NCHRP 20-44(39)

IMPLEMENTATION OF GUIDEBOOK FOR MANAGING DATA FROM EMERGING TECHNOLOGIES

FINAL REPORT

Prepared for:

National Cooperative Highway Research Program,
Transportation Research Board
of
The National Academies of Sciences, Engineering, and Medicine

Prepared by:

Kelley Klaver
Applied Engineering Management Corporation
13880 Dulles Corner Lane, Suite 300
Herndon, VA 20171-4687
July, 2023

Acknowledgements

This study was conducted with funding provided through the National Cooperative Highway Research Program (NCHRP) Project 20-44(39), "Implementation of Guidebook for Managing Data from Emerging Technologies". The NCHRP is supported by annual voluntary contributions from the state Departments of Transportation. The report was prepared by Kelley Klaver of Applied Engineering Management Corporation, with the active participation and efforts of the departments of transportation of Arizona, District of Columbia, Florida, and Rhode Island. The work was guided by a technical working group. The project was managed by Sid Mohan, (NCHRP Senior Program Officer).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. INTRODUCTION

This report presents the background, objective, approach, and outcomes of a National Cooperative Highway Research Program (NCHRP) implementation project funded through the NCHRP 20-44 Implementation Support Program to facilitate implementation of NCHRP research results. Specifically, this implementation project, NCHRP 20-44(39) Implementation of Guidebook for Managing Data from Emerging Technologies, sought to support agencies in the implementation of *NCHRP Research Report 952 Guidebook for Managing Data from Emerging Technologies for Transportation* (Pecheux, Pecheux, Ledbetter, & Lambert, 2020). This guidebook, published in 2020, is meant to help state departments of transportation (DOT) begin or further advance efforts to more effectively manage and use data from emerging technologies for operations and planning decision-making, with the long-term goal of organization-wide shifts to modern data management practices that efficiently meet current and future operational and business needs.

The research conducted for NCHRP 08-116 proved that most transportation agencies rely on traditional data management systems and practices. The research also confirmed the need to leverage data from emerging technologies to improve strategic and tactical decision-making. The modern, big data concepts and approaches presented in NCHRP Research Report 952 are critical to the ability of state DOTs to effectively manage emerging technology data. These concepts and approaches are also new to transportation agencies and run contrary to the traditional data systems and management approaches of most state DOTs. As such, an implementation project was justified to gauge the ability of state DOTs to apply the recommendations, roadmap, and tools presented in NCHRP Research Report 952 and to develop case studies of successful implementations and lessons that can support other agencies pursuing similar implementations.

Background

The modern data management framework presented in NCHRP Research Report 952 includes over 100 recommendations/guidelines across the modern data management lifecycle. While this guidance is a good start to documenting modern approaches to data management for transportation agencies, these approaches represent a drastic change for many DOTs. Implementing these approaches will involve non-trivial culture changes and serious reconsiderations of ingrained processes from information technology (IT) to procurement to human resources. Figure 1 shows the roadmap accompanying the guidance to assist agencies in implementing the modern data management framework.

Figure 1. Roadmap to Managing Data from Emerging Technologies

In September 2020, toward the end of the NCHRP 08-116 project, the research team conducted a national webinar to share research outcomes.

There were three presentations, including the research team's overview of the research products, as well as presentations from the Kentucky Transportation Cabinet (KYTC) and the Portland Bureau of Transportation (PBOT), both of which were working to implement modern data management approaches. Of the 256 webinar attendees, 90% stated they were satisfied or very satisfied with the webinar, demonstrating the interest and timeliness of the topic. Some comments from webinar participants pointed not only to their interest in this topic but also to the need for more real-world applications and examples such as those presented by KYTC and PBOT:

- "Excellent information."
- "Presentations were fantastic!"
- "The case studies were great."
- "Would like to make this a series with more general overviews first, then increasing depth."
- "The use cases implemented by KYTC and PBOT were very informative. It is good to see implemented use cases, their journey, lessons learned for such new topics, especially for agencies starting to learn about these features."

In addition, 62% of attendees responded positively to the question, "If this research reaches an implementation stage, would your agency be interested in working with the research team to apply it?" The remaining respondents reported that they were not in a position within their agency to apply these techniques.

As such, an implementation project was recommended as the next step to train and assist a few agencies in tackling the contents of the guidance. Implementing modern data management practices requires a complete paradigm shift that will not happen overnight. While one person within an organization might understand the need for and the associated benefits of the modern data approach, these champions need support in explaining the needs to management, IT, procurement, and executive leadership. Bringing together people from various parts of an organization and at various levels of management is critical to the implementation of the modern data management concepts presented in NCHRP Research Report 952.

Objective

The objective of the NCHRP 20-44(39) implementation project was to bring the recommendations/ guidelines, roadmap, and tools from NCHRP Research Report 952 to state DOTs interested in not only better managing data from emerging technologies but deriving direct benefits from using the data in ways they currently cannot, due to limitations in their existing data systems and management approaches and practices. Additionally, the NCHRP 08-116 research showed that there is often a disconnect between individual group/division needs and organization-wide IT and procurement practices. Thus, an additional objective of the project was to bring together people from different parts of an organization, provide the necessary foundational information, facilitate discussions, assist in addressing issues and concerns, and help bring about a meeting of the minds and next steps specific to the needs of the organization.

Objective of Implementation Project

Assist interested transportation agencies in progressing along the Roadmap to Managing Data from Emerging Technologies.

Methodology

The approach for the implementation project included the following:

- Identify agencies interested in participating in the implementation project.
- Develop a 1.5-day training workshop based on the content of NCHRP Research Report 952.
- Deliver a 1.5-day training workshop to each of the participating agencies.
- Work with agencies to identify a specific use case for the implementation of the NCHRP Research Report 952 concepts.
- Assist agencies in developing implementation plans based on discussions during the training workshops and their identified use cases.
- Provide technical support to agencies during the implementation project.
- Organize and facilitate a 1-day peer exchange with participating agencies approximately a year from the training to share successes and lessons.
- Analyze and synthesize the outcomes of the implementation efforts.
- Organize and facilitate a national webinar to share with other transportation agencies lessons learned during the implementation.

Organization of Report

The remainder of this report is organized as follows:

- Chapter 2. Engagement with Participating Agencies – presents information on the content, conduct, and outcomes of the training workshops with the four participating agencies; development of implementation plans for each agency, and an agency peer exchange to share information on implementation projects, successes, challenges, and next steps.
- Chapter 3. Case Study on the Arizona Department of Transportation's (ADOT) Implementation Project – provides background, an overview, outcomes, and next steps associated with the implementation efforts undertaken by ADOT.
- Chapter 4. Case Study on the District Department of Transportation's (DDOT) Implementation Project – provides background, an overview, outcomes, and next steps associated with the implementation efforts undertaken by DDOT.
- Chapter 5. Case Study on the Florida Department of Transportation's (FDOT) Implementation Project – provides background, an overview, outcomes, and next steps associated with the implementation efforts undertaken by FDOT.
- Chapter 6. Case Study on the Rhode Island Department of Transportation's (RIDOT) Implementation Project – provides background, an overview, outcomes, and next steps associated with the implementation efforts undertaken by RIDOT
- Chapter 7. Conclusions – provides conclusions of the implementation efforts undertaken during this project.

CHAPTER 2. ENGAGEMENT OF PARTICIPATING AGENCIES

Training Workshops

This section describes the content, delivery, and outcomes of the training workshops conducted as part of the implementation approach. The purpose of the workshops was to (1) train agency staff on the overall contents of the guidebook (Step 1 on the roadmap), (2) identify a use case and define the scope for the implementation effort to best utilize agency resources and appropriately reflect the implementation budget (Step 2 on the roadmap), and (3) identify key stakeholders and champions and potential roadblocks for the implementation projects. The first workshop served as an initial pilot test of the training. Feedback from participants in the first workshop was used to modify subsequent workshop deliveries.

Training Workshop Content

The research team developed, organized, and delivered training workshops that included generalized content from the guidebook, as well as tailored information specific to each agency's objectives and needs. During the training workshops, instructors shared content that helped participants:

- Contrast the traditional approach to data systems and management with the modern approach and discuss why the traditional approach will not suffice for managing data produced by emerging technologies, particularly over the long run.
- Introduce and discuss modern data management principles and industry best practices and discuss the potential challenges specific to agencies in implementing these principles and methods, as well as ways to overcome these potential challenges.
- Review the steps in the Roadmap to Modern Data Management.
- Identify champions and key players and stakeholders.
- Identify and discuss a specific use case and associated data sources for the implementation projects.
- Develop a high-level implementation plan for each agency.

This information was presented across the following six modules:

- Module #1 – What is Big Data and Why Do We Care?
    - Learn big data/modern data definitions, characteristics, and concepts
    - Understand when to pursue big data/modern data management
    - Additional resources
- Module #2 – Traditional Data Management Systems vs Modern, Big Data Management Systems
    - Discuss today's rapid growth of data
    - Compare traditional data warehouse architecture to modern/big data architecture
    - Drill down to understand the difference between traditional data management and modern data management across 13 data management lifecycle components
    - Understand the need to move toward the modern data management approach
- Module #3 – Pilot Project Implementation Plan – discussed the Who, What, When, Why, and How
    - Understand the concept of an embryotic big data test environment
    - Understand the benefits and potential barriers to cloud and cloud services
    - Discuss and solidify an achievable, modern data management pilot implementation project
    - Identify and define various aspects of the pilot project implementation plan
    - Secure buy-in for the implementation

- Module #4 – Getting Started: Establishing the Data Environment/Playground
    - Understand how to establish a big data test environment in the cloud
    - Understand how a data lake is organized and used
    - Understand how data flow in and out of modern data environments
- Module #5 – Developing the Pilot – Governance, Roles/Responsibilities, Tools, Skills
    - Understand the difference between traditional data governance and modern data governance and why a modern approach is needed for governing big data
    - Learn that how the data are governed can greatly impact the use of the data and associated outputs
    - Understand the roles, responsibilities, processes, types of tools, needs, and skillsets associated with developing a pilot project in the data playground
- Module #6 – Beyond the Pilot: Making the Case for Modern Data Management Across the Organization
    - Understand the iterative process to the growth of the modern data environment
    - Identify how to make the case for modern/big data across the agency beyond the pilot

The workshop agenda is provided in Appendix A.

Training Workshop Delivery Timeline and Participants

The team delivered four training workshops in 2021 and 2022 as follows:

- District of Columbia Department of Transportation (DDOT) held in Washington, D.C., December 6-7, 2021.
- Rhode Island Department of Transportation (RIDOT) held in Providence, RI, February 2-3, 2022.
- Florida Department of Transportation (FDOT) held in Chipley, FL, February 21-22, 2022.
- Arizona Department of Transportation (ADOT) held in Phoenix, AZ, March 8-9, 2022.

Each workshop was 1.5 days in duration, starting with an afternoon, followed by a full day.

Across the agencies, approximately 8-12 staff attended each training workshop. Groups represented in the workshops included traffic operations/engineering, planning/geographic information systems (GIS), information technology (IT), research, and performance management. Specific roles of attendees included traffic engineer, TMC manager, IT manager, applications architect, developer, research coordinator, Chief Information Officer (CIO), enterprise applications manager, program analyst, ITS network engineer, GIS manager/analyst, and planning administrator. Those attending the training sessions tended to be driven by the specific use case selected by the agency and the champions spear-heading the implementation projects (e.g., TSMO, GIS). Implementation teams tended to be a subset of those attending the training sessions. Involvement of IT in the training and implementation project was a critical factor to success for several of the agencies. Buy-in and support from the CIO was key to one agencies' success.

Training Workshop Assessment

At the end of each training workshop, the team provided participants with an evaluation form to provide their feedback on the workshop content, delivery, and suggested improvements. The evaluation form is provided in Appendix B. Figure 2 summarizes the feedback from 27 participants across the four workshops. Overall, participants strongly agreed or agreed that the objectives of the training were met, the topics covered were relevant, the materials were useful, and they were satisfied with the training.

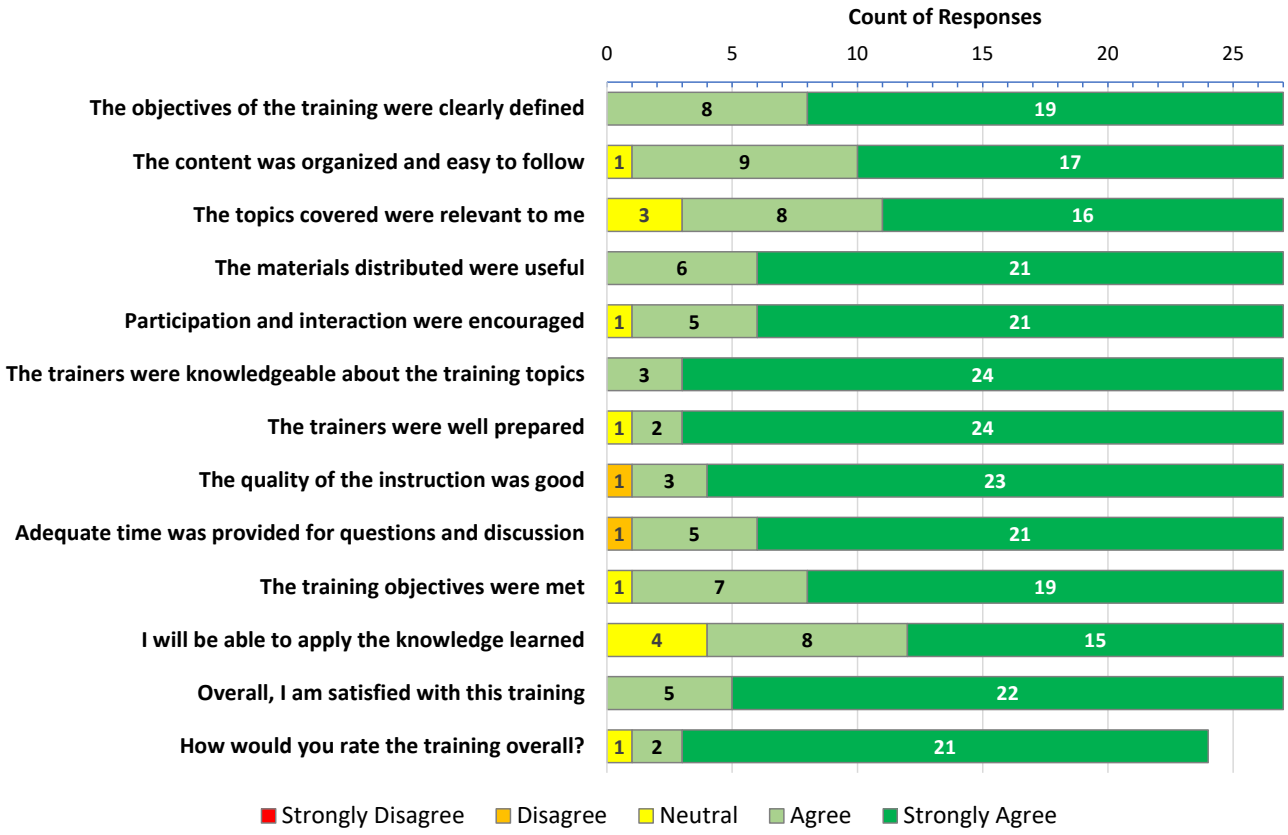


Figure 2. Evaluation Results

In addition to the Likert-scale questions, participants were asked what they liked best and least about the workshop and if they had any suggestions for improvements. Responses were as follows:

- What participants liked best about the training workshop:
  - Applicability/relevancy (4 participants)
  - Organization/clarity of course materials (4 participants)
  - Content quality (5 participants)
  - Interaction/discussion (6 participants)
- What participants liked least about the training workshop:
  - Short time period/a lot of material quickly / a lot to take in (3 participants)
  - Participants lacked experience/base information coming into the workshop (3 participants)
  - Some slides are slightly dense in material (2 participants)
- Suggestions for improvements included:
  - More use cases/examples (5 participants)
  - More time for discussion/longer training (5 participants)
  - Provide materials beforehand/prior knowledge or framing of the topic (4 participants)

Other comments provided by workshop participants included:

- “Thank you – This exceeded my expectations!”
- “Extremely relevant information, very timely to current challenges.”
- “Getting a good foundation on how to establish and leverage big data.”
- “Great training, very interactive and informative.”
- “I'd love a train-the-trainer in this information so we can do again with the agency staff.”

Agency Implementation Plans and Use Cases

Following the training workshops, the team worked with each agency to develop an implementation plan that described the implementation use case and laid out the goals/objectives, approach, schedule, datasets, and champions for the implementation project. Specific agency implementation plans are not provided in this document due to potentially sensitive nature of content; however, subsequent chapters of this report provide case studies on each agency's use case/implementation project, including:

- ADOT – Build, test, and further scale a pilot back-of-queue warning system using the ADOT cloud development environment.
- DDOT – Better understand dockless mobility device use patterns, including the factors that drive/hinder adoption of this travel mode.
- FDOT – Migrate an on-premises big data environment to the cloud to improve scalability, cost and performance and to share operational tools with other districts.
- RIDOT - Migrate RIDOT's enterprise GIS to the cloud and leverage cloud services and tools to ingest, store, enrich, process, and visualize real-time third-party data for incident and roadway management.

While the agencies selected specific use cases to explore in this implementation project, the groups discussed other potential use cases for applying the NCHRP Report 952 guidance, including:

- Storage, management, integration, and use of data from:
    - Vehicle probe data providers
    - Fleet vehicles (including images/videos)
    - Work zones/lane closures, Work Zone Data Exchange (WZDx) feed, smart work zone devices
    - Weigh-in-motion (WIM) systems
    - Radar-based vehicle detection (RVD) systems (e.g., volumes, speeds, trajectories)
    - Toll gantries
    - Construction management software (including photos taken in the field).
- Development of connected vehicle applications (e.g., work zone warning notifications, curve speed warning notifications)
- Automated data quality checking
- Integration/use of two or more datasets to provide additional insights. For instance, during a snowstorm, traffic volumes decrease, while maintenance activity increases. The integration of vehicle probe data and fleet vehicle data would help to understand how effective the agency was in getting information out to the public/getting people off the road, deploying assets for the event, improving efficiency of agency resources, and auditing payments to vendors.

Agency Peer Exchange

About one year after the training workshops, the agencies participating in the NCHRP 20-44(39) project joined for a peer exchange in Washington, D.C. on February 7, 2023. This all-day event provided the agencies with an opportunity to present their implementation projects and discuss successes and challenges with peer agencies. The agenda for the event is provided in Appendix C. The conversations and outcomes from this peer exchange are captured in the agency case studies in Chapters 3-6 as well as in the conclusions and opportunities presented in Chapter 7.

CHAPTER 3. CASE STUDY ON ARIZONA DEPARTMENT OF TRANSPORTATION'S IMPLEMENTATION PROJECT

Arizona is the sixth largest state, at 113,909 square miles, and the 14th most populated state, with approximately 7.3 million residents. Maricopa County (Phoenix metropolitan area) is the fastest-growing county in the U.S., with 4.5 million people. The state is also becoming an electric vehicle sales and manufacturing hub. As such, Arizona has and will continue to experience rapid growth and increasing traffic levels.

The overarching goal of ADOT's pilot project was to migrate a recent big data pilot application, proven on-premises, to ADOT's new cloud-based sandbox using concepts from the NCHRP Research Report 952 guidance.

Background

Arizona's Cloud First Policy

> "The initial purpose of the Cloud First Policy was to outline the state's use of cloud technologies for all infrastructure, platform, and software purchases made by state agencies."[1]

At the beginning of the NCHRP 20-44(39) implementation project, ADOT's Information Technology Group (ITG) was in the process of migrating and re-platforming its enterprise information and data system to the cloud. These efforts were being undertaken in response to Arizona's Cloud First Policy. "The initial purpose of the Cloud First Policy was to outline the state's use of cloud technologies for all infrastructure, platform, and software purchases made by state agencies. The goal was to promote and encourage the use of cloud technologies wherever possible and shut down the state's aging data centers by migrating applications to the cloud or the new Shared Hosted Data Center until agencies could complete the remaining app modernization. The state's Cloud First Policy referred to the practice of prioritizing cloud-based solutions in-lieu of on-premises solutions."[1]

Due to constraints associated with the migration timeline, the ITG was working to "lift and shift" its on-premises servers and databases to the cloud using Amazon Web Services (AWS). The ITG's roadmap for establishing a cloud-based enterprise environment involved 28 projects over 3 years (2021-2024), which included:

- Decommissioning servers
- Moving from Structured Query Language (SQL), a standardized programming language used to manage/operate relational databases, to AWS Aurora, a serverless MySQL and PostgreSQL-compatible relational database built for the cloud
- Building a data lake

An ADOT development environment had been stood up and was ready to go at the beginning of the NCHRP 20-44(39) project. This account is intended as a sandbox and is separate from the production network. It was built to allow ADOT staff to conduct research and development and learn how to use the cloud. At the beginning of the NCHRP 20-44(39) project, only IT staff were working in the AWS environment.

[1] Arizona Strategic Enterprise Technology, https://aset.az.gov/resources/cloud-resources#:~:text=Arizona's%20Cloud%20First%20Policy&text=The%20goal%20was%20to%20promote,complete%20the%20remaining%20app%20modernization

ADOT-MAG Back-of-Queue Warning Pilot Project

ADOT's System Technology Group (STG) supports Intelligent Transportation Systems (ITS) throughout Arizona, as well as Freeway Management Systems (FMS) in the Phoenix metropolitan areas. The STG aims to stay on the leading edge of technology, such as connected vehicle technology, and identifies the levels at which ADOT should invest in these technologies given cost, barriers to entry, and maintenance.

At the time of the NCHRP 20-44(39) project, the STG was leading a small-scale, connected vehicle feasibility project in cooperation with the Maricopa Association of Governments (MAG) and a 5G cellular service provider. The goal of the pilot project was to assess the effectiveness of an application to broadcast location-specific messages to travelers to improve safety and mobility along Arizona freeways. The pilot project leveraged cellular vehicle-to-everything (C-V2X) technology and multi-edge computing (MEC) to deliver traveler information messages (TIM), or roadside safety alerts (RSA), to travelers using a 5G smartphone mobile app (Figure 3). More specifically, the pilot aimed to use real-time streaming data and to dynamically transmit TIMs/RSAs to travelers one mile, two miles, and three miles upstream of the back-of-queue event using the Message Queuing Telemetry Transport (MQTT) protocol.[2] The use of the 5G cellular provider's MEC (virtual) servers replaced the need for physical roadside units (RSU). ADOT originally planned to use connected vehicle basic safety messages (BSM) from cell phones in vehicles, and they successfully tested and verified the ability to receive BSMs from they 5G cellular service provider. However, ADOT chose to make use of a probe vehicle data API for the pilot back-of-queue application instead of the cellular BSM data. The following steps represent the flow of the data for the pilot project, which is illustrated in Figure 3:

1. Probe vehicle data associated with speed and location of vehicle 1 (V1) indicates possible congestion.
2. Probe vehicle data are routed over a secure fiber backbone to a MEC server to calculate the TIM/RSA.
3. The MEC processes the probe vehicle data and automatically creates a TIM/RSA.
4. The TIM/RSA is forwarded over a secure fiber backbone.
5. The TIM/RSA is delivered to a cell phone app in vehicle 2 (V2) over a secure mobile link.
6. Event data are forwarded over a public/private network to ADOT's traffic operations center (TOC).
7. The TOC is notified and manages the event.

[2] MQTT is an OASIS standard messaging protocol for the Internet of Things (IoT). It is a lightweight publish-subscribe machine-to-machine network protocol for message queuing services, https://mqtt.org/.

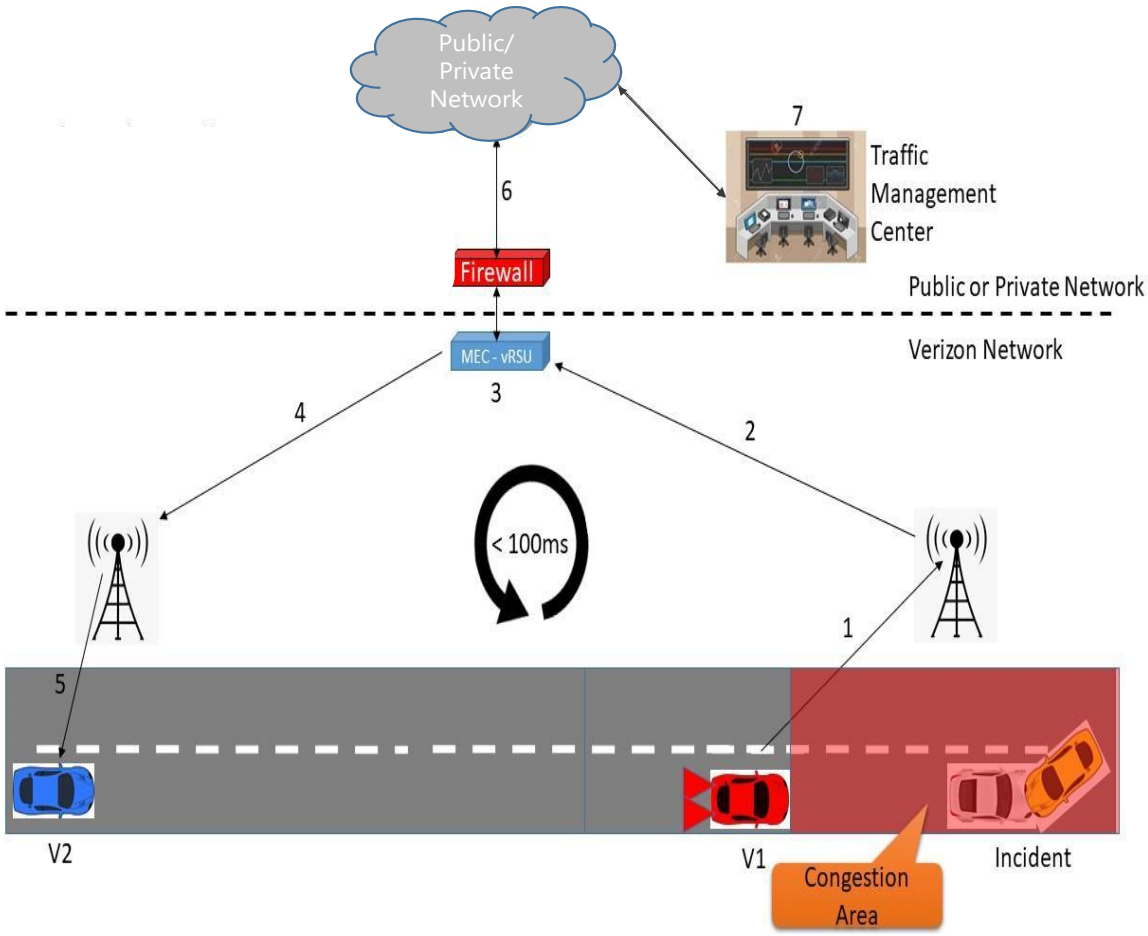


Figure 3. ADOT Back-of-Queue Warning System Feasibility Pilot

The data pipeline for the MAG-ADOT back-of-queue warning pilot project is shown in Figure 4. The pipeline involved the following steps:

- Obtain data from probe vehicle data API.
- Based on the event lat/long, use Python to calculate the geofenced areas to receive the TIMs/RSAs. Staff wrote an executable in GoLang, often referred to as Go, a programming language designed by Google. The GoLang code would execute code within the Python code and output the latitude/longitude. All Python code is processed in Golang as an executable due to its scalability and compatibility with various programming languages and graphical user interface (GUI).
- Create three messages (one-mile, two-miles, three-miles) using SAE J2735 and J2540 standards.
- Use GoLang to convert the messages to ProtoBuf format (to interface with the MQTT broker).
- Send messages to the MQTT broker.
- MQTT broker distributes messages to smartphones via the virtual RSU/AWS Wavelength managed by Verizon.

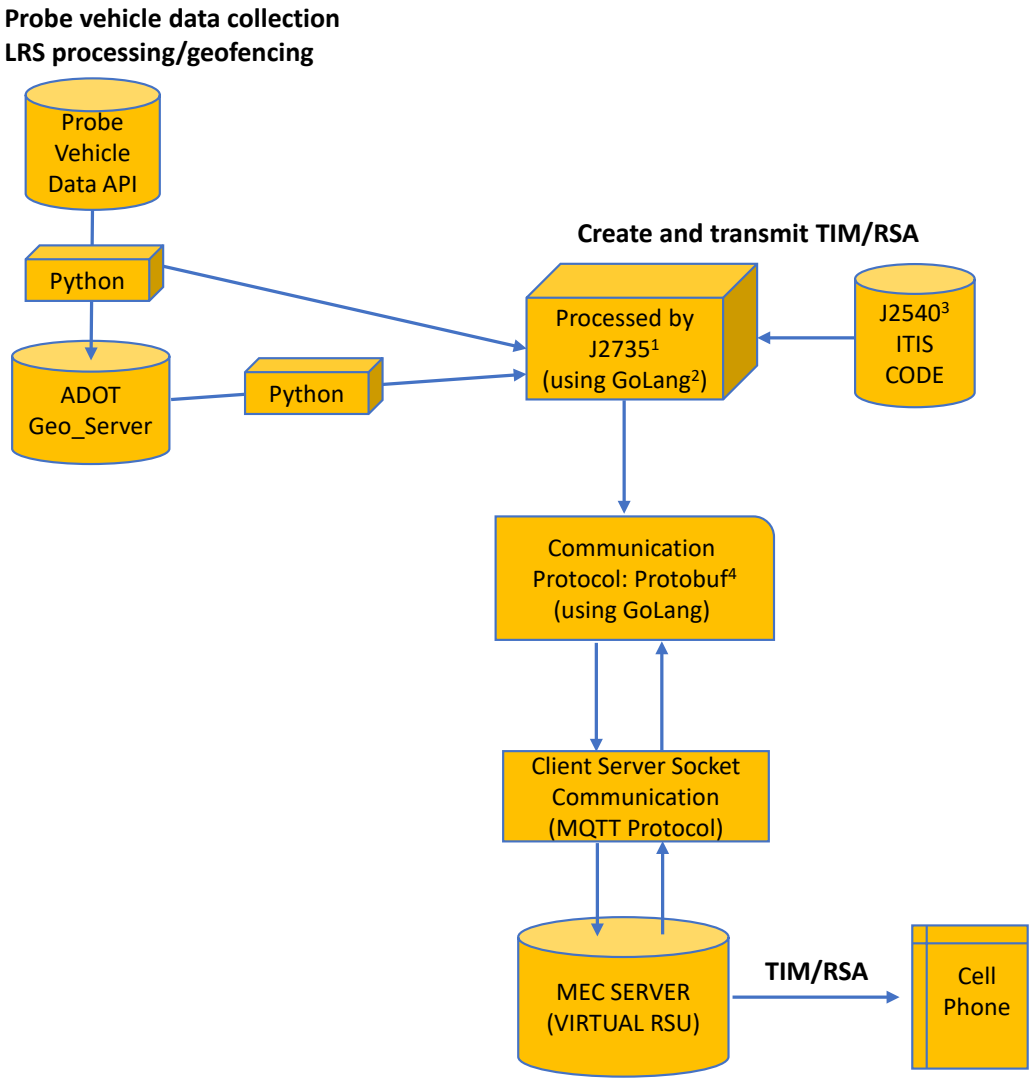


Figure 4. MAG-ADOT Back-of-Queue Pilot Data Pipeline

[1] SAE J2735 – This SAE Standard specifies a message set and its data frames and data elements for use by applications that require the use of wireless communications technologies for connected vehicles, https://www.sae.org/standards/content/j2540/2_202012/.

[2] GoLang is an open-source programming language designed by Google.

[3] SAE J2540 – This standard provides a table of textual messages meeting the requirements for expressing International Traveler Information Systems (ITIS) phrases commonly used in the ITS industry. The phrases are predominantly intended for use in the description of traffic-related events of interest to travelers and other traffic practitioners, https://www.standards.its.dot.gov/Factsheets/Factsheet/71.

[4] Protocol Buffers (Protobuf) – a free and open-source cross-platform data format used to serialize structured data, in this case, used to interface with the MQTT broker, https://protobuf.dev/.

The data are collected and processed by two AWS Lambda functions[3] in a parent-child format. The parent Lambda function collects the data and sends it to the child Lambda function for analysis, which then sends the processed data back to the parent Lambda function. The parent Lambda function then encodes the data in the appropriate format for and sends it to the MQTT broker via protobuf.

At the beginning of the NCHRP 20-44(39) project, the group was in the process of testing the feasibility of this process for a short freeway segment in Phoenix. Further, the pilot application was built using on-premises hardware (i.e., laptop computer). As such, if the STG were to confirm the pilot's feasibility and move the application into production, they would need to be in an environment that could scale. Scaling such an application to the Phoenix metro area, depending on market penetration, could require the ability to receive upwards of an estimated one million BSMs per minute. The STG was also concerned with performance, especially if the application needed to make calls to ADOT's geoserver, which hosts the linear referencing system (LRS), for the entire region during rush hour.

Description of NCHRP 20-44(39) Use Case/Implementation Project

The STG's use case/implementation project for NCHRP 20-44(39) was to build, test, and further scale the pilot back-of-queue warning system using the AWS AZDOT Development Account (i.e., the test environment established by the ITG). This was the first use case for a specific ADOT group (outside of the ITG) to have access to the AWS test environment. As such, the implementation project provided STG an opportunity to begin using/developing in the environment while leveraging the NCHRP 20-44(39) training/assistance, in conjunction with support from ADOT IT and AWS. The implementation project also provided the STG with the opportunity to test/explore various cloud services and tools and assess their corresponding budgets.

Instead of the BSM data from Verizon (as in the original pilot project), the implementation project relied on data from a probe vehicle speed data provider to identify queues forming in traffic, as well as ADOT linear referencing system (LRS) data to dynamically geofence locations one mile, two miles, and three miles upstream of the back of a queue. Additionally, ADOT decided that instead of using a Protobuf methodology they would write an API end-point so that various data consumers could perform further processing based on their specific use cases (see Figure 5).

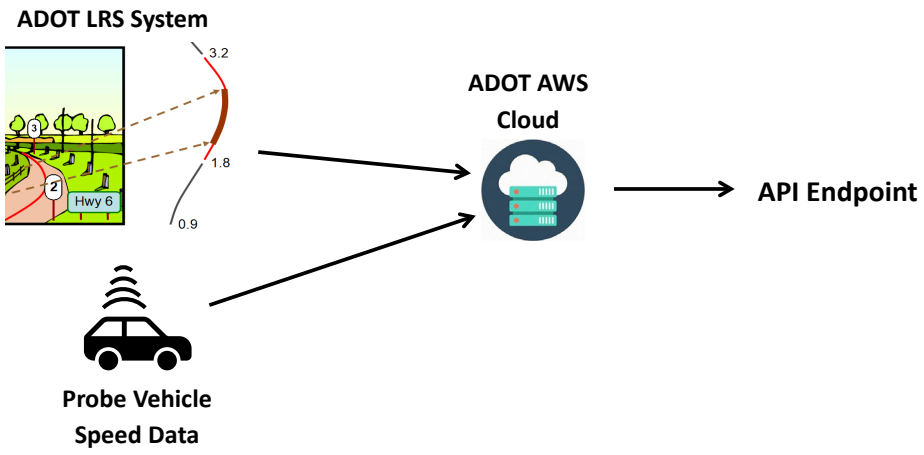


Figure 5. Back-of-Queue Implementation Project

[3] A serverless compute service that runs code in response to events and automatically manages the underlying compute resources.

Progress and Next Steps

The following bullets summarize ADOT's implementation progress for the back-of-queue traffic notification system in the ADOT cloud environment.

- Created a new virtual private cloud (VPC) in the US West using the ADOT ITG AWS account, including configuring subnets, route tables, and security groups based on ADOT ITG security requirements.
- Created an identity and access management (IAM) role specifically for the Lambda function for the back-of-queue traffic notification program.[4] This role gives permission to access the ADOT linear referencing system (LRS) and probe vehicle APIs and write the data to the AWS Simple Storage Service (S3) bucket, a resource for public cloud storage.
- Refactored the code that was working on a laptop environment in AWS Cloud9 using the same APIs. This step was more challenging due to differences in some of the Python package functions between the on-premises and the cloud environment.
- Stored this code on Github to ensure it would be available and make it easier to track changes as ADOT continues to evolve the code to cloud-based capabilities.
- Moved and tested the process using the Lambda function. Successfully ingested the LRS and probe vehicle APIs and accurately performed the calculations. The LRS data later became available in the AWS development environment, thus the Lambda code will need to be refactored to accommodate this change. Given the number of changes to the codebase, errors may exist. Hence, before moving from the pilot to a live product, ADOT will need to take on the important task of ensuring that the API endpoints are stable. Also, the code will need to be refactored taking into consideration that ADOT will not be writing to an MQTT broker but rather an API endpoint.

The next steps in the implementation will include the following:

- Once API endpoints are solidified and the code refactoring for calculating the geofences from Python into GoLang is completed, the AWS Lambda will need to be tested with live data to ensure it is working correctly. Thereafter, ADOT will monitor the Lambda function to ensure it operates within the required performance and capacity limits.
- Next, the S3 bucket will be created to store the output data from the Lambda function and the bucket policy will be configured to allow access only to authorized users and applications according to ADOT ITG security requirements.
- Once the data are saved into an S3 bucket, ADOT will use AWS Gateway to send the output data to the AZ511.com website vendor who will then write their endpoint access.
- ADOT will then configure the API Gateway to authenticate and authorize requests based on the ADOT ITG security requirements.
- ADOT will then test the scalability of the data pipeline to ensure it performs well.
- They will add appropriate logging and monitoring to the Lambda function and API Gateway to track usage and identify issues.
- ADOT will set up a Continuous Integration/Continuous Deployment (CI/CD) pipeline to automatically deploy updates to the Lambda function and API Gateway.

[4] AWS Lambda functions need permissions to interact with other AWS services and resources in the account. These permissions are set via an AWS IAM Role, which the serverless framework automatically creates for each service, and is shared by all functions in the service. The framework allows roles to be modified or function-specific roles to be created.

- To make the project easier to maintain and give greater visibility, ADOT plans to store the refactoring codebase on GitHub.
- Conduct regular security audits and penetration testing to ensure the application remains secure and adheres to ADOT ITG security requirements.
- Lastly, they will develop documentation for end users to ensure they can use the API effectively and securely.

The envisioned implementation data pipeline is shown in Figure 6.

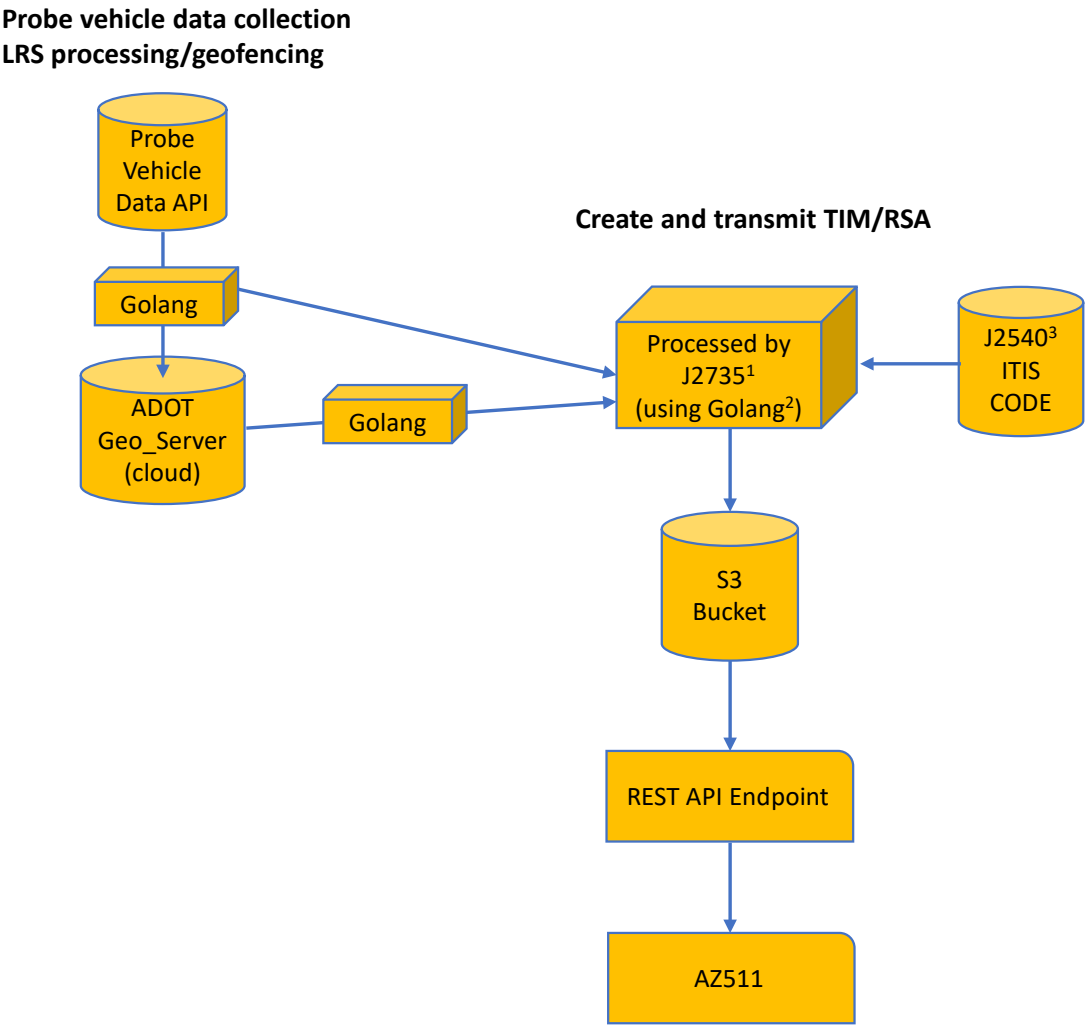


Figure 6. Implementation Data Pipeline

Accomplishments

Table 1 summarizes ADOT's accomplishments in the context of the Roadmap for Managing Data from Emerging Technologies.

Table 1. ADOT's Accomplishments Along the Roadmap to Managing Data from Emerging Technologies

	**Step 1** – Develop an understanding of big data	The state of Arizona had already recognized the need to embrace cloud technologies and had established a Cloud First Policy. ADOT's ITG was leading the charge to adopt this cloud-first approach. ADOT's STG participated in an NCHRP 20-44(39) training workshop, dedicated resources to learn more about cloud services, and participated in training with a specific cloud service provider.
	**Step 2** – Identify a use case and an associated pilot project	MAG and ADOT's STG (within TSMO) discussed creating a program that sends safety messages (based on connected vehicle BSMs and probe vehicle data) to the traveling public via cell phones. The team developed a minimum viable product (MVP) as part of a previous pilot project.
	**Step 3** – Secure buy-in from at least one person from leadership for the pilot project	ADOT's TSMO senior leadership supported the pilot project with MAG (MAG provided the funding, and ADOT provided the labor). Additionally, given Arizona's cloud-first approach, IT supported further development of the back-of-queue application in the cloud environment.
	**Step 4** – Establish an embryonic big data test environment/ playground	ADOT's ITG was already in the process of migrating and re-platforming its enterprise information and data system to the cloud in response to Arizona's Cloud First Policy. ITG's roadmap for establishing the cloud-based enterprise environment involved 28 projects over the following 3 years (2021-2024). The roadmap included decommissioning servers, moving from SQL to AWS Aurora, and building out a data lake. The enterprise cloud environment contained a development account, separate from the production network, which was intended as a sandbox to allow ADOT staff to conduct research and development and learn how to use the cloud. The ITG was the first group within ADOT, outside of IT, to have access to and begin working in the development environment.
	**Step 5** – Develop the big data project within the playground	To scale the pilot project and make the data available for various end-users, the STG successfully implemented the back-of-queue project in ADOT's cloud development environment.
	**Step 6** – Demonstrate the value of the data to other business units	Once the project is operationalized, it will be easy for other business units to see the value with dashboard displays. Other business units will also be able to easily integrate the data into their workflows, most notably the traffic operations center (TOC).
	**Step 7** – Demonstrate the value of the data to executive leadership	ADOT is actively working on this step. Upon completion of the pilot implementation in the cloud, the STG will present the project to the TSMO senior leadership to obtain approval to productionize the project.
	**Step 8** – Establish a formal data storage and management environment	ADOT's ITG is well on its way to establishing a cloud-based enterprise data and information system.

Challenges and Lessons Learned

ADOT identified the following challenges associated with the implementation project:

- **Accessing the ADOT AWS environment** – Getting access to the AWS environment was a lengthy process. The STG was the first group outside of IT to use the development environment, and it took time to go through the process.
- **Accessing cloud tools** – Getting access to some of the tools also took time (the environment was constrained). Had to go through the process of approving certain tools.
- **Refactoring code for the cloud** – Refactoring the original codebase (Windows) to make it work in AWS (Linux) was more challenging than expected.
- **Working in the cloud** – Learning a variety of new AWS modules and thinking more "distributively" was a change from the more traditional way of thinking about and executing projects.
- **Using production APIs in a cloud development environment** (due to security policies) - As some ADOT production APIs (e.g., LRS) were not available in the cloud, different approaches were needed (e.g., creating test data when production API was not made available within the test environment).
- **Meeting requirements of the Cloud First Policy** – ADOT had thousands of active Access databases. Some lacked documentation, which made it challenging to determine what data were useful.

ADOT reported that using "first principles thinking" and the NCHRP 952 Guidebook for Managing Data from Emerging Technologies for Transportation were helpful during the implementation. In addition, ADOT reported the following lessons learned:

- **There are many cloud tools available**:
    - Assessed the pros and cons of different AWS tools to deploy the project.
    - There are many tools in AWS, and documentation can only take you so far. ADOT found that some of the examples in the documentation were oversimplified, and that it would be helpful if more advanced examples were available (e.g., interfacing with other data sources). It is necessary to build the experience.
    - Gained exposure to a variety of AWS tools. The pilot staff now has the knowledge to implement other IT platforms and projects in AWS, which will reduce ramp-up time in the future.
- **There are different approaches to solving problems** – ADOT appreciated that there are different ways of solving problems. The key was to think in terms of making the project modular enough (to be cost efficient) while achieving the project goal and objectives – e.g., compartmentalized, broke things down, made codebase project so that other features/use cases/projects could be added later (the focus for implementation was just back-of-queue warning but the API endpoint can be used for other use cases).
- **Naming conventions are necessary** – Naming conventions are needed to locate and keep track of projects (which saves time and headaches down the line).
- **Documentation is extremely important** – There is a need to have a definition for every data variable in every dataset. Information gets lost as people move on, and the data are not as useful as they could be. It is important to know what you are working with.

Next Steps

One of the biggest lessons learned by ADOT in doing the implementation was that they do not want to deal with individual vendor/entity processes when developing apps like the back-of-queue warning system. Instead, ADOT intends to productize the data pipeline/process with an API endpoint that entities can ingest (from the AZ511 website) for their own needs. This approach will alleviate commercialization bias, the need to manage MQTT, etc., and will make it easier on the agency while still providing the data to those interested. As such, ADOT is working on an operational readiness document to describe what they need to focus on to get the data ready to be used by others. In other words, they have shifted the focus from what is needed to interface with a particular vendor to what they need to do as a state agency to provide the data to many users for use in different ways.

Other next steps include the following:

- Confirm the AWS data pipeline format is suitable for ADOT's needs. Change as needed.
- Work with the security group to ensure all API endpoints are secure.
- Perform stress testing on the AWS data pipeline to ensure durability under load.
- Find small ways to make the data pipeline quicker. As a pilot it is acceptable, but as data increase, how will it perform under certain conditions?
- Further modularize the codebase to use it for other traveler notification use cases.
- Gain more knowledge and exposure to AWS to help ADOT adopt the statewide Cloud First initiative.

CHAPTER 4. CASE STUDY ON THE DISTRICT DEPARTMENT OF TRANSPORTATION'S IMPLEMENTATION PROJECT

The overarching goal of DDOT's pilot project was to leverage an existing District of Columbia on-premises, big data environment for the storage and analysis of micromobility data using concepts from the NCHRP Research Report 952 guidance.

Background

District of Columbia Big Data Environment

The District of Columbia Office of the Chief Technology Officer (OCTO) developed an on-premises big data environment, "Data Lake," with the goal of bringing together DC government data and reducing agency use of various individual vendor products. The architecture diagram for the Data Lake is shown in Figure 7. The Data Lake is a Cloudera Data Platform (CDP) instance running on a cluster of on-premises bare-metal servers. Storage is a Hadoop File System (HDFS), managed through YARN, and integrated with Impala and Hive engines for SQL-like queries and Spark for distributed processing. JupyterHub, with Python and R kernels, is the primary analytics interface.

OCTO has encouraged and offered support to DDOT to adopt the Data Lake and to structure data and develop use cases within the environment. The DDOT Chief Information Officer (CIO) also expressed interest in developing a cloud strategy to overcome future scaling limitations of the on-premises system.

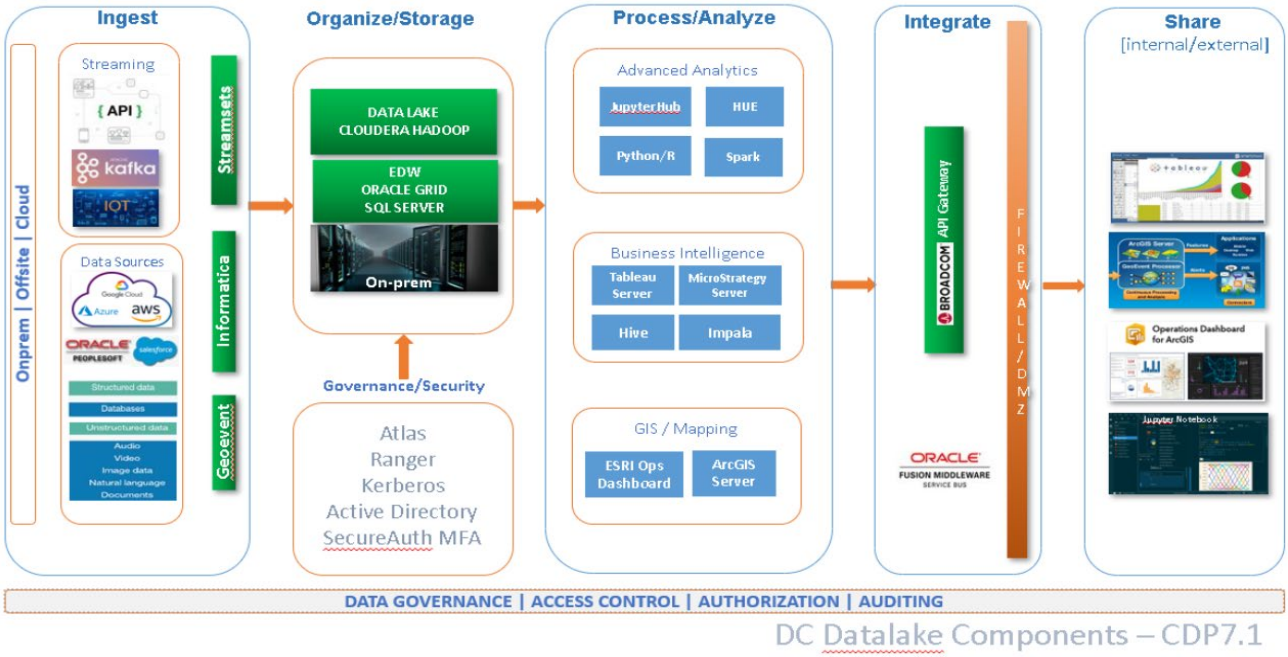


Figure 7. District of Columbia OCTO On-Premises Big Data Environment Architecture

DDOT's Focus on Micromobility

Through its long-range transportation plan, moveDC,[5] and its commitment to zero traffic fatalities (Vision Zero DC[6]), DDOT strongly encourages non-motorized mobility modes throughout the District. Vision Zero DC, moveDC, and other programs emphasize the safe and equitable use of the public right-of-way and support sustainable transportation within DC. However, single-occupancy vehicle (SOV) rides continue to be a major source of congestion within DC. Furthermore, vehicle collisions are a major source of injury and death within DC. In 2022, there were 19 pedestrian and 3 bicyclist fatalities, over 120 major injuries to pedestrians and bicyclists, an over 5,700 injuries in vehicle-related crashes.

Alternatives to SOV trips in the District include public transit (e.g., buses, Metrorail), walking, carpooling, ridesharing, and the use of personal or shared micromobility devices (e.g., bikes, scooters). Dockless scooters and bicycles are attractive alternatives to large vehicles for urban mobility. DDOT has an ongoing program to add protected bike lanes to the right-of-way, which can be used by riders of dockless mobility devices.

Dockless mobility, however, is the newest and least understood of alternative travel modes. Dockless mobility companies operating in the District are required to provide data to DDOT according to the Mobility Data Specification (MDS).[7] MDS standardizes communication and data-sharing between cities and private mobility providers, giving cities data they need to understand current and historic use patterns and the tools they need to improve the safety, equity, and quality of the mobility services on their streets (Open Mobility Foundation, 2022). These data can offer insights into the use and travel patterns of riders of dockless mobility devices.

The NCHRP 20-44(39) project provided DDOT an opportunity to use the OCTO Data Lake to store and analyze its dockless mobility data.

Description of NCHRP 20-44(30) Use Case/Implementation Project

DDOT's use case for the implementation was to understand current dockless mobility device use patterns, including the factors that drive/hinder adoption of this travel mode, that could help determine how to best encourage their increased use within DC. The primary stakeholders in the implementation project included the following:

- DDOT Micromobility Manager – As "business owner," the DDOT micromobility manager was the key stakeholder of the domain and primary consumer of the results. As such, the micromobility manager set the requirements for the project.
- DC OCTO – OCTO's role was to maintain the Data Lake, maintain system security, create and maintain the queries to the REST API of each dockless mobility service provider, and create and maintain the data pipeline to save the extracted data to the Data Lake.
- DDOT OCIO – OCIO's role was to interface with the DDOT micromobility manager to understand and support the requirements of the project, create all transformation and loading pipelines, and create all analytics pipelines.
- Micromobility Service Providers – The service providers' role was to provide data according to the MDS, maintain REST API endpoints, and grant authorization through private API keys.

[5] https://movedc-dcgis.hub.arcgis.com/
[6] https://visionzero.dc.gov/
[7] https://github.com/openmobilityfoundation/mobility-data-specification

The workplan and timeline for DDOT's implementation project are shown in Figure 8.

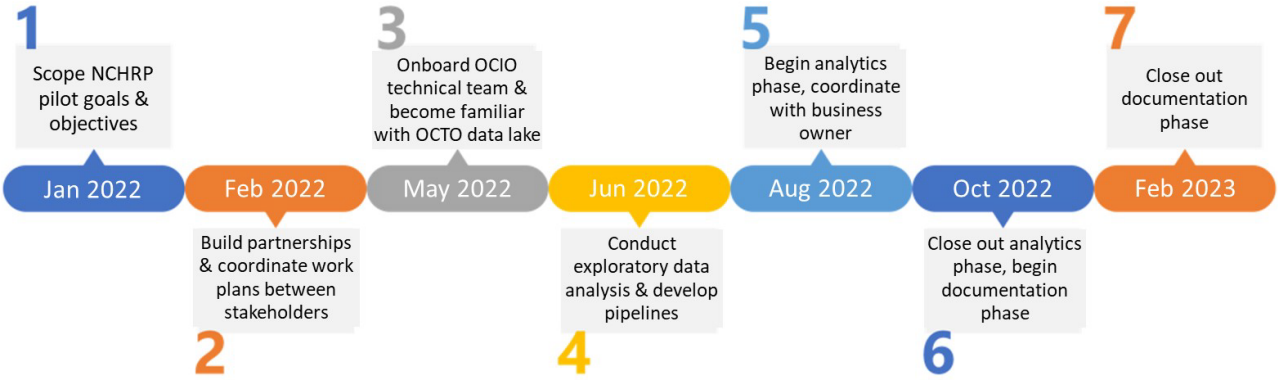


Figure 8. Implementation Work Plan and Timeline

Dockless Mobility Data Description and Considerations

All dockless mobility data used in the implementation project were pulled directly from the service providers in the District using MDS-compliant queries.

The dockless mobility data are in the form of individual status change events. Each event contains spatio-temporal coordinates for latitude, longitude, and time (x, y, t). Temporal precision is sub-second; location precision is nominally about one meter but is constrained by the urban canyon effect. Event types are numerous and include things such as agency pick-up/drop-off, trip-related events (start and end), low-battery notifications, and maintenance. Trip-related events contain paired unique trip identifiers, so the start and end events can be linked into a single origin-destination record. The team needed to consider multiple aspects of the micromobility data for the project:

- **Data volume** – Two years of DC micromobility data totals over 10 million trips.
- **Security** – The raw data do not contain user IDs or other personally identifiable information (PII); however, with enough repeated observations, individual behaviors could be extracted and potentially identifiable. Appropriate aggregation and deidentification are mandatory to safeguard individual information.
- **Business sensitive data** – There are multiple micromobility providers in the District that compete for the same customers. The set of all trips is highly business sensitive, as the raw data can reveal trade secrets. Appropriate aggregation and deidentification are mandatory to safeguard business sensitive information.
- **Missing data** – Analysis of how micromobility devices *are* used may not indicate why they *are not* used. In other words, where there is no/little use of micromobility devices, there are no/few data.

DDOT handled the data volume consideration by using the OCTO on-premises technology stack to store and process the dockless mobility data. The first step in data processing was to aggregate the data to remove service provider details. DDOT also did not use trip waypoint data (i.e., data between origin and destination), as the risk of reidentification was perceived to be too high; rather, the focus for this project was on the origin-destination behavior (when/where) without getting the specifics of trip routes. DDOT also conducted rigorous sensitivity analysis to determine appropriate levels of aggregation and spatiotemporal binning for trips. DDOT handled the missing data issue by associating regions together based on trends in usage.

Implementation Data Pipeline

The data pipeline for the dockless mobility data project is shown in Figure 9. The green objects represent OCTO's role of querying the service provider APIs and storing the extracted data in the OCTO data lake in partitioned (by date) Apache Parquet[8] files. The raw data stored in the Data Lake are immutable; the raw data can only be read (no write access).

The blue objects represent DDOT's role of processing the data for use in analysis and modelling. In the data pipeline, "bronze" (i.e., nearly raw, appended incrementally over time) Parquet files are extracted from the partitioned Parquet files in the OCTO Data Lake and transformed into "silver" (i.e., validated, enriched version of the data that can be trusted for downstream analytics) (e.g., join pick-up and drop-off in one row of the data) Parquet files for analytics.

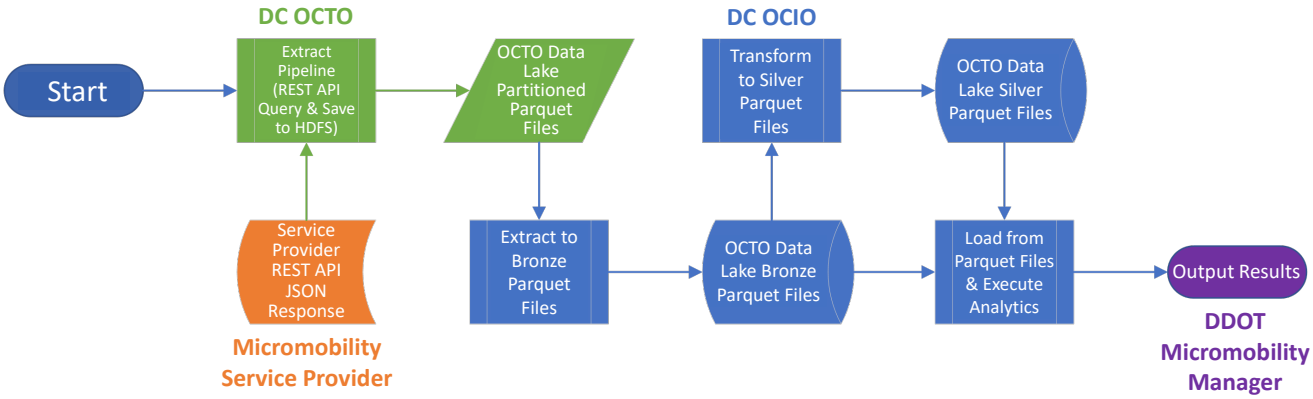


Figure 9. Micromobility Data Pipeline

Exploratory Data Analysis

As previously noted, the level at which to aggregate the dockless mobility data was a critical consideration for DDOT. A higher level of aggregation improves protections on the data but reduces the data available for analytics. As such, the team conducted an exploratory data analysis (EDA) to determine/estimate the appropriate aggregation parameters. Considerations in the EDA included the type of dockless mobility devices used and when, trip lengths in terms of distance and duration, and any time of day effects on trip distance and duration.

Figure 10 shows the number of rides by type of micromobility device. This figure shows that most dockless rides are on scooters.

Figure 11 shows the number of trips by trip duration and distance. The graphs show some oddities in the data (e.g., 0-meter trips, trips less than 5 seconds, spikes that may indicate rounding of trip durations to the nearest minute). The graphs also show that there are a lot of short trips (e.g., less than 1 minute, less than 100 meters). While these short trips may be interesting from a business perspective, they were not as relevant

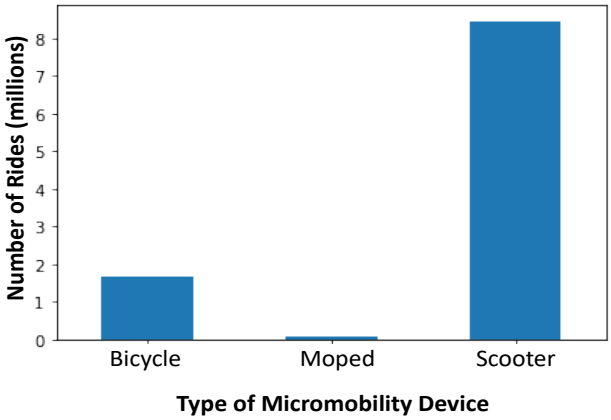


Figure 10. Number of Rides by Type of Micromobility Device

[8] Apache Parquet is a free and open source, column-oriented data file format in the Apache Hadoop ecosystem.

for this project, as they likely will not replace an SOV trip or impact how DDOT adapts the right-of-way for these vehicles. As such, these data were removed from further analyses.

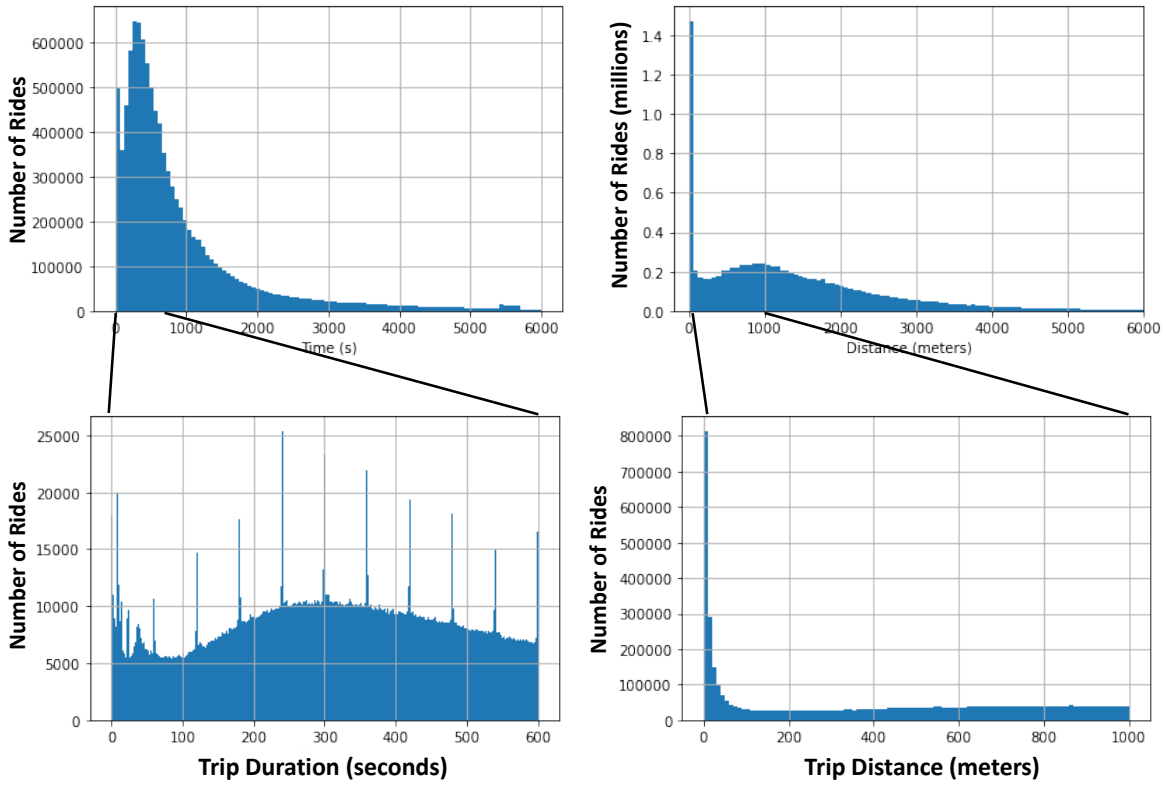


Figure 11. Number of Dockless Mobility Trips by Trip Duration and Distance

Figure 12 shows the trip duration versus distance for both dockless bicycles and scooters. The graph reflects that scooter speeds are restricted at 10 mph, and that bicycles can travel at higher speeds and are used more often for longer trips (statistically significant).

Figure 13 focuses on scooters and shows utilization by time of day and day/holiday. The graph on the left shows no clear differences in scooter trip distances and durations as function of time of day and that most scooter trips operate between 5 mph and 10 mph on average. The graph on the right shows scooter utilization profiles for a typical weekday (Wednesday), a typical weekend day (Saturday), and a holiday (4th of July). Scooter utilization on holidays is significantly different than both normal weekdays and weekends.

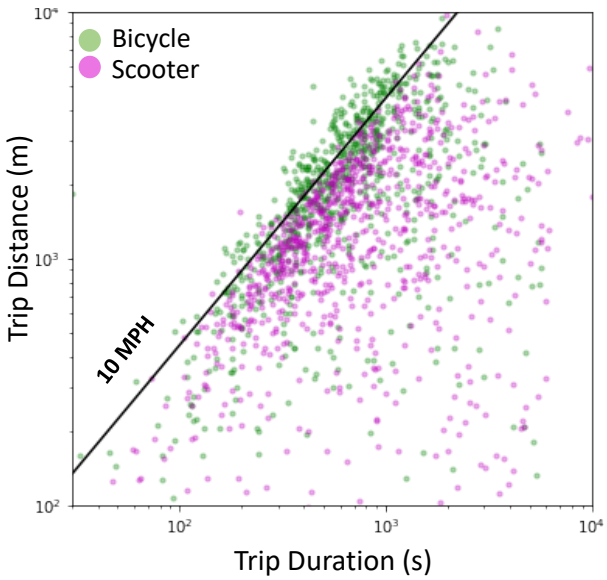


Figure 12. Dockless Utilization by Device Type

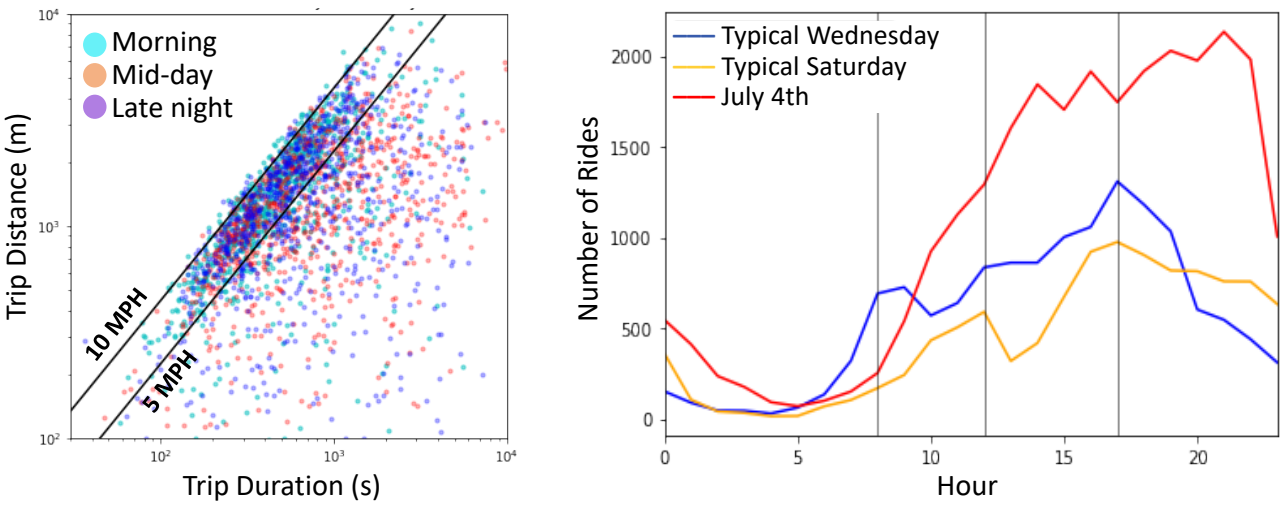


Figure 13. Scooter Utilization by Time of Day and Day/Holiday

Key takeaways from the EDA included:

- Most trips of interest are longer than 50 meters, so spatial aggregation of more than 50 meters is needed.
- Temporal aggregation at an hourly level is sufficient to capture intra-day and inter-day trends.
- Analysis must include indicators for day of week, holidays, and the type of dockless device used.

Figure 14 shows the process the DDOT team used to transform/process the raw data (bronze) into data for analytics (silver), which included joining the trip departure coordinates ($x_{1,x}$, $y_{1,x}$, $t_{1,x}$) with the trip arrival coordinates ($x_{2,x}$, $y_{2,x}$, $t_{2,x}$) through a unique ride ID ($XXXX_x$). For behavioral modeling and visualization the joined data were filtered to remove outliers and short trips, binned into a spatial grid (based on the findings from the EDA), and grouped into temporal bins. Slices of this data cube represent the density of rides that originate or terminate within each grid square in a certain period of time. The two density maps in the figure show the density of ride originations (bottom-middle) and density of ride terminations (bottom-right) on a typical weekday morning. There are three distinct behaviors highlighted by the arrows: (orange) is a mixed-use neighborhood with significant departures as people commute toward work but few arrivals; (red) is the Golden Triangle district, a commercial area that is mostly composed of offices that are the end points of commutes; and (cyan), near the Capitol South Metro station, which shows many ride arrivals as commuters use dockless vehicles as first-mile connectors to public transit as well as ride directly toward jobs in the United States Capitol complex.

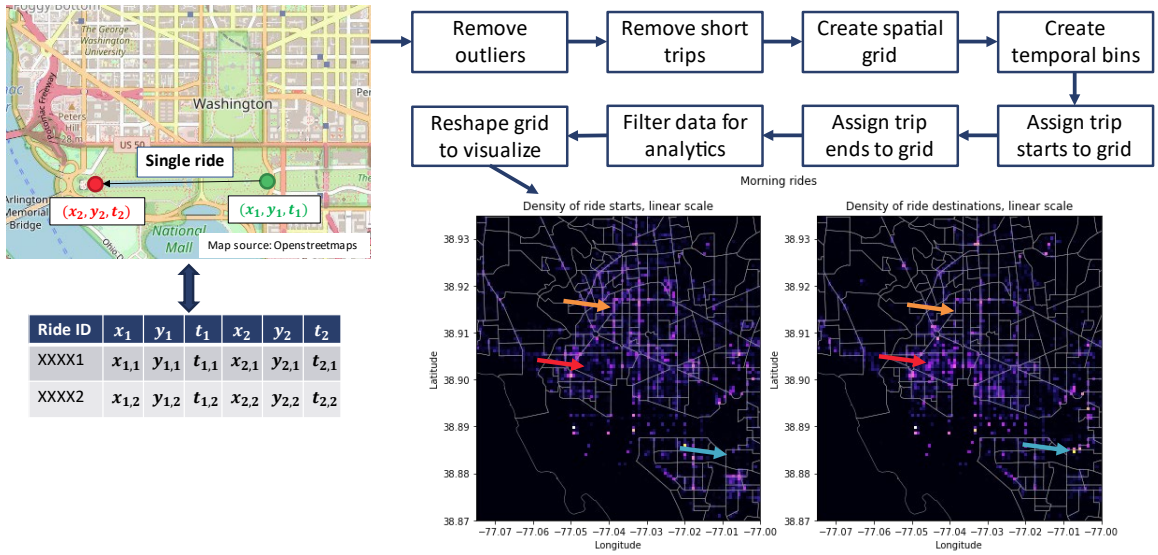


Figure 14. Data Processing for Analytics Process

Accomplishments

As the OCTO Data Lake was already established, with standard tools for big data analysis, the DDOT team was able to move quickly to Step 5 of the modern data management roadmap. The OCTO Data Lake included significant storage and computational resources. The environment easily handled the data volumes and analytics workloads associated with the dockless mobility data, and the OCTO team was knowledgeable and highly responsive to all questions and issues.

Additionally, DDOT now has a new public facing micromobility dashboard, called the Ride Report: https://public.ridereport.com/dc?vehicle=scooter&x=-77.0161804&y=38.8995737&z=12.95), which is being developed by a third-party and demonstrates potential collaboration opportunities with respect to data. A screenshot of the dashboard displaying the 2023 first quarter data is shown in (Figure 15). While there is not yet a publicly accessible version of the data, DDOT continues to explore if and how they might be able to provide the data.

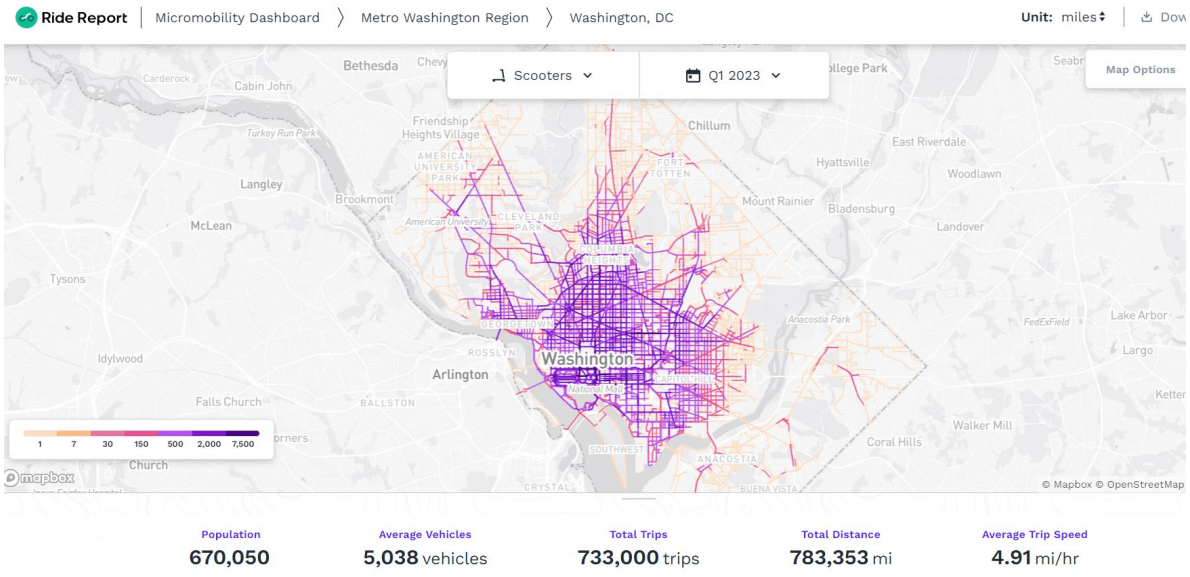


Figure 15. Screenshot of Ride Report Dashboard

Sensitivity Analysis

The team conducted a sensitivity analysis to determine what affect different levels of aggregation had on the data for analysis and to inform decision-making on how to handle the data. The team applied two techniques to aggregate the data: K-anonymization and spatial binning.

K-anonymization is a technique used to find a balance between data utility and user privacy (Herlocker, 2019). "For trip data that is sold commercially, it is common to see a K value of 5, meaning origin-destination (O-D) pairs will be removed if less than 5 trips were recorded. This technique provides the same level of protection regardless of the spatial resolution chosen for aggregation."

Figure 16 shows an example of the District scooter trip O-D pairs for two levels of aggregation (overlaid on a map of the major WMATA Metrorail lines to highlight the correlation between public transit nodes and dockless mobility usage). A level of aggregation of 5 rides per O-D pair (left) yields a well-populated map of common scooter routes, while a level of aggregation of 25 rides per O-D pair (right) highlights only the most popular routes.

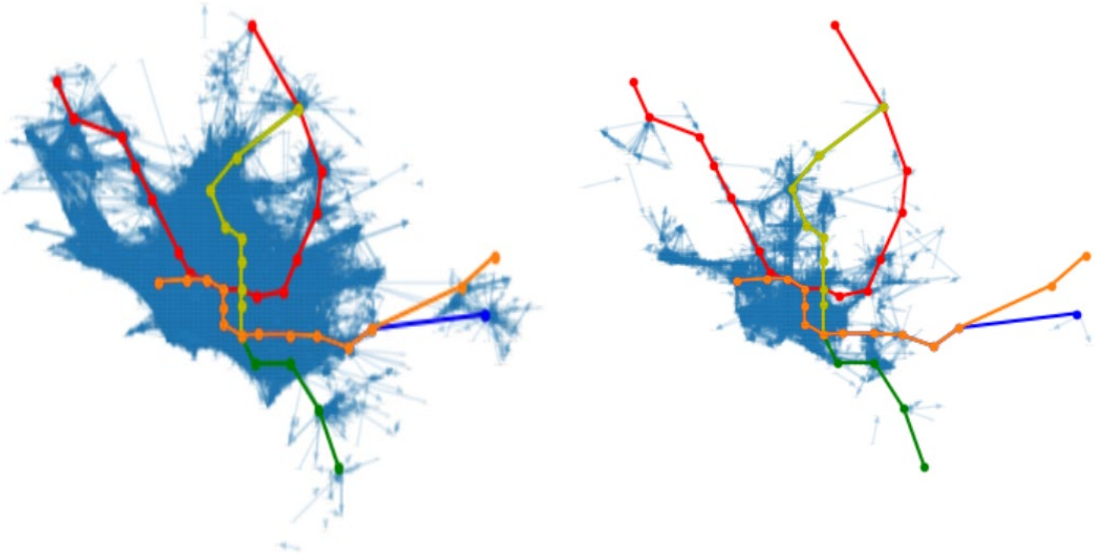


Figure 16. Comparison of Data at 5 Rides per Route (Left) Versus 25 rides per Route (Right) Overlaid on a Map of the Major WMATA Metro lines

The second technique used by the team for aggregation was spatial binning. A bin is a grid square on a map. A trip starts in one bin and ends in a different bin. Spatial binning only retains routes with a specified minimum number of bin pairs. As such, the larger the bins, the more trips will originate/terminate in each bin, thus more routes/ride data are retained for analysis.

As example is shown in Figure 17. The figure shows rides/trips made between two 300-meter by 300-meter bins (shown in green) versus 18 bins measuring 100 meters by 100 meters (shown in purple). With a minimum number of five trips to retain routes (i.e., ride data) for analysis, a route between bin 1 (left) and bin 2 (right) would be retained with the 300-meter by 300-meter bins. However, no routes would be retained for analysis with 100-meter by 100-meter bins.

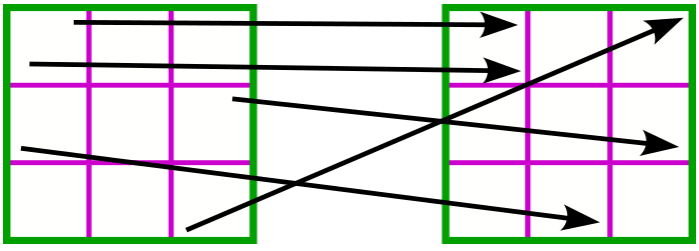


Figure 17. Trips Between a 300 X 300 Meter Zone Versus 100 X 100 Meter Zones

Figure 18 shows the results of combining both the K-anonymization and the spatial binning for two different specified minimum rides per route and four different bin sizes. The top graphic represents the results when selecting a minimum of 5 rides per route (more routes retained), and the bottom graphic represents the results when selecting a minimum of 25 rides per route (fewer routes retained). The colors on the charts and maps represent the bin sizes, with purple representing the largest bins (300 meters by 300 meters), red representing the smallest bins (50 meters by 50 meters), and blue and orange representing the medium sized bins (200 meters by 200 meters and 100 meters by 100 meters, respectively).

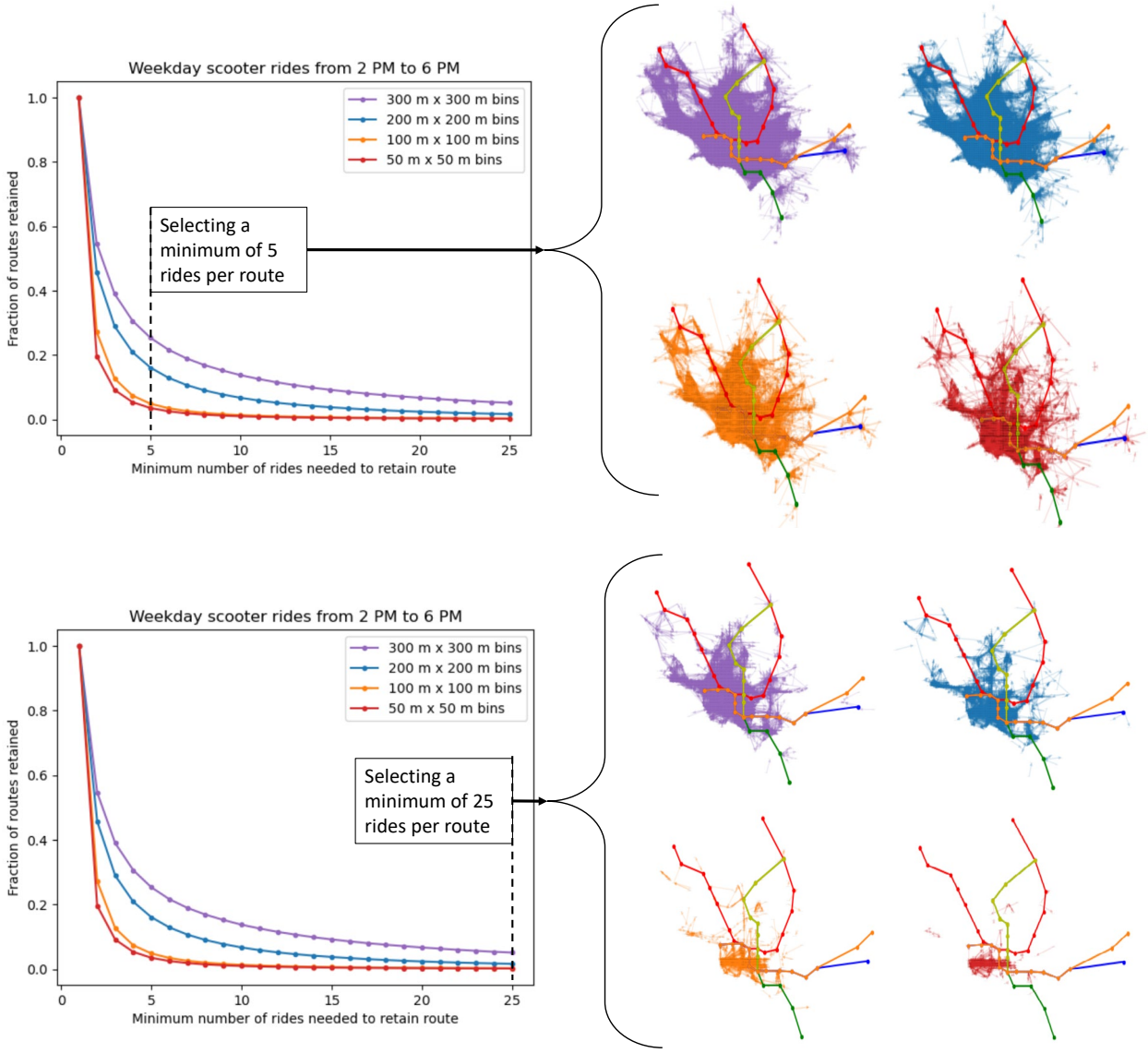


Figure 18. Route/Ride Data Retained Based on Level of K-Aggregation and Size of Spatial Bins

With larger bins and a smaller number of minimum rides, the most data are retained. As bin sizes decrease and the number of minimum rides increases, data are lost quickly due to aggregation, leaving fewer routes/data for analysis. While larger bins retain more routes/data for analysis, this aggregation comes at the cost of analytic fidelity. A 300-meter by 300-meter bin represents several city blocks, while a 10-meter by 100-meter bin represent a portion of a city block. As such, larger bins do not provide as much fidelity with respect to where the trips originated or terminated.

The maximum proportion of routes retained (with 300-meter by 300-meter bins and a minimum of 5 rides per route) is about 25 percent, while the lowest proportion of routes retained (with 50-meter by 50-meter bins and a minimum of 25 rides per route) is less than 0.1 percent. The latter represents the most well-defined corridors with the heaviest scooter use, which mostly coincide with the National Mall, Georgetown, Downtown, and Howard University areas of Washington, DC.

Opportunities and Next Steps

Figure 19 shows some initial results of behavioral similarity clustering (i.e., clustering grid cells together based on patterns of use to highlight common behaviors across the District). In this map, blue squares are generally centered on the National Mall, white squares are generally centered on residential areas, and green squares are generally centered on areas with business/government offices. The dark areas of the map lack data about how dockless mobility devices are used. The edges of the District are generally less dense; the bottom right and top right of map areas of the map are historically disadvantaged areas of DC. There are large blanks where the road and trail network is absent or closed to public use (e.g. large parks, bodies of water).

This analysis begins to explore the question on how to predict the utilization of a system that does not exist. The first step is to categorize the dockless mobility behaviors as a function of geography. The next step would be to add other datasets (e.g., census/demographics, zoning, Capital Bikeshare) to predict usage (what factors are linked to dockless mobility usage) and to create more suitable infrastructure to support non-SOV trips. In the next phase of this work, DDOT would like to better understand behavior and begin to fill in the dark areas on the map.

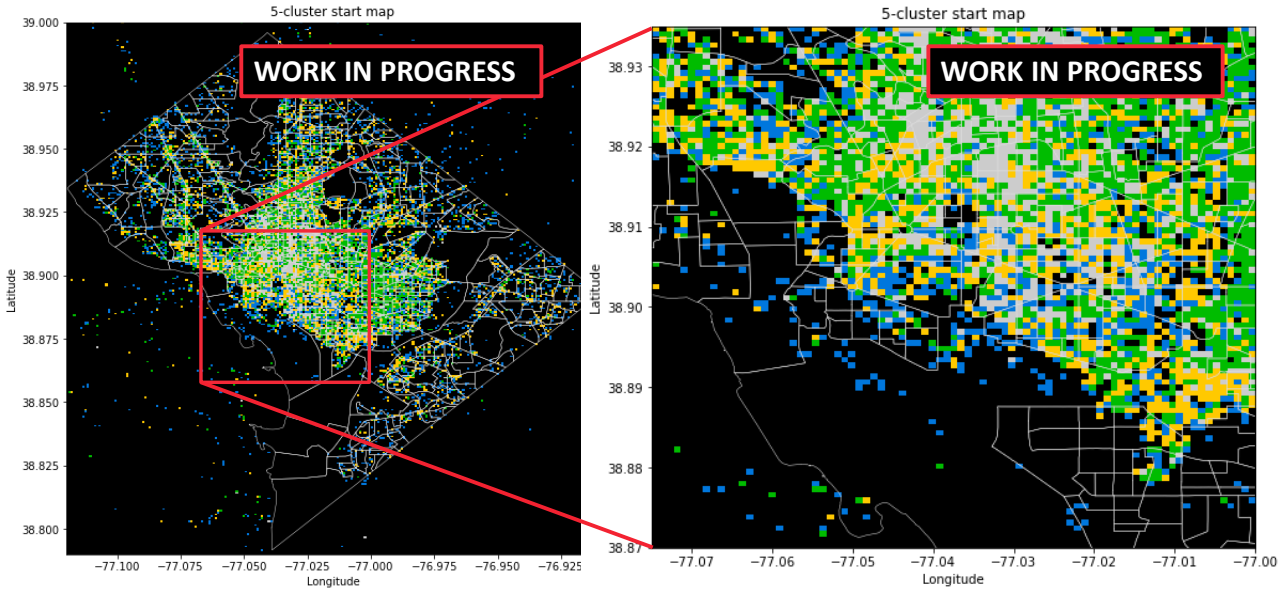
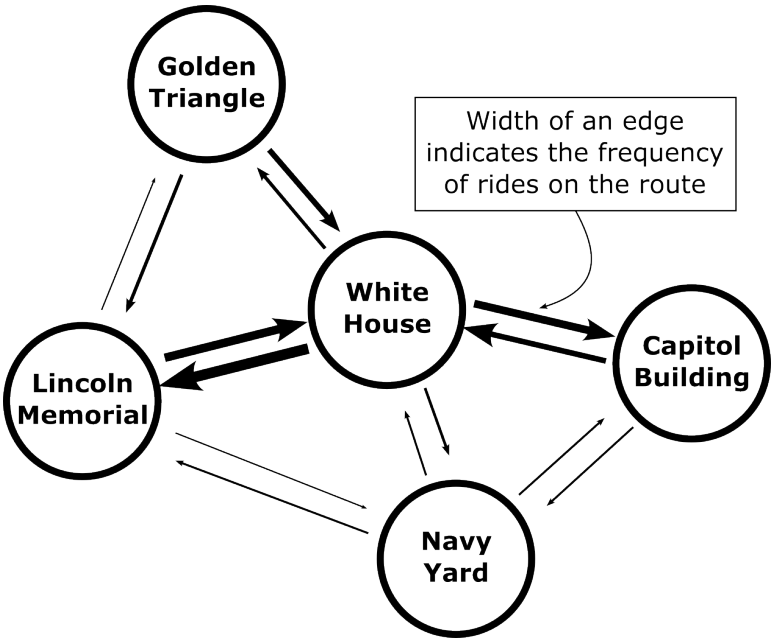


Figure 19. Behavioral Similarity Clustering

Additionally, beyond ride *volume*, DDOT would like to capture ride *direction* using directed networks, as shown in Figure 20. Networks *nodes* (departure and arrival locations) and *edges* (individual trips between nodes) contain metadata such as landmarks, distance, and duration. Well-established network and graph theory could be applied to provide information about where riders choose to go, which indirectly provides evidence about where riders choose *not* to go.

This step would go beyond looking at where dockless mobility devices are picked up/dropped off to including information on bidirectional travel to assist DDOT in how they might modify the public space/network and to encourage micromobility use in other areas.



Notional example of a directed network for landmarks within the District of Columbia, with edge weights (directional volume of rides) depicted as line widths. Not to scale. Weights are schematic, not real data.

Figure 20. Network and Directional Modeling

Challenges and Lessons Learned

While the OCTO data environment is a reliable analytics platform for pilot work, the team was presented with some challenges associated with the environment:

- The system functions more like an analytics sandbox/computing environment than a complete production cloud system (i.e., commercial cloud service providers). It's a powerful computational environment/resource designed for analytics, but it's not good for general cloud use:
    - New machines cannot be provisioned (e.g., if something needs to be done in a different computational environment).
- It is challenging to add additional applications natively into the data lake. It would be challenging to implement, integrate, and manage production pipelines with other systems (e.g., integrate with automated ETL to pass data back and forth between systems).
- A data catalog exists, but there it contains limited information (hard to know what data are there).

As such, as DDOT extends this data work further, they believe they will need to move beyond the OCTO data environment.

Additionally, DDOT's sensitivity analysis revealed how privacy concerns can be lessened by aggregating over broad spatial ranges or by only retaining routes that contain a minimum number of trips. Heavy aggregation strongly selects for trips that start and/or end near areas of high activity. Single trip routes cover the entirety of DC. Removing these routes reduces available trip data by over 70 percent. Minimum thresholds are likely 5-10 observations per reported cohort. DDOT continues to review best practices in anonymization.

CHAPTER 5. CASE STUDY ON FLORIDA DEPARTMENT OF TRANSPORTATION'S IMPLEMENTATION PROJECT

The overarching goal of FDOT's implementation project was to migrate an existing FDOT District 5, on-premises, big data environment to the cloud using concepts from the NCHRP Research Report 952 guidance.

Background

FDOT District 5 Data Fusion Environment (DFE)

The FDOT District 5 Data Fusion Environment (DFE) is an on-premises big data environment. The DFE architecture is shown in Figure 21. In 2016, the original use/test case for the environment was for centralized planning data and documents to be contained within a big data platform.

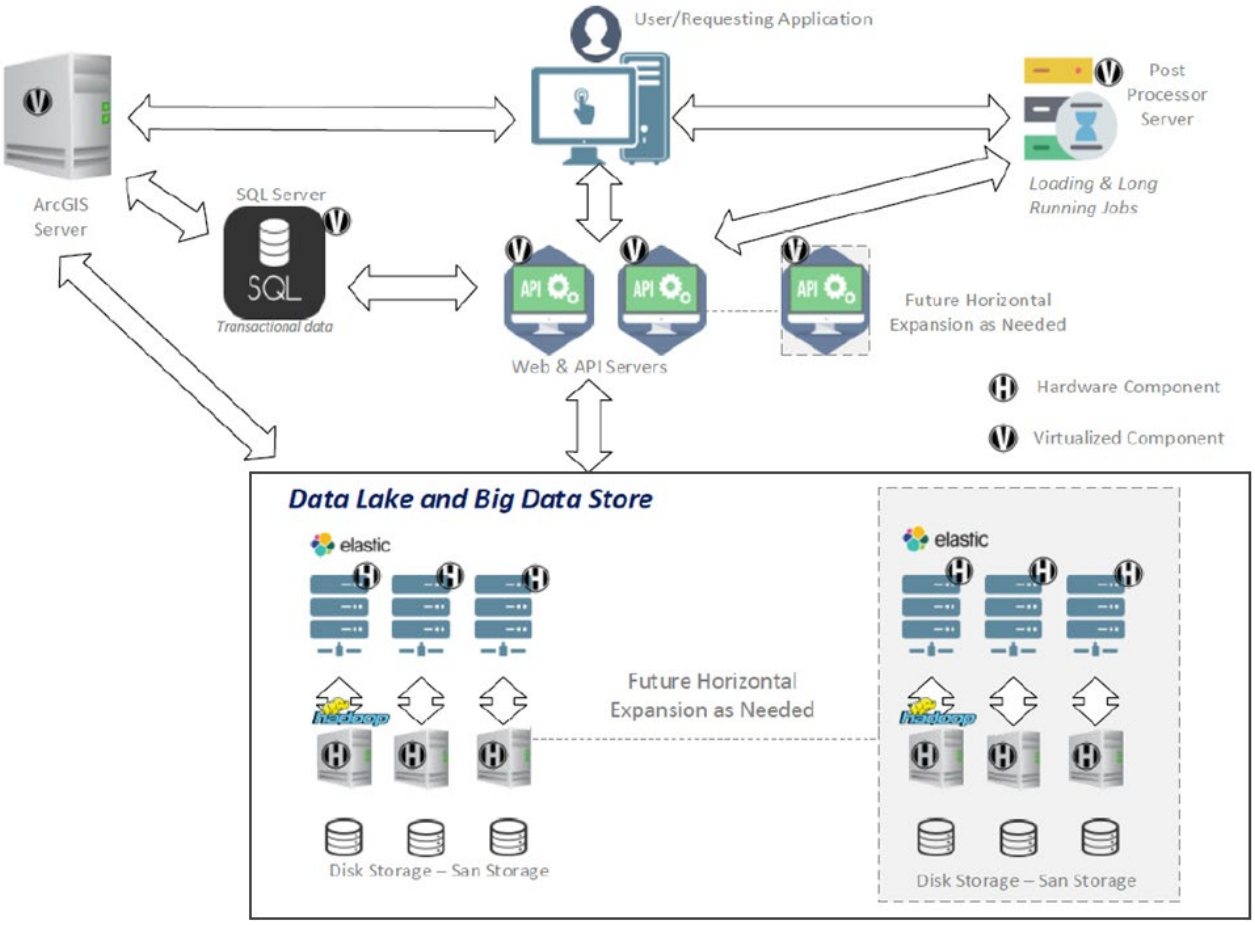


Figure 21. FDOT District 5 DFE Architecture

The DFE is comprised of a big data store running an Elasticsearch cluster, an ESRI server, a Microsoft SQL Server, and various servers hosting data processing, web APIs, and web applications. This environment allows the district to host and query large volumes of data that are consistently being indexed. The Elasticsearch cluster is built using 15 specialized server nodes, with each node having a particular job and working with the other nodes to manage the district's data. Data that are more static (not updated regularly) are indexed into the district's Microsoft SQL server. Data that are constantly updated are

indexed in the district's Elasticsearch cluster. The district's big data environment resides on a set of virtual machines (VM) hosted by the district. The data in the big data environment are housed with the district's SAN (storage area network) storage. The district delivers the needed data from the Elasticsearch cluster and Microsoft SQL server through various restful APIs. These APIs are built using NodeJS and the Express framework. A mix of coded .NET indexers running on a schedule, along with Windows services, serve as indexer services to insert data into the big data environment.

As the sheer amount of data has grown within the District 5 DFE, scaling the environment had become problematic. With the amount of data being indexed daily, the district could easily be required to run twice as many nodes than it was running. Also, the SAN storage space was limited, and the district was already forced to remove older and less used data to make room for new data to keep the system operational. The district realized that these limitations would become a major issue moving forward, as the system would not be able to operate normally with the anticipated growth in data.

Connected Vehicle Data Exchange Platform (CV DEP)

Additionally, FDOT is making investments in the cloud, with the Florida Department of Management Services preselecting approved cloud services. The FDOT Office of Information Technology (OIT) uses Microsoft Azure, and FDOT is using Amazon Web Services (AWS) to build the Vehicle-to-Everything (V2X) Data Exchange Platform (DEP). A key objective of this project is to standardize the collection, analysis, and sharing of data from several proprietary systems, all of which have different coding and encryption methodologies, and to make additional considerations for privacy and safety. The DEP will lay the foundation for FDOT to send alerts to drivers and traffic managers to coordinate routing, road closures, and emergency response.

Description of NCHRP 20-44(30) Use Case/Implementation Project

As a participant in the NCHRP 20-44(39) project, FDOT District 5 was looking to migrate the on-premises DFE system to the cloud. The cloud environment would provide not only the scalability needed, but it would also provide cost and performance gains (e.g., reduce maintenance and use more recent hardware) in comparison to running the system on physical servers. Additionally, District 5 had developed a speed/congestion dashboard, within the on-premises DFE, that used free navigation system data, center-to-center (C2C) data, and probe vehicle speed data, but the dashboard was inefficient due to the limitations of the DFE. By migrating to the cloud, the dashboard could be provided as a service for other districts to run (with specific interest from District 3). Having the environment in the cloud would also allow the districts to create more curated datasets so that information could be pulled more quickly from the raw data being collected. Finally, with changes since the District 5 DFE was originally designed/built (6+ years ago), migrating the system to the cloud would allow FDOT to take advantage of new/built-in services and tools that were not available at that time.

FDOT District 5 had to develop multiple plans to accomplish its objective for the implementation project. Their initial plan was to coordinate internally to utilize the CV DEP being developed in the cloud. For various reasons, including project schedule and institutional issues, District 5 needed to develop a "Plan B." This approach involved installing Apache Kafka[9] on-premises, implementing the existing probe vehicle data pipeline, and synching the data to a cloud service provider (see Figure 22). The pipeline is meant to be real-time with minimal latency in order to get the data quickly to where they are needed.

[9] An open-source, distributed event store and stream-processing platform, which provides a unified, high-throughput, low-latency platform for handling real-time data feeds, https://kafka.apache.org/.

District 5 worked with the DFE vendor to configure the data pipeline to enable future connections to the statewide cloud efforts. The data pipeline included the following steps:

- Step 1 – The probe vehicle data were downloaded every 5 minutes to the on-premises District 5 processing server.
- Step 2 – The District 5 processing server performed an extract-transform-load (ETL) on the probe vehicle data (taking "messy," raw data files and transforming them into a usable format). In this step, the ETL is done as the data come in, and "cleaned" JSON-formatted data are passed to the D5 Kafka Producer application.
- Step 3 – The Kafka Producer application took the probe vehicle JSON data and pushed/wrote the data into a Kafka Topic in the District 5 on-premises Kafka environment (which allows users to read the data).
- Step 4 - From the Kafka Topic, the data were automatically synched with the new District 5 cloud environment for storage (Amazon S3) and for API query functionality (MongoDB) for dashboarding. A topic can be considered an end-staging point for the data where they are ready for consumption. Users can read the data directly from the topic. Or in the case of District 5, users with access to the cloud can access the data in MongoDB for query and dashboarding.

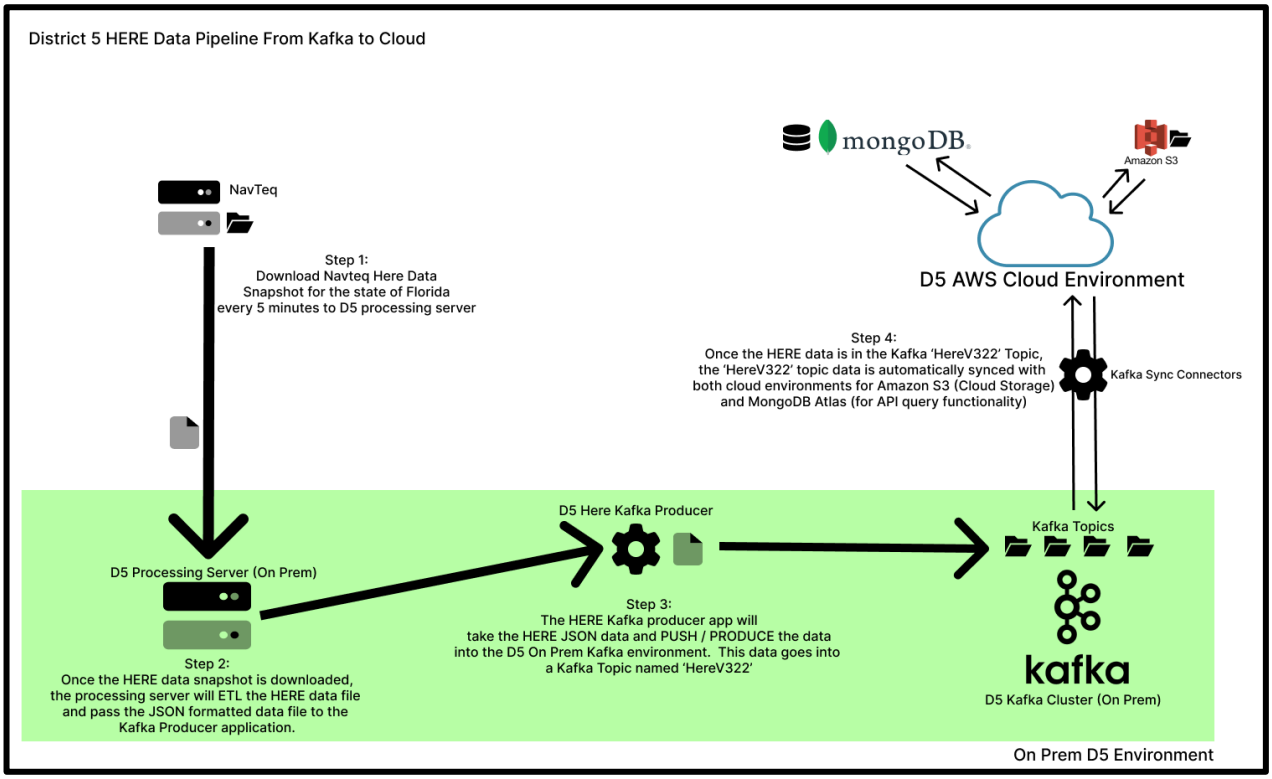


Figure 22. Data Pipeline – Plan B Approach

Figure 23 shows what the probe vehicle data look like following the ETL process when they are made available in the Kafka topic to query.

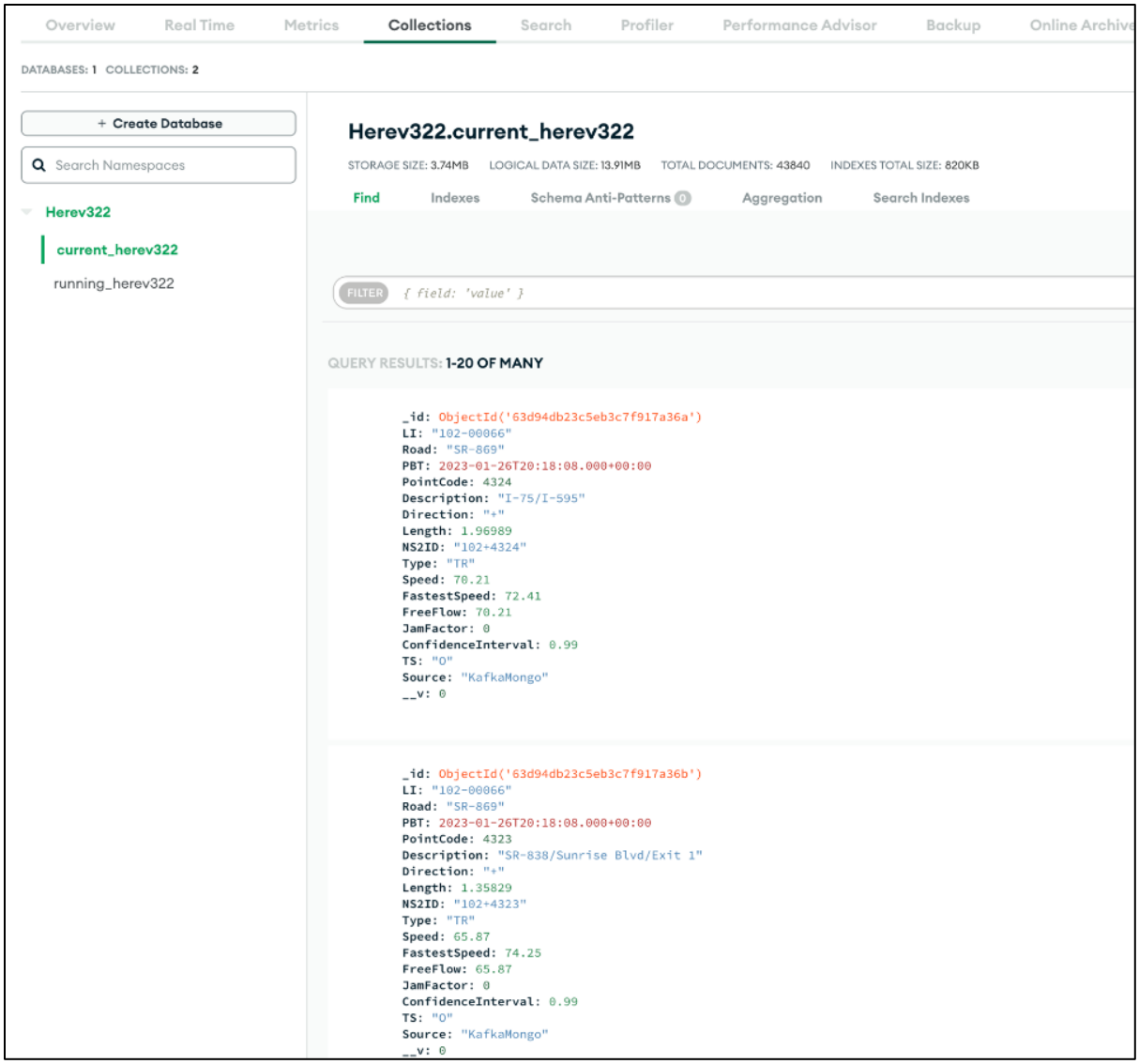


Figure 23. "Cleaned" Data in Kafka Topic

While D5 set up Kafka and built the data pipeline for the probe vehicle data, the workflow is the same for any dataset/API and provides District 5 with the workflow mechanism to make any dataset in the future available via Kafka topics. This process greatly reduces the number of steps required to deal with these datasets. This was an "aha" moment for District 5. Rather than multiple users hooking up to the API and doing the ETL themselves, this is a pipe that is ready to use; users just need to connect to it.

While FDOT District 5 expected plan B to move forward, they realized there was a better approach to migration. Plan B involved taking probe vehicle data natively in the cloud, bringing them down locally, then pushing them back to the cloud. As such, "Plan C" involved moving the ETL work that was being done on-premises to the cloud and running the pipeline as a service (i.e., picking-up data, running ETL, dumping transformed data directly into Mongo, making it immediately available for query). This process bypasses the District 5 on-premises pipelines and ingests the probe vehicle data directly into MongoDB. The D5 on-premises pipelines are left in place to process data that are needed for their on-premises solutions and the Kafka environment. The data pipeline for the Plan C approach is shown in Figure 24.

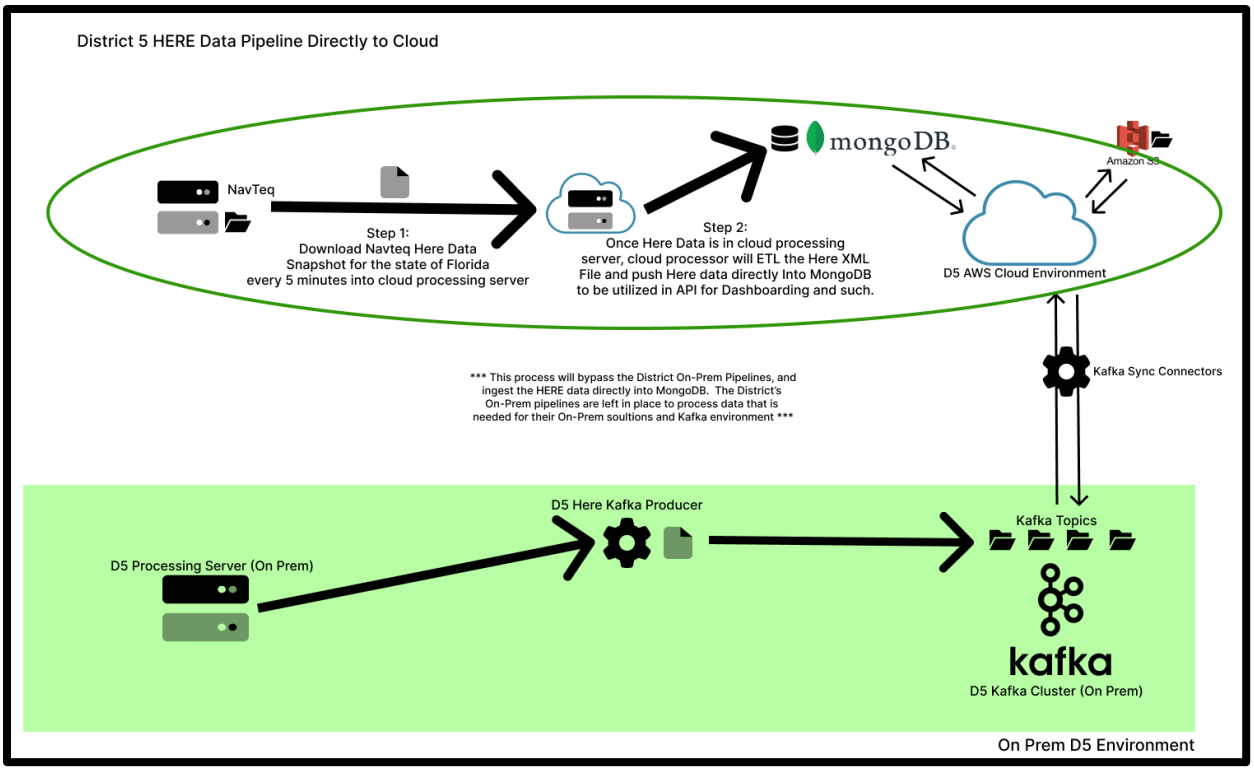


Figure 24. Data Pipeline – Plan C Approach

Accomplishments

Before the NCHRP 20-44(30) project, FDOT District 5 had an on-premises, district-specific big data environment and an operations application that looked for unusual traffic patterns in the district. During the NCHRP implementation project, District 5 moved the data pipeline and application to a cloud environment, and the application is now available statewide across the internet for other districts to use. FDOT District 3 wanted to use the probe vehicle data for operations but did not have the infrastructure to manage the data. District 3, which also participating in the training workshop, is now able to use the application developed and made accessible by District 5. Additionally, FDOT District 5 has seen a significant reduction in the duplication of effort, doing the ETL on the probe vehicle data once and providing a single interface of consistent topics for others to tap into rather than data users tapping different sources and decoding/cleaning the data for every project.

The implementation project has also simplified processes from an IT standpoint. Rather than creating firewall openings to pipe various datasets to various third parties/apps one by one, and in the format they want the data, data consumers can all get the data from one source that has been secured.

The FDOT District 5 implementation pilot demonstrated the success of moving an existing product to the cloud – storing the data, accessing the data, creating dashboards, and making the data and application available statewide – that can be tied into FDOT's broader efforts to move to the cloud.

Challenges and Lessons Learned

While FDOT District 5 initially felt the implementation project would run smoothly, they did experience the following challenges, most of which were institutional in nature:

- **Engaging other districts** – There was an initial concern about the adoption of the cloud environment due to hurricane resiliency concerns for critical need functions/recoverability (e.g., when the internet connection goes down, FDOT private systems continue to function).
- **Shared resources mean overlapping interaction with projects** – After months of trying to get different contractors/developers to work in a shared space, FDOT set up a parallel structure (i.e., two cloud environments) using the same technologies so that they can be merged in the future.
- **Guarding the data/access to the data** – While FDOT has taken a step forward, sharing data is still a work in progress.
- **Multiple clouds** - As a state, FDOT uses Azure for everything on the IT side of the house but uses AWS for operations. At some point, this may be a conflict.

FDOT District 5 reported the following primary lessons learned during the implementation project:

- Experienced an “aha” moment when they realized that the pipeline built for the probe vehicle data could be used for other data sources.
- When ingesting the probe vehicle data, District 5 initially planned to parse the data down to only those for the district; however, they determined it was faster and easier to pull in all the data for the entire state of Florida. Because the segments are constantly changing and being updated, all the data/updates are needed so as not to inadvertently omit district segments. As district roads and traffic do not stop at county lines, attempts to break down the data by district throw off the numbers. For example, the highway between Orlando and Tampa experiences a lot of congestion, which crosses a district boundary. Problems occur in District 1, but the traffic backs up into District 5. District 5 reported that they did themselves a favor by including all the data so that operators could look across district lines.
- Consultants and industry are ready for the pivot to the cloud (FDOT is approximately 10 percent state employees, 90 percent contractors). Most consultants are already using the cloud, so their teams know the technology and are ready to be offered from the "handcuffing" of being on-premises (where systems cannot expand fast enough and are often no longer properly supported).
- They did not experience as much concern about data security as they imagined. The data pipeline is not inside of the DOT network, and with proper disaster recovery and backup, IT was not as heavily involved.
- Doing the ETL once up front in the pipeline removes licensing issues and PII and helps address Sunshine laws. And it provides one consistent source of the data (via Kafka topics and the cloud), which reduced redundant efforts to clean the data for every project.
- Moving into Kafka was incredibly efficient. It reduced the number of APIs that FDOT had to stand up and hit individually.
- While FDOT District 5 expected Plan B to move forward, they realized there was a better approach and shifted accordingly by moving the ETL from on-premises to the cloud and running it as a service in AWS (rather than bringing down data natively in the cloud, processing on-premises, and pushing back up to the cloud).
- When taking parts of the data pipeline to the cloud, the data are no longer controlled at the operations center. This introduces a potential point of failure for operations. District 5 needed

to understand where in the pipeline they could put data out to the cloud to be used/useful versus where they needed to have data readily at hand. They realized they needed to map out the data flows for critical datasets and the associated business processes to understand where and when these processes need certain datasets to operate effectively, with the goal of minimal latency.

Opportunities and Next Steps

Opportunities and next steps include the following:

- Apply the data pipeline to other datasets such as the Intelligent Transportation Systems Integrated Quality and Analysis (ITSIQA), and the Automated Traffic Signal Performance Measures (ATSPM).
- The outcomes of this implementation project will improve the efficiency of the university research process moving forward. The old process involved FDOT giving data to a university on an external storage device or via email. The university developed an application (e.g., crash prediction/detection, computer vision) in the cloud, and FDOT got the application code from the university and refactored the code with support from the university (because the application was developed in a different environment), and then operations requesting enhancement after the first use of the application. The new process will give universities access to the cloud to develop applications, allow for FDOT testing and feedback during the process, and have the project being accepted deployed directly from the cloud. FDOT is working on an approach to handle payments (i.e., cloud costs) so that there is an equitable risk share between FDOT and universities/vendors.

CHAPTER 6. CASE STUDY ON RHODE ISLAND DEPARTMENT OF TRANSPORTATION'S IMPLEMENTATION PROJECT

The overarching goal of RIDOT's pilot project was to develop a cloud-based data environment in which to migrate the enterprise GIS system and build a data lake to support TMC's day-to-day operations.

Background

RIDOT's data have almost exclusively been stored on-premises on siloed servers, and before the NCHRP 20-44(39) project, little had been done to move to the cloud despite new types of big data becoming available. For example, staff at the RIDOT Transportation Management Center (TMC) wanted to get the highest value from emerging crowdsourced and other big datasets to improve situational awareness and ultimately traveler safety and travel reliability. While RIDOT ingests and stores some free navigation data, it struggles with how to effectively make use of the probe vehicle data purchased by the agency. These probe vehicle data are hosted by the University of Maryland CATT Lab as part of RITIS but have only been used on a minimal or trial basis for real-time benefit by/to/for the TMC. As such, the RIDOT TMC was not using the probe vehicle data in any formal way, and although the free navigation data source had been used successfully in a traveler information system as well as a project monitoring application, RIDOT felt confident that both data sources could be used to provide additional and up-to-date valuable insights and improved situational awareness, at least at the TMC level of operations. At the same time, RIDOT also noted that they desired assistance with the backend management of the enterprise geographic information system (GIS). They had not been able to support or provide on-premises equipment in the same manner or with as much capacity or resources as they had in the past and faced issues with upgrading systems as well. Moving to the cloud was increasingly enticing as a means to help alleviate issues with the backend management of GIS.

Description of NCHRP 20-44(30) Use Case/Implementation Project

RIDOT's implementation project was two-fold. First, the implementation involved migrating RIDOT's enterprise GIS to the cloud. RIDOT's intent was to develop a plan that reflected a new way of thinking about data management (e.g., conduct an inventory of traditional and emerging data sources, establish data retention policies, move to the cloud, and leverage cloud services), and follow through with, iterate, and learn from the implementation of the plan. Second, the implementation involved establishing a means and method for RIDOT TMC operators and staff to expand the use of free navigation and probe vehicle data to help primarily with real-time incident and roadway management, but also with planning and performance management for improved transportation systems management and operations (TSMO). As such, RIDOT sought to leverage cloud services and tools to ingest, store, enrich, process, and visualize these big data sources, overcoming limitations of being on-premises and gaining flexibility and scalability. Additionally, this work would help to establish the business case for RIDOT to continue to purchase probe vehicle data in the future. The GIS group and the TMC were to be used as test cases to demonstrate early wins from the use of cloud services and a data lake environment.

RIDOT's existing GIS on-premises production environment is illustrated in Figure 25, and RIDOT's proposed cloud-based GIS environment is illustrated in Figure 26.

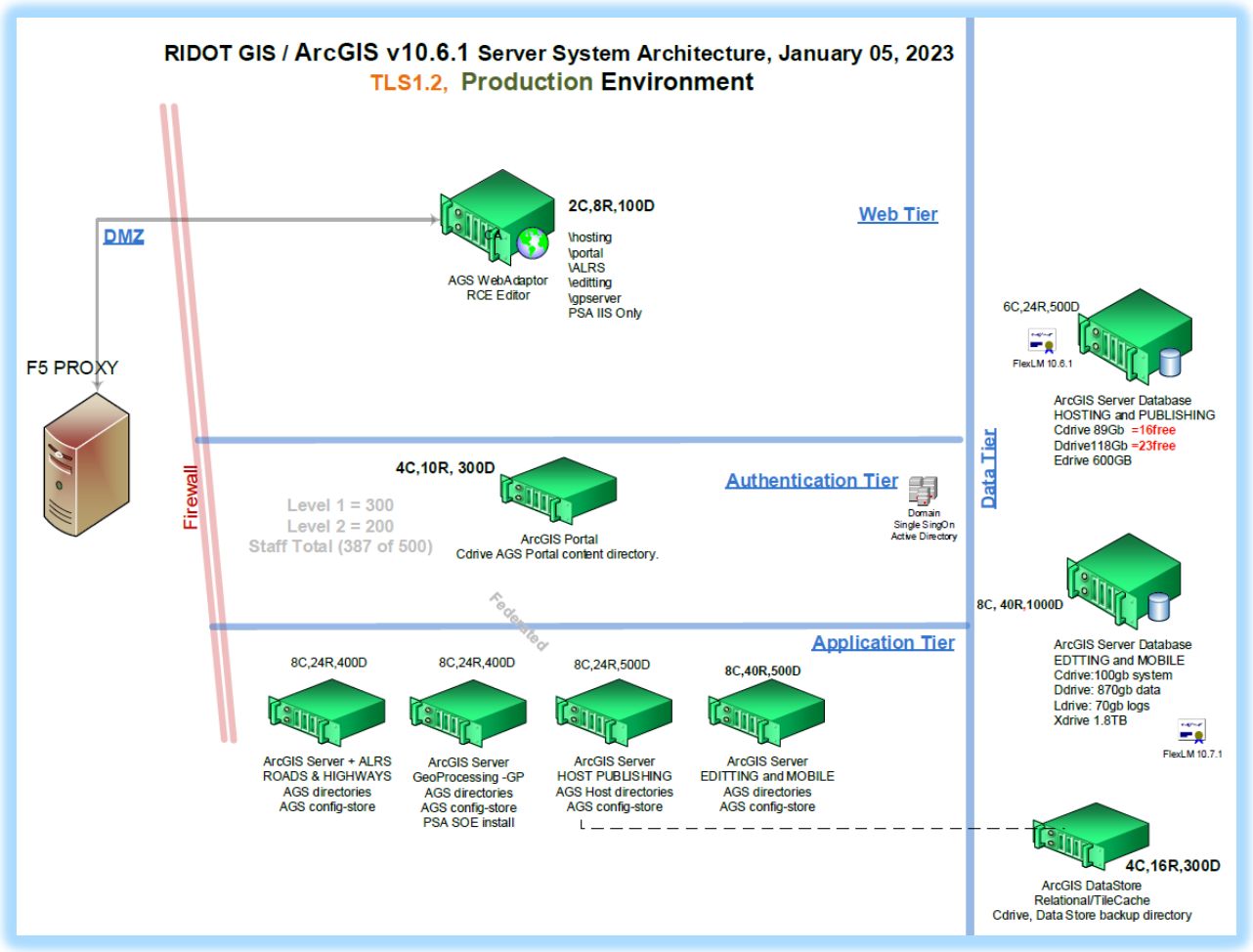


Figure 25. RIDOT's Existing GIS Production Environment

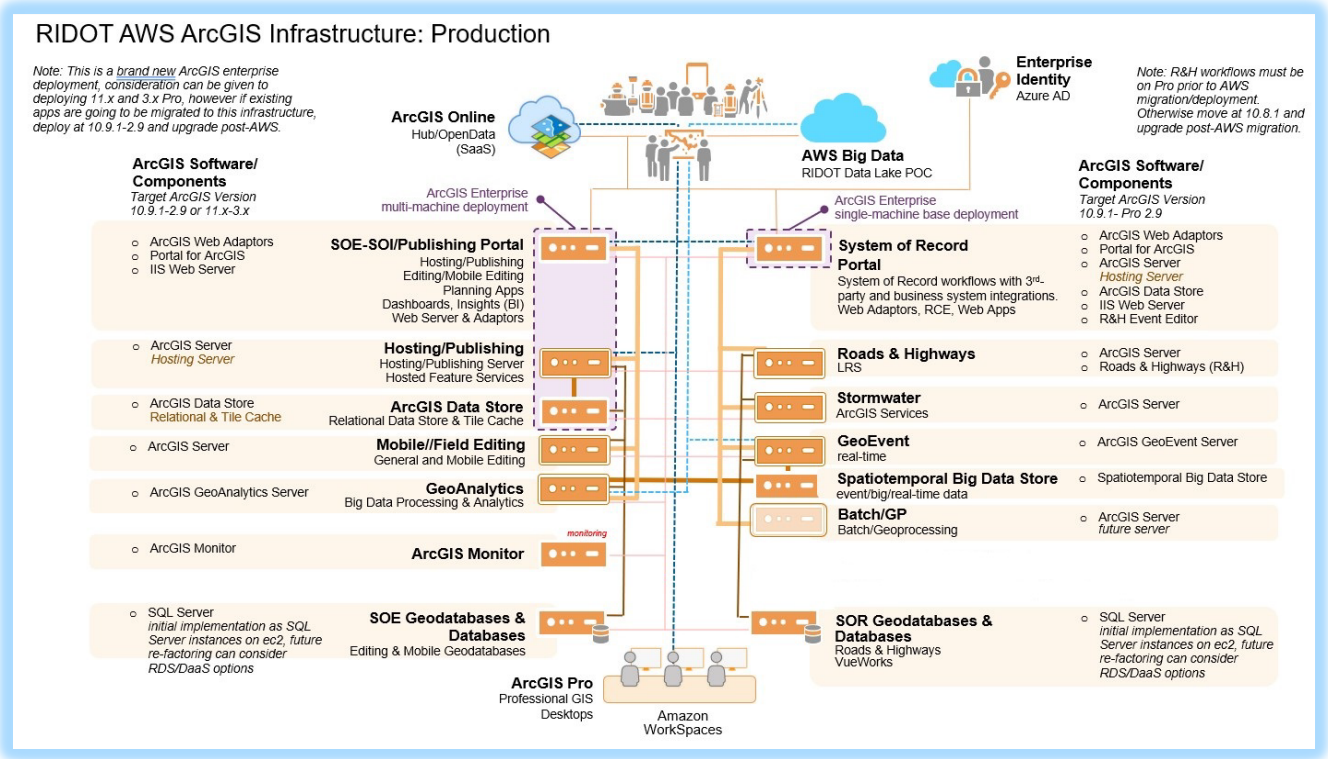


Figure 26. RIDOT's Proposed GIS Environment

Accomplishments

By participating in the NCHRP 20-44(39) project, RIDOT completed Steps 1-3 on the Roadmap to Managing Data from Emerging Technologies. They received approval to move the enterprise GIS to the cloud and to build a cloud data lake. They secured approximately $1,000,000 in funding for cloud services (to include cloud hosting and enterprise licensing for GIS) for three years, which is a large amount for RIDOT.

At the end of the NCHRP 20-44(39) project, RIDOT was working on Step 4, preparing to establish the proposed cloud/environment and data lake. To accomplish this step, RIDOT developed a scope of work (SOW) for cloud migration of the enterprise GIS, development of the big data environment/data lake, and hosting/managing the infrastructure, to include transparency associated with usage, billing, and performance. RIDOT solicited proposals through a Master Price Agreement (MPA) in December of 2022; however, they did not receive a response that met their needs. RIDOT is now preparing an RFP with an expected award in October 2023.

Challenges and Lessons Learned

While RIDOT made a good deal of progress during the NCHRP 20-44(39) project, this progress did not come without challenges. A big hurdle was getting buy-in from the Rhode Island Division of Information Technology (DoIT) to move to the cloud. There was no existing formalized cloud transition plan in place at the time of the RIDOT GIS Office's request to pursue this initiative. Additionally, the governance process was not modernized for the cloud. The RIDOT implementation team had to make a good case for their projects. This turned out to be a much longer process than expected.

Before submitting a formal project proposal, the RIDOT implementation team engaged DoIT to start a dialogue and get guidance on how to move forward. Initially, DoIT was not fully open to pursuing the project, as they were concerned about the costs of the cloud being too high, and they wanted to protect RIDOT from incurring costs they did not need to incur. The RIDOT implementation team negotiated with DoIT regarding how they could approach this project as a pilot program only and let DoIT know that RIDOT was prepared to incur costs as part of the pilot in order to learn.

Working with the information from the workshop and guidebook, as well as with the GIS vendor, the RIDOT implementation team created an illustrative example (covering pain points or problems), which was well documented and diagramed, and the team created a one-pager to explain visually some technical concepts (Figure 27).

Once DoIT realized the size of the data and that the existing architecture was not set up in a way to provide ideal functionality, they became more open to the ideas proposed by RIDOT. Several weeks later, DoIT asked RIDOT to re-present the proposal. RIDOT addressed DoIT's questions on both efforts. They had several positive conversations and meetings to share with and get DoIT to understand the "what," "why," and "how." In between some meetings, limits on DoIT resources seemed to result in periods of delay and/or lack of input, all resulting in a slowdown of progress. This process took multiple iterations, as DoIT needed to assess the impact of the proposed projects.

The proposal was eventually reviewed with chief architects at the state. It also made its way up to the Chief Information Officer (CIO), got reviewed as part of the annual capital budget, and was approved to move forward for further discussion. From here, instead of being discussed as part of the capital proposal, meetings were scheduled to focus on RIDOT's proposal as standalone projects. Ultimately, the chief architects and the CIO recommended that both the GIS migration and the big data pilot were good candidates for cloud implementation, and approvals were set in motion.

RIDOT's primary lessons learned from this effort include the following:

- Deciding to include all the stakeholders from the beginning was key. In particular, close coordination with DoIT was critical to RIDOT's success. While some hesitated to engage DoIT in the process, the team chose to involve them, and it was the right approach.
- The implementation plan developed following the training workshop was well worth the effort. It contained an entire page of bullets of other use cases for the cloud and the use of emerging technology data, and this list helped establish buy-in for the pilot project by executives at both RIDOT and DoIT.

> "Just by going through the implementation project, we've learned so much about how to evolve as a technical unit working within a larger agency."

- The RIDOT implementation team was clear and well-versed in why they were moving to the cloud and how this approach fits into the larger picture for the agency going forward, which proved to be important for responding to anyone questioning, “why change, and why now?”
- The RIDOT implementation team was fully transparent with DoIT that the existing state architecture was not structured in a way to support the ingestion or analysis of big data. They shared the implementation plan and vision and communicated openly with DoIT. After participating in multiple meetings, asking (and answering) questions, and reviewing the implementation plan, DoIT determined it was in RIDOT’s best interest to move to a fully-hosted environment outside of state IT/infrastructure.
- Other agencies and RIDOT groups were also (and remain) interested in moving to the cloud. For example, the Rhode Island Executive Office of Health and Human Services (EOHHS) has expressed interest in big data solutions, and there have been considerations and requests to move the RIDOT pavement management system (PMS) and the bridge management system (BMS) to the cloud. Furthermore, new systems for construction and project management are being rolled out natively in the cloud. DoIT was either going to need to make the investment to support these requests to modernize/expand or they were going to have to let the agencies move forward. This decision took some time, as DoIT had to determine if they needed control over it (how should DoIT approach the issue as a policy-setting agency for the state). But in the end, DoIT was supportive of the RIDOT implementation team moving to the cloud, and the RIDOT implementation team was key to communicating the need and driving the decision.

Next Steps

Towards the end of the NCHRP 20-44(39), the RIDOT implementation team noted that they had a lot of momentum and that they were ready to get started on deployment. RIDOT is already sharing experiences with peer agency Massachusetts DOT regarding migrating the enterprise GIS system to the cloud. MassDOT shared similar pain points but ran into production issues, which caused frustrations and delays. RIDOT is hoping to avoid these issues by engaging a cloud infrastructure support contractor from the beginning.

RIDOT plans to leave the existing system and data on-premises until they are happy with the new cloud environment, at which point they will move to production. They will develop documentation on the environment and open it to other operating units within RIDOT to demonstrate the value. RIDOT also wants to make use of cloud functionality (serverless, Kubernetes, a user has one login, etc.) to be fully embedded in the cloud.

RIDOT feels like its journey to the cloud has just begun. Most new vendor systems that are procured by RIDOT are cloud-based, and this is not expected to change going forward. An important procurement expected in the coming years is a new advanced traffic management system (ATMS) for the RIDOT TMC, which is all but certain to use and rely on cloud-based systems and data. When the enterprise GIS systems have been established, the RIDOT TMC still plans and looks forward to trialing/using its data lake to support improvements in the TMC’s day-to-day operations.

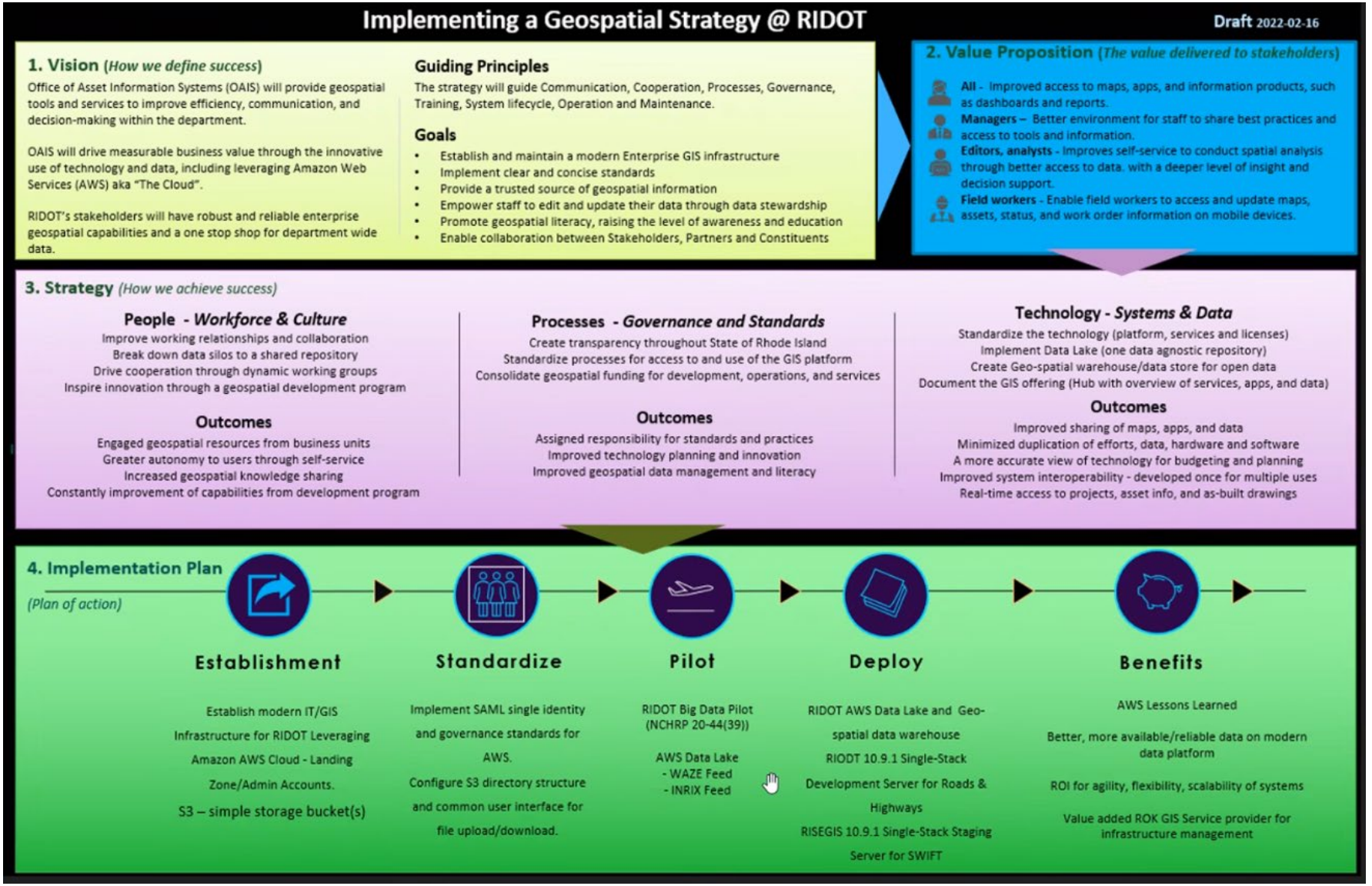


Figure 27. RIDOT One-Pager – Business Case for Moving to the Cloud

CHAPTER 7. OUTCOMES AND OPPORTUNITIES

Successful NCHRP Implementation projects have evolved beyond NCHRP to take hold at the state and local levels as a common practice. For example, the TRB Strategic Highway Research Program (SHRP) on Traffic Incident Management (TIM) Responder Training evolved to become an FHWA National Highway Institute (NHI) course and an FHWA Every Day Counts Round 2 (EDC-2) innovation, with training continuing to be strong in nearly every state to date. Other NCHRP projects have evolved from implementation to FHWA-funded capability maturity curricula and pooled fund study programs.

Each participating agency in this implementation project demonstrated significant progress in one or more ways, including the following:

- Shifting traditional attitudes regarding data management.
- Advancing knowledge and understanding of modern data management practices within the agency.
- Applying modern data management practices through the use of modern systems, services, and tools.
- Advancing the use and understanding of particular emerging technology datasets.
- Learning from their implementation projects and shifting as necessary to more efficient or effective approaches and processes.
- Positively impacting other divisions or groups within the agency through their implementation projects.

Based on the outcomes of this implementation project, NCHRP Research Report 952 has the potential to shift the data management paradigm within state and local agencies, realizing modern data management as the state of practice over the coming years rather than the coming decades.

Discussions with agency participants during the peer exchange uncovered needs and potential opportunities for advancing modern data management practices for transportation agencies. These needs and opportunities are discussed in this final chapter of the report.

Open Source Community

Amongst 50 state transportation agencies, and many more regional and local transportation agencies, all share similar data-related issues and challenges moving into the future. Many agencies are investing in the same or similar projects to address these issues and challenges, while some do not have the time or resources to make data-related investments. A primary topic of conversation amongst the agencies participating in this implementation project was the need and opportunity for an open source community or communities to share ideas, data, code, and innovative emerging data projects and products so that transportation agencies can maximize their cross-agency resources as a “team.” Other noted challenges that might be addressed with an open source community on modern data management include:

- Personnel turnover, which leads to a loss of knowledge, momentum, traction, etc.
- Internal procurement processes, which limit and delay the implementation of innovative ideas.
- Challenges with determining the “right” repository in the “right” location rather than connecting the people and the code in the same space.
- Reliance on closed, proprietary systems, which are slow and expensive to modify to address changing needs.

An example of how an open source approach can benefit transportation agencies is the Waze Travel Times Poller, which was developed by the Lake County Division of Transportation, Illinois, and made available on GitHub.[10] The purpose of this program is to help members of the Waze for Cities (W4C) program process raw travel time data provided by Waze. The program gets data from the Waze Travel Times feed, processes and archives the data, and sends email alerts for any road segments that are "congested." FDOT (as well as several other transportation agencies) picked up the code from GitHub and had the program running the same day.

Other suggestions surrounding an open source community for data management included establishing a potential pooled fund study, encouraging DOTs to commit resources (formal or informal), embracing a "less is more" attitude (e.g., developing a list of common issues, asking people to contribute, sharing code for common DOT problems that can be built from), and involving cities and counties.

Other Key Considerations

Other key considerations discussed by participating agencies in advancing data management practices within transportation agencies that warrant further discussion include the following:

- **Estimate and document cloud costs** – Document the cost of managing on-premises servers versus using the cloud (aggregate by year for improved understanding). Include factors that influence the fluctuating costs of the cloud (e.g., similar to a carsharing model – we can get anything we want when we need it, but we don't have to pay for the best one when we don't need it). Consider the following:
    - How much data do we have (straightforward)?
    - What other data are we going to ingest (a little harder, but still relatively simple)?
    - Ballpark estimates are good enough.
    - Present the business case and costs.
        - If the cloud is not on the roadmap, what is the plan for supporting the growing data now and into the future?
        - Consider capital and maintenance costs.
        - Outline the broader benefits to the agency.
    - Keep it simple/non-technical.
    - Be prepared to answer questions.
    - Consider different audiences/conversations:
        - IT groups – focus on the big picture and future outlook.
        - Executives – focus on use cases, not the unknowns.

> "We underestimated the cost of "petting" our on-premises system (e.g., patch management, maintenance). We spend so much time managing the system and not enough time managing the data. When comparing on-premises versus cloud, we underestimated by 2-3 times the cost to manage our own system."
>
> – *Agency participant*

- **Use a well-architected framework** – Cloud service providers have well-architected frameworks, which include key concepts, design principles, and architectural best practices for designing and running workloads in the cloud. Cost optimization is one of the best practices associated with a well-architected framework. Cost optimization is an ongoing process that can be built into the

[10] https://github.com/lakecountypassage/WazeTravelTimesPoller.

training/education of project managers. Costs can increase if no one is paying attention and will continue if pipelines are not shut down when they are no longer needed. Get project managers in the habit of looking at the monthly bill and re-evaluating/shifting parts of the architecture. There are multiple different ways to do things in the cloud depending on the scale of the data. A well-architected framework is a great place to start.

Areas in need of improvement include:

- **Communication** – The same data are sometimes purchased by different groups within the same agency.
- **Internal Knowledge Management** – How do agencies handle internal knowledge management? Do they have the right platform for doing so?
- **Centralized Data Architecture** – Data silos are a real challenge.
- **Documentation and Metadata** – Data and data processes lack sufficient documentation. Every variable within the data should have a definition. There are tools that will document as the work is being done and as new data are brought into the data lake (e.g., Amazon Glue Data Catalog).
- **Talent** – How do agencies grow and foster talent?
- **Procurement** – The ability to pay-as-you-go with the cloud can be a challenge (e.g., credit card needed versus an up-front fixed amount).

> "We work with the CIO to make sure that all the groups working with data are communicating."
>
> – *Agency participant*

> "I'm sitting on datasets that if I leave, people don't know anything about."
>
> – *Agency participant*

Agencies articulated their takeaways from the implementation project, how it impacted their agency/group, their visions for the future, and recommendations for other agencies. Responses included the following:

- Understanding how access to this new cloud environment opens things in the future.
- Data security issues (e.g., minimizing re-identification) need further investigation (e.g., hashed data).
- Within the agency itself, participation in this project broadened the number of people that want to engage in the conversation and led to interest/participation from other districts.
- Has demonstrated that this approach (i.e., use of cloud, big data) is accomplishable by multiple states and is not just something that is talked about but never acted on and is something that we can and should be doing.
- Now have a "play space" that is used across districts and consultants in a meaningful and intentional way (as opposed to just these people doing these projects).
- Getting out of Step 5 (develop a project in the playground) and into Step 6 (demonstrate value to other business units) is to have that "thing" that can be demonstrated, and it was only possible because of this work in the cloud (was not possible to do on a laptop because of security concerns).
- Learned a lot about how to work in the cloud and how we will do things differently next time around.
- Do not have the fear I had 5 years ago that I was about to delete something. I know the raw data is there, and I can rebuild what I need in a few lines of code.

- Can evangelize with new people joining the journey.
- Looking at what seems to be innocuous data, but needing to keep in mind security/privacy.
- Teaches you things you didn't know about the data that you know now. Has informed next steps.
- Getting buy-in to move to the cloud took a long time, but we were successful; however, a big takeaway is that we are behind other agencies. But moving forward, we have a huge opportunity to truly do things statewide. When we look at steps, we might be able to get to Step 8 sooner than a larger state DOT.
- Don't try to boil the ocean; it is hard to make headway when taking that approach. We were pretty narrow when starting our data lake with a specific goal.
- Over-documentation is not a bad thing. There is a lot to be learned when you document everything you are doing. This is particularly true when working with metadata (can tag columns, code, what can/can't be combined). You want to know about the things you need sooner rather than later.

APPENDIX A TRAINING WORKSHOP AGENDA

NCHRP 20-44(39) Implementation of Guidebook for Managing Data from Emerging Technologies

TRAINING WORKSHOP AGENDA

Day 1

1-1:30	Welcome, Introductions, and Objectives/Overview of Guidebook and Workshop
1:30-2:30	Module #1 – What is Big Data and Why Do We Care?
2:30-2:45	*Break*
2:45-4:30	Module #2 – Traditional Data Management Systems vs Modern, Big Data Management Systems

Day 2

9-10:30	Module #3 – Pilot Project Implementation Plan Brainstorming Activity – 4 Ws and 1 H: Who, What, When, Why, and How
10:30-10:45	*Break*
10:45-12	Module #4 – Getting Started: Establishing the Data Environment/Playground
12-1	*Lunch*
1-2:30	Module #5 – Developing the Pilot – Governance, Roles/Responsibilities, Tools, Skills
2:30-2:45	*Break*
2:45-3:45	Module #6 – Beyond the Pilot: Making the Case for Modern Data Management Across the Organization
3:45-4	Closing Remarks and Training Assessment

APPENDIX B TRAINING EVALUATION FORM

Training Evaluation Form

Implementation of Guidebook for Managing Data from Emerging Technologies

Agency/organization _____

Job title/role? _____

Please indicate your level of agreement with the following statements:	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
1. The objectives of the training were clearly defined.	❑	❑	❑	❑	❑
2. The content was organized and easy to follow.	❑	❑	❑	❑	❑
3. The topics covered were relevant to me.	❑	❑	❑	❑	❑
4. The materials distributed were useful.	❑	❑	❑	❑	❑
5. Participation and interaction were encouraged.	❑	❑	❑	❑	❑
6. The trainers were knowledgeable about the training topics.	❑	❑	❑	❑	❑
7. The trainers were well prepared.	❑	❑	❑	❑	❑
8. The quality of instruction was good.	❑	❑	❑	❑	❑
9. Adequate time was provided for questions and discussion.	❑	❑	❑	❑	❑
10. The training objectives were met.	❑	❑	❑	❑	❑
11. The training met my expectations.	❑	❑	❑	❑	❑
12. I will be able to apply the knowledge learned.	❑	❑	❑	❑	❑
13. Overall, I am satisfied with this training.	❑	❑	❑	❑	❑
14. What did you like most about the training?					
15. What did you like least about the training?					
16. What aspects of the training could be improved?					
17. Other comments?					
How would you rate the training overall?	❑ Excellent	❑ Good	❑ Average	❑ Poor	❑ Very Poor

THANK YOU FOR YOUR PARTICIPATION!

APPENDIX C PEER EXCHANGE AGENDA

NCHRP 20-44(39) Implementation of a Guidebook for Managing Data from Emerging Technologies for Transportation

State Participant Peer Exchange

February 7, 2023
8 am – 4:30 pm
District of Columbia DOT, Washington, DC

Agenda

8-8:30	Welcome and introductions	
8:30-9	Overview of project and workshop goals	
9-12	State presentations and discussion on pilot/implementation projects	
	9-9:40	Arizona DOT (testing connected vehicle back-of-queue pilot project in AWS)
	9:40-10:20	District DOT (using dockless mobility data in the District's on-premises big data environment)
	10:20-10:30	*BREAK*
	10:30-11:15	Florida DOT (shifting District 5 on-premises big data system to the cloud)
	11:15-12	Rhode Island DOT (implementing a data lake in AWS for real-time crowdsourced data)
12-1	*LUNCH*	
1-2	Discussion on challenges (All – roundtable, 15-min each)	
2-3	Discussion on next steps (All – roundtable, 15-min each)	
3-3:20	*BREAK*	
3:20-4	Recommendations for other transportation agencies (All – roundtable, 10-min each)	
4-4:30	Discuss webinar Closing remarks	