipment (i.e. in-pavement loop detectors, radar, and video).

Outside of aggregated link-speeds and potential incident reports that are currently provided by commercial sources to DOTs, there are potentially other rich data on driver and vehicular behavior, which can benefit DOTs (and thus travelers through enhanced information sharing and traffic management strategies). As of June 2016, HERE has published a connected vehicle data sharing standard [9], which may greatly accelerate the availability of trajectory-based commercial connected vehicle data to DOTs, assuming the information can be suitably anonymized. To date, eleven major automotive and supplier companies have already joined the SENSORIS Innovation Platform now under the coordination of ERTICO. They are: AISIN AW, Robert Bosch, Continental, Daimler, Elektrobit, HARMAN, HERE, LG Electronics, NavInfo, PIONEER and TomTom [9]. Momentum is also building for using commercial aftermarket devices for setting insurance premiums and road mileage rates as an alternative to gasoline taxes [8]. Traffic management agencies and TMCs will likely have new sources of trajectory-based vehicle data available from private companies in a very short time.

U.S. DOT has supported the development of applications to improve the safety and mobility of connected vehicles. These applications rely on DSRC technology to generate, send and receive Basic Safety Messages to other vehicles and to roadside equipment (RSEs) at high frequency (10 times per second) and with very low latency (50 ms from transmission to receipt). Low latency and high frequency are both critical in the exchange of BSMs for V2V and V2I safety applications. Traffic management agencies and TMCs will soon be able to process, store, and use BSMs for improving traffic management functions. In the next ten years as the number of vehicles with this technology increases, traffic management agencies will need to develop new tools and methods to process, store, and use the new information.

In addition to BSMs, a Probe Data Message (PDM) encapsulates a string of "snapshots" (a more comprehensive data element than the BSM) to provide vehicle trajectory information over a longer time frame than the local trajectories shared by the BSMs. Both messages are shared with RSEs when the vehicle is within range, although there are alternative transmission methods being considered for transmission of PDMs through cellular wireless. As of December 2016, National Highway Traffic Safety Administration (NHTSA) has issued the Notice of Proposed Rulemaking that DSRC devices must begin to be installed in passenger vehicles in 2019 and will be mandatory on all passenger vehicles manufactured or made for sale in the U.S. by 2023. Thus, agencies will have the opportunity to collect, process and use data disseminated by from connected vehicles and devices to potentially support a wide range of traffic management and traveler information related functions discussed in this report.

Connected Infrastructure

ITS Infrastructure devices that provide information to the traffic management agency and TMC include public infrastructure assets such as traffic signals, ramp meters, closed-circuit television (CCTV), vehicle detection systems, Road Weather Information Systems (RWIS), flood warning devices, high wind warning

devices, among others. These sensors and data sources have enabled traffic management agencies and TMCs for over 35 years with information on traffic volume, incidents, device malfunctions, speeds, and environmental conditions. Agencies will continue to deploy more infrastructure sensors and devices for traffic and transportation management, thus continuing to increase the load on existing processing, storage, and analysis systems in the TMC. While some connected infrastructure devices and systems may become obsolete as data from connected vehicles and connected travelers provides the same information (at potentially higher resolution), connected infrastructure is still needed since the penetration rate of connected vehicles will not reach 100% for more than 20+ years.

Currently, most agencies delete or archive old ITS infrastructure status information simply because existing database and processing system technologies are expensive and agencies lack the business case to expand to handle the volume. Traffic management agencies and TMCs can use big data tools and technologies to gain analytical insights if more information from connected infrastructure could be retained and stored.

Connected infrastructure devices for traffic management may eventually evolve to use standard Internet-of-Things (IoT) protocols as IoT technologies continue to mature. IoT and big data tools and technologies are integrally linked together for the connectivity of other types of sensors and mobile devices in a wide range of markets including wearables, household appliances, and commercial equipment. In aggregate, these markets are much, much larger than transportation management. Migration of connected infrastructure to IoT technologies will require traffic management to adapt to new systems and technologies.

Other Potential Sources of Data

Other sources include the following:

- **Mobile sensors**: A mobile sensor is defined as a device that records data about the environment where the vehicle is traveling through (rather than data related to the vehicle's status and performance). This includes Light Detection and Ranging (LiDAR) point cloud and 3D camera data from connected vehicles. This data is collected to support automated vehicle operation, but could be valuable to traffic management functions, particularly incident management. Collection and processing of such feeds to an external consumer (i.e. from a vehicle to the TMC) is not currently possible in real-time, although technology continues to evolve. LiDAR point clouds are currently available in de-facto standards such as LAS (LASer file format) but end-user consumption, storage, display, etc. appears to still be product-specific. The Open Geospatial Consortium (OGC) is currently developing a set of LiDAR exchange standards to better enable sharing of Point-cloud data as well as 3D videos [15]. Within 10 years, it may be commonplace for a TMC to receive real-time streaming 3D video from incident response vehicles, for example. This will require new technologies at the TMC.
- **High-definition maps**: High-definition maps are maps that include lane-level accuracy geometry, accurate placement of all traffic control signs and advisories, allowable traffic controls at intersection junctions, and major street furniture. Such high-definition (HD) maps are foundational data to support automated driving and connected vehicle functions. HD maps are likely to be provided by private companies to traffic management agencies, and thus could be considered a data source. The connected vehicle "Map" message, or geometric intersection description (GID) file, is also an important element that traffic management agencies and TMCs will need to manage. This data describes, at high resolution, the geometry of an intersection or traffic facility (such as a work zone) and is necessary for many V2I functions. With tens of thousands of intersections in a given jurisdiction, the management of the map data will be a non-trivial process, perhaps requiring big data tools and technologies.
- **Transactional data**: Examples include commodity-specific, county-level cross-modal global freight flow data, supply chain and logistics management, purchasing behaviors, real estate marketing and

valuation, and other economic transactions. Each of these may be useful to traffic management agencies and TMCs to gain a better understanding of travel patterns.

As with any statement about the future, much is uncertain. While these data types are anticipated based on what we know now, many different industries are collecting and analyzing data, which is moving the big data and IoT industries forward. Also, sensors are becoming ubiquitous and less expensive, which will make additional data sources possible, many of which we can't foresee at this time. Regardless of any additional sources that may emerge and be available, the information that may be available from connected vehicles when penetration levels increase offers public agencies with a unique opportunity to benefit from collecting and using this information. New technologies and tools will be needed to ingest, process, and store these data and make it ready for use by traffic management functions. These tools and technologies are generally described as big data systems. The next section describes the state of the practice in big data tools and technologies.

1.2.2 Big Data Tools and Technologies

The volume, variety, velocity, and veracity of connected traveler, connected vehicle, and connected infrastructure data will require traffic management agencies to adopt modern methods of processing and storing the information. Even without the connected vehicles and connected traveler data, most agencies are not utilizing the information they currently collect from infrastructure (particularly traffic signal systems) in meaningful ways since these systems were not designed to take advantage of tools and technologies that did not exist 10 or more years ago. Changes to IT infrastructure, such as relational database management systems (RDBMS), software, computing capabilities, and server architectures will need to be made as well to take advantage of new technologies. RDBMS systems as typically used in traffic management systems are not suited for processing and storing connected vehicles data on the scale that is anticipated over the next 10 years. TMC and traffic management staff will also need new resources to take advantage of these evolving big data tools and technologies.

There are a wide range of tools and technologies available now for data acquisition, marshalling, and analysis of huge data sets that are proven in a variety of use cases and markets outside of traffic management and system operations. There is no single best tool, technology, or provider for a particular service or function for managing huge data sets. There are however recommended industry practices and common core components of many commercial off-the-shelf products or systems which are available for agencies to consider integrating into their current systems. Specific tools and technologies should be selected based on the appropriateness for a system or function and the type(s) of data to be processed and used. Additional factors for selecting the right tools and technologies for a particularly traffic management function include the maturity and stability of the software and provider, are they open source or proprietary, easy to implement and maintain, and the ability for TMC and traffic management staff to properly use the system.

Key Characteristics of Big Data Tools and Technologies

Some general characteristics of these big data tools and technologies are discussed in the following sections. These characteristics summarize the key elements of big data as applied to the use of emerging data sources for traffic management functions.

Scalability. One key attribute of big data tools and technologies is the general ability to scale from small test environments to large deployments without a large amount of agency effort or reprogramming, redesign, or reconfiguration. Scalability includes considerations for:

- Higher levels of data volume.

- Increases in the speed at which the data is presented.
- Larger numbers of processors for larger and larger data sets.
- Larger numbers of processors to complete a process in a timely fashion (either due to the size of the data or a requirement on the time to produce a result).

Scalability for data storage is usually achieved by simply adding storage and processing capacity (database and computing servers). This is sometimes referred to as “horizontally scalable.” Cloud services are available to make scalability even easier, by enabling the server provisioning to happen automatically when additional capacity or performance is necessary. This scalability is critical as the volume of data from connected vehicles and other emerging sources will rapidly rise from the small trickle of data from equipped vehicles and travelers over the next few years to the flood of information available when 20%, 50%, and finally close to 100% of the vehicles in the U.S. fleet may be transmitting basic safety messages. Traditional RDBMS and processing systems that existing traffic management systems are dependent on are not as flexible as these newer systems designed for massive amounts of data.

Relational database management systems are not designed for big data. There are a few main shortcomings of RDBMS that big data tools and technologies attempt to address:

- **The data is too big to reside in one place**. The big data solution is to distribute it to multiple storage hosts. Sophisticated data harmonization strategies are employed by big data solutions to enable fast queries on huge data stores.
- **Reading the data sequentially takes too long or the processing steps can't all fit in memory of a single machine**. The big data solution is to split the data into chunks and read and process it in parallel on separate nodes, aggregating the intermediate results to a final response. Parallel processing or multi-threading can be done on a single machine to speed things up or swap to disk when the data set doesn't fit in memory, but the power of big data is realized by using a networked cluster of machines to store and process the data.
- **The data is received at a rapid rate and must be processed in real-time to support a particular function**. The big data solution is to use a software system designed for "*streaming data*," which will store data in short-term memory, process it, report on it and then delete it from memory once used. Such systems also simultaneously feed the data into longer-term storage for functions that are not time-critical.
- **The data may not conform to a fixed database schema where all the column names and data types are known in advance**. The big data solution is to use a not only structured query language (*NoSQL*) database which has this sort of flexibility. These systems require a different way of thinking about storing and querying data. New tools and technologies that are needed to process NoSQL come with a learning curve, but they are sometimes the only way to retrieve a query response on a huge data set in a reasonable amount of time.
- **Distributed storage includes the Hadoop Distributed File System** (HDFS), which stores data across hosts in a redundant way similar to a Redundant Array of Independent Disks (RAID) drive-based storage. If one host fails, no data is lost as all data is stored in at least two places. This has the added benefit of allowing one to use less expensive storage media. Big data also makes use of distributed processing. Computations using Hadoop send code chunks of the data, run computations on the distributed hosts, gather the intermediate results and aggregate them at the master host. Storing the data in a distributed manner with redundancy reduces the amount of movement of the data, making solutions more scalable as the dataset gets larger. Hadoop is typically considered more of an “ecosystem” of open-source tools and systems rather than a single technology. HDFS is just the base of that ecosystem. Over the last few years, additional tools have been developed to work with HDFS to schedule jobs for multiple users, make querying easier

(using structured query language (SQL)), and to store the data in different formats. Supporting SQL, in particular, provides significant value. From a workforce perspective, there is a broad base of software developers and data scientists that already have skills in traditional SQL programming languages. By supporting SQL queries and functions, big data tools have a shallower learning curve and are being adopted much more readily.

Tools for processing *streaming data* are an important class of big data tools relevant to traffic management practices. Streaming data is data that is generated continuously by a variety of data sources, which typically send in the data records simultaneously and in small sizes (order of Kilobytes). In other industries, streaming data includes data such as log files generated by customers using mobile or web applications, e-commerce purchases, in-game player activity, information from social networks, financial trading information, geospatial services, and telemetry from connected devices or instrumentation. This data needs to be processed sequentially and incrementally on a record-by-record basis or over sliding time windows. The information is used for a wide variety of analytics including correlations, aggregations, filtering, and sampling. Streaming data is stored and processed in rapid succession. The processing layer consumes data from the storage layer, running computations on that data, and then notifying the storage layer to delete data that is no longer needed or sending it to a lower cost, higher latency storage array.

Unstructured or semi-structured data requires NoSQL tools. NoSQL is used by programs to store and retrieve data in RDBMS. RDBMS are rigid in their structure, with a fixed *schema*. A schema refers to the column names and types of data (integers, text, etc.) in each table of the database. There are many benefits to RDBMS and the fixed schema structure provides predictability, enforces data integrity and can speed up queries. However, their fixed structure has limitations and drawbacks as discussed above. Modern applications that store information like images, videos, documents, and internet uniform reference locators (URLs) are not supported well by relational databases. Over the last 15 years, the internet has driven the development of NoSQL databases which have flexible data models, offer scalability and performance, and offer redundancy.

Traffic management functions will benefit from the flexibility and scalability of big data technologies when connected vehicle and traveler data are commonplace in the TMC. Advanced traffic management systems (ATMS) applications are typically written in object-oriented programming languages such as C++, Java or C#. “Objects” (such as the attributes of a traffic signal, including the timing data) are saved to and read from an RDBMS by the client application (e.g. changing the values of the cycle, split, and offset of a traffic signal plan). As new features are added to the traffic signal controller (e.g. new functionality for transit signal priority is added to the traffic signal control firmware), a new version of the RDBMS is required. As we look ahead to connected vehicle and connected traveler data, it is difficult to predict what new data will be emitted from vehicles or from the infrastructure. This will undoubtedly evolve over time to support new functions. Schemas will need to change as new data elements are introduced to support as-yet unimagined functions. Fixed schemas and hard-coded data description protocols pose a constraint on this evolution. This can be avoided with NoSQL.

NoSQL is also designed to be scalable, which is particularly important for the data coming from connected vehicles. At the outset, while there are few connected vehicles on the road, the back-end systems can start small. As vehicle penetrations increase, NoSQL systems can scale accordingly. Auto-scalability also helps to avoid large up-front costs since currently it is uncertain how quickly emerging data sources will grow.

Technology continues to mature and staff skills will need to grow. As time passes, new and increasingly sophisticated methods and tools will be developed to deal with increasingly bigger datasets and analysis suites will become more mature. To make sense of these large datasets, traffic management agencies will need individuals and teams with skill sets that span software development, database administration, IT, statistical analysis and modeling, and interpersonal communication. Just as importantly,

these individuals must also have domain knowledge in traffic management to understand the information, perform meaningful analyses, and effectively communicate results. These individuals are now known as data scientists. Just as Traffic management has emerged over the last 20 years as a functional area within DOTs and local agencies, data science will need to be an important area of focus over the next ten years.

Outsourcing to vendors or contractors may be required for sophisticated big data tools and technologies. Big data tool vendors and providers continue to offer more and more holistic solutions ("as a service") for increasingly complex data processing and management systems. An agency not having the technical personnel to initiate or maintain an system may consider going with a vendor or contractor to manage the process for them. This includes not only the implementation of the functions, but the hosting of hardware, software, and data. This concedes some control from the agency to the vendor and includes long-term recurring costs, but may offset other potential risks of managing on-premise and in-house implementation of these complex systems.

1.2.3 Issues affecting the ability of Emerging Data Sources and Big Data Tools and Technologies to enhance traffic management and Traffic Management Centers

As discussed in the previous section, there are many important attributes of big data tools and technologies that are relevant to improving ATMS. In this section, the following topics are introduced:

- Issues related to ATMS system architecture and integration of big data tools and technologies.
- Issues related to ATMS system architecture and integration of data from emerging data sources.
- Opportunities to enhance traffic management functions with using data from emerging sources.

A baseline of the functions and technologies of advanced transportation management systems is first outlined. This section is for general overview of the typical architecture of most ATMS and the range of issues related to existing and future software and the use of various big data technologies. Chapters 2 and 3 will then discuss the details of how emerging data sources (Chapter 2) and big data tools and technologies (Chapter 3) can be used to enhance ATMS.

An ATMS has three basic components—a database or databases, a software service or set of services (sometimes called *servers*), and a user interface (sometimes called the *client* application). Figure 1 illustrates a generic architecture for the application of big data technologies to traffic management functions. In this figure, we use color to indicate a "data temperature" or need for real-time, near-real-time, and offline data transmission and processing. Components shown in red ("hot" data) are real-time components. Red lines and red boxes represent software systems and data flows that process real-time data and have control actions that need to be taken within seconds. Orange systems and arrows are near-real-time systems, or functions with execution times in minutes. Blue arrows and software systems are offline elements. They do not have the time urgency of real-time or near-real-time functions and can ship and process data on-demand, in batches, or even in daily updates. "To-be-determined *(TBD) data processing and storage*" represents the new elements of the big data ecosystem. These are new technology elements that will be necessary to harness the power of the emerging data sources for traffic management and TMCs. These systems are specifically designed for handling and processing of massive data sets. While these new TBD components are shown inside of the TMC, they might be deployed off-premise in a cloud architecture, or the agency may purchase the resulting data/information from a third-party solution provider.

In Figure 1, we depict an ATMS as a single entity. In practice, a typical TMC includes multiple systems that perform specific functions such as a Dynamic Message Sign (DMS) management system, a CCTV

management system, a traffic signal control system, and a freeway management system. All of these might be considered an ATMS. There are notable examples of integrated control systems that perform multiple functions (CCTV, DMS, Traffic signals, and freeway device management) with one integrated user interface and one integrated RDBMS. ATMS systems typically connect to field devices in real time or near-real time using standard or device-specific protocols and message definitions. One or more RDBMS are used for data storage with strict schema adherence rules and functions. Status and control data may be shared with peer agency systems typically through standardized data descriptions and protocols (e.g., such as Traffic Management Data Dictionary (TMDD) or National Transportation Communications for Intelligent Transportation System Protocol (NTCIP)), or using legacy methods such as File Transfer Protocol (FTP) file transfers.

Most ATMS are proprietary software supplied to DOTs by vendor companies, although there are examples of open-source ATMS and systems that have been developed and maintained by agencies with in-house staff. Few existing ATMS have any current capability to ingest or process connected vehicle or connected traveler data. The Connected Vehicle (CV) Pilots on-going in New York, Tampa, and Wyoming are demonstrating a wide variety of CV applications which includes the collection and storage of CV data at TMCs. Few agency systems have any decision support capability, although notable exceptions do exist in the University of Maryland's Regional Integrated Transportation Information System (RITIS, http://www.ritis.org/.), California DOT's Performance Measurement System (PeMS, http://pems.dot.ca.gov/), and others.

In Figure 1 we have represented the points of technical integration of big data tools and technologies with existing elements of the TMC as "*TBD interfaces*". These to-be-determined technology linkages and components will be critical in leveraging emerging data sources for enhancement of traffic management and TMCs. Issues related to these points of technical integration are discussed in the following paragraphs. Chapter 3 of this paper discusses these data processing and storage components in detail, and presents the same figure with more details filled in.

Issues in ATMS Architecture for Integration of Big Data Tools and Technologies

Lack of direct compatibility of NoSQL and RDBMS. ATMS use traditional relational database management systems such as Oracle or Microsoft SQL Server. Some systems use open source databases such as PostgreSQL or MySQL but commercial RDBMS are more common because they are often subject to IT department standards which tend to favor commercial products for a variety of good reasons including support. As discussed in the previous section, these database tools are not scalable to meet the demands of storing large volumes of connected vehicles and traveler's data. New data storage systems and processing will be required, and strategies for integrating these new tools and technologies with legacy RDBMS will be needed. Note that most NoSQL products now have substantial support for traditional SQL processes (where just 2-3 years ago, this was much less common). A key barrier to adoption is the lack of experience with big data solutions in the IT department. This is likely to change, but will take collaboration and careful planning.

ATMS typically cannot, and do not, store all relevant data due to RDBMS limitations and a lack of a clear business case. ATMSs process data from field devices in real-time using protocols such as NTCIP or device-specific proprietary protocols. Devices are typically polled by the ATMS every 1-60 seconds (once per second for most modern signal systems) and data is stored in the RDBMS. Many systems do not store second by second data because the expense of using RDBMS for this sort of application. After some period of time (e.g., a few weeks or a few months), data is either deleted or moved from the live real-time database to an archive, which is typically a separate database that can be queried without taxing the real-time system. Some combination of the archive database and the live database is typically queried for reports. With big data tools and technologies, storage and merging of archived and online data from separate RDBMS may

no longer be necessary. In addition, discarding high-resolution, real-time data may no longer be necessary. Strategies for integration of NoSQL and traditional RDBMS systems will be required.

ATMS databases have inherent limitations in storage of unstructured information and data sharing. ATMS systems also store incident, construction and special event data as entered by Traffic Management Center operators. These data, while not as "big" as device data, can be challenging to store, process and use because of their many data fields and inconsistency of data entry. Further, most ATMS RDBMS schemas typically don't conform exactly to TMDD or other standards, making them difficult to share with other systems and agencies without a data translation software service. New methods will be needed for *merging* existing incident management data onto new big data tools.

Client software is not typically able to interact with big data tools and technologies. ATMS client software connect to real-time software services and the RDBMS to present real-time, near-real-time, and historical status information for operators and analysts in the TMC. This client software, whether web-based or installed on each workstation, typically will have no capability to consume or process data from big data tools or technologies without significant software development. New methods will be needed to present the data, any actionable insights, and any derived analytics from emerging data sources to the users of ATMS systems. End-user client systems will need additional interfaces or new modules to display information from big data technologies.

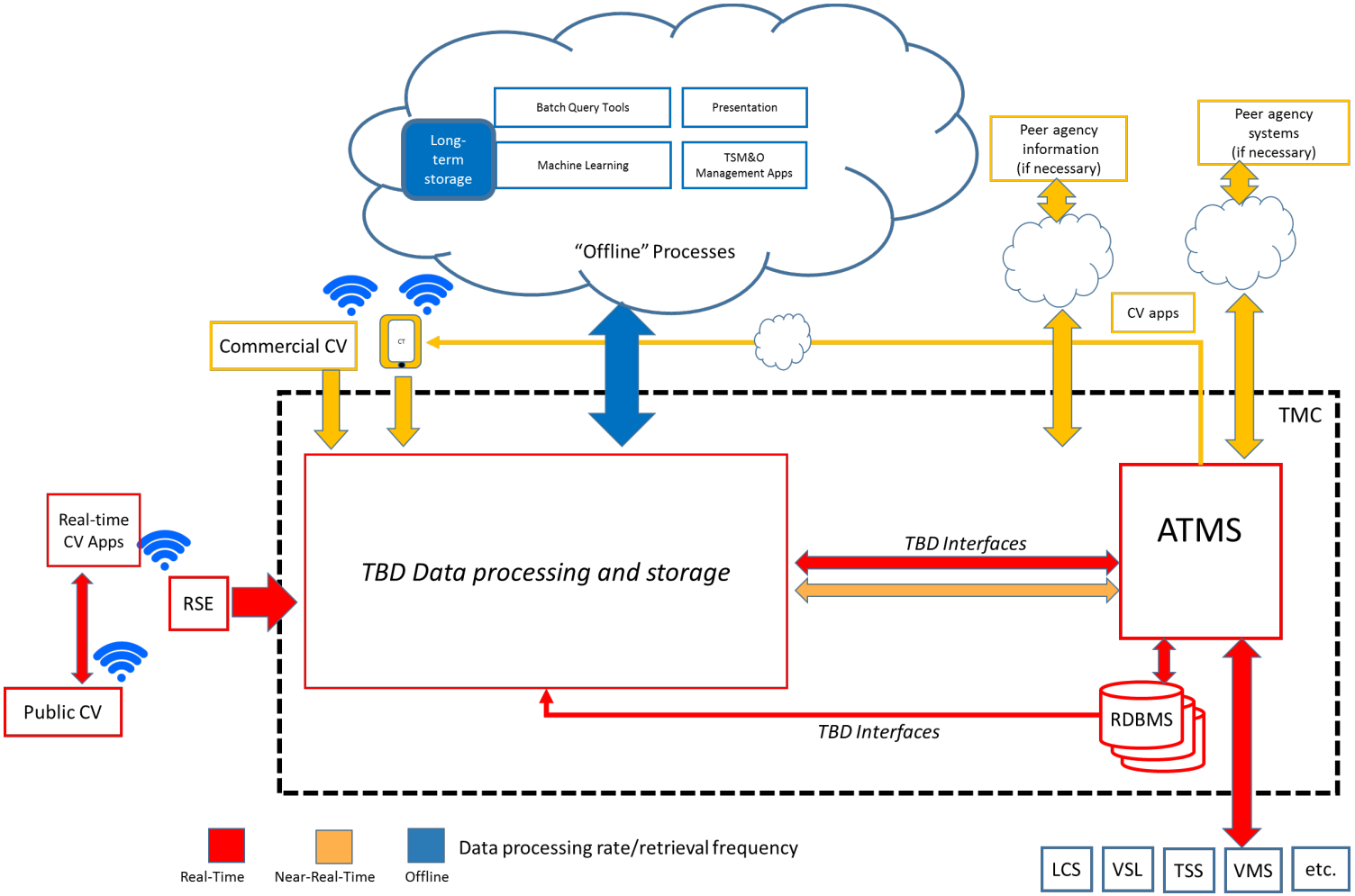


Figure 1. Diagram. Transportation Systems Management and Operations reference architecture.
(Source: Kimley-Horn and Associates, Inc., 2016.)

Issues in ATMS Architecture for Integration of Emerging Data Sources

Standard approaches for acquisition of real time emerging data sources have not been identified. While standards for connected vehicle messages have been developed for V2V (BSM), V2I (BSM, PDM, Signal Request Message (SRM), etc.) and I2V (Traveler Information Message (TIM), MAP (a message containing roadway geometric information), Signal Phase & Timing (SPaT), etc.), there are no defined standards for how an existing ATMS would connect to a real-time software service(s) for presenting real-time, near-real-time, aggregated, and historical status information to operators and analysts in the TMC. Prototype systems have been developed in the Safety Pilot, Multi-Modal Intelligent Traffic Signal System (MMITSS), and other research including the Connected Vehicle pilots in New York, Florida, and Wyoming. New methods will be needed to acquire the emerging sources data when penetration levels of connected vehicles are significant.

Standard approaches for integration of real time emerging data sources with existing ATMS processes have not been identified. Once the emerging data is brought into the TMC, there are no defined standards or structure for how an existing ATMS would merge this information with existing data from connected infrastructure for enhancement of traffic management functions. New methods will be needed to compile and process the emerging sources data for use in existing ATMS functions and for provision of new services (e.g. roadway hazard warning).

Summary of Issues in ATMS Architecture for Integration of Emerging Data Sources

To enhance existing traffic management functions with emerging data sources, it is clear that big data tools and technologies will need to be integrated with legacy ATMS. First, the data from the emerging data sources must be ingested for use by the ATMS. Given the data velocity and volume expected when connected vehicles have significant penetration, this will require new technologies. Next, the data from the emerging data sources must be stored. Given the volume of data, this will require new NoSQL storage tools since RDBMS are not suited for the required scalability. Finally, the data must be processed for use by the legacy ATMS systems to make better, faster, and more informed decisions. This again will require new tools and technologies and enhancements to the ATMS systems to interface and use these new capabilities. New decision support functionality will also be enabled by the use of large volumes of emerging data sources, coupled with the legacy connected infrastructure data. This implies that the legacy connected infrastructure data will need to be stored with the emerging data. Methods for integrating existing RDBMS systems with new NoSQL storage systems will be required.

In this chapter, topics related to the state of the practice in big data tools and technologies and the characteristics of emerging data sources were discussed. The reader should understand that:

1. Connected vehicle, traveler, and infrastructure data could grow many thousands of terabytes by 2026 if all the data is retained by a typical agency.
2. Big data tools and technologies will be required for data storage and management on this scale of volume and velocity. Legacy RDBMS are not suited for this purpose.
3. Big data tools and technologies have been developed to overcome the limitations of RDBMS and software that relies on RDBMS for processing.
4. Existing traffic management systems currently have limited capabilities to integrate with big data technologies or ingest emerging data sources without significant modifications.
5. Taking advantage of emerging data sources will require development of new systems, interfaces, and merging of data stored in legacy RDBMS with new data storage methods based on NoSQL.

In the next chapter, we present common traffic management and TMC functions and detail the following characteristics for each function:

- How emerging data from connected vehicles, connected travelers, and connected infrastructure can enhance each function.
- What penetration rates of connected vehicles and connected travelers are needed to achieve enhanced functionalities.
- Which agencies should be involved in the deployment of individual functions (as this impacts technical solutions.
- What big data tools and technologies could be considered to enable these enhanced functions.
- What are the requirements for data processing (e.g., speed).

Similarities among functions are then identified to aggregate functions into three sets of categories: real-time, near-real-time, and offline. Chapter 3 then discusses the big data tools and technologies that may be applicable to each category of traffic management function. The notional architecture shown in Figure 1 is populated with these tools and technologies for processing, storage, and use of real-time, near-real-time, and off-line emerging data and integration with existing traffic management systems. For example, real-time functions will need *stream processing* tools to handle the volume and velocity of emerging data. Near-real-time functions will need *data aggregation* tools to consolidate the real-time streams into meaningful summaries. Off-line functions will need *data analytics* tools that can process and query the SQL and NoSQL data repositories to derive performance metrics and insights.

Chapter 2. How Emerging Data Sources Will Affect Traffic Management and Traffic Management Centers

The purpose of this chapter is to identify how emerging data sources can enhance current State and local traffic management and TMC activities and functions. This chapter introduces categories of functions that are performed by traffic management systems and the types of data that feed different functions, services, actions, or activities supported by the agency or other peer systems or stakeholders. After reading this chapter, the reader will understand the types systems, functions and services that may be enhanced by the use of emerging data sources that will be available over the next ten years. These functions will be grouped into real-time, near-real-time, and offline categories. Each category of function implies a different process flow from the big data technologies toolkit and implications are discussed in this section.

Chapter Objectives:

- Identify high-priority functions that are enhanced by emerging data sources.
- Identify how agencies could assess the potential influence of emerging data sources on the need for big data tools and technologies
- Identify categories of TMC functions that are associated with real-time, near-real-time, and offline processing of data to inform the requirements for common tools and technologies.

Chapter 3 will then identify an architecture for the integration of big data tools and technologies with existing systems to leverage the emerging data sources. Data acquisition, marshalling, storage, and analysis components are identified for each functional category. Differences of data volumes and velocities are identified for the small, medium, and large systems and deployment footprints are analyzed based on regional deployment differences and the need to share data outside the TMC.

After reading this chapter, the reader will understand:

1. Ways in which emerging data sources from connected vehicles, travelers, and infrastructure will enhance common traffic management functions and potentially enable new services or capabilities.
2. How much connected vehicle and traveler data (i.e. penetration rate) is necessary to enable a traffic management function.
3. Implications for use of big data tools and technologies to capitalize on the emerging data for enhancing each traffic management function.
4. Implications for collection and sharing of emerging data sources among agencies for regional, corridor, and statewide functions.
5. Ways in which traffic management functions can be categorized with respect to processing speed for identification of common requirements of big data tools and technologies.

2.1 Traffic Management and Traffic Management Centers Functions

Moving Ahead for Progress in the 21st Century Act (MAP-21) defines transportation system management and operations (TSMO) as:

> *integrated strategies to optimize the performance of existing infrastructure through the implementation of multimodal and intermodal, cross- jurisdictional systems, services, and projects designed to preserve capacity and improve security, safety, and reliability of the transportation system…including actions such as traffic detection and surveillance, corridor management, freeway management, arterial management, active transportation and demand management, work zone management, emergency management, traveler information services, congestion pricing, parking management, automated enforcement, traffic control, commercial vehicle operations, freight management, and coordination of highway, rail, transit, bicycle, and pedestrian operations; and coordination of the implementation of regional transportation system management and operations investments (such as traffic incident management, traveler information services, emergency management, roadway weather management, intelligent transportation systems, communication networks, and information sharing systems) requiring agreements, integration, and interoperability to achieve targeted system performance, reliability, safety, and customer service levels.*

TSMO is the active management of the transportation network by collecting data on system performance and making adjustments to real-time controls, information, and demand-management strategies. TSMO does not involve the implementation or construction of new facilities or rebuilding existing facilities and a variety of other functions of Departments of Transportation. Traffic management and TMC operations are a discipline within TSMO. This report is focused on traffic management activities within the TMC environment. Other functions of DOTs will likely be affected by new emerging data sources, including the data sources identified in this report, but are not discussed here. In addition, those additional DOT functions may very well be improved by application of the same data acquisition, marshalling, and analysis tools and platforms, but are not the focus of this report.

Emerging data sources from connected travelers and vehicles can improve a wide range of traffic management and TMC functions in the following ways [3]:

- Incident and event management—improved queue detection, verification, incident response, on-site monitoring, and management.
- Road hazard warnings—higher fidelity location information, more accurate confirmation of hazard types, more timely warnings.
- Speed warnings—specific recommendations to different vehicle types based on roadway conditions, including weather; more timely and location-specific warnings.
- Traffic signal timing—better operation in oversaturated conditions, provision of priority for emergency and transit vehicles, more timely updates to fixed timings, broad-based adaptive controls, reduced reliance on physical sensor devices and maintenance, shift towards in-vehicle data delivery, performance monitoring of signals with no physical links to DOT communications infrastructure.
- Freeway ramp metering—more accurate and coordinated corridor metering algorithms.

- Variable speed limits/recommendations and lane-use control strategies—more accurate and coordinated responses, shift towards in-vehicle signage reducing needs for infrastructure investments.
- DMS displays—more accurate messaging, shift towards in-vehicle signage for more personalized recommendations, reduced need for infrastructure investments.
- Work zone implementation—higher safety for workers and drivers, higher resolution maps of work zone geometries, real-time information on new zone locations, less need to manually update locations.
- Broadcasted and Personalized Traveler information—higher fidelity information, more accurate and timely information, personalized recommendations.
- Congestion pricing, road user fees, and tolls—more granular toll rates, more accurate congestion prices, personalized tolls, and road user fees.
- Performance measurement, including weather and emissions monitoring—higher fidelity analysis, more comprehensive coverage of geography, reduced need for infrastructure investments.
- Asset management and maintenance—reduced need for infrastructure investments, faster detection and response to equipment failures and sub-optimal roadway infrastructure conditions.

In the next section, we discuss in more detail how these functions are improved by use of emerging data sources and the need for big data tools and technologies to enable these improvements. The characteristics of the data *temperature*, volume, and velocity then inform how tools and technologies might be deployed to manage the emerging data sources for enhancement to traffic management and TMC functions.

2.2 How Emerging Data Sources will affect Traffic Management and Traffic Management Centers Functions

This section describes how emerging data sources will affect each traffic management function. This categorization of traffic management and TMC functions is derived from [3]. The following categories of issues are discussed for each traffic management function:

- **Roles**: Roles for agencies involved in the function.
- **Impacts and Data Elements**: Impacts of emerging data from connected vehicles, connected travelers (and new types of data from connected infrastructure, where appropriate) to each function and what types of data are relevant.
- **Equipped Vehicles**: Penetration rates of equipped vehicles needed to achieve enhanced functionalities.
- **Infrastructure Requirements**: Infrastructure requirements for collection of emerging data sources and dissemination of information back to connected travelers.
- **Analytics and Systems**: What big data technologies should be considered for analytics, processing, storage, integration with existing traffic management systems, and any requirements for data processing speed.

At the conclusion of the discussion of each traffic management function, key characteristics of emerging data impacts are presented. These key considerations include:

- What significant enhancements to the function can emerging data sources provide.
- The geographic extent of infrastructure (RSEs) and CV data required to enable the function.
- The penetration level of CVs necessary for meaningful enhancements to the function.

- The importance (or lack of importance) of inter-agency coordination of emerging data sources.
- How many TMCs are expected to be involved (either inter-agency or intra-agency).
- The importance of data processing speed and location of data processing assets (i.e. in the field, at the TMC, at a centralized processing location).
- The importance of integration of big data systems with legacy ATMS and other TMC systems.

Table 1 summarizes these characteristics for each of the functions. This table provides a high-level assessment of the viability of emerging data sources to impact that TMC activity. The next chapter will discuss the technology architecture to manage functions with varying "data temperatures" or need for real-time, near-real-time, and offline data transmission and processing. Quantitative estimates of data volumes for a small, medium, and large system are provided for context of the truly "big" nature of the data processing and storage challenge facing TMCs in unlocking the enhancements provided by use of emerging data sources.

Traffic Incident Management

Roles—Traffic incident management is primarily performed by State and local police with coordination and support by TMCs at the State and local level. State and local agencies are, also commonly involved in management of planned events, such as sporting events, concerts, and parades.

Traffic Management Impacts—Emerging data from connected vehicles and video analytics can improve traffic incident management through more accurate identification of incident location and severity, upstream impacts (queue lengths), and real-time situational awareness. Connected vehicle trajectories have the potential to be used to identify rapid deceleration, unusual lane changing behavior and uneven lane utilization—together these may suggest (for example) that a lane blockage has occurred. This data could be used to identify the location and extend of the incident (e.g., one or multiple lanes, location) and support the dissemination of event specific messages to connected vehicles via RSEs in the field or through a centralized cellular based capability, much like existing 511 services. On-site incident management activities can be enhanced through V2I data from equipped agency and first responder vehicles (including the future potential for streaming video, including 3D video).

Analytics and Systems Impacts—Before the prevalence of cell phones, TMCs used algorithms based on loop detector data to identify incidents. 911 calls quickly became a much more reliable method for TMCs to learn of incidents once most drivers had cell phones. However, sudden braking, rapid lane change movement, air bag deployment or even LiDAR point data could make automatic incident detection even faster and more accurate than cell phone calls from motorists. Further, the lane utilization patterns of connected vehicles may allow TMCs to know when and where lanes are blocked without human surveillance from CCTV feeds. This data can be fused with data from connected travelers, police, social media and traditional detectors. This information could then be disseminated to targeted travelers based on their location and heading.

Equipped Vehicles—Moderate penetration of CVs can enable enhanced incident management practices, including incident detection, incident data dissemination, and incident resolution. Traffic management systems disseminate information to many other peer agencies, including law enforcement, towing, and emergency medical services. Equipping these first responder service vehicles with connected vehicle and other enhanced data collection systems can enable many incident management functions. New sources such as on-site streaming 360-degree cameras or streaming LiDAR feeds or snapshots could provide improvements to situational awareness at major crash scenes.

Infrastructure Requirements—Incident location can be identified from basic safety messages collected from CVs by traffic management systems through RSEs. To enable such functions, dense deployment of RSEs would be required.

Key characteristics of emerging data impacts on incident management include:

- **V2I Enhancement potential**: V2I data can provide high resolution incident timing and location information.
- **I2V Enhancement potential**: I2V information can target travelers by location, heading and expected incident duration.
- **Geographic scope**: Incidents occur at a specific location. Only a limited number of RSEs (say, less than five) and CV trajectories (only the vehicle trajectories that are nearby that location) will be relevant to each event. Aggregation of trajectories across wide areas of the transportation network is not required.
- **Emerging data volume needed**: A reasonable penetration level of connected vehicles can provide enhanced incident management functionality.
- **Interagency coordination**: With integrated corridor management, incident management is more commonly including diversion strategies from freeways onto parallel arterials. Coordination of emerging data sources with adjacent agencies will be important.
- **Number of TMCs involved:** Incidents are typically managed by coordination across public safety, State, and local TMCs.
- **Data processing speed**: Incident detection needs to be fast and management actions need to be taken as early as possible. Processing of data for use by TMC operators will need to be timely (e.g. on the order of less than one minute delay between data collection and data presentation).
- **Integration of big data systems**: Integration of existing RDBMS and traffic management systems with new processing of emerging data sources is important.

Roadway Hazard Warnings

Roles—Existence of hazards is especially critical on high-speed roadways (and thus to State DOTs more so than local agencies) and dangerous to a very specific region around the hazard. Information on road hazards is less important to be disseminated across vast geographic regions.

Traffic Management Impacts—Emerging data from connected vehicles is highly important in improving hazard warnings (including work zones) by collecting reports of transient hazards from the vehicles themselves. Similar to incident management, connected vehicle trajectories can be used to identify rapid deceleration, unusual lane changing behavior and uneven lane utilization which together may suggest the presence of a roadway hazard. The presumed hazard location could then be rapidly disseminated upstream via RSEs in the field, potentially with edge processing, or through a centralized cellular based capability, much like existing 511 services. (The “edge” refers to the closest point between the vehicle and the infrastructure. In the case of DSRC applications for active safety, this is at the RSE. Since the processing must be done in microseconds, for, say, red-light running warning, it is not likely that the V2I data could be shipped to a TMC and then back to the vehicle in a timely manner. Those applications require deployment of the application on the RSE) Key elements of big data technologies relevant to roadway hazards include analysis methods for deducing the existence of a transient hazard and its geo-location from the trajectory information and embedded elements of the individual snapshots. Emerging sources like 3D LiDAR streaming/snapshots or 360 video/snapshots or even 2D mobile photos could provide precise information on the type of hazard at a particular location, and help to distinguish between hazards and traffic incidents.

Analytics and Systems Impacts—Key elements of big data technologies relevant to roadway hazards include analysis methods for deducing the existence of a transient hazard and its geo-location from the trajectory information and embedded elements of the individual snapshots. TMC software would need to be able to process trajectories in real-time and combine them by location to determine whether traffic is swerving or braking to avoid a hazard. Video analytics are unreliable in outdoor environments. Traveler information systems are not able to send targeted alerts in an immediate area. Roadway hazards are typically not broadcast over large areas via systems like 511. Agencies are typically aware of hazards reported by the public. Dissemination of roadway hazard data through connected vehicle technologies is a new service capability for TMCs.

Equipped Vehicles—Induced location of roadway hazards from atypical CV trajectories (i.e. swerving, partial lane changes) would require at least a reasonable number of repetitions to avoid many false positives.

Infrastructure requirements—Hazard location can be identified from CV trajectory data from public sources faster than commercial sources. Dissemination could occur through RSEs or via cellular channels.

Key characteristics of emerging data impacts on roadway hazard warnings include:

- **V2I Enhancement potential**: V2I data are necessary to enable roadway hazard warnings.
- **I2V enhancement potential**: I2V connections are necessary to enable rapid dissemination of roadway hazard data.
- **Geographic scope**: Roadway hazards happen at a specific location. A limited number of RSEs and CV trajectories are relevant to each event. Travelers far away from the hazard do not necessarily need to be warned.
- **Emerging data volume needed**: A reasonable penetration level of connected vehicles required for enhanced functionality.
- **Interagency coordination**: Coordination with adjacent agencies is not critical as hazard data is only relevant within a limited distance upstream of the hazard.
- **Number of TMCs involved**: Sharing of hazard data with other TMCs would not likely be necessary.
- **Data processing speed**: Avoidance of hazards by approaching vehicles requires extremely responsive data processing (e.g. on the order of seconds of delay between the collection of trajectory data indicating a hazard exists and the dissemination of the hazard location to other approaching connected vehicles).
- **Integration of big data systems**: Processing at the edge implies *limited* need for big data tools and technologies at the TMC except for analysis of performance.

Speed Monitoring and Warning

Roles—Promoting safe driving during adverse conditions (e.g., fog, rain, glare), and/or protecting pedestrians and workers is particularly important on high-speed roadways and sensitive areas such as school zones and thus applies to State and Local agencies alike. Speed warnings are agency-specific and collaboration with law enforcement and safety management staff may be important.

Traffic Management Impacts—Speed warnings are valuable to a very specific region and less valuable when disseminated across vast geographic regions. Emerging data from connected vehicles is highly important in improving speed warnings by collecting reports of adverse conditions from sensors on downstream vehicles. Key elements of big data technologies relevant to speed warning areas include analysis methods for deducing the existence of a transient environmental condition and its geo-location from

the trajectory information and embedded elements of the individual snapshots. Precise geo-referencing of speed warnings is not possible without the trajectory data from connected vehicles and travelers. Emerging sources like 3D LiDAR streaming/snapshots or 360 video/snapshots or even 2D mobile photos could provide precise information on the type of speed issue at a particular location (weather, work zone, incident, school bus, etc.).

Analytics and Systems Impacts—TMC software would need to be able to identify adverse weather either from CVs or roadside environmental sensor stations and provide that information to motorists in the vicinity. While current systems can use environmental sensor stations (ESS) and variable speed limit (VSL) signs to attempt to slow traffic, they are expensive and limited in their coverage. Furthermore, it is not possible to target drivers with individualized messages should they be exceeding a safe speed. Software would need to blend data from a platoon of vehicles to estimate conditions accurately and to provide accurate warnings to motorists.

Equipped Vehicles—speed warning from RSEs can be effective at very low penetration levels of connected vehicles.

Infrastructure requirements—RSEs can be prioritized for placement at critical locations, such as high crash (and run-off-road) locations, sensitive zones, and so on. One RSE location and one CV is all that is required for a successful application.

Key characteristics of emerging data impacts on speed warnings and need for big data tools and technologies include:

- **Enhancement potential**: V2I data are necessary to enable speed warnings.
- **Geographic scope**: Speed warnings occur at a specific location. A very limited number of RSEs (i.e. one) and CV trajectories (i.e. one) are relevant to each event.
- **Emerging data volume needed**: Only a low penetration level of CVs is required for enhanced functionality.
- **Interagency coordination**: Coordination with adjacent agencies is not critical.
- **Number of TMCs involved:** Speed warnings occur at specific locations. Only one TMC will be involved.
- **Data processing speed**: Speed warnings need to be disseminated to the vehicle as quickly as possible. Data processing is critical at the edge (order of less than one second from data collection to warning message dissemination).
- **Integration of big data systems**: Processing at the edge implies limited need for big data tools and technologies and integration with legacy traffic management systems.

Intersection Collision Avoidance

Roles—Traffic signal control and un-signalized intersection operations are performed by State, County, and Local agencies. Intersection collision avoidance is a localized function (I2V and V2V) and does not require coordination among agencies except possibly at agency boundaries.

Traffic Management Impacts—Emerging data from connected vehicles and connected travelers is critically important for intersection collision avoidance. While red-light running detection systems have existed for many years, they have primarily been deployed for enforcement rather than prevention. Intersection collision avoidance services are focused on arterial and local roads. Intersection collision avoidance systems (notably including un-signalized intersections on divided high-speed rural highways) can be a key component of the societal benefits provided by connected vehicle services. At low penetrations, emerging

data can warn equipped vehicles. As more vehicles are equipped, there is more opportunity to communicate warnings from both the infrastructure and vehicles to other vehicles, pedestrians, and cyclists.

Analytics and Systems Impacts—Intersection collision avoidance is not a function of current TMCs. The challenge for TMC software will be to quickly sense unsafe conditions and provide targeted warnings with low latency, while keeping a low rate of false positives. Intersection collision avoidance systems could also identify incidents should collisions occur. Arterial incident management could be enhanced.

Equipped Vehicles—Basic intersection collision avoidance (red-light running warning at a signalized intersection, for example) can be enabled by just one RSE and one CV. Higher-level functions of intersection collision avoidance systems such as left-turn and crossing crash protection, and pedestrian and cyclist warnings, require significant levels of CV penetration and connected travelers.

Infrastructure requirements—RSEs can be prioritized for placement at critical locations, such as at high crash locations and sensitive zones.

Key characteristics of emerging data impacts on intersection collision avoidance and need for big data tools and technologies include:

- **Enhancement potential**: V2I data are necessary to enable intersection collision avoidance.
- **Geographic scope**: Only the RSE at the intersection where the warning occurs is relevant.
- **Emerging data volume needed**: Only a low penetration level of CVs is required for enhanced functionality, which increased levels of equipped vehicles and travelers provides additional capabilities.
- **Interagency coordination**: Coordination with adjacent agencies is not critical.
- **Number of TMCs involved**: Intersection collision avoidance events happen at specific locations. Only one TMC will be involved.
- **Data processing speed**: Collision warnings need to be sent to CVs and connected travelers as quickly as possible. Data processing critical at the edge (order of less than one second).
- **Integration of big data systems**: Limited need for direct integration with legacy traffic management systems except for status monitoring, reporting, and performance measurement.

Traffic Signal Control

Roles—Traffic signal operations are performed by State, County, and Local agencies. Collaboration is key at agency boundaries. Many integrated corridor management (ICM) deployments and coalitions have been formed and many more are emerging.

Traffic Management Impacts—Emerging data from connected vehicles can be significant in improving traffic signal operations. Traditional detection is difficult for most agencies to keep adequately maintained for optimal performance. Trajectory information in and around the traffic signal can indicate performance such as vehicle speed (and lack of vehicle motion when compared to traffic signal status) and location (including specific lane orientation).

Analytics and Systems—Operators in TMCs sometimes make spot signal timing adjustments in real-time, but it is very difficult for humans to anticipate the system implications of localized changes. As a result, most TMC operators will only select between carefully tuned and tested signal timing plans. Adaptive systems are able to make real-time adjustments, but they rely on extensive and reliable detection. Signal timing algorithms will require changes to use connected vehicle data. In the near term, there will be some benefits but also challenges to use of both traditional detection data and data from connected vehicles while the

percentage of CVs in the vehicle fleet is relatively low [10-12]. It may also be possible for big data tools and technologies to operate as a decision support system. Such a system could assist TMC operators in predicting the impacts of localized changes on system performance. Big data tools and technologies also pose an opportunity in traffic signal control systems for long-term trend analysis, asset management, and performance metrics. As current common practice is to discard most data from signal-related connected infrastructure (because of RDBMS limitations), saving that information for much longer periods of time could enable analysis activities that were impossible before.

Equipped Vehicles—At low percentages of connected vehicles in the vehicle fleet, the benefits of emerging data may lie in the ability to observe trends. Areas with worsening trends can be investigated further using conventional methods. Broadcast of signal status (i.e. SPaT) can enable eco-driving and automated vehicle operations. As the penetration level increases (including data from connected travelers such as pedestrians and cyclists), there will be more opportunity for real-time decision making and control using the emerging data sources.

Infrastructure—It is not likely that traditional detectors will be decommissioned for a very long time. However, even with a small percentage of connected vehicles in the vehicle fleet, information on travel times and wait times may be a valuable input into future adaptive traffic signal control systems. Arterial streets experience incidents that are typically unmanaged by any TMC staff. Data from emerging sources of connected vehicles offers the ability for rapid arterial incident detection, which can be fed into adaptive traffic management systems or sent to TMC operators as alerts to adjust timings manually.

Key characteristics of emerging data impacts on traffic signal control and need for big data tools and technologies include:

- **Enhancement Potential**: V2I data can improve traffic signal control through higher-resolution information.
- **Geographic scope**: Traffic control strategies can be implemented for isolated locations, but in cities it is more common that signals are operated in groups. A significant number of RSEs and CV trajectories will be relevant to each system.
- **Emerging data volume needed**: A low percentage of CVs in the vehicle fleet is sufficient for some functionality. Additional capabilities become available with higher percentages of CVs.
- **Interagency coordination**: Most traffic signals are operated by a single agency. Coordination (i.e. sharing of connected vehicle and traveler trajectories) with adjacent agencies is not critical, except at agency boundaries and crossings. Coordination of signal timing parameters such as cycle times and offsets is critical at agency boundaries to maintain efficient traffic flow.
- **Number of TMCs involved**: traffic management is typically centralized in one TMC.
- **Data processing speed**: Data processing rate is critical for timely changes to traffic control parameters (order of less than one minute) and less critical for offline trend analysis and decision support.
- **Integration of big data systems**: Offline analysis will require integration of big data tools and technologies with legacy ATMS.

Probe Data Collection

Roles—All agencies will be involved in collection of probe data to the extent that RSEs are deployed in specific jurisdictions. Link speeds currently available from commercial probe data providers are not high enough resolution to enable enhanced functions described in other sections. The individual trajectory information from V2I is necessary to enhance other TMC functions. Some agency coalitions may emerge to

simplify data sharing by developing memoranda of understanding (MOUs) related to ownership of hardware/data and roles for operations and maintenance of equipment and applications.

Traffic Management Impacts—Probe data collection (which includes the collection of BSM trajectories as vehicles pass RSEs and trajectory data from commercial CV providers) is the cornerstone of how V2I data are used for other all other functional packages. Big data tools and technologies are critical for processing, storage, analysis and sharing of the information from the trajectory data.

Analytics and Systems Impacts—It is possible TMCs will use probe data for network monitoring but the factors that will make this possible for TMCs (high penetrations of CVs) will also make commercial providers' data better. Given market forces beyond transportation, it is likely private companies will continue to lead in this area. However, private companies limit data sharing, which may spur TMCs to aggregate CV data to share with neighboring jurisdictions. How the trajectory data is processed and shared across regions with multiple jurisdictions and collection points is a key issue for institutional integration as well as technical software and system integration issues. Big data tools and technologies are integral to the enabling of most other traffic management functions.

Equipped vehicles—Benefits are directly proportional to the penetration level of data from CVs.

Infrastructure—Benefits are directly proportional to the deployment level of RSEs for functions that require RSEs.

Key characteristics of emerging data impacts on probe data collection and need for big data tools and technologies include:

- **Enhancement potential**: V2I data are the foundation of wide-spread probe data collection.
- **Geographic scope:** All RSEs and CV trajectories are relevant.
- **Emerging data volume needed**: Functionality improves as penetration rate increases.
- **Interagency coordination**: Coordination with adjacent agencies is critical for cross-agency functions.
- **Number of TMCs involved**: Distribution of TMCs as aggregation points for data is a critical consideration.
- **Data processing speed:** Data processing rate is critical for timely distribution to time-critical functions (order of less than one minute) and less critical for offline analysis such as emissions monitoring.
- **Integration of big data systems**: Critical for aggregation of trajectory information from multiple agencies.

Traffic Metering

Roles—Metering is typically a function of a State DOT. Coordination of metering rates with adjacent interchange and traffic signal timing is an important enhancement that could be achieved with fusion of CV-trajectories across agency boundaries. A single TMC may be involved if the agency controls both the interchange traffic signals and the ramp meters. Two TMCs may be involved if a local agency manages the arterial and the State manages the ramp meters. Three or more TMCs may be involved if the freeway is a boundary line between two agencies or if the corridor crosses through multiple jurisdictions.

Traffic Management Impacts—Emerging data from connected vehicles is critically important for freeway ramp metering. Traditional detection is difficult for most agencies to keep adequately maintained for optimal performance, and regional coordination of metering rates is quite difficult using traditional detection

methods. Vehicle speeds and locations can indicate the presence of queues that may impact upstream traffic signal operations.

Analytics and Systems Impacts—Similarly to signal timing, CV data could improve metering algorithms. It is likely ramps will retain traditional vehicle detection while CV penetration rates are low. Metering algorithms could take advantage of corridor-level freeway conditions data rather than relying on fixed detector locations. Further, mainline detectors are often not located optimally. Big data technologies can analyze the statistics of vehicle performance for adjusting timings in real-time. While some research indicates that lower levels of penetration of CV-equipped vehicles may result in adequate information for meter rate tuning, heavy penetration levels are required before traditional detection can be decommissioned at the freeway meter itself [10-12]. Advanced coordinated methods that adjust metering rates in a corridor using simulation models can be enabled with provision of high-resolution traffic conditions through connected vehicle data. Similarly, diversions due to corridor management strategies can more easily be assessed with CV trajectories. Big data tools and technologies are relevant for data sharing of trajectories with multiple TMCs.

Equipped vehicles—Many metering rate determination algorithms could be improved with a modest level of CVs in the vehicle fleet.

Infrastructure—Many improved metering methods could be implemented with commercial CV data. Decommissioning detection at ramps would require close to 100% penetration of CVs and deployment of RSEs at every ramp. RSEs at ramps could also make bus-on-shoulder operations safer by holding the meter in red when the bus is in the conflict zone on the freeway shoulder.

Key characteristics of emerging data impacts on traffic metering and need for big data tools and technologies include:

- **Enhancement potential:** V2I can improve ramp metering operations, particularly for corridor management algorithms.
- **Geographic scope**: Ramp meter management is more effective when considering a series of ramps in a corridor together as a system. A significant number of RSEs and CV trajectories will be relevant to each system.
- **Emerging data volume needed**: A reasonable penetration level of CVs would be required for enhanced functionality.
- **Interagency coordination:** Coordination with adjacent agencies is important since ramp meters are typically the jurisdictional boundary between local agencies and State DOTs.
- **Number of TMCs involved**: Ramp metering may involve (1) only the State DOT, (2) only the State DOT and one local agency, or (3) State DOT and two local agencies (one on each side of the freeway corridor. one, two, or three TMCs may be involved.
- **Data processing speed**: Data processing rate is critical for timely control application (order of less than one minute). Integration of big data systems: big data tools and technologies are likely needed to aggregate and share CV data when multiple TMCs are involved in ramp metering operations.

Lane Management

Roles—Lane management is typically handled by the State DOT on freeways in isolation. Diversions from the freeway to parallel arterials may include coordination with local agencies. Some arterial management agencies have deployed arterial lane management systems, but currently such systems are not common. A single TMC may be involved in lane management if diversions are not considered. Two TMCs may be involved if a local agency manages the arterial and the State manages the freeway systems. Three or more

TMCs may be involved if the freeway is a boundary line between two agencies or if the corridor crosses through multiple jurisdictions.

Traffic Management Impacts—Emerging data from connected vehicles is critically important for lane management (Active Traffic Management (ATM) and Variable Speed Limits). Traditional detection is difficult for most agencies to keep adequately maintained for optimal performance, and determination of lane closure recommendations and speed advisories (or limits) is difficult using traditional detection methods. CV trajectories can better inform agencies of lane utilization.

Analytics and Systems Impacts—Current lane management systems rely on point detection and overhead signage at ¼- to ½- mile spacings. In addition, compliance is difficult to monitor as these systems tend to be advisory rather than regulatory. TMC algorithms could use CV data to better tune these systems based on actual compliance. They may even be able to target specific non-compliant drivers with warnings. For scenarios with reversible lanes, TMCs be able to detect wrong-way drivers much faster both from wrong-way vehicle trajectories and from trajectories of nearby vehicles. Key elements of big data technologies relevant to lane management include analysis methods for statistical assessment of vehicle performance for adjusting lane controls and speed advisories in real-time.

Equipped vehicles—Some research indicates lower levels of penetration of CV-equipped vehicles may result in adequate information for lane management [9-11]. Combined with other third-party traffic conditions data from commercial sources (e.g., HERE, Google, TomTom, etc.), traditional freeway detection stations will become decreasingly necessary over the next 10 years.

Infrastructure—At low penetrations, connected vehicle data could be used to fine tune lane assignment parameters based on a sample of equipped vehicles. At higher penetration levels the data could be used for control, i.e. assigning specific trajectories to individual vehicles in a coordinated fashion to improve bottleneck performance.

Key characteristics of emerging data impacts on lane management and need for big data tools and technologies include:

- **Enhancement potential**: V2I data can improve lane management strategies.
- **Enhancement potential**: I2V information can inform connected vehicles and travelers of downstream issues without need for field infrastructure (gantries, lane control systems (LCS), and VSL).
- **Geographic scope**: A significant number of RSEs and CV trajectories are relevant to each system.
- **Emerging data volume needed**: A reasonable penetration level of CVs would be required for enhanced functionality.
- **Interagency coordination**: Coordination with adjacent agencies is not important, except with diversion strategies.
- **Number of TMCs involved**: Typically managed by only one TMC.
- **Data processing speed**: Data processing rate is critical for timely control application (order of less than one minute from data collection to control actions).
- **Integration of big data systems**: Big data tools and technologies are relevant for statistical analysis and processing of probe trajectories.

Electronic Payments / Fee Collection

Roles—Fee collection and electronic payments are typically managed by Toll Road Authorities. Migration of DOT revenue mechanisms to road-user fees from gasoline taxes may bring other agencies into the role of electronic payment collection from road users (Department of Motor Vehicles (DMV), DOT, or other).

Traffic Management Impacts—Emerging data from connected vehicles can be used by toll authorities for setting congestion prices and billing road users for miles-driven versus traditional gasoline taxes. Traditional detection is difficult for most agencies to keep adequately maintained for optimal performance.

Analytics and Systems Impacts—TMCs could use CV data to improving pricing models to control demand and need not be limited to specific facilities with expensive gantries, tag readers and license plate recognition (LPR). Key elements of big data technologies relevant to fee collection include analysis methods for statistical assessment of vehicle performance for adjusting prices in real-time.

Equipped vehicles—While some research indicates lower levels of penetration of CV-equipped vehicles may result in adequate information for pricing, heavy penetration levels are required to collect fees (such as road-user fees and congestion charges) via the connected vehicle equipment itself [9-11].

Infrastructure—Adequate alternative technologies already exist (i.e., license plate reading) for assessment of fees, so connected vehicle data is most relevant for *informing* pricing strategies or future methods of road user charging.

Key characteristics of emerging data impacts on electronic payment and need for big data tools and technologies include:

- **Enhancement potential**: V2I can improve real-time congestion pricing strategies.
- **Geographic scope**: A significant number of RSEs and CV trajectories are relevant to determining congestion prices.
- **Emerging data volume needed**: A reasonable penetration rate of CVs is required for enhanced functionality.
- **Interagency coordination**: Coordination with adjacent agencies is not important.
- **Number of TMCs involved**: Congestion prices are typically managed by only one TMC.
- **Data processing speed**: Data processing rate is semi-critical for timely fee determination (order of minutes from data collection to changes in prices).
- **Integration of big data systems**: Big data tools and technologies are relevant for long-term trend analytics.

Traffic Information Dissemination

Roles—Traffic information is typically disseminated by State DOTs or regional Metropolitan Planning Organization (MPOs) through 511 phone, web, and app services. A handful of County and Local agencies provide similar services for limited jurisdictions.

Traffic Management Impacts—Emerging data from connected vehicles can be used by agencies to enhance traffic conditions information, particularly for anomalies such as roadway hazards and weather. Key elements of big data technologies relevant to traffic information are the coordination of the information across wide geographic areas and multiple agencies. Emerging data sources improve traffic information by increasing the fidelity of data available in each roadway segment. Current sources provide only speed. Emerging data from connected vehicles can provide additional metrics such as precise locations of incidents and hazards, on-site photos of incidents, etc.

Analytics and Systems Impacts—TMC software will likely need to be updated to make targeted traveler information services based on driver location, heading and their location relative to congestion, an incident or a road hazard. Key elements of big data technologies relevant to traffic information include analysis methods for statistical assessment of vehicle trajectories for identifying hazards, incidents, etc. As has been demonstrated by 511 deployments, regional coordination across metropolitan areas is more common and desirable.

Equipped vehicles—Pathways directly back to the connected vehicles and travelers either through RSEs or cloud connections can provide targeted traffic information tailored for each individual user rather than generic traffic conditions data that may or may not be relevant to a driver's current location or status.

Infrastructure—Pathways directly back to the connected vehicles and travelers either through RSEs or cloud connections can provide targeted traffic information tailored for each individual user rather than generic traffic conditions data that may or may not be relevant to a driver's current location or status.

Key characteristics of emerging data impacts on traffic information dissemination and need for big data tools and technologies include:

- **Enhancement potential**: V2I data can augment existing traffic information dissemination.
- **Geographic scope**: A significant number of RSEs and CV trajectories are relevant to dissemination of traffic information.
- **Emerging data volume needed**: Only a low penetration level of CVs is required for enhanced functionality.
- **Interagency coordination**: Coordination with adjacent agencies is important.
- **Number of TMCs involved**: Traffic information dissemination is typically consolidated across multiple TMCs for Statewide dissemination through 511.
- **Data processing speed**: Data processing rate is critical for timely information dissemination (order of less than one minute from data collection to information dissemination).

Emissions Monitoring and Management

Roles—Emissions monitoring and modeling is typically the responsibility of the MPO or adjacent agencies such as the department of air quality.

Traffic Management Impacts—Emerging data (trajectory information of vehicles with high-resolution of speed vs location including vehicle type, make, model, year, engine type, etc.) from connected vehicles can be used by agencies to increase the accuracy of emissions monitoring by much higher resolution modeling. Traditional emissions detection equipment is highly localized. There is a host of work ongoing in development of lower cost air-quality sensors, which would be emerging as new types of connected infrastructure over the next ten years. Key elements of big data technologies relevant to emissions monitoring include the statistical analysis methods assembling the data for existing emissions calculation models that use trajectory data and distributed point-measurements from air quality sensors as inputs.

Analytics and Systems Impacts—Emissions monitoring is not a current function of TMCs but with CV data they could assist with ground-level air quality assessments and warnings. Near-real-time air quality alerts could be coupled with information dissemination strategies to encourage more eco-friendly routes or to discourage driving altogether.

Equipped vehicles—Some research indicates that lower levels of penetration of CV-equipped vehicles may result in adequate information for emissions estimation [12-14].

Infrastructure—Synthesis of regional-scale emissions models is only possible with collection of vehicle trajectory data (and/or low-cost air quality sensors at fixed points) over vast geographic areas. This is a new functional area for TMCs enabled by emerging data sources from connected vehicles.

Key characteristics of emerging data impacts on emissions monitoring and need for big data tools and technologies include:

- **Enhancement potential:** V2I data are critical for improving emissions monitoring through regional modeling.
- **Geographic scope**: Emissions monitoring has a wide geographic purview. A significant number of RSEs and CV trajectories will be relevant to each system.
- **Emerging data volume required:** A reasonable penetration level of CVs is required for enhanced functionality.
- **Interagency coordination**: Coordination with adjacent agencies is important.
- **Number of TMCs involved**: Consolidated across multiple TMCs for Statewide evaluation.
- **Data processing speed**: Data processing rate is not critical (order of days-weeks).
- **Integration of big data systems**: Big data tools and technologies are critical for large scale analysis and processing of trajectory information for use in regional emissions models.

Road Weather Monitoring and Management

Roles—Road weather monitoring is typically performed by State DOTs and some County and regional agencies for specific functions, such as high water detection in flood-prone cities (e.g. Houston, San Antonio, New Orleans, etc.)

Traffic Management Impacts—Emerging data from connected vehicles can be used by agencies to increase the accuracy of road weather monitoring. Traditional road weather detection (i.e., surface temperature, icing) equipment is highly localized. Road weather monitoring from connected vehicles provides high-resolution surface temperature and conditions data from long stretches of roadway. Atmospheric weather monitoring via satellite cannot measure with close enough detail road surface issues that can be assessed using vehicle-based sensors.

Analytics and Systems Impacts—TMCs currently monitor weather conditions for traveler information purposes. TMC software could utilize CV data to augment physical sensors and weather forecasts and target alerts. Key elements of big data technologies relevant to weather monitoring include the statistical analysis methods for assembling the data for weather models that use trajectory data as inputs. Some research indicates that lower levels of penetration of CV-equipped vehicles may result in adequate information for road weather monitoring [9-11].

Equipped vehicles—Low penetration levels could inform agencies if weather is present on certain roadways.

Infrastructure—Commercial sources (when/if available) may provide higher coverage than public collection via RSEs.

Key characteristics of emerging data impacts on road weather monitoring and need for big data tools and technologies include:

- **Enhancement potential**: V2I data can improve road weather monitoring.

- **Geographic scope:** a significant number of RSEs and CV trajectories are relevant to each system, although road weather warnings can be localized in nature.
- **Emerging data volume needed:** Only a low penetration level will be required for enhanced functionality.
- **Interagency coordination:** Coordination with adjacent agencies will be critical for holistic modeling.
- **Number of TMCs involved**: Road weather data would likely be consolidated across multiple TMCs for Statewide evaluation.
- **Data processing speed:** Data processing rate is semi-critical (order of minutes from data collection to information dissemination).
- **Integration of big data systems**: Big data tools and technologies will be important for consolidation of road weather data across regions.

Asset Management

Roles—Depending on the asset type, asset management is performed by State, County, Local, and regional agencies. Coordination and collaboration across agencies is not critical unless responsibility for actions are shared (such as memorandums of understanding, MOUs, regarding ownership and O&M).

Traffic Management Impacts—Emerging data from connected vehicles can be used by agencies to increase the accuracy of asset management, particularly pavement conditions assessment. Traditional pavement assessment is highly resource-constrained involving individual trailer units dispatched on a rotating basis throughout a region allowing a measurement of pavement quality perhaps at best once per year. Pavement assessment measurements from connected vehicles could increase ability of repair crews to fix ailing locations sooner. Particularly in connected infrastructure functions, such as traffic signal systems, the traditional practice in most ATMS is to archive status information for later retrieval or simply throw the data away after a few months of storage in the RDBMS. Device status trends could be deduced if the data was stored for years using the power of big data tools and NoSQL data stores. Additional data from non-traditional sensors such as bridge condition monitors open up new areas of asset management.

Analytics and Systems Impacts—TMC software could use sensor data to detect device health and status. As more devices become connected, it will be more feasible to locate devices and track status in real-time. As devices are upgraded and replaced, they can self-identity. TMC software can also incorporate surface roughness through motion sensors on vehicles. Key elements of big data technologies relevant to asset management include the statistical analysis methods for assembling the data for pavement condition models that use trajectory data as inputs. Big data tools for data analysis of large scale data sets are critical for processing and using trajectory information from connected vehicles for asset management. Big data databases and processing tools open up new insights in field device performance evaluation.

Equipped vehicles—Some research indicates lower levels of penetration of CV-equipped vehicles may result in adequate information for pavement conditions monitoring [9-11]. Low penetration levels could inform agencies about pavement quality on certain roadways, better coverage of regions is directly proportional to penetration level.

Infrastructure—Commercial sources (when/if available) may provide higher coverage than public collection via RSEs.

Key characteristics of emerging data impacts on asset management and need for big data tools and technologies include:

- **Enhancement potential:** V2I data can improve asset management by providing higher fidelity information on road conditions.
- **Geographic scope:** Entire agency jurisdictions are involved. A significant number of RSEs and CV trajectories will be relevant to each system.
- **Emerging data volume needed**: Only a low penetration level of CVs is required for enhanced functionality.
- **Interagency coordination:** Asset management is agency-centric. Coordination with adjacent agencies is not significant.
- **Number of TMCs involved**: Consolidate data across multiple TMCs will be likely for Statewide evaluation.
- **Data processing speed:** Data processing rate is not critical (order of days-weeks).
- **Integration of big data systems:** Big data tools and technologies for large scale data analysis will be highly relevant.

Parking Management

Roles—Urban parking is typically managed by local agencies, in public-private partnerships, and by the private sector. Many State agencies manage truck parking facilities on Interstate roads.

Traffic Management Impacts—Emerging data from connected vehicles can be used by agencies to increase the accuracy of parking management. Traditional parking utilization detection is difficult for agencies to maintain, particularly for State-managed freight facilities. However, a wide variety of new connected infrastructure approaches to parking management may emerge over the next ten years lowering the cost. Local agencies with agency-managed parking facilities may find value in use of connected vehicle data to inform travelers of parking availability for events, or in urban centers.

Analytics and Systems Impacts—TMCs don't typically manage parking but some have truck parking systems off Interstates. TMC software could be used to predict parking shortages and target communications to truck drivers, optimizing available parking capacity and helping to ensure all drivers can safely park when their time limitations are reached. Offline parking management functions, such as building models of parking lot utilization may be enabled by storing vehicle trajectories over long periods of time.

Equipped vehicles—Connected vehicles both commercial and public can provide new ways for TMCs to distribute parking availability information that is geo-referenced to the user's current position. Emerging data sources of connected vehicle trajectories may not directly improve parking management until high levels of penetration are achieved in the fleet.

Infrastructure—RSEs at truck parking areas are probably necessary for truck parking management. For other urban parking applications, RSEs at key lots may track activities, but direct improvements in parking management are likely only with high levels of penetration.

Key characteristics of emerging data impacts on parking management and need for big data tools and technologies include:

- **Enhancement potential**: V2I data can augment existing parking management systems.
- **Geographic scope:** Parking information is regional in nature and only relevant near where a CV is currently or where it is destined. A significant number of RSEs and CV trajectories are relevant to each system.

- **Emerging data volume needed**: A high penetration level of CVs is required for enhanced functionality.
- **Interagency coordination:** Coordination with adjacent agencies is not significant.
- **Number of TMCs involved**: Parking data would typically be managed by only one agency TMC.
- **Data processing speed**: Data processing rate is semi-critical (order of minutes from data collection to information dissemination).
- **Integration of big data systems**: Big data tools and technologies will be important for data processing and long-term data analytics.

Performance Measures

Roles—All agencies are involved in some level(s) of performance measurement and monitoring.

Traffic Management Impacts—Emerging data from connected vehicles can be used by agencies for more accurate performance measurement.

Analytics and Systems Impacts—TMCs are currently challenged with assimilating data from a variety of sources and deriving measures of traffic management performance. The proliferation of big data will only increase the demand for detailed reporting while it makes more data available to calculate meaningful measures. Key elements of big data technologies relevant to performance measurement is the marshalling of the information in a flexible manner so that a wide variety of calculations and queries can be accomplished by agency data scientists in a reasonable amount of time and with a minimum amount of data manipulation. New and unique performance metrics will be available for TMC operations when a significant number of emerging data sources from connected vehicles are stored over time.

Equipped Vehicles—Connected vehicles both public and commercial can provide new ways for TMCs to evaluate performance. Emerging data sources from connected vehicle trajectories may be able to monitor performance measurement with low levels of penetration in the fleet, much like the 2-5% penetration rates of Bluetooth and Wi-Fi travel time measurement field device systems. Some new performance measures may be identified as the percentages of CVs in the vehicle fleet continues to rise.

Infrastructure—Fidelity of performance tracking is directly related to deployment penetration of RSEs. Commercial data acquisition (if/when available) may provide results sooner.

Key characteristics of emerging data impacts on performance management and need for big data tools and technologies include:

- **Enhancement potential**: V2I data can improve performance measurement.
- **Geographic scope:** Performance measures are inherently applicable to entire agency jurisdictions and multi-agency coalitions. A significant number of RSEs and CV trajectories are relevant to each system.
- **Emerging data volume needed**: Only a low penetration level of CVs is required for enhanced functionality.
- **Interagency coordination:** Coordination with adjacent agencies is significant.
- **Number of TMCs involved:** Performance measures data would typically be consolidated across all TMCs and processed centrally.
- **Data processing speed:** Data processing rate is not critical (order of days-weeks).
- **Integration of big data systems:** Big data tools and technologies will be critical for large-scale data analytics.

Table 1. Key characteristics of how emerging data sources will affect Transportation Systems Management and Operations functions.

	Traffic Incident Management	Roadway Hazard Warning	Speed Warning	Intersection Collision Avoidance	Signal Control	Probe Data Collection	Metering	Lane Management	Electronic Payments	Traffic Info	Emissions Management	Weather Management	Asset Management	Parking Management	Performance Management
Potential to improve performance of functions	- Rapid identification and impact assessment - Rapid dissemination via RSEs	- Rapid identification and impact assessment - Rapid dissemination via RSEs	- Rapid identification and warning of unsafe speed - Vehicle-specific and targeted	- Ability to improve safety at intersections	- At low penetrations, can be used to fine tune timing plans - At high penetrations, can be used for real-time control	- BSMs and PDMs used to estimate traffic flows on the network	- At low penetrations, can be used to fine tune metering parameters - At high penetrations, can be used for real-time control	- At low penetrations, can be used for feedback to ATMS - At high penetrations, can be used for lane control	- Can be used in lieu of transponders for toll payments - Can be used to pay for parking or for road-user charging	- Can be used for targeted traveler information based on vehicle location and trajectory relative to incidents	- Even at low penetration, mobile emissions testing can provide better air quality data over a region and targeted emissions enforcement	- Can be used to supplement ESS for wide area road weather monitoring	- Pavement condition monitoring is a substantial cost to agencies and could be supplemented from vehicle data	- At high penetrations, parking availability will enable better utilization of parking spaces and better real-time information	- Even at low penetrations, a sample of vehicle data can supplement traditional sources to measure performance
Number and location of RSEs and CVs to enable function	Low—initial deployments can focus on key corridors	Low—initial deployments can focus on key corridors	Low—location specific	Low—initial deployments can focus on key intersections	Low—initial deployments can focus on key intersections	High—wide coverage needed to improve upon existing sources	Low—location specific	Low—initial deployments can focus on key corridors	Low—limited to toll roads and other pay facilities	High—wide coverage needed to improve upon existing sources	Low	High	High	Low—limited to toll roads and other pay facilities	High
Penetration rate of CVs to enable function	Medium	Medium	Low	Low	Low	Medium	Low	Medium	Medium	Medium	Low	Medium	Low	High	Medium
Need for agency to agency data sharing	Critical	Not critical	Not critical	Not critical	Not critical	Critical for cross-agency functions	Not critical	Not critical	Not important	Important	Not important	Important	Not important	Not important	Important
Processing location needs	One TMC	One TMC	One TMC	One TMC	One TMC	One or multiple TMCs	One TMC	One or multiple TMCs	One TMC	Multiple TMCs	Multiple TMCs	Multiple TMCs	Multiple TMCs	One TMC	Multiple TMCs
Processing speed requirements	Very Fast (<1 minute)	Very Fast (seconds)	Very Fast (<1 second)	Very Fast (<1 second)	Fast (<1 minute)	Fast (<1 minute)	Fast (<1 minute)	Fast (<1 minute)	Medium (minutes)	Fast (<1 minute)	Slow (days/weeks)	Medium (minutes)	Slow (days/weeks)	Medium, Slow (minutes, trends—days/weeks)	Slow (days/weeks)
Big data tools and technology implications	Interfaces with traffic management functions; fusion with traditional sources	Edge function; Limited except for data analysis and performance monitoring	Edge function; Limited except for data analysis and performance monitoring	Edge function; Limited except for data analysis and performance monitoring	Real-time stream ingestion, NoSQL storage, data aggregation and processing, interfaces with traffic management functions	Critical for real-time streaming ingestion, NoSQL storage, data aggregation and processing	Real-time stream ingestion, NoSQL storage, data aggregation and processing, interfaces with traffic management functions	Real-time stream ingestion, NoSQL storage, data aggregation and processing, interfaces with traffic management functions	Data analysis	Real-time stream ingestion, NoSQL storage, data aggregation and processing, interfaces with traffic management functions	Data analysis, NoSQL storage, data aggregation	Data analysis, NoSQL storage, data aggregation	Data analysis, NoSQL storage, interfaces with traffic management systems	Data analysis, NoSQL storage, data aggregation	NoSQL storage, data aggregation and processing, interfaces with traffic management functions, data analysis

(Source: Kimley-Horn and Associates, Inc., 2016.)

2.3 Functional Characteristics of Traffic Management and Traffic Management Centers Functions When Enhanced with Emerging Data Sources

Traffic management functions can be enhanced with big data tools and emerging data sources from connected vehicles, travelers, and infrastructure. A summary of the characteristics of each functional area in a number of big data characteristics is presented in Table 2. In Table 2 we also introduce the concept of "data temperature". Data is **hot** if it should be processed in real-time to be useful (i.e. less than one second from receipt to action). *Real-time* traffic management functions require hot data and processing of trajectories as close to the edge as possible. These functions are shown with red color for data temperature. Data is **warm** if must not necessarily be processed in real time, but actions should be taken reasonably soon to provide benefits (i.e. less than a few minutes from receipt to action). These functions are categorized as "*near-real-time*" functions and shown in orange color. Slow processes can work on **cool** data. Performance metric calculations, emissions model calculations, and asset management actions, for example, can operate with data that has been cleaned, processed and stored. These functions are categorized as "*offline*" functions and shown in blue. Actions on cool data might be taken every few hours, once a day, once a week, or for even longer time horizons.

Volume is the total amount of data required to make a function effective. Infrastructure scalability becomes paramount as data volumes increase exponentially. Data aggregation strategies must be developed. Priorities for what to keep and what to delete may have to be developed.

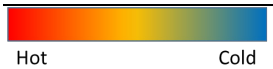
Velocity is the rate at which data is generated and the rate at which the data should be processed for the function to be effective. There are primarily two categories of data processing, *batch* and *streaming*. Batch processing is for analysis done after-the-fact and in large chunks at a time. Data that does not require immediate action can be analyzed independently from the real-time performance of the system. Stream processing enables real-time decision making and alerts by analyzing the data as soon as it arrives. Deciding between streaming and batch analyses depends on the function. In Chapter 3 we discuss the need for both approaches.

Variety refers to the number of different data sources (e.g., Connected Vehicles, Connected Travelers, and Connected Infrastructure) and types of data being generated (e.g., trajectories, status reports, video streams, etc.) relevant to the function. High variety may require considerations for data governance and storage capabilities for structured, unstructured, and semi-structured data. More advanced analysis techniques may be needed to make use of complex data (e.g., unstructured image files).

Veracity refers to the quality of the raw data being received. Biases, noise, abnormalities, or general inaccuracies can introduce significant challenges for data processing. Veracity can play a particularly important role in automated decision making without human interaction and intervention (e.g., adaptive traffic control, automated incident alerts, or future concepts for regional congestion pricing or road user charging). Such functions will have to be tolerant to quality abnormalities of connected vehicle and traveler data.

Table 2. Summary of "big data" characteristics of Transportation Systems Management and Operations functions.

	Volume	Velocity	Variety	Veracity	Current Technical Maturity	Future Analytical Complexity	Data Temp
Traffic Incident Management	Low	Low	**High**	**High**	**High**	Low	Med
Roadway Hazard Warning	Low	**High**	Low	**High**	Low	Low	Hot
Speed Warning	Low	Low	Low	Low	Low	Low	Hot
Cooperative Intersection Collision Avoidance Systems (CICAS)	Low	**High**	Low	**High**	Low	**High**	Hot
Signal Control	**High**	**High**	Low	**High**	**High**	**High**	Med
Probe Data Collection	**High**	**High**	Low	**High**	Low	Low	Hot
Metering	**High**	**High**	Low	Low	Low	**High**	Med
Lane Management	**High**	**High**	**High**	Low	Low	**High**	Med
Electronic Payment	**High**	**High**	Low	**High**	**High**	**High**	Hot
Traffic Information	**High**	**High**	**High**	**High**	**High**	Low	Med
Emission Monitoring	**High**	Low	Low	Low	Low	**High**	Cool
Weather Monitoring	**High**	Low	**High**	**High**	Low	**High**	Med
Asset Management	**High**	Low	**High**	**High**	Low	**High**	Cool
Parking Management	**High**	Low	**High**	**High**	Low	Low	Med
Performance Measurement	**High**	Low	**High**	**High**	Low	**High**	Cool



(Source: Kimley-Horn and Associates, Inc., 2016.)

Current technical maturity refers to the robustness of the current function in most traffic management agencies and TMCs. For example, incident management is a popular and well-established practice in many if not all traffic management organizations at the State DOT level. Intersection collision avoidance or road hazard warning functions, on the other hand, are not common since they are enabled by emerging data sources.

Future analytical complexity refers to the potential for the function to increase in technical sophistication with the use of emerging data sources and increased availability of volume and variety of new information.

Note that not all real-time, near-real-time, and offline function categories have the same set of attributes for volume, velocity, variety, etc. of emerging data sources.

These TMC functions are categorized in terms of processing speed requirements ("real-time", "near-real-time", and "offline") are summarized as follows in Table 3.

Table 3. Groupings of Transportation Systems Management and Operations functions by data temperature.

Real-time Functions	Near Real-time Functions	Offline Functions
Hazard Warning	Incident Management	Emissions Monitoring
Speed Warning	Signal Control	Asset Management
Intersection collision avoidance	Metering	Performance Measurement
Probe Data Collection	Lane Management	
Electronic Payment	Traffic Information	
	Weather Monitoring	
	Parking Management	

(Source: Kimley-Horn and Associates, Inc., 2016.)

In this Chapter we have described how each traffic management function will be enhanced with the use of emerging data sources from connected vehicles and travelers, connected infrastructure, and emerging data from mobile video. Real-time functions include probe data collection, hazard warning, speed warning, and intersection collision avoidance. These functions will likely operate on the. These functions are less reliant on big data tools for data collection at the TMC, except for analytics on the performance of the functions. Probe data collection is the cornerstone of all other functions. To collect the massive amounts of data from connected vehicles and travelers, the probe data function will need tools designed for rapid data processing and ingestion of massive data streams and NoSQL storage methods to store and share the collected information with other processes. Near-real-time functions will build upon the probe data collection module needs with similar tool requirements, but with less stringent requirements on the timeliness required to process the new emerging data sources. Offline functions need massive amounts of information, collected over longer time frames. These functions need significant storage of emerging data sources but do not require the data be processed in real-time. Big data aggregation and analytics tools will be critical to offline functions.

After reading this chapter, the reader should understand:

1. Emerging data sources from connected vehicles, travelers, and infrastructure will enhance common traffic management functions and enable new traffic management services at TMCs.
2. Most TMC functions will enhanced through moderate penetration of CVs. Some functions are enabled with very low levels of CV deployment and very few require near 100% CVs.
3. Probe data collection is the cornerstone of most other TMC functions. When significant levels of CVs are in the fleet, big data tools and technologies will be necessary to acquire, store, process, share, and analyze the data for use by all of the other functions.
4. Some real-time functions may not need big data tools if they are deployed directly on the RSE. Aggregation methods will likely be necessary to streamline the sharing of data collected on the RSE with the TMC.
5. Near-real-time functions have moderate needs for high-throughput processing of CV trajectories. Processing emerging data sources within minutes will be necessary to turn the data into information and take timely traffic management actions.
6. Offline functions will require significant data storage and data analysis capabilities. Big data tools and technologies will be critical for successful implementation of offline functions.

In the next chapter, we will describe a generic data processing architecture using big data tools and technologies that uses the emerging data sources to enhance each of the three categories of TMC functions. Specific tool types for each big data technology category are also discussed.

Chapter 3. How Big Data Tools and Technologies Can Enhance Traffic Management and Traffic Management Center Functions

Chapter Objectives:

- Identify what big data tools and technologies may be appropriate for different traffic management or TMC functions
- Identify how agencies could assess the potential volume, velocity and implications of collecting, compiling and using connected vehicle and traveler data on the performance of traffic management systems and TMCs
- Provide examples of how big data tools and technologies could influence changes to the architecture for data sharing among agencies.

The purpose of this chapter is to identify how big data tools and technologies could be deployed to enhance the traffic management systems and TMC functions. After reading this chapter, the reader will have an understanding of the characteristics of the big data tools and technologies to be considered for integrating the information from emerging data sources with ATMS in TMCs. The data temperature categories as introduced in Chapter 2 (real-time, near-real-time, and offline) are used to categorize the types of big data tools and technologies for different types of traffic management functions. A system architecture for deployment of big data technologies in a TMC is then presented.

After this discussion, typical emerging data volume and velocity estimates are presented for small, medium, and large agency systems. This discussion also provides a framework for any agency to estimate their data storage and processing needs from some basic assumptions about the size of the region (population, number of traffic signals, etc.) and the expected deployment of V2I devices. Finally, data sharing with peer and partner agencies is discussed. Implications for changes to the deployment of big data tools and technologies due to the needs for data sharing are introduced for a variety inter-agency structures. After reading this chapter, the reader will have an understanding of the data volume and velocity estimates and implications for data processing hardware and storage requirements of big data tools and technologies for typical small, medium, and large agencies responsible for TMC functions.

3.1 Integrating Big Data Tools and Technologies with Traffic Management Systems

In Chapter 1, Figure 1 illustrated the basic concept of integration of big data tools and technologies with traffic management systems in a TMC. Figure 2 expands upon Figure 1 by illustrating the tools and technologies required to acquire, store, and analyze emerging data sources and integrate with existing traffic management systems. In Figure 2, color is used to indicate data temperature or need for real-time, near-real-time, and offline data transmission and processing. Components shown in ***red*** are *real-time* components. (A component can refer to a software service, an interface, an application, or a database.) Red lines and red boxes represent software components and data flows that process real-time data and have control actions that need to be taken within seconds. ***Orange*** components and arrows are *near-real-time* components. These elements have need to process data quickly and take actions on the order of minutes. ***Blue*** arrows and software components are *offline* elements. These elements can process data on-demand, in batches, or even in daily updates. Components shown with the **yellow elephant icon** (the mascot of the Hadoop Distributed File System (HDFS)) are members of the big data ecosystem. These are the new technology elements that will be necessary to harness the value of the emerging data sources. These components are specifically designed for acquiring, storing, and processing massive data sets.

The components of this conceptual system will be discussed starting from left to right. As is typical in current intelligent transportation system architectures, data from field equipment is usually brought back to the agency's Traffic Management Center through wire-line and wireless communication networks. It is assumed in this deployment analysis that agency-owned RSEs will be connected over high-speed communications to the TMC. Connected vehicles will exchange data with agency RSEs and consume information from real-time connected vehicle functions. Basic Safety Message and Probe Data Message information will be collected through the RSEs and transmitted to the TMC as they are received. We assume that *connected traveler* and *commercial connected vehicles* data will be delivered through cloud Application Programming Interfaces (APIs) to the TMC. These data will be acquired in near-real-time. We represent this data being acquired by the same component as the data from the RSEs.

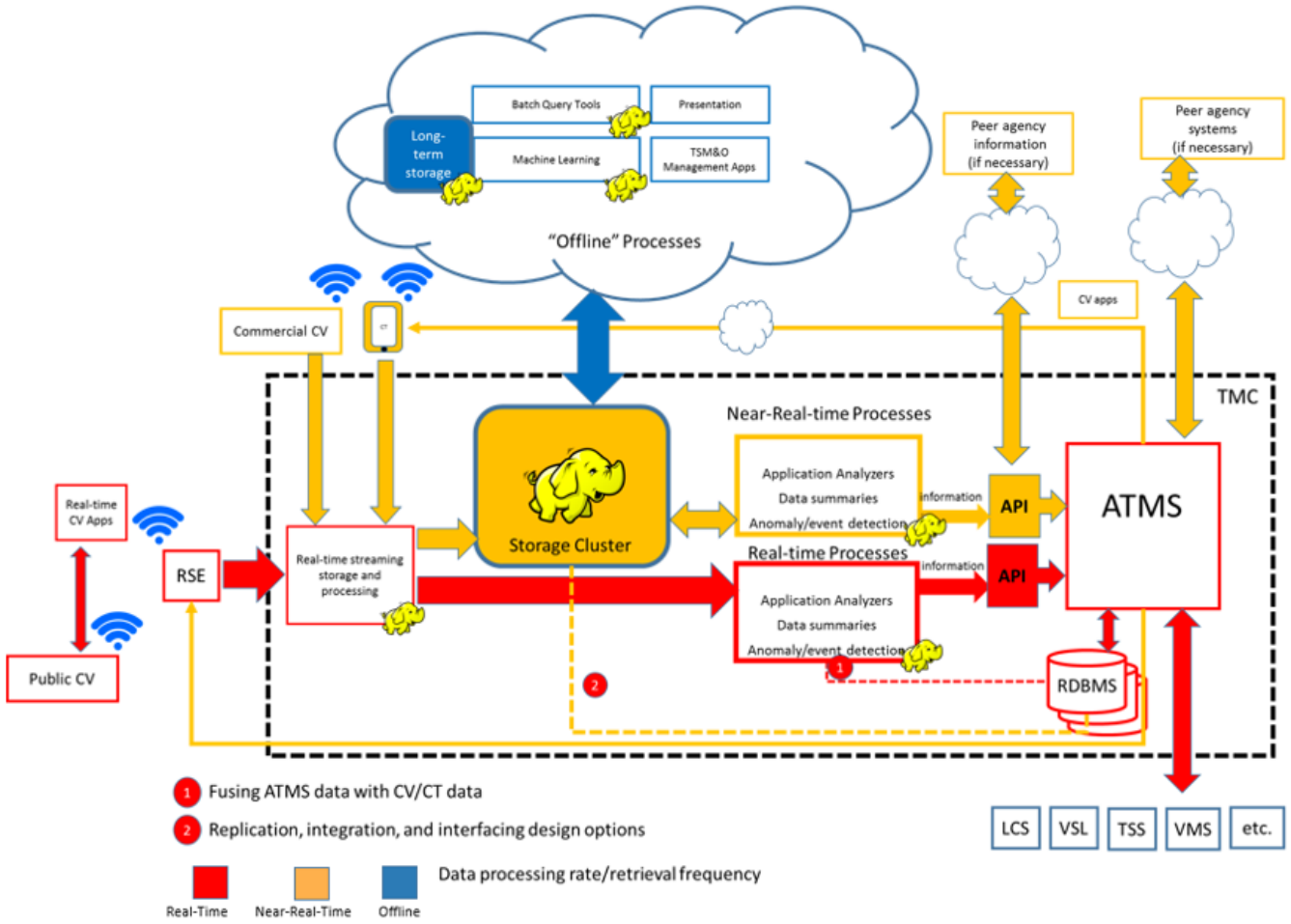


Figure 2. Diagram. Big data architecture for a single Traffic Management Center.
(Source: Kimley-Horn and Associates, Inc., 2016.)

For the purpose of this illustration, we will assume that existing ATMS in a TMC are represented as a single component. While many agencies have multiple, separate applications for specific purposes (such as having one system that manages ATM devices, another that manages traffic signals, another than manages dynamic message signs, and so on), the principles remain the same whether the agency has one ATMS or multiple systems.

The solution architecture depicted in Figure 2 includes the following components:

- Public connected vehicles providing trajectory data of their status through RSEs.
- Commercial connected vehicles and connected travelers providing trajectory data of their status through cloud-based APIs.
- A data acquisition component that ingests high-velocity, high-volume trajectories in real-time.
- A NoSQL component for storing the trajectory data and derived real-time analytics.
- A *real-time* analysis and data evaluation component that interacts directly with the real-time streaming acquisition component for processing real-time data as quickly as possible.
- A *near-real-time* analysis and data evaluation component that interacts with the NoSQL component for near-real-time functions.
- A method by which the NoSQL component and the ATMS RDBMS are merged and used together for data analysis of real-time, near-real-time, and offline functions.
- An API interface from the real-time analysis component to the ATMS.
- An API interface from the near-real-time analysis component to the ATMS.
- A set of *offline* analysis components that reside in a cloud system for trend analysis of the data in the NoSQL store, in conjunction with the traditional ATMS data and meta-data, to provide new services and TMC functions.
- The existing ATMS system(s), comprising both Graphical User Interface (GUI) functions and back-end services and processing components.
- Field device components that display traffic controls and advisories (LCS, VSL, traffic signal systems (TSS), variable message signs (VMS), etc.).
- The existing ATMS RDBMS database(s).
- Interfaces from the ATMS to peer agency systems (through wide-area network connections of various types—cloud, fiber, wireless) to coordinate control actions.
- Interfaces from the results of near-real-time emerging data analysis components to peer agency systems (through wide-area network connections of various types—cloud, fiber, wireless).
- CV apps that send information back to Commercial CV and connected travelers (through wide-area network connections of various types—cloud, fiber, wireless).
- CV apps that send information back to RSEs for local broadcast of information and controls (through wide-area network connections of various types—cloud, fiber, wireless).

The characteristics of big data tools and technologies for each component of the system are discussed in the following sections.

3.1.1 Field-to-Traffic Management Center Data Collection

As shown in Figure 2, agency CV data in the form of trajectories of BSMs and trajectories embedded in PDMs are forwarded from the field RSEs in a real-time manner. We assume in this architecture that the communications bandwidth from each RSE to the center is adequate for real-time forwarding of each trajectory as soon as it is received. With a 20% penetration level of CVs in the vehicle fleet (FHWA-JPO-16-424), the volume of data transmission from all of the RSEs in a typical region will be in the range of 100 Mbps—5 Gbps. Commercial CV and connected traveler trajectories will increase the data rates by about another 20% (see further analysis the *Deployment of Big Data Tools and Technologies for a Typical Agency* section of Chapter 3). While these rates are sustainable by current but very expensive RDBMS technology, the RDBMS cannot sustain the growth rate of the individual tables and maintain indexes required for processing and analysis of the information in real-time. A big data tool will be required for data acquisition of the high-volume, high-velocity trajectory data.

3.1.2 Data Acquisition and Analysis

With the trajectory data now streaming in from RSE feeds and the commercial cloud APIs at rapid rates, a number of technical options are available to integrate the information into existing ATMS functions. The trajectories will be stored in the NoSQL for longer-term persistence, analysis and use. For immediate real-time application of the trajectories and other emerging data sources, it will be necessary to use a *real-time streaming* component. This component will then provide the data directly into a real-time analysis component that is built to analyze the data as it arrives; bypassing the storage and re-extraction of the relevant data from the NoSQL data store. There is a slight difference between real-time processing and stream processing which is described in the next sections.

Real-Time Functions

Stream processing technically has no time limitations on when outputs have to be generated and storage has to be considered. In the event that the input data is coming in too fast, in stream processing components the data can be queued for processing. *Real-time processing* has a hard time limit on output generation that may constrain the ability of the component to produce results, in some cases. To appropriately handle streaming connected vehicles and connected traveler data a real-time streaming processor will likely be needed. **Tools capable of meeting such functionality include: Apache Storm, Apache Kafka on Storm, Amazon Kinesis, SQLstream Blaze, and Spark**. New tools are also continuing to be released frequently based on developer experiences with standard tools. With a large number of RSEs for collection of probe data, it may be necessary to pursue use of the fastest tools on the market (as of 2017), such as **in-memory analysis tools like Spark or SAS LASR**.

For these potential real-time use cases, the agency may consider implementing a series of** disk based, in-database, or in-memory analysis tools **befitting the various types of analyses to be conducted. Tools exist on the market that address the needs of analyzing structured or unstructured data, as well as data coming in and needing to be analyzed at varying speeds (including real-time). The distinction between these tools must continue to be influenced by considerations around the agency's existing or projected technical capabilities, budgetary restrictions, computing options and functions, preferences around system latency, and projected requirements for statistical analysis (e.g., machine learning).

Near-Real Time Functions

For near-real-time functions, the streaming processor may not need to process the individual trajectories on-the-fly but can likely store the data in the NoSQL first as shown in Figure 2. A separate analytics system can

then extract the relevant data from the NoSQL and process the data in bulk for deriving information relevant for near-real-time TMC functions. Still a set of purpose-specific analysis components is likely necessary that have Hadoop ecosystem capabilities to turn the real-time trajectory data into information that can be fused with ATMS data for traditional functions on the order or seconds to minutes. **Tools capable of meeting such functionality include Apache Storm, Apache Kafka on Storm, and Spark**.

3.1.3 Data Storage

Information from the existing ATMS database is necessary to make sense of the trajectory data and its relevance to each application type. For example, map-matching, correlation of the data with certain corridors, traffic signal control parameters, device failure conditions, and other similar data and meta-data will need to be merged. This may require the ATMS data to be replicated in Hadoop, pulled from the ATMS database(s) by the analysis components directly through traditional connections, or some other architectural solution. The decision to integrate existing systems into the Hadoop cluster or run them along-side the cluster is an important one. There are three possible ways to go about working with legacy databases in concurrence with NoSQL:

- Migration.
- Replication.
- Interfacing.

Each of these will be discussed in subsequent sections.

Migration involves the deprecation of all existing RDBMS systems and the relocation of this data into the new NoSQL platform. This would allow all of the incoming and existing data to cohabitate in a unified data store thereby reducing bottlenecks to outside systems, but this may also be an expensive option as it would require substantial migration planning to ensure existing applications can utilize the new data store, as well as to ensure minimal disruption to existing operational practices. **Tools capable of facilitating this migration process include Hbase and Hive**.

Replication, on the other hand, involves the duplication of all new data in the RDBMS systems on some set interval (i.e., minute-by-minute, hourly, daily, monthly, etc.) into the new NoSQL platform “data lake”. This has similar benefits to the migration option as all of the data (incoming and legacy) would eventually come to cohabitate in the same NoSQL environment, albeit at some delay (at the agreed replication interval, for example, hourly). This option also has the added advantage of not requiring any changes to the ATMS applications currently in production as they may continue to read from and write to the systems they have always interfaced with. Some possible disadvantages to this option include the added resources needed to maintain two systems at once and the recurring requirement of ensuring the veracity of both systems. The latter may be accomplished with well written archival processes that continually compare and verify data veracity across the two platforms. **Tools to consider for the replication options include Hbase, Hive, and Sqoop/Flume**.

Interfacing is the third possible option for running NoSQL data stores alongside legacy systems. Interfacing requires some means of connecting the legacy system with the NoSQL system in real-time or near real-time. In this approach, the analysis components for real-time and near-real-time processing would be able to simultaneously query both systems at once to extract an answer. This process is advantageous as it would not require significant expansion to the NoSQL solution (unless the analysis is being conducted on the NoSQL nodes and not the legacy nodes) and it would allow the systems to interface in real-time or near real-time. Some disadvantages of this option include the added resources needed to maintain two systems at once, as well the need to set up the data communication, querying, and transfer bridge between the two

systems. Tools that are capable of facilitating the process of interfacing between the NoSQL and legacy RDBMS systems include Hive and Impala.

3.1.4 Interfacing Data Processing Tools to Advanced Traffic Management System Functions

The real-time and near-real-time analysis components supply the summaries, aggregations, trends, and other collections or derived statistics to the ATMS through defined interfaces. These interface components are depicted as red and orange APIs in Figure 2. The ATMS then uses these new sources of information for existing control system decisions, such as setting traffic signal timings, ramp metering rates, or lane control designations. The ATMS also feeds appropriate information back to the RSEs for edge functions such as roadway hazard warning, speed warnings, work zone status, and so on. Information can also be shared back to connected travelers and commercial connected vehicles through new functions. Development of these APIs will be system-specific as ATMS functions have been developed over time using a variety of languages, tools, and architectures.

Some ATMS may be ready for emerging data sources, particularly by 2021. The CV Pilots and the second wave of pilots that will are likely to be started in 2018 may develop features and functions that utilize emerging sources and build upon big data tools and technology. Analysis components shown generically in Figure 2 as new elements based on the Hadoop ecosystem may already be included in the ATMS umbrella of functionality by 2021. Similarly, ATMS systems may evolve to deal directly with NoSQL data stores instead of RDBMS in order to take advantage of many of the benefits of the structured and non-structured storage, speed, and so on. ATMS integration issues will be explored further in the *Plans and processes to enhance current transportation management systems* report of this project.

In the next group of sections, specific issues related to each group of traffic management functions are discussed.

3.2 Benefits of Big Data Tools and Technologies for Real-Time Functions

The real-time group of traffic management functions includes the following:

- Hazard warning.
- Speed warning.
- Intersection Collision Avoidance.
- Probe data collection.
- Electronic payment.

Real-time functions process emerging data sources as they are collected as shown in Figure 3. The individual event data may be passed directly to the appropriate ATMS elements, or simply fed into the NoSQL database after vetting and cleaning. Hazard warning, speed warning, intersection collision avoidance, and electronic payment functions are best deployed at the edge (on the RSE). These functions must provide timely information dissemination as a vehicle or traveler approaches a roadway location that requires a warning. Intersection collision avoidance functions will run strictly on an RSE. Intersection collision avoidance violation events will be transmitted back to the system as soon as they occur. For hazard

and speed warnings, as more CVs encounter the same hazard in a certain location (as evidenced by swerving and severe braking) evidence will accumulate that can be used by the ATMS to generate a warning message. Probe data collection is the foundational emerging data enabling enhancements to all of the near-real-time functions and thus the collection and processing of the probe data must be done in real-time. **Tools capable of meeting such functionality include Apache Storm, Apache Kafka on Storm, Amazon Kinesis, SQLstream Blaze, and Spark**. New tools are also continuing to be released frequently based on developer experiences with standard tools. (http://www.infoworld.com/article/3075501/open-source-tools/twitter-open-sources-heron-for-real-time-stream-analytics.html.) With a large number of RSEs deployed for collection of probe data, it may be necessary to pursue use of the fastest tools on the market, such as in-memory analysis tools like Spark or SAS LASR.

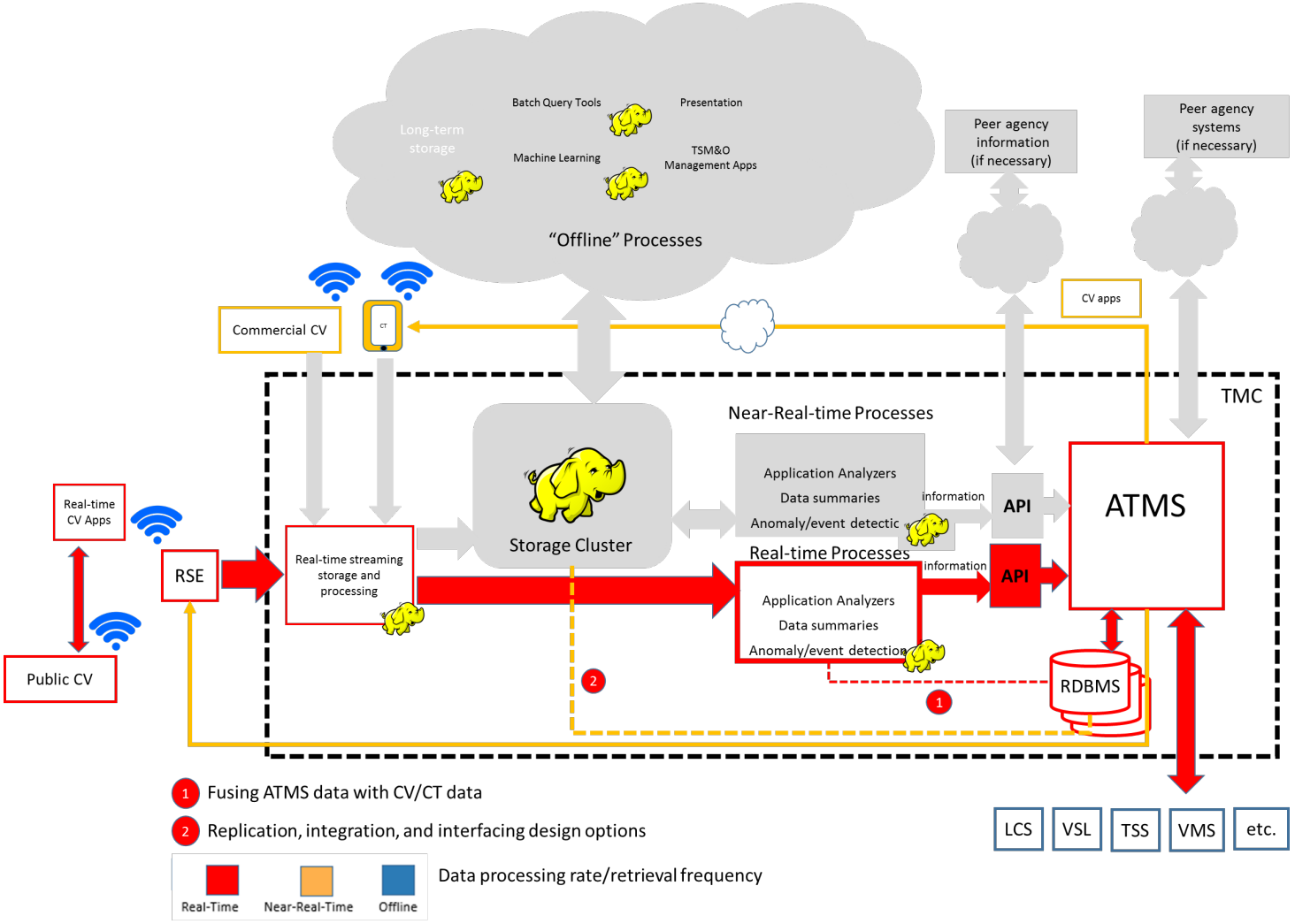


Figure 3. Diagram. Real-time components.
(Source: Kimley-Horn and Associates Inc., 2016.)

3.3 Near-Real-Time Functions

The near-real-time group of traffic management functions include:

- Incident management.
- Signal control.
- Metering.
- Lane management.
- Traffic information.
- Weather monitoring.
- Parking management.

As highlighted in Figure 4, the likely data flows for near-real-time functions is to store the emerging data sources in HDFS from the real-time streaming acquisition tool. The near-real-time analytics tool extracts the appropriate data elements from the HDFS and the ATMS database(s) to perform analysis for a specific function. For example, a data analysis process may collect all the trajectories at a specific traffic signal from the HDFS over the last 90 seconds. These trajectories are aggregated and processed by phase to calculate metrics such as delay or estimated queue length. These metrics are then transmitted to the appropriate adaptive traffic control algorithm via API. The adaptive traffic control methods calculate new timing parameters for that intersection and downloads the new settings to the field controller.

In another example, a near-real-time analysis process may extract all of the trajectory data relevant to a traffic commuting corridor with ramp metering over the last three minutes. These trajectories are synthesized into traffic conditions estimates (speed, density, etc.) along the corridor and pass this traffic conditions data into an adaptive ramp metering algorithm. This new information from the emerging sources can supplement (or replace) the traditional loop detection inputs of the ramp meter rate determination algorithm. The ramp metering rate determination algorithm then downloads more accurate metering rates to the corridor meters that reduce delay. Over time, it may even learn from the effects of past days to optimize the control function.

In a similar manner to real-time functions, near-real-time functions may transmit commands and information back to connected vehicles and travelers through RSEs or directly to their mobile apps or commercial connected vehicle systems. The information on updated control commands may be communicated to peer or partner agencies through the ATMS interfaces or the derived statistics, summaries, or events may be shared from analysis component at one agency to the analysis component at another agency. For example, a peer agency local traffic control system may take into account the new ramp metering rates on a corridor in determining the interchange traffic control settings. The peer agency analysis component may also use the freeway conditions data (speed, density, etc.) directly in its calculation of signal timings. **Tools capable of meeting such functionality include Apache Storm, Apache Kafka on Storm, and Spark**.

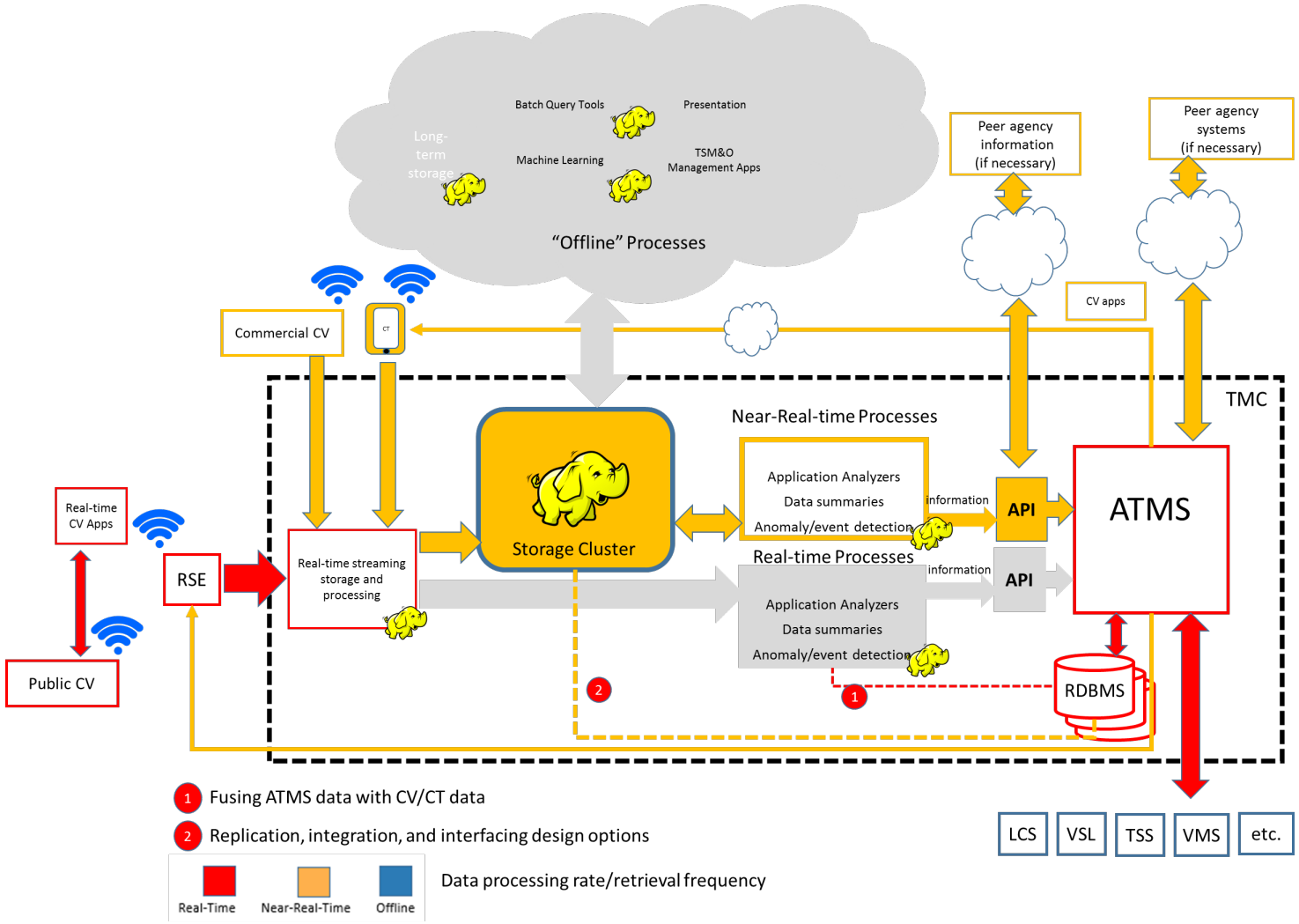


Figure 4. Diagram. Near-real-time components.
(Source: Kimley-Horn and Associates Inc., 2016.)

3.4 Benefits of Big Data Tools and Technologies for Offline Functions

The offline group of traffic management functions includes:

- Emissions monitoring.
- Asset management.
- Performance measurement.

As highlighted in in Figure 5, the likely data flows for offline functions is to store the emerging data sources in NoSQL from the real-time streaming acquisition tool and then periodically ship that data to long-term storage for offline processing. The offline analytics tool processes the data from the long-term storage system. This may include retrieval of data from multiple years in the past, and comparison of long-term trend and statistics. For example, a data analysis process may collect all the trajectory summaries (phase delays and queue lengths by cycle) at a specific traffic signal from the NoSQL over the last year in 2 hour bins. These performance metrics are aggregated into a pivot table for comparison of day of week, time of day, and month of year. Data analysis tools are used to automatically apply the same analysis to all traffic signals in the system and flag any locations with queues that show a trend of queues growing by 100% or more, on average, over the last year. Graphs and charts illustrate trends and problem locations are identified on a regional map. Offline functions may generate alarms and alerts. Offline applications for traffic management may inform other agency activities and analyses in other areas including finance and budgets, human resources, and staffing and organization.

Some offline functions may run continuously on the archived data and identify anomalies in conditions data, identify any negative trends (congestion is getting worse), and flag problem areas. A host of other functions can only be imagined once the emerging data sources become available. Tools with point and click capabilities could offer a sufficient and sustainable solution as they require minimal technical maturity and are easy to personalize. **Some tools that fall under this categorization include SAS Visual Analytics/Visual Statistics, Tableau, Informatica, and Gephi, Clik, Pentaho, Looker, Mahout, and ML Lib among many others**.

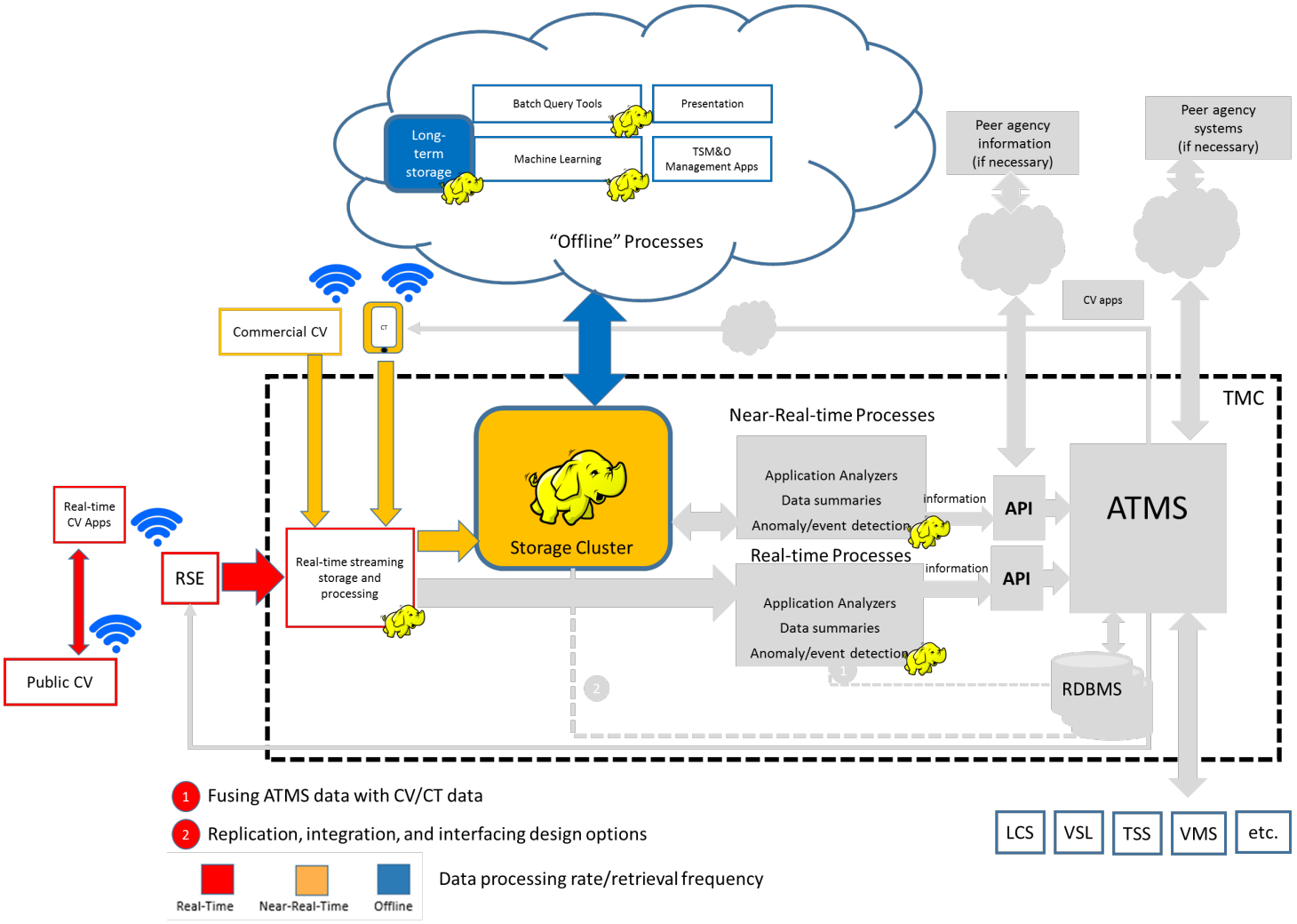


Figure 5. Diagram. Offline components.
(Source: Kimley-Horn and Associates Inc., 2016.)

3.5 Deployment of Big Data Tools and Technologies for a Typical Agency

In this section, we estimate the size and scale of a big data NoSQL and analysis tools deployment for a typical agency. The reader will understand after reading this section how the data loads result in requirements for big data storage and processing systems. For this representative situation, we will estimate the size of an agency jurisdiction by the number of roadway segments they manage and the population of adult travelers. We also assume that this representative agency has three TMCs for three major metropolitan areas. The TMCs operate regionally (using three separate high-speed, high-bandwidth field-to-central networks), and feed a central data warehouse for performance measurement and reporting. We will assume one TMC covers a large metropolitan area and the other two TMCs cover smaller regions.

Large TMC are defined as having:

- Regional population under jurisdiction: 1,500,000 adult travelers.
- Regional population of vehicles: 1,000,000 registered vehicles.
- Regional population of connected infrastructure: 300 devices, 300 closed-circuit televisions (CCTVs).
- Regional road network: 100,000 segments.
- Connected travelers (20% in 2021): 300,000 travelers.
- Connected vehicles (20% in 2021): 200,000 vehicles.

Small TMCs (each) are defined as having:

- Regional population under jurisdiction: 500,000 adult travelers.
- Regional population of vehicles: 250,000 registered vehicles.
- Regional population of connected infrastructure: 100 devices, 100 CCTVs.
- Regional road network: 50,000 segments.
- Connected travelers (20% in 2021): 100,000 travelers.
- Connected vehicles (20% in 2021): 50,000 vehicles.

Using the assumptions for data volume and velocity discussed in FHWA-JPO-16-424, the resulting levels of data for our representative State agency are shown in Table 4. For comparison purposes, one streaming HD two-hour movie requires approximately 6 gigabytes (GB). Thus for example all of the BSM and PDM data for 200,000 connected vehicles is equal to processing and storing approximately 140 streaming movies per day (noting that the bulk of the data is transmitted during the peak travel periods).

Table 4. Data loading analysis for a typical agency.

LARGE TMC					
quantity		rate			
1,500,000	travelers				
1,000,000	vehicles				
300	field devices	5.8	MB/day/ea	1,740	MB/day
300	CCTV	4	GB/day/ea	1,200,000	MB/day
200,000	connected vehicles (20%)	4.1	MB/day/ea	820,000	MB/day
	subtotal field-to-TMC			2.02	TB/day
100,000	roadway segments	1.44	MB/day/ea	144,000	MB/day
300,000	connected travelers (20%)	500	kB/day/ea	150,000	MB/day
	subtotal cloud-to-TMC			0.29	TB/day
	Total for Large TMC			2.316	TB/day
SMALL TMC					
quantity		rate			
500,000	travelers				
250,000	vehicles				
100	field devices	5.8	MB/day/ea	580	MB/day
100	CCTV	4	GB/day/ea	400,000	MB/day
50,000	connected vehicles (20%)	4.1	MB/day/ea	205,000	MB/day
	subtotal field-to-TMC			0.61	TB/day
50,000	roadway segments	1.44	MB/day/ea	72,000	MB/day
100,000	connected travelers (20%)	500	kB/day/ea	50,000	MB/day
	subtotal cloud-to-TMC			0.12	TB/day
	Total for Small TMC			0.728	TB/day
	Total for 3 TMCs			3.77	TB/day

(Source: Kimley-Horn and Associates, Inc., 2016.)

3.5.1 Conceptual Big Data Technology System

This scenario presents a case of an agency with multiple data sources, regional administrations, and technical requirements with a need to collect, store, and analyze data at two levels of administration. An illustration of this is shown in Figure 6. The regional TMCs own and collect all of the field data within their jurisdictions. This distinction is important to make as the statewide data warehouse is not collecting any field data of its own, but rather acquiring data through the regional TMCs. Given the technical and operational needs of this agency, the implementation will likely need a two-level system building upon the conceptual architecture shown in Figure 2. This system will likely combine (1) on-premise storage and processing hardware at each of the TMCs for the functions that require "fast" processing of the emerging data sources, and (2) a connection from those TMC data centers to a cloud infrastructure that allows consolidation, analysis, and reporting on a statewide basis. The functions in the cloud system would be those that data processing speed is not time-critical such as performance metrics, asset management, and emissions monitoring. Note that this is not intended to be a prescriptive architecture. Any flavor of on-premise and as-a-service IT models can be used by an individual agency based on their evaluation of the risks and benefits as discussed in Section 3.5.3.

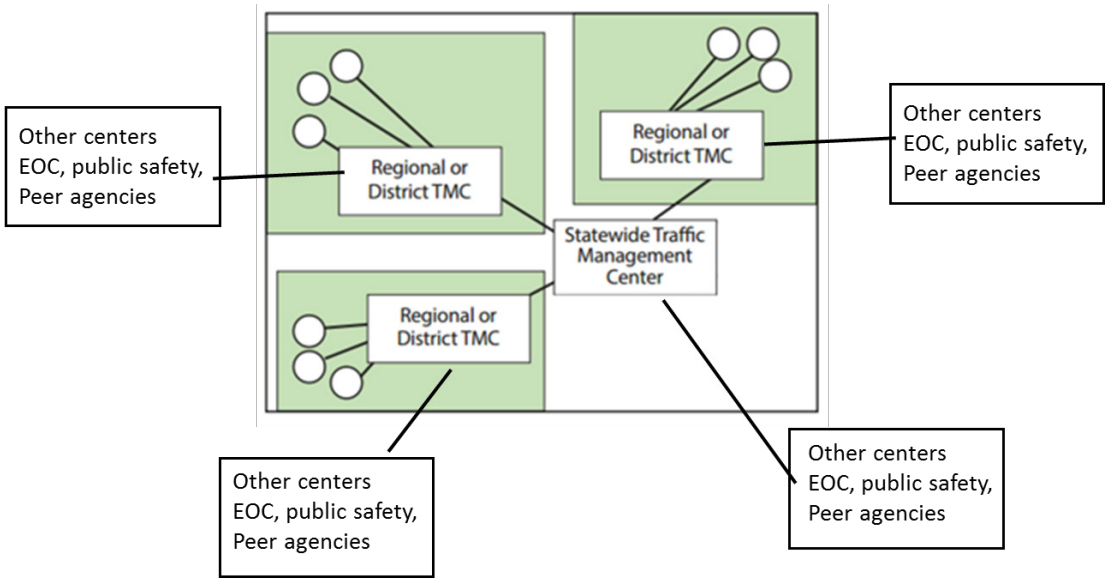


Figure 6. Illustration. State traffic management center fed by data streams from three regional traffic management centers.
(Source: Kimley-Horn and Associates, Inc.)

In this system deployment, each of the three regional TMCs would have a self-managed data warehouse that would act as a data collection, storage, and analysis center. Based on the size and velocity of the emerging data sources, this data warehouse must be based on a NoSQL/HDFS platform. The three centers will be staffed by TMC personnel or contractors working under the management of TMC personnel. As the TMCs cover geographies of varying sizes and populations, their hardware solutions will vary accordingly, as will their deployment and maintenance costs.

The conceptual solution presented for the regional TMCs is based upon a distributed on-premise HDFS platform that would follow the approach of a multiple centralized deployment model as defined by the *Guidelines for Virtual Transportation Management Center Development* report [12]. Hosting the platform on-premise allows the maintenance staff greater flexibility in installing, managing and maintaining all aspects of

the deployment. The added tangibility of the data and solution allows for more control over the security, up-time, and access to the infrastructure while ensuring organizational compliance every step along the way. The managing staff are able to support and monitor exactly what is of consequence to the agency with knowledge of all of the dependencies that could impact the most important services. Regional agency leadership will know with certainty that they have direct control over their regional resources, and statewide functions are left to statewide personnel. While this option may be more costly to initiate as it requires investments in hardware, software, climate controlled space, and technically trained hardware personnel, it allows for a great level of control over the system and the data.

Collectively, the three regional TMCs cover the entire geography under the jurisdiction of the State DOT. Each regional agency meets most or all of its high-priority functional needs from RSEs, probes, and data feeds without the need for data sharing across regions (noting the potential for commercial CV and connected travelers data to be delivered centrally and distributed to each region as needed). As such, one approach the State DOT may take in their statewide data solution is an Infrastructure-as-a-Service (IaaS), cloud-based Hadoop platform that mimics the individual data warehouses in structure and functionality. However, it would primarily serve to aggregate and analyze data from each of the TMC data centers to provide the visibility and analytical capabilities needed to conduct statewide analyses for “slow” functions such as emissions monitoring that do not require real-time return of quick analyses and decisions for real-time application (i.e., within less than one minute from data receipt to control function). A cloud-based system would provide the State DOT headquarters with faster and more dynamic flexibility in scaling resources up or down as needed than the on-premise deployments, and would reduce the implementation level of effort for headquarters staff. DOT personnel and/or contractors would be responsible for managing applications, data, runtime, middleware, and operating systems, but would not need to manage or house the hardware upon which the solution is built. Functionally, the big data technology system would mimic the hardware and organizational setup of the regional TMC big data technology systems.

3.5.2 Big Data Technologies System Sizing Estimation

Table 5 summarizes the data volumes for the typical agency with two small TMCs and one large TMC. These regional locations then feed their summary data to the State DOT repository.

Table 5. Data volume, delivery, and variety by traffic management center type.

	Data Volume (per day)	Data Delivery Frequency	Data types
Small Regional TMC	0.728 terabytes (TB)	Real-time (< 1 s), Near-real-time (< 1 min)	Connected vehicle trajectories, connected traveler trajectories, video, infrastructure status
Large Regional TMC	2.316 TB	Real-time(< 1 s), Near-real-time (< 1 min)	Connected vehicle trajectories, connected traveler trajectories, video, infrastructure status
State DOT	3.77 TB	Daily (batch, interval)	Summaries of trajectories, video analytics, infrastructure status summaries

(Source: Kimley-Horn and Associates, Inc., 2016.)

At the large regional TMC level, an average day may require the ingestion and processing of as much as 2.3 TB of data per day. This data is fed to the TMC from RSEs, CCTV feeds, existing field devices, and third party data providers. As discussed in FHWA-JPO-16-424, 80% of the information is assumed to be comprised of BSMs (PDMs) and CCTV images. These data sources are reporting data continuously (e.g., using a firehose API). At the State DOT level, an average day will require the ingestion and processing of all or nearly all data collected by the three regional TMCs. If it is desirable to store all of the 3.7 TB of raw information at the common repository, it need not be replicated to the statewide system in real-time. A more likely scenario is that trajectory summaries will be stored at the statewide repository rather than wholesale replication of the raw data in the HDFS at the State repository. However, there may always be the potential for new functions and analyses that will require the raw data to be retained.

For these requirements, the big data technology system must be capable of ingesting, replicating, and securing large amounts of data at each velocity. The big data technology system must operate seamlessly with a Hadoop platform deployed with both on-premise and cloud infrastructures. The big data technology system would also need to ensure that the right data is acquired along with all necessary metadata as discussed above with regard to Figure 2. The potency and size of the acquisition tool may be chosen through a careful consideration of these variables, in addition to the overarching organizational requirements surrounding budgeting, staffing, security, and the like.

Data Storage and Processing

Given that the data being collected by the regional systems is coming in various formats (e.g., tabular, video, text, etc.) and types (e.g., structured and unstructured), the process of cleaning, organizing, storing, and managing this data will be of major consequence to the operational success of the regional systems and, by extension, the State DOT repository. The universal requirements for all data stored and processed in this scenario will be that the data be secured, fault tolerant, scalable, and backed up. Additionally, there may be a requirement at the regional TMC level of compressing and archiving data locally while the responsibility of long term data storage lies at the statewide level.

Hadoop systems use *data replication* as a form of redundancy to create a fault tolerant system. To meet the demands of 0.728 TBs of data arriving daily from the two smaller regional systems, a storage capacity must be maintained to hold at a minimum three times that (with the default data replication factor of three) resulting in a total of 2.2 TBs of data per day. Assuming an initial data size of 100 TBs (data that already exists within these regional systems) and a daily intake of 2.2 TBs of data, a minimum storage capacity of approximately 600 TBs with an increase of 300 TBs approximately every six months would be required. The larger TMC in this scenario has approximately three times the data of the smaller in all respects, and consequently requires three times the storage as well (i.e., 1.8 petabytes (PB)). It is worth pointing out here that these rough calculations are for storage of all of the raw data from the emerging sources. **This clearly points to the need for aggregation and edge processing of the information in the real-time streaming analytics before it is stored to HDFS.** If this were the case, the storage requirements of the HDFS systems at all regional systems, and the statewide repository will be reduced. Significant system resources (random-access memory or RAM and disk) for processing raw data into aggregations and summaries will still be required.

A Hadoop system provides an innovative method of storing data in a relatively cheap environment using commodity hardware; however, because it stores all data in triplicate to meet fault tolerance requirements, even cheap storage will still generate significant cost and it can add up quickly. As mentioned earlier, aggregating, managing, and pre-processing data on the edge is a method of reducing the impact on the centralized system, and beginning to be employed more heavily in many IoT platforms, as discussed in FHWA-JPO-16-42. Aggregating data at the edge may prove to be a valuable approach, particularly for data

being prepared for offline analysis (i.e., data that isn't required for frequent or immediate analysis; can be easily compressed and stored for potential future needs).

The other half of determining system sizing capacities and resource requirements depends on the analyses to be performed with the incoming data. The more high-churn, disk-level analyses required for a particular agency's functions, the more disk resources that will be required. Additionally, analyses that require significant real-time or even near real-time speed will require significant availability of RAM. A complete solution would need to undergo proper capacity and sizing exercises to understand the full ingestion (e.g., speed of the data coming in at peak times and speed with which it needs to be processed at the peak times as well), storage (e.g., how much data exists and is anticipated to grow and how long does it need to be stored), and analytical (e.g., how quickly does the data need to be analyzed and how much RAM versus disk space is required to meet those needs) needs of the system.

Translating data storage and RAM requirements into nodes, or number of servers, can be challenging as it depends on several variables. As of writing this report in 2017, there exists no industry standard for the definition of a "big data" node's configuration. While some vendors may have developed definitions and configurations of their own that best fit their pricing, marketing, and procurement models, these definitions are often not pertinent to other vendors or the industry at large. Most nodes however are typically a low-cost commodity device that stores and processes several hundred gigabytes of data with typical RAM. It could also be an entirely virtualized system, in which each node could be an arbitrary set of resources. The industry best practice in addressing this question is to approach a hardware vendor (in the case of the on-premise system) with an estimation of how much information is to be stored and the degree of processing expected to be carried out with the system. With all of these caveats in mind, a "node" referenced in this report will refer to a low-cost commodity server with 128 GB of disk space and 4 GB of RAM. The authors of the report came at this configuration as a result of researching a number of configuration benchmarks available on the market as of this report's writing. For this agency where there is an identified need to store 1.8 PB (as outlined above), a bare minimum of 15 nodes would be required for the first year of operation (for storage alone).

Data Analysis and Processing

With the traffic management functions and the scale of data volume being ingested and processed in mind, this typical agency will need to conduct process-intensive analyses in real-time, near-real-time, and offline. Real-time processing with process-intensive analyses will dictate the amount of memory and the right selection of processing tools for the functions. It is difficult to reach an exact estimate of how much overhead is needed to 'cover' an analysis, with processing overhead here being defined as the amount of additional disk space needed to process some amount of stored data.

For example, if a server holds 100 GBs of data, it may be estimated that an additional 50% of storage is needed to cover the overhead of analyzing this same data, so the final server size recommendation will come to approximately 150 GBs of storage. The metric of 50% is merely an estimate here as this number will differ based on the analysis being conducted, and calculating the exact overhead requires much more detailed trial and error benchmarking on the DOT's actual data and processes.

However, if we were to apply this estimation of overhead to our existing hardware storage recommendation for one of the smaller TMCs (and keeping in mind the previously discussed caveat that the storage recommendation will be *heavily* impacted by data management practices and archival preferences), the actual storage needs for each of the smaller TMCs then balloon to 900 TBs (600 TBs + 50%) with an increase of 450 TBs (300 TBs + 50%) every 6 months. If through on-site benchmarking the DOT finds that the overhead estimate of 50% is too high, the storage estimates will scale down accordingly, and vice versa

if the benchmarking finds a higher than 50% overhead buffer is needed. These kinds of impacts can only be known after the system is deployed and begins operation.

Given all of these considerations, the DOT in this scenario will benefit from using commodity hardware for the on-premise systems. This option is low cost and may already fit well within the DOT's existing procurement agreements as commodity hardware can be any mass-market stand-alone computer with no set standard on storage or memory capacity. Furthermore, developing a system with 'smaller' (in disk and memory storage) nodes translates to a higher level of flexibility in scaling and maintaining the overall system. For example, a failed commodity node with a storage capacity of 500 GB is cheaper and easier to replace than a failed server with a storage capacity of 10 TBs.

As discussed earlier, integration of existing ATMS database(s) in the Hadoop system is important to note. The TMC or DOT may have a system in place that does not need changing or upgrading at the current time. The system may be a recent investment, a significant multi-year investment, or a well-used and well-managed one.

Most ATMS-enabled agencies also have geospatial, or Geographical Information Systems (GIS), systems that are used in time critical capacities with a high level of institutional and maybe even public visibility. Several tools have already been developed for the integration of geospatial systems into Hadoop, such as Hadoop GIS and GIS Tools for Hadoop. These tools currently operate to expand the application of existing geospatial and geo-processing capabilities to very large datasets, but it is not unrealistic to assume that geospatial applications of big data will grow in complexity and utility with the reduced limitations on processing power.

All three of the options for integration of ATMS databases with the Hadoop repository (discussed in the Data Storage section of Chapter 3) would entail an expansion to the hardware estimation. For example, if 100 TBs of legacy data needed to be migrated or replicated into the Hadoop system, then 450 TBs of additional Hadoop storage would be needed (100 TB x 3 for the Hadoop replication factor + 50% previously used estimate for processing overhead). Even in the example of interfacing where the legacy systems' data would continue to live in the legacy RDBMS, there would need to be additional storage capacity for processing overhead integrated into the system.

The decision of integrating legacy traffic management system data stores with HDFS is an important consideration. This decision should take into account the initial and maintenance investments of the existing systems as well as the return on investment (ROI) of streamlining the systems into the cluster systems compared to the cost of maintaining both systems in tandem and developing a means to bridge information and data between the two. With time, and in particular as the emerging data sources penetration rates grows and IoT platforms mature, this decision will cease to be relevant as more and more ATMS platforms will migrate entirely to big data tools as they become the norm and all new development will occur natively on these platforms.

3.5.3 Other Considerations

A number of considerations need to be made by traffic management system decision makers when choosing big data tools and technologies for their regional and State deployments.

Scaling up to meet future demands. When selecting a given solution, a DOT needs to consider the impact of scaling this very system up to meet future demands on the budget, staff, available rack space (in the case of on-premise systems), and technical capabilities. For example, having limited or no access to climate controlled rack space would lend an agency to make the decision to host their system in the cloud.

Outsourcing to vendors or contractors. Furthermore, an agency not having the technical personnel to initiate or maintain an system may consider going with a vendor or contractor to manage the process for them. As critical as those considerations are to make, the decisions that need to be made around an agency's data needs are of even greater importance. The needs of any given agency in a world with connected vehicles will almost certainly change over time, and a system lacking in these considerations may be a costly one to reform or mitigate down the line.

While big data remains an emerging field and the costs of mitigating defects at various stages in the deployment cycle has not been calculated across multiple platforms as of yet, we may extrapolate from the cost of mitigating defects in various stages of the software development lifecycle that it is sigfinicantly cheaper to invest the time and resources in planning, designing, and constructing a well-researched and well-thought out long term data system rather than mitigating the short-comings of a hastily put together one. The consequences of deprioritizing these crucial early steps may be deterimental to the agency's abilities to carry its operational duties as it may lead to a non-functioning or under-functioning system.

Statistical and analytical requirements. Similarly, when choosing the tools upon which an agency's analytical capabilities will be built, agency personnel will need to consider the level of statistical and analytical sophistication required to meet decision makers' needs, as well as the methods of day-to-day reporting and visualization operations staff would benefit from the most. For example, conducting a process-intensive longitudinal study such as a corridor wide, multi-year network analysis will necessitate an enormously different set of technical requirements than visualizing the location of assets and personnel in a storm evacuation. A more intensive process may need to be conducted on a faster in-memory tool and more RAM, whereas a less intensive process can be sufficiently powered with traditional disk-based analytics with less RAM.

Additionally, the data mangement practices of the agency will be of significant consequence to defining the size and long term maintenance of the system. What data to store, how long to store it, and at what level of granularity to store it (raw vs. aggregated) are all critical considerations to make. Furthermore, addressing how the data coming in will be used (in real-time, near real-time, or archivally) and designing the storage capability around those individual velocities will make for a more efficient data processing infrastructure, as well as more responsive operational decision-making.

On-premise vs. hosted systems. Another critical consideration are the trade-offs between on-premise and hosted systems and deployment models. An on-premise system, such as the one discussed as a potential system for the regional TMCs, translates to a higher level of control and security for the technical team managing the infrastructure, but also demands a higher level of effort and technical skill. In this case, the team is managing every aspect of the system, as opposed to the State DOT statewide cloud system where the team would not have to manage hardware (see Figure 7). The required level of staff involvement for a given system is an important early consideration to make when choosing and deploying a big data system as it has the power to gravely impact existing resources.

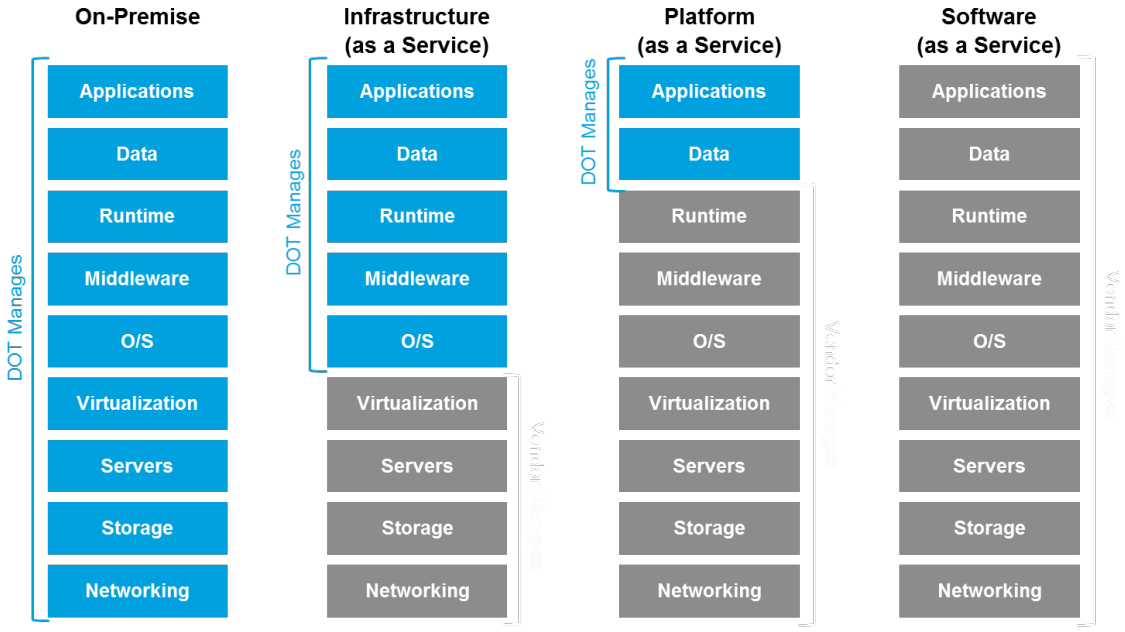


Figure 7. Chart. Information Technology considerations for various system delivery options.
(Source: Deloitte, 2016.)

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

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

Finally, one of the most important considerations to note in-line with the location of a system is the security and privacy of an agency's information. An on-premise deployment may yield a DOT complete control over the security and privacy of the data, but the maintainance of that security and privacy will be the responsibility of the DOT staff. A cloud deployment may be seen as a less customizable privacy system or a less secure system overall, however many cloud vendors have pursued industry certifications in security and privacy, and often have staff dedicated entirely to ensuring their networks are not compromised. As a general rule of thumb, any agency lacking in the technical capability to ensure its data and network security would benefit more from a full or hybdrid cloud deployment.

3.5.4 Estimates for Data Loading Differences for Different Size Systems

Based on the information provided in Table 4 regarding a "typical agency," this report provides estimations for the differences in data loading of several generic system sizes (small, typical, and large). Typical agency data was determined based on the information provided in Table 4 using simple assumptions of data volumes and velocities. Table 6, below, builds on the assumptions of the former table to provide estimates of total data volume, average data velocity, and a few comments on the functional complexity of the typical agency.

The typical agency considered throughout this report is the combined "system" of two small regional TMCs, one large regional TMC, and an aggregated State DOT. Based on the estimated data volumes from Table 5 the total volume of the combined system can be estimated as 7.54 TB with 3.77 TB being ingested at real-time speeds (across three regional TMCs, two small and one large) and 3.77 TB being ingested concurrently on a daily basis (at the State DOT). Dividing the total system volume by 24 hours and then by 60 minutes, provides an average rate of ingestion (0.0052 TB/minute or 5.23 GB/minute) the entire system experiences over the course of one day.

Also of consideration is the expected work load and online/offline needs of the data. For the sake of simplicity, here we assume a correlation between system size and functional complexity where an increase in the former leads to a more complex instance of the latter. We estimate small systems to be three times smaller than a typical system and a large system to be three times larger than a typical system.

The total data volume and average rate of data ingestion are not intended to provide complete capacity and sizing estimations for a given agency, nor are they intended to be used without understanding the context with which they were determined. They are intended to serve as broad estimations that lend an understanding to the magnitude that many systems may need to deal with. For example, the typical system model described above may be ingesting 7.54 TB daily; however, the small regional TMC will need to ingest approximately 0.51 GB/minute in real-time (or near real-time), each large regional TMC will need to ingest approximately 1.61 GB/minute in real-time (or near real-time), and the State DOT will need to ingest all 3.77 TB at once (and those regional TMCs will need to provide it all).

Each of these scenarios presents an entirely different problem, and will certainly need to be examined when designing a final system; however, for the purposes of this section, each system size uses the following rough estimations to provide context to any discussions that follow. In addition to the system's ingestion and real-time, near real-time, and offline needs, there may also be administrative and privacy needs that could impact the estimated size of the system. For example, a geographically large system may involve multiple administrative organizations each with their own policies on data security, archiving, and privacy. To accommodate such a system, complex processes of redundancies and data management may evolve requiring new storage requirements and greater or, possibly, less storage needs.

Table 6. Daily data volume, velocity and functional complexity by system size.

	Data Volume (per day)	Data Velocity (per minute)	Functional Complexity
Small System	2.51 TB	1.75 GB/min	Few critical functions, minor needs for data sharing
Typical System	7.54 TB	5.23 GB/min	Some critical functions and real-time data, some needs for data sharing
Large System	22.63 TB	15.72 GB/min	Many critical functions and mostly real-time data, complex sharing needs

(Source: Kimley-Horn and Associates, Inc., 2016.)

In the next section, we discuss the implications of needs for data sharing on the big data tools and technology deployment for particular functions.

3.6 Impacts on Big Data Tools and Technology Deployment Due to Agency Needs for Data Sharing

Traffic management agencies and TMCs continue to evolve towards more and more cross-jurisdictional data sharing functions and coordination with peer and partner agencies through both technical systems and communication methods. As the connected traveler and connected vehicle data sources emerge, the needs to share information with partners will never be greater. As has been true in traffic management for years, travelers do not perceive crossing of jurisdictional boundaries; they simply expect systems, functions, and services to work regardless of the jurisdiction. In particular, when dealing with trajectory data from PDMs, commercial connected vehicles, and connected travelers; starting and ending points of trajectories will invariably fall outside of agency boundaries for a significant portion of trips. Holistic views of systems such as regional emissions models will invariably be made better with regional data sharing. In this section we discuss some of the issues related to data sharing and big data tools and technologies with respect to several types of institutional relationships including:

- Multi-regional State DOTs.
- Multi-agency coalitions.
- Joint operations centers.
- Local agencies.

3.6.1 Conceptual System(s)

Figure 8 illustrates the conceptual relationship between partner agencies for data sharing between big data tools and technologies. Data sharing tools have become more and more streamlined as big data systems have evolved. As illustrated in Figure 2, a particular agency's big data system for traffic management and operations will likely comprise five major components:

1. A real-time processing component for ingest of the emerging data sources.
2. A HDFS instance for storing the emerging data sources and related ATMS legacy data.

3. A real-time analytics component for streaming analysis.
4. A near-real-time analytics component for streaming analysis.
5. A long-term HDFS repository for offline storage and assessment.

In Figure 2, we illustrated connections between the live HDFS and the long-term repository and (potential) connections between the near-real-time streaming analysis component and other peer systems. In Figure 8 we represent these connections using our typical agency organization with three TMCs and a statewide headquarters. In this case, each of the regional HDFS systems synchronizes data with the regional repository on a scheduled basis. Tools for analysis of various traffic management and operations functions would be available at the statewide level for either statewide analysis or regional systems. In such an institutional arrangement, it would be wasteful to develop systems at the regional level that would not be available to other regions, or deploy three additional instances of long-term repositories for each region separately. Peer data sharing of near-real-time emerging data and derived information is accomplished through a centralized data distribution system. Peer connections between ATMS and other systems (such as a regional ramp metering control system with a local traffic signal control system) are still accomplished through existing legacy or dedicated system-to-system connections. Similar architectures apply for other institutional arrangements as discussed briefly in the following sections.

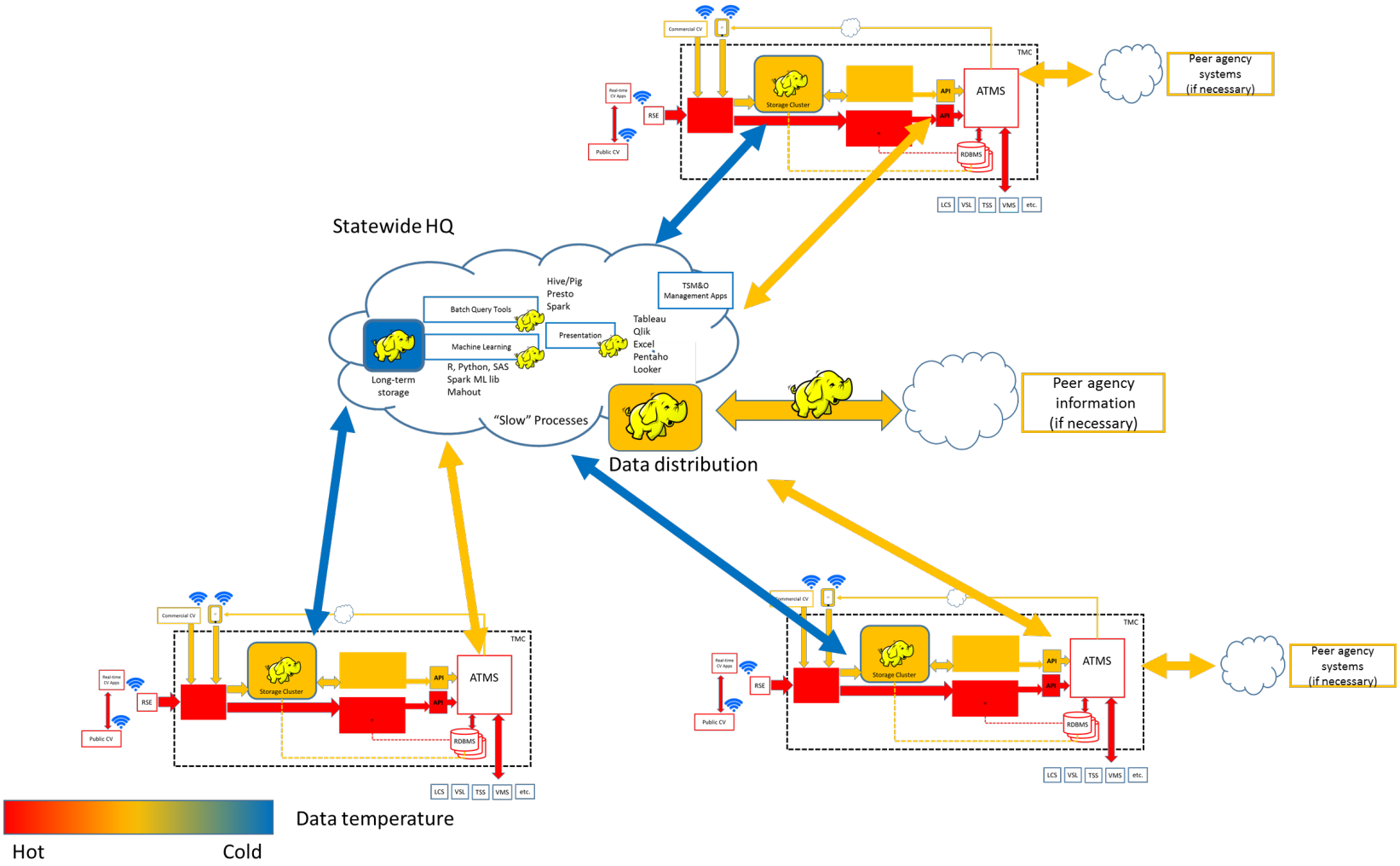


Figure 8. Diagram. Visual representation of the three traffic management centers and the statewide cloud data warehouse.
(Source: Kimley-Horn and Associates, Inc., 2016.)

3.6.2 State Departments of Transportation with Multiple Regions

State DOTs with freeway management (and arterial management, in many cases) responsibilities typically have large a geographic footprint. Typically, a State DOT has more than one TMC strategically deployed across the State for regional management and a headquarters which is responsible for statewide coordination. The headquarters facility may be co-located with one of the regional centers but typically staff is allocated to either regional responsibilities or statewide coordination, but not both. In this situation, emerging data will be collected by the regional TMCs through their RSEs and brought together in the HDFS statewide repository as discussed in the previous section for statewide analysis. Individual regional TMCs may coordinate with local agency partners directly on real-time traffic management issues and share regional emerging data on a peer to peer basis. Commercial connected vehicles and connected traveler data is more likely provisioned to each region through a common API for the State, or APIs tailored to each region. Management of the statewide repository is best allocated to the staff at headquarters.

Key Characteristics include:

- Large geographic footprint, potential challenges with many high-speed RSE connections for real-time data retrieval to TMCs.
- Region-to-region sharing and region-to-State consolidation.
- Individual regions coordination with local agencies on arterial-freeway interactions.
- Clear organizational structure.
- Standardized approach more likely than multiple disparate products in each region.

3.6.3 Joint Operations Centers

Some regions of the U.S. have co-located State DOT and City/County freeway and arterial responsibilities into joint operations centers. In many cases, these joint facilities enable “swivel chair” coordination more so that direct technological integration of management systems, although there are several examples of such technology integrated regions such as the integrated corridor management (ICM) system in San Diego. Having a smaller jurisdictional footprint than that of an entire State, a regional joint operations center typically has robust communication connectivity to field assets. These agencies have a large number of activities that can benefit from the emerging data sources, but a smaller geographical region to manage, so a joint operations center can focus more effort on specific functions in a coordinated manner. Integration of the emerging data sources from local and State field assets becomes easier if only because the agencies are located in the same building. Integration of data and functionality for regional transit operations may also be included.

Key Characteristics include:

- Limited geographic footprint; minor challenges with high-speed connections to RSEs.
- Agency data sharing technology (HDFS repository hosting) may conflict with statewide technology deployment planning.
- Highly coordinated local operations; lead and follow roles in IT responsibilities will require agreements.

3.6.4 Multi-State or Multi-Agency Coalitions

Some regions of the U.S. have formed coalitions among multiple State DOTs such as the I-95, I-80, and I-15 corridor coalitions. These coalitions share data in both technological and institutional ways and sometimes include field operational control functions where one agency can manage the field assets of another in an

emergency, however this is not common for States (it is however more common for multi-agency agreements among local agencies, where they exist). Each State DOT system is managed separately, typically with some kind of over-arching web-based system that collects information from all of the States into one "map" that shows situational awareness and device status across the corridor. Architecturally, this is perhaps the most challenging situation for procurement and installation of big data technologies and tools since it may include the integration of multiple vendor systems and encompasses a wide geographical footprint, with a limited number of agency TMCs in each State being relevant to the corridor operations. Also, only a small portion of each State is relevant to the coalition, so extracting only relevant data for the coalition purposes is an additional challenge. For multi-agency coalitions among local agencies, many local jurisdictions may cede control and management to other larger agencies due to lack of resources.

Key Characteristics include:

- Enormous geographic footprint; consolidated HDFS repository would be unwieldy without significant data curation.
- Challenging to get high-speed connections to all RSE assets on a corridor.
- Complex regional and corridor-specific issues.
- Codified coordination among agency partners.
- Multiple TMCs, including only some TMCs/regions of a partner agency that are relevant to the coalition. Complex data sharing algorithms for extracting or sharing only some of a statewide repository or regional HDFS store.

3.6.5 Local Agencies

Local cities or counties in urban and suburban regions have bordering jurisdictions and cross-coordination of TMC activities can be challenging at times, both with their local partners and State DOTs. Local agencies typically have only one TMC where all field assets are managed and communications is terminated. Geographic regions are typically limited, but can be extensive in many metropolitan regions. Some larger local jurisdictions have more challenging traffic management issues than many States with limited population. Urban and suburban jurisdictions also have strong needs to coordinate and share data with transit agencies and other civil services.

Key Characteristics include:

- Limited geographic footprint, typically good coverage of high-speed connections to RSEs.
- Large number of high-priority CV-enabled functions.
- Opportunities for coordination with adjacent agencies, but uncodified.
- Only one consolidated TMC facility.

3.7 Summary of System Architecture for Integrating Big Data Tools and Technologies with Traffic Management Systems

With the emergence of connected vehicle and connected traveler data, TMCs will need additional communications bandwidth (both field-to-center and internal networking capabilities) hardware and data storage, but also specific tools and technologies for processing and handling the volume and velocity of

the data streams. The magnitude needed by individual TMCs and partnerships will vary across regions. Regardless of the size of system needed, however, the architecture of the big data tools and technologies system is likely quite similar from deployment to deployment. This integration architecture was illustrated in Figure 2.

After reading this Chapter, the reader should understand:

- Public connected vehicles exchange data with agency RSEs and consume information from real-time CV functions.
- BSM/PDM information on public connected vehicles is collected through the RSEs and transmitted to the TMC as they are received in real-time.
- A real-time streaming acquisition system processes the trajectories and passes the information to an analysis component(s) which performs real-time anomaly checks, aggregations, and other transformations for *real-time* functions.
- At the same time, the data is stored in HDFS and extracted by *near-real-time* processes that use more trajectory data over small portions of time to send near-real-time data to ATMS functions.
- The HDFS database is synchronized with a data repository that stores emerging data and connected infrastructure information over very long time periods for offline functions and analysis. Methods are in place to synchronize and harmonize the data from ATMS RDBMS for use by each of the real-time, near-real-time, and offline functions.
- A typical agency will be faced with 2-22 TB of data a day. Data aggregation and analytics strategies will be needed to bring these rates to reasonable levels.
- Individual agency functions that utilize emerging data sources are coordinated with peer and partner agencies through technical data sharing connections. While a common data sharing model likely includes a synchronized data repository and a data distribution service that shares related information between relevant partners, regional and institutional differences may result in different types of deployments as technology is deployed over the next 10 years.

Chapter 4. Opportunities for Integration of Emerging Data in Traffic Management and Traffic Management Centers Using Big Data Tools and Technologies—Next Steps

As discussed in Chapter 2, emerging data from connected vehicles and travelers can enhance traffic management and TMCs across many functional areas. In order to take advantage of these data sources, big data tools and technologies will be needed for data acquisition, storage, processing, and analytics. A notional system architecture was identified in Chapter 3 that illustrates the types of tools and interfaces that will be needed. Chapter 3 provided examples of the data processing and storage requirements for a typical agency when connected vehicle, traveler, and infrastructure data is being transferred to the TMC at significant levels. Implementation of new technologies to take advantage of the significant opportunities will require careful planning. After reading this Chapter, the reader will understand the key questions to be addressed in developing a plan for leveraging the emerging data sources with big data tools and technologies. Since there are many possible combinations of functions, tools, agency capability maturity, existing legacy system hardware, software, and communications network characteristics, a plan cannot be "one size fits all". However, common themes have been identified in a number of areas.

Emerging data will impact traffic management and TMC functions in a variety of ways as discussed in Chapter 2. Some of the highest impacts of the new information will likely be found in the following areas:

- Improvement of **situational awareness for incident management** through video analytics and high-resolution vehicle trajectories.
- Provision of new services for **speed warning, hazard warning, and Intersection Collision Avoidance**.
- Improvement of ramp metering and **traffic signal control operations** through reduced reliance on in-pavement or non-intrusive point sensors.
- Improvement of lane management and **variable speed limit systems** through reduced reliance on in-pavement or non-intrusive point sensors.
- Provision of new services for **emissions and road-weather modeling and estimation**;
- More **cost-effective** and **timely asset management**.
- New insights in **performance measurement** through deeper coverage of the network by high-resolution vehicle trajectories.

To achieve these results, agencies will need to develop a comprehensive plan for leveraging emerging data sources and incorporating big data tools and technologies into their traffic management practices and TMCs. An agency plan need not address all of these functions, but any comprehensive plan should discuss the following topics in following the general systems engineering process:

- High-priority functions that align with agency goals and objectives.
- Regions and sites for deployment of field equipment and communications.
- TMC equipment upgrades, changes, and augmentations.
- Changes to staffing, education, and organization.
- Plans for collaboration, partnering, and data sharing and exchange.

The following sections outline the key elements of each of these five topical areas.

4.1 Identifying High Priority Functions According to Agency Goals and Objectives

The plan should start first by identifying a concept of operations for how emerging data sources will be used to augment existing and provide new traffic management functions. Most agencies have many strengths and some weaknesses in their current delivery of traffic management and TMC functions. Emerging data and big data tools might be used to *enhance current strengths*, *improve current weaknesses*, and, in many cases, *provide new services* that the agency is not providing today. Some questions to be answered include:

- What *real-time* functions identified in Chapter 2 are most important to the agency? Which functions align best with agency objectives? Most real-time functions will require *roadside equipment deployment*.
- What *near-real-time* functions identified in Chapter 2 are most important to the agency and align best with agency objectives? Most near-real-time functions will require big data tools for *acquisition* and *storage*.
- What *off-line* functions identified in Chapter 2 are most important to the agency? Most off-line functions will require big data tools for *storage* and *analytics*.
- What services in Chapter 2 are provided by the agency today that could be further improved through emerging data sources? Refer to Table 1 for qualitative assessment of potential improvements.
- What services are *not* provided today that the agency could start providing when emerging data sources are readily available? Refer to Table 1 for potential new services.
- How should the enhancements be implemented over time? Which services should be provided in the next 0-5 years? 5-10 years? 10+ years? Functions that require higher levels of emerging data (more travelers, more vehicles, and larger coverage areas) will not be as successful if implemented too early.
- How does the provision of enhanced functions in Chapter 2 align with the need for high levels of connected vehicles and connected travelers? Near term services should be those that require only low levels of connected vehicle and traveler data to be effective.
- Should the plan rely on *public* connected vehicles, *commercial* connected vehicles, or *both*? Can connected traveler data sources provide the improvements to traffic management and TMC functions? If so, how will the data sources be procured or developed? Many functional enhancements and new services could be provided without extensive field device deployment if commercial sources are available that provide similar information fidelity.

This concept of operations should articulate how the high-priority functions meet will serve as the framework for identifying what big data tools and technologies will be needed to meet the agency goals and objectives of utilizing emerging data sources to improve traffic management and TMC functions.

4.2 Identifying Sites and Regions for Field Equipment Deployment

If the concept of operations includes traffic management functions that require field elements for data collection and data dissemination (in most cases this will include the real-time functions summarized in Chapter 2), there are a host of additional issues to be addressed in this section of the plan.

- How does the size of the field deployment relate to the needs for each high priority function in the concept of operations? How many RSE locations are needed to that function to be effective? Refer to Table 1.
- What regions of the agency jurisdiction would best benefit from deployment of each high-priority function? Identify corridors and areas with traffic safety, travel time reliability, recurrent congestion, and environmental issues.
- How can deployment of field equipment for collection of emerging data sources be aligned with other technology projects? Replacement of traffic signal controllers? Deployment of ATM equipment, DMS signs, CCTV cameras, or other field devices? It may be cost effective to link field equipment deployment with other device replacement and upgrade activities.
- What locations and regions should be targeted for deployment in 0-5 years? 5-10 years? 10-15 years? How do these deployment footprints align with available funds? Can a plan for additional funds for equipment procurement, deployment, and on-going operations be developed? Compare the agency size and functions to the example data loading analysis provided in Chapter 3. Is the existing communications network adequate to handle the data volume and velocity? What changes and augmentations are needed to make to the communications links and backbones ready for the data loading in 0-5 years? 5-10 years? 10+ years? Adequate communications network capacity is a key element of a successful strategy.
- If field equipment deployment is *not* desired, is there a plan for acquisition of emerging data from commercial sources that can provide real-time functions? Internally-developed applications? What agency regions would be desirable for these data sources to cover?

4.3 Planning for Traffic Management Centers Equipment Upgrades, Changes, and Augmentations

Once the high-priority functions are identified and field deployment plans are conceptualized, the next step of the plan will be to identify the necessary changes and additional big data tools and technologies in the TMC. Refer to Figures 2, 3, 4, and 5 for the necessary big data tools and technologies related to real-time, near-real-time, and off-line functions. It will be important to estimate how the expected agency data acquisition rates compare to the example analysis provided in Chapter 3. This section of the plan should address the following questions:

- Pre-planning
    - Is the agency comfortable with open source products or is there a preference for commercial products? Refer to FHWA-JPO-16-424 for a discussion of the state of the practice in big data

systems. In many cases, commercial systems are combinations of open source tools with proprietary enhancement services and user interfaces.

- What are the current barriers to adoption of new software? Procurement procedures? Standardized policies? Staff skills? Big data tools and technologies may need to be procured as a combination of off-the-shelf products and customized integration services.
- At what level of maturity would you classify your organization's technical capabilities?

  Will the solutions be built, managed and maintained by internal staff, contractors, or a combination of the two? Software delivery models may differ (see Figure 7) depending on the desired level of responsibility between system vendors, integrators, and agency staff.
- Can any existing agency hardware, software, or systems be leveraged?
- What big data tools and technologies do you already use for ATMS functions? Which systems are integrated and which are separate? Do your current ATMS vendors or developers already have systems of plans for including big data processing of emerging data sources? To integrate emerging data with legacy systems and functions, existing ATMS integrators and vendors should be brought into the conversation as early as possible.
- What funds are available for technology upgrades in 0-5 years? 5-10 years? 10+ years? Big data tools and technologies are not inexpensive. Even use of free open source tools can come with significant costs for setup, integration, operations, and data storage.

- Acquisition
  - According to the concept of operations, do we have enough capacity for data acquisition rates in 0-5 years? 5-10 years? 10+ years? Refer to Chapter 3 for example data rates of a typical agency.
  - If adequate storage capacity is not available on-site, what is the agency's preferred method of procuring services among the options in Figure 7? Each option comes with strengths and weaknesses as discussed further in FHWA-JPO-16-424.
  - Of the tools listed for *real-time* acquisition in Chapter 3, do we have experience with any of these? Many more real-time stream processing tools are emerging and maturing at a fast rate in 2017.
  - Of the tools listed for *near-real-time* acquisition in Chapter 3, do we have experience with any of these? Many more stream processing tools are emerging and maturing at a fast rate in 2017. (https://thenewstack.io/apache-flink-addresses-continuous-stream-processing/)
- Marshalling and Storage
  - Where does the agency plan on hosting the data - locally or in the cloud? Are there any barriers to using private or public clouds? Among the procurement options in Figure 7, what is the agency's preferred procurement method, today? In 5 years?
  - According to the concept of operations, what technical issues with acquisition and storage will need to be managed? What variety of data does the agency expect for each traffic management function? Connected vehicle data may be strongly standardized today, but may expand to include additional data elements in the future. Scalability is an important consideration for use of HDFS/NoSQL.
  - How often does the agency anticipate needing to scale the solutions up (or down)? Emerging data is expected to climb continuously as more and more vehicles are equipped with the broadcasting technology.
  - What level of latency is the agency comfortable with in accessing the data for each type of function? Has the agency identified the necessary components from Figure 2 for real-time and

near-real-time processing? If traffic management decisions are expected within minutes or perhaps seconds of data arrival, the system design must accommodate this.

- For each functional enhancement, what *unstructured* and *semi-structured* data would be necessary? Notably video and images require much different software processing systems than trajectories of status information from travelers or vehicles.
- How long does the agency wish each type of data to be retained? Can a strategy for archiving or aggregation of aging data be developed that minimizes costs? A future report in this series will propose some aggregation schemes. What data can be offloaded to long-term storage to reduce online data costs?
- What agency security and data governance standards will apply for the data collected? What protections for personally identifiable information (PII) will need to be ensured? Refer to other guidance from U.S. DOT related to the CV Pilots.
- What is the acceptable availability of the new functions and augmentation of existing functions? How long could the new data sources and software services be "off line" and still provide acceptable traffic management services to the public?
- What RDBMSs are currently in use? Are they compatible with NoSQL and Hadoop systems already? If not, how can they be made compatible?
- What data from the legacy RDBMS will be needed to be shared with the emerging data sources to achieve the functions in the concept of operations? At minimum, to geo-locate the emerging data sources the regional asset metadata will be required to be merged. (Data such as road names, signal phases and movements, intersection and segment geometry, and timing plan parameters among many others.)

- Analysis
  - What languages, analyses, and tools/technologies does the agency currently use or have experience with? Many big data tools are based the Linux operating system and use Java, Scala, Python, R, and others.
  - How complex are the agency's current analyses and how frequently is data processed for traffic management functions?
  - What analytical skills is the agency interested in expanding or willing to expand to?
  - How advanced will the data analysis procedures be? Is machine learning needed or anticipated? Machine learning systems continue to advance although truly self-service machine learning systems are still developing.
  - Will media (e.g., video, imagery, etc.) analysis, text analysis, or other types of unstructured data processing be needed to provide the high-priority functions?
  - Does the agency want to maintain manual control over analytical algorithms and procedures? Are you willing to allow traffic management functions to operate without user intervention when considering emerging data sources? It may be prudent to pursue a two-phase approach where data from emerging sources are used to synthesize recommendations for TMC staff initially, transitioning to on-line, closed-loop operation after a learning period.
  - What is the criticality of the interruption of data analyses? If data query responses to emerging data sources were slowed for some reasons, would this affect traffic management operations? While the system should be designed and tested from the beginning for scalability to realistic data flow rates, until significant levels of emerging data are flowing it may not be completely known if the processing can be done in a timely fashion.

- Action
  - What interfaces currently exist to the agency's ATMSs? What interfaces will need to be developed? Refer to Figure 2.
  - What ATMSs do not currently exist that will need to be procured to provide a traffic management function? Can they be procured "from the ground up" to include capabilities and features of big data tools and technologies?
  - What data sharing mechanisms currently exist? Will the interfaces be adequate to share emerging data sources? What new interfaces will need to be developed? It is more likely that new interfaces will need to be developed, but also possible that standards will be developed in the next five years as the sources continue to grow.
- Maintenance and security
  - Would the agency prefer a platform with on-site technical support? Remote support? Transfer of support to agency staff? Initial procurements of new technologies, or technologies new to an agency, are not commonly procured without maintenance contracts. In 10 years, if such products become commodity elements, it would be more likely that agency staff could provide maintenance and support.
  - How sensitive might the emerging data sources be or become? How will the agency control and restrict access?
  - What regulations and standards will the new software need to meet? What new standards could be needed in 0-5 years? 5-10 years? 10+ years?
  - How frequently will the data and product life cycles reset? Will systems need replacement or upgrade in 5 years? 10 years? Big data tools and technologies are continuing to develop at a rapid pace. Systems that required manual coding and integration in 2021 are likely to be "off the shelf" by 2026.
  - How rigorous of a backup and recovery process will be required? Will backup and recovery be managed in-house or via service contract?

4.4 Planning for Staffing, Education, and Organizational Changes

The previous sections of the plan focus on technologies and systems. Improvements to traffic management functions and TMC operations are not only brought about through new physical components, hardware, and software but also through people—their training, skills, and core competencies. It may perhaps be even more important to develop staff capacity and education in these growing technology areas than purchase and integration of new software. This section of the agency plan should address the following questions:

- How prepared are agency TMC users for emerging data (e.g., their challenges, technical skills, etc.)?
- What is the expected user experience for agency operators? User experiences are as important as data and systems integration in the success of new software functions.
- What are the desired outcomes of the end user's tasks (e.g., will they impact other aspects of the organization or other users)? Designing the user interfaces and business process flows to be intuitive is an important element in software adoption.
- How will agency users access the big data systems and tools (e.g., local client software, intranet, internet)? Are there any barriers inherent to agency information systems and communications

systems architectures preventing access? Many big data tools and technologies are usable through web-browser interfaces. Depending on the deployment choices selected from Figure 7, providing access to user software may require internet access or other IT configurations specific to an agency network.

- Does the agency have a champion for big data tools and technologies? If not, who is a likely champion? If there is no internal champion, are there champions at partner agencies? Advocacy and project ownership by specific individuals are strong determinations of project success.
- Does our organization embrace technology and innovation? Is there a commitment from leadership to advance technology, including connected vehicles, connected travelers, and other innovating data sources?
- What groups within the agency will have involvement in deploying, operating or maintaining data collection, data processing, analytics, and integration technologies and systems?
- What skills in databases, software installation, software maintenance, application scripting, dashboarding, and analytics does the agency already have? How can we obtain personnel with such skills?
- What specific technical areas do we have that can support emerging data and big data tools and technologies?
    - System engineering.
    - Design.
    - Deployment/Integration.
    - Data management.
    - Operations.
    - Maintenance.
    - Analytics.
- Is there flexibility to acquire agency staff with these skill sets (i.e., redefine roles, expand technical staff groups)?
- Do we have a mechanism to obtain these skills if they cannot be addressed by current staff or roles (i.e., contract/outsource, training)?
- Are there any operational or policy limitations to our agency deploying data collection devices, apps, or other services?
- Is the agency active in national traffic management organizations (or activities) where we are hearing about peer agency programs and experiences, national trends, and emerging technologies?
- Are there any barriers to participating in these organizations?

4.5 Planning for Collaboration, Partnering, and Data Sharing

Planning for partnering and collaboration with other agencies is as important as planning for internal staff skills and capabilities. Many traffic management and TMC functions can be enhanced by sharing data across agency boundaries:

- Is there regional interest or established goals that will be achieved through collection of emerging data sources? Are multiple agencies interested in advancing capabilities for processing and

analytics of connected traveler and vehicle data? Based on costs, it may be necessary to partner with multiple agencies in procurement of new tools and technologies.

- Have potential roles and responsibilities been identified for implementing or piloting data collection, processing, storage, and analytics?
- Is there a forum for partner agencies to collaborate / discuss / obtain consensus on potential data collection, processing, storage, and analytics capabilities?
- Are there opportunities to leverage existing processes among agencies (business processes, planning, procurement, system engineering, operations) to initiate data collection, processing, storage, and analytics capabilities?
- Do some partner agencies have fewer barriers to certain processes?
- Are there regional processes that would need to be factored in to piloting data collection, processing, storage, and analytics capabilities (i.e., Transportation Improvement Plan, programming cycles, flexibility to fund near-term improvements)?
- Are there partner agencies with staff who have skill sets that would align with data collection, processing, storage, and analytics capabilities?
- How aligned are partner agency missions with regard to traffic management and TMC operations related to the functions in Chapter 2? Is there a consistent appetite and leadership support among partner agencies to pursue enhancement to TMC functions through use of emerging data sources?
- How involved is the private sector in different traffic management initiatives at the agency? Are there barriers to engaging the private sector in some capacity as part of a big data tools and technologies deployment pilot? What are those barriers?

4.6 The Future of Advanced Traffic Management System Integration with Emerging Data and New Big Data Tools and Technologies

The future of ATMS integration with emerging data and big data tools and technologies is (1) direct integration of the big data technologies into the ATMS, and (2) edge processing and distributed computing. This report suggests integration between legacy systems and emerging technologies through newly-developed interfaces. In the future, the approach will likely be to overhaul the way ATMS RDBMS connections are made to utilize HDFS storage directly. Integrating analysis tools and decision support system functions into ATMS will also be a trend that is enabled with big data tools and technologies. As “off the shelf” big data analysis suites become more robust, it will be easier to incorporate these tools directly into ATMS over the next 5-10 years.

After reading this Chapter, the reader should understand:

1. Taking advantage of emerging data sources to enhance traffic management and TMC functions will require using big data tools and technology.
2. Implementing the big data tools and technologies will require careful planning.
3. A concept of operations should be developed that encompasses a number of key topics.
    a. A concept of operations for the high priority traffic management functions to be enhanced with emerging data sources.

   b. A plan for field equipment deployment, if field equipment is required for the traffic management functions.
   c. A plan for procurement, deployment, integration, and operation of big data tools and technologies in the TMC.
   d. A plan for development of staff skills and modifications to organizations roles and responsibilities.
   e. A plan for collaboration and data sharing with agency partners.

4. The future of ATMS may directly integrate big data tools and technologies into their systems. In the next 0-5 years, integration of new tools with legacy systems and databases will be required to take advantage of the new data sources.

Appendix A. References

1. U.S. Department of Transportation, ITS Joint Program Office, “Big Data’s Implications for Transportation Operations: An Exploration”, Publication No. FHWA-JPO-14-157, December 2014.
2. McKinsey Global Institute, “Big Data: The next frontier for innovation, competition and productivity”, May 2011. Accessed at: http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.
3. Kimley-Horn and Associates, Inc., “Traffic Management Centers in a Connected Vehicle Environment”, TMC Pooled Fund Study, March 2014.
4. U.S. Department of Transportation, ITS Joint Program Office, “Big Data and ITS”, White Paper, October 2013. Accessed at: http://connectedvehicle.itsa.wikispaces.net/file/detail/ITS+and+Big+Data+White+Paper+Final+Draft+10_2+%282%29.docm.
5. International Transport Forum, “Big Data and Transport: Understanding and assessing options, 2015”. Accessed at: http://www.trb.org/Main/Blurbs/172646.aspx.
6. U.S. Department of Transportation, ITS Joint Program Office, “Estimate Benefits of Crowdsourced Data from Social Media”, Publication No. FHWA-JPO-14-165, February 2015.
7. AASHTO, “National Connected Vehicle Field Infrastructure Footprint Analysis”, Publication No. FHWA-JPO-14-125, June 2014.
8. Oregon DOT, “MyOReGO | A new way to fund roads for all Oregonians”, Accessed May 13, 2016, http://www.myorego.org.
9. HERE, “HERE, automotive companies move forward on car-to-cloud data standard”, Accessed July 1, 2016, https://lts.cms.here.com/static-cloud-content/Newsroom/290616_HERE_automotive_companies_move_forward_on_car_to_cloud_data_standard.pdf.
10. Goodall, Noah, Brian Smith, and Byungkyu Park. "Traffic signal control with connected vehicles." *Transportation Research Record: Journal of the Transportation Research Board* 2381 (2013): 65-72.
11. Hawkes, Allen. "Traffic Control with Connected Vehicle Routes in SURTRAC." (2016).
12. Day, Christopher M., and Darcy M. Bullock. "Opportunities for Detector-Free Signal Offset Optimization with Limited Connected Vehicle Market Penetration: A Proof-of-Concept Study." *Transportation Research Record* (2016).
13. Lukasik, Dan, et al. Guidelines for Virtual Transportation Management Center Development. No. FHWA-HOP-14-016. 2014.
14. NCSA, “About Blue Waters”, Accessed October 19, 2016, http://www.ncsa.illinois.edu/enabling/bluewaters.
15. http://www.opengeospatial.org/docs/is.

Appendix B. List of Acronyms

AASHTO	American Association of State Highway and Transportation Officials
API	Application Programming Interface
Apps	Applications
ATM	Advanced Traffic Management
ATMS	Advanced Traffic Management System
BSM	Basic Safety Message
CCTV	Closed-Circuit Television
CICAS	Cooperative Intersection Collision Avoidance Systems
CV	Connected Vehicle
DMS	Dynamic Message Sign
DMV	Department of Motor Vehicles
DOT	Department of Transportation
DSRC	Dedicated Short Range Communications
ESS	Environmental Sensor Stations
FHWA	Federal Highway Administration
FTP	File Transfer Protocol
GB	Gigabyte
GID	Geometric Intersection Description
GIS	Geographical Information Systems
GPS	Global Positioning System
GUI	Graphical User Interface
HD	High-Definition
HDFS	Hadoop Distributed File System
I2V	Infrastructure to the Vehicle
IaaS	Infrastructure-as-a-Service
ICM	Integrated Corridor Management
IoT	Internet of Things
IT	Information Technology
ITS	Intelligent Transportation Systems
LiDAR	Light Detection and Ranging
LCS	Lane Control System

MAP A message containing roadway geometric information
MAP-21 Moving Ahead for Progress in the 21st Century Act
MMITSS Multi-Modal Intelligent Traffic Signal System
MOU Memorandum of Understanding
MTA Metropolitan Transportation Authority
NHTSA National Highway Traffic Safety Administration
NoSQL Not Only Structured Query Language
NTCIP National Transportation Communications for Intelligent Transportation System Protocol
OGC Open Geospatial Consortium
PB Petabyte
PDM Probe Data Message
PeMS Caltrans Performance Measurement System
PII Personally Identifiable Information
RDBMS Relational Database Management System
RAID Redundant Array of Independent Disks
RAM Random-Access Memory
RITIS Regional Integrated Transportation Information System
RSE Roadside Equipment
RWIS Road Weather Information System
SPaT Signal Phase & Timing
SQL Structured Query Language
SRM Signal Request Message
TB Terabyte
TBD To-Be-Determined
TIM Traveler Information Message
TMC Traffic Management Center
TMDD Traffic Management Data Dictionary
TSMO Transportation Systems Management and Operations
URL Uniform Reference Locators
V2V Vehicle-to-Vehicle
V2X Vehicle-to-Other Objects
VSL Variable Speed Limits

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

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

FHWA-JPO-18-625

U.S. Department of Transportation