White Paper

Big Data Workgroup

How Toll Agencies Can Make Best Use of Big Data

Abstract

The future of tolling depends on the data and analytics capabilities we build and scale. Big data can produce a lot of value for tolling agencies, but only if we know how to claim it. But big data's value doesn't come from the collection of information—that's just the starting point. The real value comes from our ability to combine tolling data with other data repositories and use that stored and/or real-time information to uncover new insights with big data analytics, and then present those ideas to promote better business decisions. The intent of this white paper is to assist tolling agencies make more informed business investment decisions and collect better business insight and intelligence by using big data appropriately.

.

Working Group Chair: *Marwan Madi*

Authors: *Marwan Madi, Rob Marsters, Matt Usher, Said Majdi, Pat Louthan, Ken Juengling, Steve Novosad, Matt Putterman, Suresh Karkala, Joseph Soliz, Titus Moore, Benton Tempas, Michael Davis, Charlotte Frei, Alex Beata, Stephen Kyriakos, Ranjith, Nair*

Committee Members: *Marwan Madi, Jeff Dailey, Michael Davis, Ken Juengling, Suresh Karkala, Pat Louthan, Said Majdi, Matt Milligan, Titus Move, Steve Novosad, Matt Putterman, Joseph Soliz, Benton Tempas, Matt Usher, Neil Gray, Patt Jones, Frank Velez,*

Committee Chair Liaison: *Frank Velez*

IBTTA Representative: *Neil Gray*

Table of Contents

Introduction

With increased connectivity and communications among vehicles, organizations, systems, and people, unprecedented amounts of data are being generated. However, despite its abundance, access to this data and the ability of transportation agencies across the United States to translate it into relevant and valuable information remains limited, particularly as it relates to mobility, programming and planning, and investment decisions, as well as performance evaluation.

Tolling has evolved from an all cash ticket system, to all electronic tolling due to advancements in technology. However, the system integrators have built the systems on a “piece meal” basis in response to the technology changes. All electronic tolling has made tolling more transaction-based starting with pre-paid customer accounts using vehicle transponders to post pay utilizing billing systems where drivers are billed and sent an invoice for payment. Reporting needs have also drastically changed from providing a simple report of expected revenue based upon vehicles paying cash tolls and cash received to inventory management, road side transactions, back office received and processed and finally, a billing/accounting system. Reports often are the last thing system integrators provide, and if provided, it is long after system go-live. Toll agencies are often operating in an informational vacuum working with limited reports.

Scope of White Paper

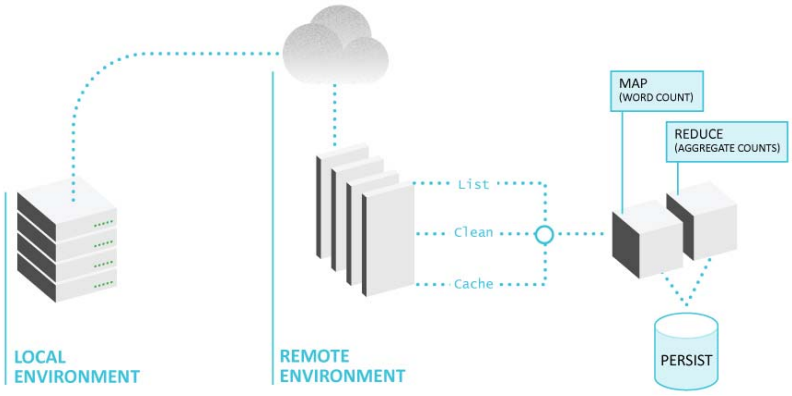
The future of tolling depends on the data and analytics capabilities we build and scale. Big data is defined as extremely large data sets from varied sources that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. Big Data can produce a lot of value for tolling agencies, but only if we know how to claim it. But Big Data’s value doesn’t come from the collection of information; that’s just the starting point. The real value comes from our ability to combine tolling data with other data repositories and use that stored and/or real-time information to uncover new insights with Big Data analytics, and then present those ideas to promote better business decisions.

The intent of this white paper is to assist tolling agencies make more informed business investment decisions and collect better business insight and intelligence by using Big Data appropriately.

Toll agencies invest in volumes of data. It is time to turn this business investment into value. This paper explores the Big Data infrastructure that supports new opportunities, cost savings, transformation, and return on investment (ROI) in your data by addressing the questions raised in the reminder of this white paper.

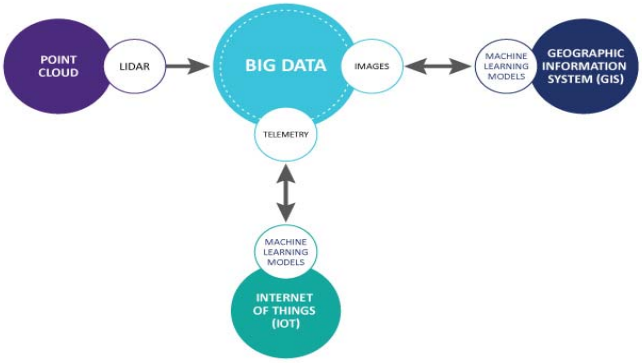
Section 1. What is Big Data?

Big Data is defined more by how it is processed than by raw numbers – but today, data sets of 500 GB and above would be considered Big Data. If datasets exceed the capacity of conventional computing platforms, then *distributed cluster processing* is the 'go to' technology and that essentially defines the problem as a Big Data problem. Big Data programs start with migrating locally produced data to cloud-based storage where it is loaded onto a distributed cluster of machines. The data is prepared for usage and then a process – MapReduce – is used to map a transformation over all elements of the distributed data and reduce those transformations to results or a summarization.



What Interactions Does Big Data Have with Other Technologies?

Beyond the obvious and original connection of Big Data to the web, there are new connections being formed with emerging data sets including GIS, IoT, and Point Cloud (remote sensing) technologies.



Where Is Big Data Headed?

The surprising increase in scale for Infrastructure as a Service (IaaS), cloud vendors are the most obvious indication that a landslide of new data is on the horizon. The more subtle change is that how the data is used is going to change in equally dramatic fashion. There are a few noteworthy changes suggested below.

Machine Learning Takes MapReduce to New Levels

Just as MapReduce provided a distributed solution to very large datasets, Machine Learning (ML) models will become the functions that perform both a map and reduce step to produce fully transformed and summarized data. These models will be managed, and as their

performance degrades, they will be replaced by a pipeline of new models. The complexity of the transformations that ML models bring on-line should not be underestimated as input vectors to Deep Neural Networks (DNN's) climb beyond 100,000 elements feeding more than 1,000 hidden layers to produce results.

Use Case: Geospatial Data. 3D GIS leads the pack for applications of Big Data particularly as point cloud generation continues to make inroads into highly realistic digital renderings of environments. Look for Big Data processing to solve the computational burden of object detection, classification, and for computing interacting geometries.

Nowcasting Rather than Forecasting

With the advent of extremely large data sets in streaming form, e.g. motion video from a car-mounted video camera, it is likely that "in stream" processing will become the norm. This will fuel a predictive analytics cycle that provides immediate estimates of future states based on huge quantities of current data.

Use Case: Video Stream Processing. Remote sensing by our connected machines like connected autonomous vehicles will require real-time processing before storage to present detected objects to the control systems. This places the burden on the stream to perform the processing for everything from inbound streaming video to vehicle-to-vehicle communications.

Streaming Data and ML will Fuel the Semantic Web

While people perform searches, IoT generates data, and remote sensing systems digitize the world, they are building relationships between digitized objects. These relationships are multi-dimensional including spatial orientation, compositional makeup, ownership, and preference just to name a few. The digital ecosystem will not only receive and store raw data but will also produce more data by making inferences about these relationships. This presents a new frontier of data to be mined, understood, and utilized.

Section 2. Who Should Care About Big Data

Changes in roadside and in-vehicle technologies, as well as changing vehicle ownership are all opportunities for the toll industry to look at how, when, and where we participate. The opportunity to put these changes and evolutions in buckets and analyze their effects on our industry is a good first start.

The immediate bucket that will affect users is the method used to pay for these. While vehicles are evolving, the need to pay for tolls remains unaffected. No matter what you are driving, or who is driving, or even if no one is present in the vehicle, the need to pay for tolls is still required. Who gets charged and how is a question we should look to solve—now. Connected, automated, and/or electric vehicles will not change this aspect of toll road usage.

Internal --- Tolling Industry

In general, there are several major strategic drivers of the tolling industry that could be supported by accurate and thorough use of available Big Data. They are:

- Safety – to provide the safest possible environment
- Customer – to meet and exceed customer expectations
- Financial – to maintain a sound financial position
- Infrastructure – to make sound investments in new assets and maintain existing assets
- Mobility – to maintain an accessible, reliable, and available travel system

Managers, analysts, operators, enforcement officials, customer support and maintenance personnel, training staff, and developers should all be aware of and knowledgeable about various aspects of Big Data – its availability and how to leverage it. Managers should use Big Data to make better decisions faster, enabling all the strategic drivers above. *For example, adding weather data into trip data could show differences in rain vs. travel by time of day. This would allow for the staging of safety vehicles and assist in provisioning dynamic messaging to customers.*

External --- Customers

The major drivers for customers are accuracy, speed, safety, and relevance.

External --- Government/Regional Transit and Local Partners

Major drivers for governmental agencies, regional transit and local partners are:

- Accuracy, flexibility, and scale of analytics
- Accurate application of regulatory compliance
- Optimization of capital expenditures

Predictive analytics allow agencies to investigate uses of their systems and identify areas of fraud, waste, and abuse simply by investigating outlying or emergent data. With Big Data analytics, agencies are past the point of trying to use static reporting to understand complex data patterns.

Section 3. Data Management

Introduction

There is data being generated and recorded all around us in nearly every industry. Utility industries use the data being generated from the various metering to record peaks and valleys in usage. This allows the particular company to ensure that sufficient capacity is available during the hours, and days for their customers. This meter data from the sensors, on the sides of buildings, may not tell the complete story. There may be other data sources to consult. The weather report could affect the usage for a particular time period. The number, and type of

construction permits in an area would give string indicators of usage increasing, or not given a geographical area. All of these data points can be leveraged, with the idea in mind of trying to give the customers the best experience, while making the best decisions for the business. What is your line of business? What are the factors that influence it? Are there auxiliary activities that depend on it? What auxiliary activities do you depend on? Is there a single governing body? Is there a committee that votes? Are there parallel activities that are similar to yours? Once you have identified one, or more of these areas, now how do you find data from these processes. This may take a sustained effort of patience, and following a process, rather than a single action. Not every action or activity will have publicly exposed, freely available data. The more data you gather, the more possibilities exist from its possible combinations. This can be a good thing, and it can be an impediment. When the amount of data overwhelms the person's, or people's ability to sift through it, Big Data just becomes big confusion.

Approach

The goal is to choose the right data to be analyzed for the best results every time. This is much easier said than done. One of the best ways to ensure failure in Big Data and Analytics, is to attempt a project beyond the scope of understanding of the team that will be working with it. It is far more effective to choose a smaller, simpler dataset that more people understand. Complexity, and larger datasets, can always be added at a future time when people have a mastery of the simpler ones. Individuals and teams who work closely to the source of the data are very valuable sources of knowledge. They may have no idea what is going to be done with the data they produce, but they can explain the data is being produced. This information should be recorded and placed with the data as it is stored for future use.

As the demand grows for greater amounts of data, and as the number of data sources increases, the responsibility for the collection and distribution becomes increasingly important. As we become more dependent on data to make decisions, we must either take responsibility for getting the data we need, and then depositing it in a local storage location. Or we must get another team involved, who specializes in the care and feeding of the data, in another location. That location can vary based on the technology choices that each group makes. Their involvement in the data management can be very useful, as standards and policies need to be enforced.

Many policies and standards for how to treat data will likely exist in your direct organization, or extended organizations. This does not mean that you should shy away from trying to use data to get insights on how to improve aspects of your business. It does mean, however, that you will need to familiarize yourself and teams of any risks that each data set has. Some data will have no restrictions on it. Some data will be highly restricted due to its sensitive nature. You will be responsible for identifying the sensitivity level of the data and using it properly within the bounds of your permissions. The central organization that handles the data operations will have to grant you access, that is appropriate for your job. Do not assume that just because you have permissions to view it, that the data is fit for public release. There will be times that you are not sure of the sensitivity of data you are working with. Reach out to the subject matter

expert or technical lead on the project and inquire. If that person is not available, reach out the organization closest to the source of the data that you have access to.

Securing data for internal and external usage can be very detail oriented, but can also help protect your organization, and others from unintended consequences. The details of implementing an all-encompassing security policy are to too numerous to list in this document, but we will attempt give some guidelines to help start the conversations around the topic for your organization and others. The security options will be different based on access to each resource as well. When we give access to outside individuals, we need to have agreements in place to protect both sides. This will likely have to come from the legal department where you are. A few questions the organization will need to answer are: Is this data ok for the public to see? How long can the data be available? Is there an access control mechanism? Who owns each piece of the data? Is there a certification process defined for adding new data?

Standards, and interoperability can be divided into three areas of concern, storage, format, and transport/availability. These three areas are related and can affect one another.
The first area to address is storage. Where and how you store your data can greatly affect the other two areas. A common question to ask is whether you should store the data on premises or in one of the major cloud providers. If you have a team of IT professionals to take care of the data and the hardware it resides on, then you may want to leverage that investment and in-house expertise. If you do not have such a team or your team is too busy, one of the major cloud providers would likely be better for the time being. The cost is usually very attractive, and do not hesitate to ask the provider of your choice about their pricing, and what they are willing to do to win your business. This cost will be an on-going operational expense so that will need to be considered in the current budget and the following budget planning sessions.

The second area to address is the format with which to store your data. Three popular file formats which are universally accepted are CSV (comma separated values), XML (extensible markup language), and JSON (JavaScript object notation). Nearly every data processing technology will have the ability to ingest these file formats. There are a few others that have been designed especially for Big Data analysis, such as Avro or Parquet. If your organization already has these Big Data formats being used for information processing, then you can build from the current examples being used. If not, it is recommended to start with more general file formats, and further specialize when situations demand it. Another advantage to storing your data in a generally understood and consumable format, is for future projects. Technology moves at a very rapid pace, but the need for data is going to increase as time goes on. We cannot predict the future, but we can do our best to remain as flexible as possible. It will typically be a built-in feature of the platform to take a general data format and transform it to a future specialized format. It will likely be more effort to go from current specialized format to future specialized format.

More than likely, organizations will have some form of database, relational, or non-relational where some of their data resides. If this is the case, this organization should inquire with the team in charge of the data that resides on the database server. The group in charge of the databases can usually back up the data to a file, after which it can be catalogued and stored.

There may be extreme circumstances where data should not be exported. One of these cases might be when Personally Identifiable Information is present, always check your organizations data privacy policies before storing any data. Once the data is in file format, it can be compressed and stored in the location selected. If the amount of data is going to be cost-prohibitive to store for some reason, you can reduce the historical length of time, or which fields to export, until you have achieved the right balance of cost versus data.

The third area of concern to address is the transport/availability of the data. Once the data is obtained and stored, it needs to be used to justify the effort and cost. Some data will be fit for general public consumption, some will not. For data that is cleared for public use, it is recommended to provide an online portal where users inside and outside of your organization can discover it via search engine and download their own copy to work with. Even though it is a public download it is best practice to meet with your legal representation and ensure that a user license agreement is included where appropriate. When data is meant for a bigger audience, but not everyone, you can put a login in front of the data to ensure that those looking at the data have traceable login credentials. Once again, check with the legal team to ensure everything is above board before asking people to give any of their information. Once the overall logins are established, more sensitive data can be restricted to those with logins for the appropriate scope of usage. If one does not exist, a user management portal with the ability for users to create or request accounts will need to be deployed or created.

Section 4. Big Data Strategies

From Data to Business Intelligence/Information

Business Intelligence should not be confused with Big Data Analytics. We can make some abstractions and talk of business intelligence as the function of applying some analytic tools to enterprise data with the purpose of gaining business insights. Both Business Intelligence and Big Data Analytics do that. That's where the similarity stops. When we talk about Business Intelligence as an application, we can find many differences comparing Business Intelligence and Big Data Analytics; mainly, (a) the type of data that Business Intelligence can handle is a subset of what Big Data Analytics can handle, (b) Business Intelligence and Big Data Analytics have different implementation architectures, and (3) the technology stack used in Business Intelligence cannot handle Big Data.

Once interesting facets of Big Data are found, how that transition to an actual contribution to an agency's future decisions is key. That's where Business Intelligence comes in.

Business intelligence is used by many people to mean many things, but for this section, we will keep the definition relatively simple. It will mean the effort of taking acquired data, spelunking through it, and finding actionable items to present and discuss with your organization. To apply business intelligence, it requires knowledge of the domain from a particular business domain and a willingness to learn about others. There are many opportunities for business intelligence in each role of an organization, because each role has a different perspective. The data itself will not change but public interpretation will. One example of using data sources, to aid in the health of a company would be pricing of a particular good or service. The sales team may be

looking at data to figure out how sales are doing at the moment. The customer service team may be looking for a price point with which more people end the call and give great survey answers. The finance team may be looking at the price point that will allow them to keep the company in good shape financially. All of these questions can be handled with as complex of data analysis as you would like to do. You could start with yesterday's totals versus the day before totals. Then look at the hours during the day. Was there a high or a low point? Was the slope drastic or gradual? Did both days have a similar shape of points? If so, you might add another day. Did it have the same shape? What is going on during those high and low points? You can take several next steps. There is likely more data that can be gathered. You may or may not have it though, as each one of these business intelligence quests has the possibility of finding something that no one had thought of. More research may be needed. This is why the business domain knowledge is very important. That knowledge, combined with the technologies, methodologies, and tools can have drastic effects on your overall business. Once you know what to look for and how to measure it. These known data points then become part of the research for other business intelligence challenges. Data analysis, business intelligence, and investigation are not perfect—you will not always find something very useful. Once interesting points of the data are found, though, future decisions can be greatly influenced due to the effectiveness of other data analysis and business intelligence.

Section 5. Big Data Capability

Agency leaders and managers need to build Big Data Capability within their organization to be able to make smart decisions using Big Data Analytics and drive business transformation. Instead of relying on their intuition, their experience, anecdotal evidence, or their "gut", they can base important decisions on data and facts. The decision to embrace Big Data Analytics does come with one big challenge—establishing the required capabilities.

In this section, we provide a four-layer capability model to help leaders and managers chart a strategic path forward toward building that capability.

Before we introduce the capability model, we first take a closer look at the analytics lifecycle to provide some guidance for the required organizational collaboration and the roles and responsibilities with the associated skills and interests of the various stakeholders.

The Analytics Lifecycle

*Figure 1. **The Analytics Lifecycle***

Discovery

Discovery focuses on the following activities:

- Gaining a good understanding of the business domain and the business processes.
- Capturing the most important business questions that the business users are trying to answer.
- Assessing the available resources and framing the business problem as an analytic hypothesis.

Data Preparation

Data preparation focuses on the following activities:

- Provisioning an experimental analytic workspace, where the data scientist can work without the contraints of a production data warehouse.
- Acquiring, cleansing, and analyzing the data using techniques such as data visualization to gain an understanding of the data.
- Transforming the data using techniques such as logarithmic and wavelet transformations to address potential skewing in the data. Other tools used to transform data consist of extract, transform, load (ETL) tools, and SQL and Java.

Model Planning

Model planning focuses on the following activities:

- Exploring and testing a few analytical models to determine which one will yield the best predictive results.
- Determining correlation between variables to select key variables in model building. Since correlation does not necessarily mean causation, care must be taken to select and quantify cause-and-effect variables.

Model Building

Model building focuses on the following activities:

- Preparing the data sets for testing, training, and production.
- Assessing the quality and reliability of the data to use in the analytic models. Different transformation techniques could be used to see if the quality of the data can be improved.
- Developing, testing, and tuning the analytic models. Testing helps to determine which variables and analytic models yield the most predictive and actionable insights.

Communicate Results

The communicate results step focuses on the following activities:

- Ascertaining the quality and reliability of the analytic model and the statistical significance and actionability of the resulting insights.

- Developing charts and graphics to communicate the analytic model insights and recommendations. Business stakeholders should be able to understand and buy into the resulting analytic insights.

Operationalize

The operationalize step focuses on the following activities:

- Delivering the recommendations, reports, code, and technical documentation.
- Implementing the analytic models in the operational environment. Working with the application and production teams to determine how to set up the analytic models to run on a regular basis.
- Integrating analytic scores (KPIs) into executive and operational dashboards and reporting systems.
- Optionally, running a pilot to verify the business case and the financial return on investment.

Big Data Capability Model

The Big Data Capability Model provides a basis to systematically develop the necessary capabilities for the adoption and strategic usage of Big Data Analytics. The capability development and adoption encompass the acquisition of sufficient knowledge of how to extract business value from Big Data Analytics and the application and management of the underlying technologies. The capabilities are organized in four groups or layers.

Business Layer

The business layer addresses the innovative usage of Big Data Analytics in line with the business strategy and considering the changing trends. This affects the ability to develop new Big Data innovations that provide new value to the organization. At the same time, trends must be detected early and monitored.

Organizations find use cases by discovering new patterns in previously unassociated datasets and developing predictive models that yield key operational advantages.

Building a complete view of data, internal and external customers, and technology enables agency leaders and managers to address this large opportunity. This approach allows an organization to predict likely outcomes based on various what-if scenarios, and possibly influence the outcome.

Because of its disruptive nature, Big Data Analytics requires organizations to transform current ways of work and possibly their business models. Essential to that transformation is the inherent value of data, where data is not viewed as a by-product of the value creation process, but a source of value for the organization, and therefore, for its customers. The competence to transform and adapt to environmental dynamics is critical to the successful implementation of Big Data Analytics. This transformation depends on a unified vision and integrated strategy that originates from the leadership of the organization.

Analytical Layer

The analytical layer addresses the data pipeline: Data collection, storage, processing, and analysis. These data pipeline steps enable speed and access to quality data and analytics.

Data Collection

How an organization collects data is a primary indicator of its Big Data Analytics capability. Typically, organizations invest a major effort in collecting primarily structured data to use in evaluating business performance. It is critical to include unstructured data that is usually ignored or discarded because its size or format does not fit traditional data models.

Data Storage

Traditional enterprise data warehouse infrastructures come with limited storage capacity. This results in what's called "dark data"—data that remains unused because of lack of visibility or access until it is discarded. A successful Big Data Analytics implementation takes advantage of the potential future value of storing data even if its value cannot be determined initially.

Data Processing

A fully operational Big Data Analytics engine uses multiple data types and sources, including real-time streaming data. Establishing an enterprise Data Lake (a system or repository of data stored in its natural/raw format) based on a unified enterprise data architecture is necessary for creating a shared data service that is open to users across the organization.

Data Analysis

The data analysis that consists of using data to report on basic business key performance indicators (KPIs) for the purpose of performance management can be greatly enhanced. New kinds of data and sources enable deeper analytic exploration based on what-if scenarios. This analytic practice evolves with the implementation of Big Data Analytics to extend to predictive modeling and real-time analysis. The result is consistently high quality, availability, and value across multiple areas of the organization.

Capabilities Layer

A McKinsey study found that some organizations have an easier time implementing Big Data Analytics initiatives than others, depending on four critical factors: (1) talent, (2) IT intensity, (3) data-driven mindset, and (4) data availability. Talent being the main barrier.

The ability to achieve transformational results relies on having a sound approach to identifying the right organizational structure and staffing model.

Table 1 depicts the skillsets and roles needed to enable new analytics-based processes in the organization.

Table 1. Big Data Analytics Skillsets and Roles

		ROLES						
		Data Scientist	Business Stakeholder	Software Engineer	Database Admin	Systems Admin	Business Analyst	Graphic Designer
	Advanced Analytics	✓						
	Application Development			✓				
	Collaboration						✓	
	Creativity	✓	✓				✓	✓
SKILLS	Data Analysis	✓			✓			
	Database Management	✓			✓			
	Domain Expertise		✓					
	Programming			✓				
	System Integration					✓		
	Data Visualization	✓			✓			✓

Technology Layer

The technology layer focuses on the adoption of a hosting strategy, analytic tools, and level of integration to enable shared access to key analytics capabilities.

Hosting Strategy

Hybrid hosting scenarios (on-premises and cloud) maximize data access across the organization. Security concerns prevent many organizations from exploring cloud hosting solutions. Ultimately, as demand grows, organizations will have to design hosting environments that optimize availability and speed. Hybrid hosting solutions deliver high availability, reliability, and security across cohesive public/private cloud and on-premises implementations.

Analytics Tools

Initially, analytic tools support department specific requirements for reporting on business performance. As analytic tooling matures, multiple technologies will enable convergence in an enterprise Big Data Analytics platform. The organization-wide presence of analytics is what will allow transforming the way to do business by creating a culture where analytics resources are connected to every business decision.

Integration

Integration is a key factor in unifying deployed technologies around a common architecture. When implemented successfully, integration enables the creation of a single view of key

business functions and entities and delivers a centralized Big Data Analytics capability to the whole organization.

Section 6. New Business Models for Big Data

Introduction

What needs to be in place to implement Big Data and how do you build a data driven culture? Organizations need to plan how they use data so that it's handled consistently throughout the business to support business outcomes.

Data-driven decision management (DDDM) is an approach to business governance that values decisions that can be backed up with verifiable data. The success of the data-driven approach is reliant upon the quality of the data gathered and the effectiveness of its analysis and interpretation.

One of the most potent choices a company can make is to focus on data and develop a strategy of data-driven decision making. Experience and inference are powerful leadership tools, but data can guide CEOs and other executives to the best decisions possible.

Organizations who successfully do this consider the who – what – how – when – where and why of Big Data to not only ensure security and compliance but to improve business performance by extracting value from all the information collected and stored across the business.

Governance

The who – what – how – when – where and why of Big Data can be defined as Data Governance, a set of principles and practices that focus the implementation and ensure high quality through the complete lifecycle of your data as shown in Figure 2 below.

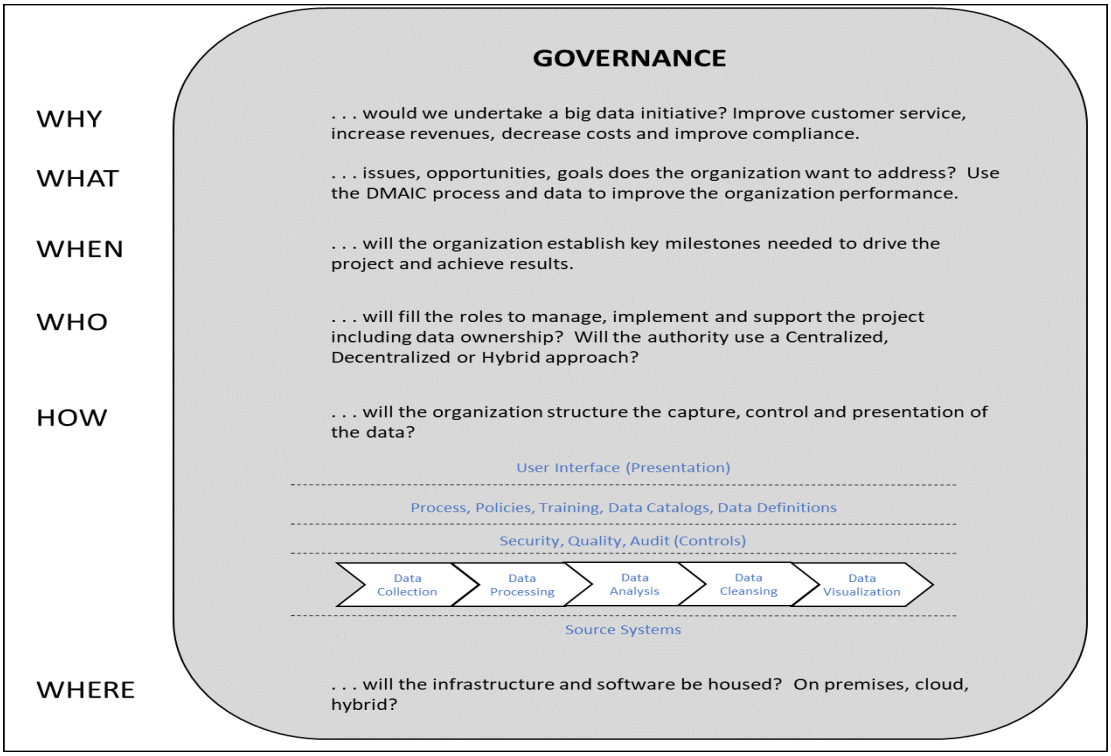


Figure 2. Governance

In this section, we want to focus on the critical success factors that ensure a sustainable and prosperous Data Governance discipline within the organization. As much as we try to focus on the longevity of the business and program, the reality is, it is quite common to launch Data Governance multiple times. Data Governance is a cultural shift and needs to align with your organizational culture, and there is no magical way to be successful. The bottom line is that most companies or organizations don't find success in their first or second efforts.

Best Practices

You can learn a lot from others who have been on a Data Governance journey. However, every organization is different, and you need to adapt the Data Governance practices all the way from the unaware maturity phase to the nirvana in the effective maturity phase.

Below is a collection of best practices that will apply in general:

1. Start small. Strive for quick wins and build up ambitions over time.
2. Set clear, measurable, and specific goals. You cannot control what you cannot measure. Celebrate when goals are met and use this to go for the next win.
3. Define ownership. Without business ownership, a Data Governance framework cannot succeed.
4. Identify related roles and responsibilities. Data Governance is teamwork with deliverables from all parts of the business.

5. Educate stakeholders. Wherever possible use business terms and translate the academic parts of the Data Governance discipline into meaningful content in the business context.
6. Focus on the operating model. A Data Governance framework must integrate into the way of doing business in your enterprise.
7. Map infrastructure, architecture, and tools. Your Data Governance framework must be a sensible part of your enterprise architecture, the IT landscape, and the tools needed.
8. Develop standardized data definitions. It is essential to strike a balance between what needs to be centralized and where agility and localization works best.
9. Identify data domains. Start with the data domain with the best ratio between impact and effort for rising the Data Governance maturity.
10. Identify critical data elements. Focus on the most critical data elements.
11. Define control measurements. Deploy these in business process, IT applications, and/or reporting where it makes the most sense.
12. Build a business case. Identify the advantages of rising Data Governance maturity related to growth, costs savings, risk, and compliance.
13. Communicate frequently. Data Governance practitioners agree that communication is the most crucial part of the discipline.

Section 7. Cost to Build, Operate, and Maintain Big Data Capabilities

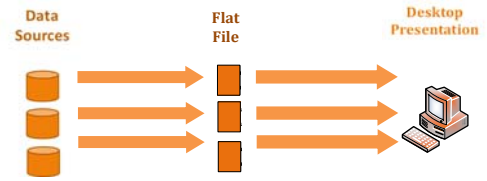
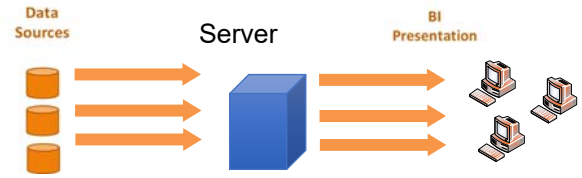
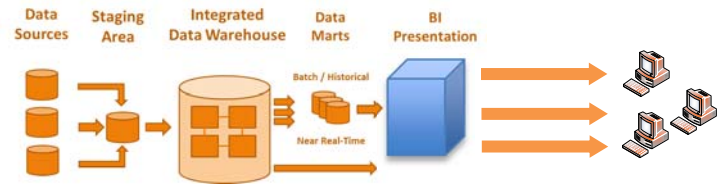
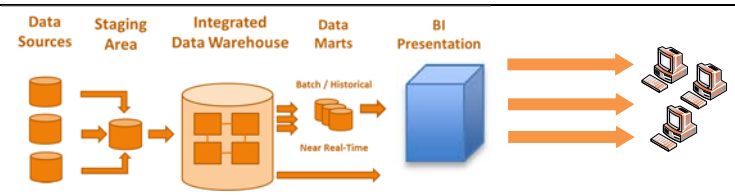
Estimating the cost of a Big Data implementation is challenging due to the wide range of factors that impact cost. To provide some structure, this section organized costs around the size of data being analyzed, the type of data being analyzed, the method of capturing and storing the data, the number of users producing or consuming the analysis, and the complexity of data analytics being performed. Table 2 provides an overview of the segmentation used to organize the costs in Table 3.

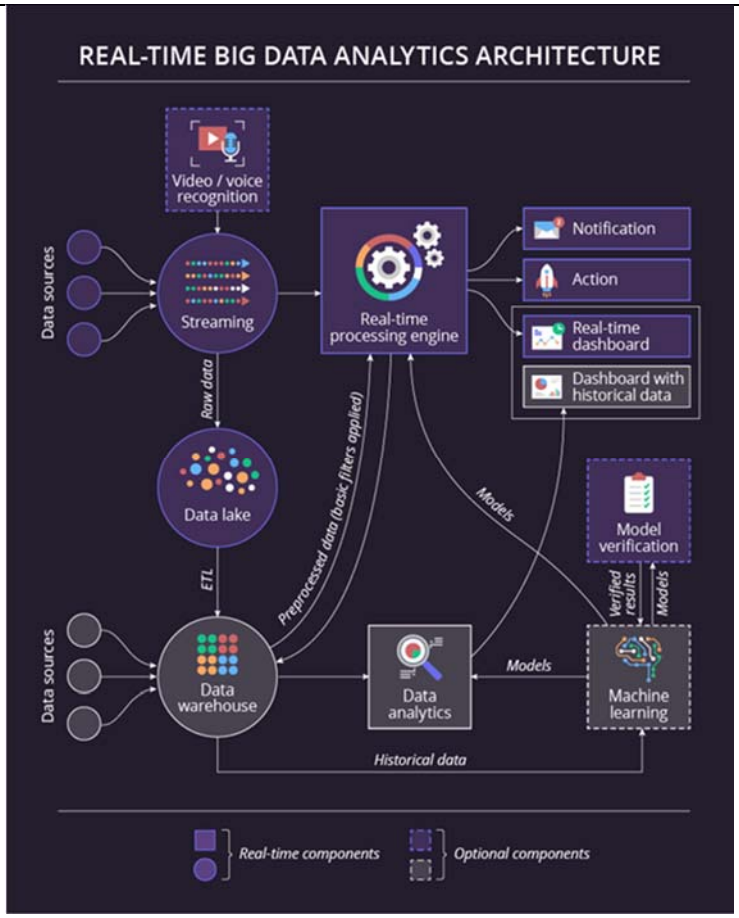
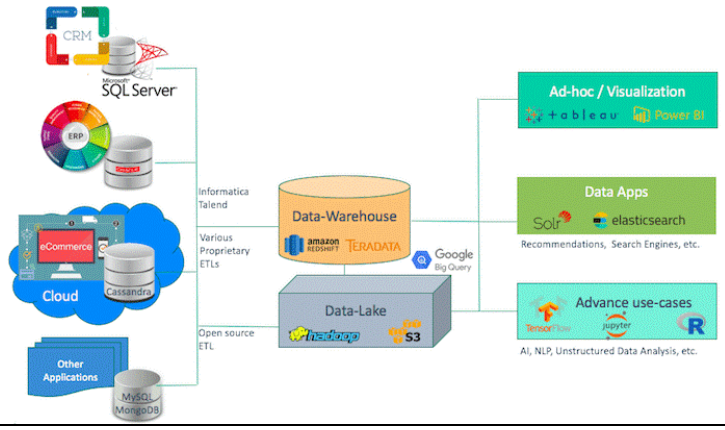
Because of the cost involved in implementing a fully comprehensive Business Intelligence program and the historical failure rate (50% or greater according to various organizations), it is recommended to start small with an “Investigation” implementation as is included in Table 2. It is highly recommended to use this initial implementation type to gain an understanding of the relationships data fields in your systems and to establish a governance approach prior to launching a full-scale implementation.

While broad cost numbers are provided, it is recommended that consulting services are sought during the initial planning stages to help ensure all the facets of establishing a Business Intelligence environment are considered prior to launching an initiative.

Although storing data is not a significant amount of the cost of an overall Business Intelligence implementation, it is indicative of the complexity of the system that would be needed to handle the daily data capture, processing and presentation of the information and therefore included in Table 3.

Table 2. Business Intelligence Segmentation

Segmentation		
Investigation		 Single-user system with multiple data sources fed into a Desktop presentation tool (Tableau, Power Business Intelligence) fed by flat files
Traditional Business Intelligence	Simple	 Multi-user system with multiple data sources fed into a server-based Presentation Tool (e.g. Tableau, MicroStrategy). Data feeds are managed by the presentation tool.
	Medium	 Multi-user system with multiple data sources fed into a standardized, server-based Enterprise Data Warehouse (EDW) using a formalized Extract, Transformation, Load (ETL) process and software tool. The server-based presentation tool now typically feeds from the EDW only to preserve the integrity of all data analyses.
	Complex	 This is the same infrastructure as the Medium Tradition Business Intelligence architecture except that the scale increases with additional data sources, data processing, end users, number of reports, complexity of analyses, etc.
Advanced Analytics		

		 (Source: ScienceSoft – Alex Bekker) Advanced and Real-time analytics uses a much different physical infrastructure to provide specialized functionality like Artificial Intelligence (AI) and Machine Learning (ML) to perform Data Discovery and create an “in memory” system upon which analytics and modeling is performed. Typically, there is no formalized EDW or ETL; however, processes like Data Streaming, and Data Lakes take their place.
NoSQL (Hadoop)		

		Within the Business Intelligence industry, the term “Big Data” typically references a system infrastructure built to process and analyze disparate data (i.e. database data, image data, video data, CSV data, PDF data, etc.). Again, non-traditional Business Intelligence tools come into play like NoSQL databases, Data Lakes, and the Hadoop framework and utilities to make analyses possible.

Table 3 provides cost estimates by segmentation and the categories of Infrastructure, Software, and Personnel.

Segmentation, see Table 2.

Infrastructure is the hardware and associated licenses required to store the data and the applications needed to process the data.

Software is the cost of the applications used to capture, process, and display the data.

Personnel is the different positions required to build and maintain the Business Intelligence system.

Depending upon the existing capabilities of the agency implementing the Business Intelligence system, personnel may already reside at the agency and can cover some of the roles required to implement and maintain the Business Intelligence system. External contractors or consultants may also fill roles until internal resources are developed to transition the roles.

Cost estimates for Advanced Analytics and NoSQL (Hadoop) systems are not included in Table 3 because these are very specialized applications, custom-built to deliver specific results. Artificial Intelligence and Machine Learning are typically factored into the design of these systems significantly impacting cost. As such, a generalized estimate of cost to build these systems would be inaccurate, if not misleading. What can be suggested is that these applications might more economically be built in the cloud, where these specialized tool sets are readily available, building the system a la carte rather than investing in custom development.

Table 3. Cost Estimates by Segmentation and the Categories of Infrastructure, Software, and Personnel

Segmentation	Size	Infrastructure		Software		Personnel	
Investigation Time to Initial Results: 2M – 6M	< 1TB	**CPU** Laptop/Desktop **Storage** Laptop/Desktop **Licenses (NA)**	$1K - $3K Per Computer See Above $ 0	**Presentation** Desktop Business Intelligence **ETL (NA)**	$1K - $2K Per Seat $0	**Business SME**	$40K - $100K Per FTE

Segmentation	Size	Infrastructure		Software		Personnel	
		Cloud (SaaS) NA	NA	NA	NA	NA	NA
Business Intelligence – Simple Time to Initial Results: 6M – 1YR	1TB – 5TB	**CPU** Server **Storage** Included w/ server **Licenses (OS/DB)**	$20K - $30K Per Server $0 $5K - $15K Per Server	**Presentation** Enterprise **ETL NA**	$3K - $8K Per Person $0K	**Business SME(s)**[1] Assumes existing IT resources	$40K - $100K Per FTE
		Cloud (SaaS)	See Software	1 Power Business Intelligence User (Azure Pricing Tool)	$4K - $5K Per Month	See above	See above
Business Intelligence – Medium Time to Initial Results: 1YR – 2YR	5TB – 50TB	**CPU** MPP Appliance (not fully racked) **Storage** MPP Appliance **Licenses (OS/DB)**	$0.3M - $.5M See Above $0.5M - $1M	**Presentation** Enterprise **ETL** **Dev Ops**	$2K - $5K per seat $100K - $400K for the enterprise $50K - $100K	**Manager** **Team Lead** **Business Intelligence Architect** **ETL Developer** **Data Modeler** **Presentation Admin** **Presentation Developer** **Business SME(s)**	$80K - $150K Per FTE

Segmentation	Size	Infrastructure		Software		Personnel	
		Cloud (SaaS)	See Software	Approximately 15 Users	$15K - $40K Per Month	See above	See above
Business Intelligence – Complex Time to Initial Results: 1YR – 2YR	> 50TB	**CPU** MPP Appliance (fully racked) **Storage** MPP Appliance **Licenses (OS/DB)**	$1.0M - $2.0M See Above $1.5M - $5M	**Presentation** Enterprise **ETL** **Dev Ops**	$2K - $5K per seat $400K - $1M $50K - $100K	**Manager** **Team Lead** **Business Intelligence Architect** **ETL Developer** **Data Modeler** **Presentation Admin** **Presentation Developer** **Business SME(s)**	$80K - $150K Per FTE
		Cloud (SaaS)	See Software	Approximately 15+ Users	$40K+ Per Month	See above	See above

Segmentation	Size	Infrastructure		Software		Personnel	
Cost Driver(s)		CPU Storage Licenses	Size of Data # of Concurrent Users Data Complexity Data Processing Frequency Response Time Size of Data Data Growth Rate Redundancy Disaster Recovery Number of CPUs or Cores Number of Users	Presentation ETL Dev Ops	Single or Multi-User Environment Frequency of Update # of Connections # of Transformations # of Environments # of Batch Jobs Scheduling of Batch Jobs Automated Alerts Automated System Checks	FTE Type	Implementation: Type Complexity Size of Team

Section 8. How to Assess Agency Readiness for Big Data?

To get started with a basic assessment of readiness to implement Big Data Analytics, leaders and managers must answer a few questions regarding the way they run their organization's business. This is a self-evaluation. After answering the questions under each of the seven categories, select a score from 1 (Strongly Disagree) to 5 (Strongly Agree).

Each category is scored. A tally of the seven scores gives an indication of the readiness level of the organization to implement Big Data Analytics. Deficiencies must be addressed first in order to increase the chances of a successful implementation.

Readiness Assessment

Strategic Alignment

a. We are aware of the environmental factors that drive the business.
b. We have a clear, actionable business strategy.
c. We have implemented key management and business processes that reinforce each other.
d. We have implemented key management and business processes that effectively execute business strategy.

1	2	3	4	5

Continuous Improvement Culture

a. We consistently measure business factors (costs, quality, outputs, …).

b. We continuously drive change to core business processes.
c. We apply data-driven improvement techniques (Six Sigma, TQM, …).
d. We replace maturing best practices over time.

1	2	3	4	5

Information Usage Culture

a. We use historical performance information for planning (forecasts, budgets, plans, …).
b. We have enough relevant information to make fact-based decisions.
c. We use quantitative methods (linear programming, optimization, modeling and simulation, collaborative filtering, …).
d. We use metrics and KPIs.

1	2	3	4	5

Program Management

a. We identify Big Data opportunities within key business functions.
b. We manage Big Data opportunities as a program (portfolio of projects).
c. We invest in core Big Data competencies.
d. We use new Big Data applications to improve business performance.

1	2	3	4	5

Decision-Making Processes

a. We have well-established decision-making processes.
b. We practice formal collaboration in decision making.

1	2	3	4	5

Data Warehousing

a. We have effective data warehousing processes.
b. We have data quality skills.
c. We have Extract, Transform, Load (ETL) skills.
d. We have query and reporting skills.

1	2	3	4	5

Business-IT Partnership

a. We have business leaders and managers who are IT-savvy.
b. We have IT leaders and managers who are business-savvy.
c. We allow IT to contribute at a strategic level.
d. We have instituted effective IT governance mechanisms.

1	2	3	4	5

Scoring

For each of the seven categories above, you have selected a score from 1 to 5. The following example shows how to take the seven individual scores and combine them to determine the overall score for your organization's readiness to transform into a data-driven organization.

Scoring Example

Let's assume an organization has the following scores for the seven categories of the readiness assessment:

Strategic Alignment	**5**
Continuous Improvement Culture	**4**
Information Usage Culture	**5**
Program Management	**1**
Decision Making Processes	**3**
Data Warehousing	**2**
Business-IT Partnership	**4**

Next, sort the individual scores in ascending order (lowest to highest):

1 2 3 **4** 4 5 5

The median number, 4 in this example, is the overall score for this organization's readiness. This organization is mostly ready to go big with Big Data.

Section 9. Big Data Case Study

Pennsylvania Turnpike Commission

Vehicle to Infrastructure Broadcasts
Vehicles in small quantities are beginning to produce valuable data streams based on vehicle telemetry data. The Pennsylvania Turnpike Commission (PTC) has invested in the Road-Side Unit (RSU) hardware necessary to collect data being broadcast from DSRC-enabled vehicles. The IEEE 1609 Wireless Access in Vehicular Environments (WAVE) specification (variously referred to as J2735, 802.11p, or DSRC) provides both a method and means for producing and consuming vehicle telemetry data that includes useful vehicle performance information like location, instantaneous speed, braking behavior, and acceleration variables. DSRC data is unique in that it is crowdsourced, real-time, and relatively inexpensive considering the cost of the infrastructure required to utilize it. As DSRC adoption continues, it will become a ubiquitous 'Big Data' source for traffic management suitable for both 'nowcasting' and the production of analytics with longer time horizons.

PTC performed an exploratory analysis of DSRC data in a small study area with RSU's positioned on either side of the Susquehanna River near Harrisburg, Penn. The layout of the study area is shown in Figure 3. The figure shows the location of slower speed vehicles primarily on the entry/exit ramps located near the headquarters of the PTC as they enter and exit tolling areas. While the scope of the effort was limited, it was immediately apparent that this real-time data could support a number of uses including: 1) validating other road network data from third party vendors, 2) supporting real-time data analytics for detecting macro-level events like incidents, 3) providing a faster means of incident detection via vehicle telemetry, for instance by watching for acceleration vectors that were statistically significant.

As an example of the data available, Figure 4 shows vehicle behavior at a toll booth as vehicles approach and begin decelerating (North bound) or begin accelerating away from the system (South bound). DSRC has a unique strategy for data retention in its message specification. It gathers similar messages from the same broadcaster ID (anonymized) and increments a message count every 10 msec.. Consequently, very accurate telemetry is available in the DSRC stream without totally overwhelming the ability of the data infrastructure to keep pace.

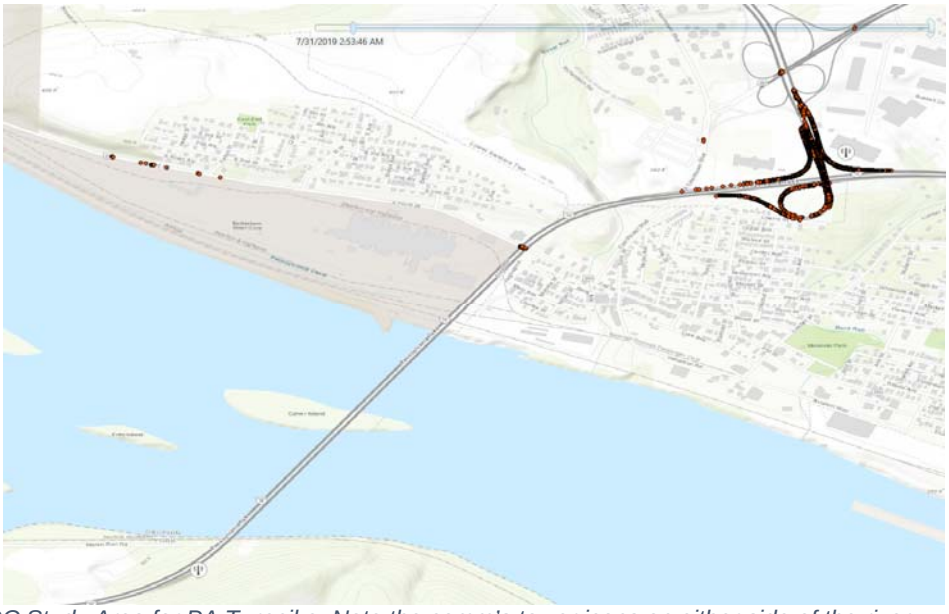


Figure 3. The DSRC Study Area for PA Turnpike. Note the comm's tower icons on either side of the river.

Once DSRC data is received, it is published to a specific endpoint. In the case of PTC, this infrastructure relies on the Cisco/Kinetics architecture to ingest, store, and re-publish the data to an endpoint. That endpoint can then be subscribed to in the same manner as any other feed for instance from an HTTP endpoint, RESTful service, or Kafka broker.

Figure 4. Braking vehicles arriving and departing from a tolling station on the PA Turnpike

A SaaS Big Data analytics package is used to ingest and process the DSRC data. Figure 5 below indicates the 24-hour profile of average speed < 50 mph received by the RSU and published at the endpoint. Most of this data is ramping and tolling acceleration and deceleration. This kind of profiling could potentially be performed across the roadway by highway segment (supplied by INRIX data) and used to compare a learned model of vehicle behavior against observed values over some time period. Variance to the model would provide the basis for real-time nowcasting, such as vehicle incident at location, dispatch emergency responders to scene, and prepare for delays > 1 hr.

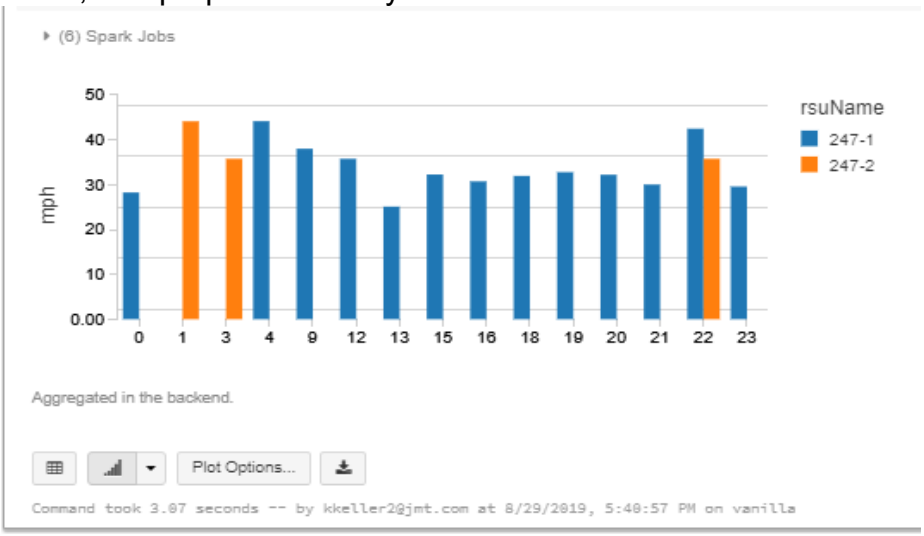


Figure 5. Vehicle Speeds <50 mph observed by roadside units

In addition to basic summarizations of DSRC data, more exotic analyses can be undertaken by using a Big Data engine. For instance, tracking DSRC vehicles for origin-destination travel times is of real interest as a road network performance metric. Below in Figure 6 is an example of computing the tracks for a given set of DSRC data. Although the tracks are close to what the raw DSRC data would suggest in Figure 5, Figure 6 uses an algorithm to tie together data points based on their spatial-temporal proximity. This is a complex algorithm the implementation of which uses the Spark parallel processing library to accomplish in reasonable time.

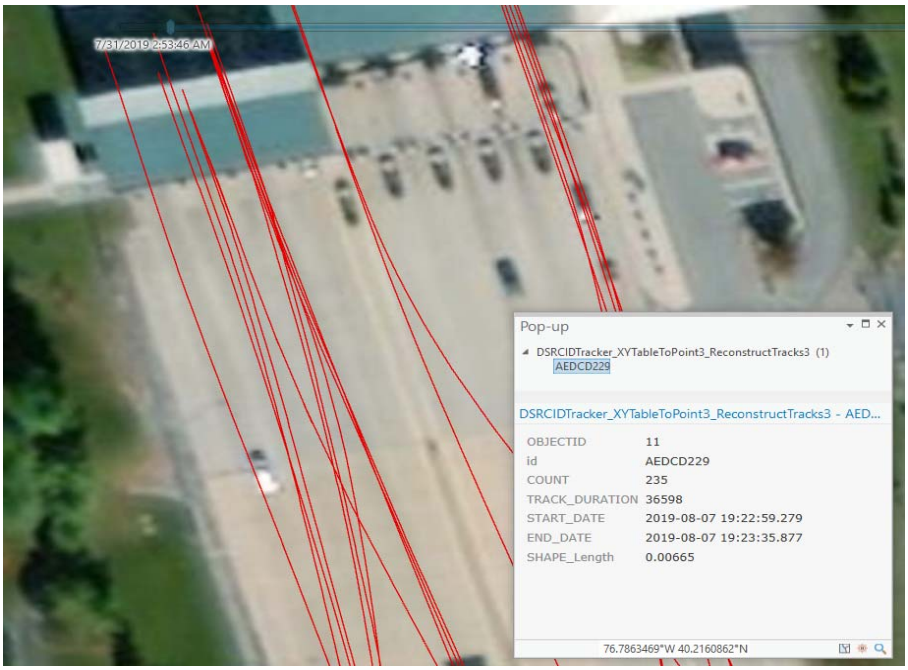


Figure 6. Tracking DSRC vehicle telemetry through the tolling station

While the promise of leveraging this kind of real-time data is significant, there are several requirements for an architecture to be able to handle the V3 aspects of the data.

IoT frameworks provide a general-purpose solution to this data 'supply chain' problem. IoT frameworks can organize the provisioning and secure communication of devices in large-scale environments. They support communications both from (data) and to (control) devices. In addition, they move data to the cloud where cloud-based services and storage assist in analyzing and persisting data. Finally, they provide message hubs that can be used to publish the data to interested parties.

The parties and their various interests in traffic data and the associated analytics are:

- Traffic Managers
    - o situational awareness (assessing the current state of the network, understanding the reasons behind it, and predicting future states and necessary responses to ensure safety and mobility)
    - o publishing (making that 'awareness' available to and consumable by other parties)
- Value-Added Service Providers

  - timely data (making the sensed data available in a usable timeframe)
  - enriched data (using other data to expand the offering)
  - reliable data (reinforcing uncertain data with supporting data)
- The general Public
  - situational awareness (understanding what is happening or about to happen that will affect safety or mobility)
  - options generation (improving their current or predicted situation)
  - value-added services (getting more out of the travel scenario)

Each of these stakeholders is potentially well-served by an infrastructure that can securely and quickly organize and enrich data to enhance its reliability.

Section 10. Where Do We Go From Here?

Recommendations and Next Steps

The Big Data subcommittee recommends the following items for further investigation:

- Further investigate the source(s) of data to support a bid data enterprise.
  - How to collect data?
  - How to process data?
- Further investigate issues related to interoperability, security, privacy, using blockchain, and keeping data up to date as it related to building a Big Data platform for toll agencies.
  - For examples, explore whether agencies will need to develop security protocols and adopt new policies to ensure the protection of sensitive technical systems and appropriate use of the data they generate.
- Further investigate how Big Data can be used to improve Data Analytics.
  - How does Big Data translate into decision support systems?
  - How does Big Data benefit end users/customers and third-party providers?
- Consider inviting Big Data vendors to present to IBTTA and Toll Agencies the variety of sophisticated ways other industries are:
  - Adapting existing platforms to accommodate the scale of Big Data.
  - Transitioning and modernizing their systems and using Big Data platforms.
  - Warehousing/storing data.
  - Using portals to help guide business intelligence and investments in maintenance and operations.
  - Consider doing a Big Data pilot with one or more tolling agency.

Appendix A. Who Should Care About Big Data?

Internal --- Tolling Industry

In general, there are several major strategic drivers of the tolling industry that could be supported by accurate and thorough use of available big data. They are:

- Safety – to provide the safest possible environment
- Customer – to meet and exceed customer expectations
- Financial – to maintain a sound financial position
- Infrastructure – to make sound investments in new assets and maintain existing assets
- Mobility – to maintain an accessible, reliable, and available travel system

Managers, analysts, operators, enforcement officials, customer support and maintenance personnel, training staff, and developers should all be aware of and knowledgeable about various aspects of big data – it's availability and how to leverage it. Managers should use big data to make better decisions faster, enabling all the strategic drivers above. *For example, adding weather data into trip data could show differences in rain vs. travel by time of day. This would allow for the staging of safety vehicles and assist in provisioning dynamic messaging to customers.*

Value of Emerging Technology and Connected Data

There are multiple big data analytical techniques that can be useful to the above areas. They are:

- Basic Analytics – for producing summary data utilizing MapReduce style aggregations of data. *For instance, to summarize the traffic volumes throughout the day by hour for a given week, month, or year.*
- Predictive Analytics – for forecasting or nowcasting by leveraging various forms of regression including linear, logistic, and decision trees. *For instance, to predict the expected traffic volumes based on current trending data.*
- Classification – for identification and detection of common objects of interest in the domain. *For instance, whether a vehicle is a motorist, law enforcement, or an emergency responder.*
- Clustering – to identify related groups of objects. *For instance, understanding where incident hotspots are in the network to make targeted safety enhancements.*
- Network and Graph Analysis – to understand the interconnectedness of objects in the domain. *For instance, a road network as a set of intersections and road segments used for the purposes of re-planning detours.*
- Machine Learning – to provide a mechanism for learning about new objects in the domain and to apply the learned model to identification or prediction tasks. *Such as, what business indicators suggest that an increase in revenue for a location would be expected.*

Impact of Emerging Technology and Connected Data

There are four potential impacts of utilizing emerging technology: handling greater scope, saving time, reducing cost, and improving quality. Any or all can lead to significant disruptions of the status quo. In some cases, adoption of emerging technology can lead to overhauling long existing practices and procedures that might be overly time-consuming and even outdated, resulting in eliminating the need for those practices.

The primary benefit to the tolling industry of these technologies is to provide better KPI's that demonstrate the objectives of the strategic drivers are being met. By learning from significant datasets and incorporating that into programs, the time required for gaining insights and applying them to current operations is shortened.

Another practical benefit of adopting big data is the potential for reducing or eliminating backlogged analyses. Since analysis has traditionally been time consuming and talent intensive, often engaging external groups, tools that reduce both the time and the level of expertise required for sophisticated analyses can eliminate both outside engagement of experts and encourage analyses that have been neglected due to time or resource constraints.

Future Use of Emerging Technology and Connected Data

While streaming of entertainment media is now commonplace, the future of streaming data of all kinds seems likely. That capability is being demonstrated in the real-time streaming video of semi-autonomous vehicles, smart homes, and even doorbells. Consequently, as the volume of data increases, the processing paradigm will fully shift from gather, store, and process to simply gather and process, with storage being an option after the fact. The ability to work with streaming data that is continuously produced will become paramount to the agile organization.

In addition to a data streaming environment, multi-cloud operations seem inevitable. While cloud and other Infrastructure as a Service (IaaS) aka "cloud" vendors are currently battling for near exclusive commitments from customers, the logical alternative is that customers choose the best infrastructure, platforms, and services for their applications.

As autonomous vehicles are developed and gain market share, the traveling behavior of customers is likely to change. Autonomous vehicles show promise for reducing the number and severity of incidents as well as more efficient capacity utilization of the road network. There will be opportunities for investment in areas such as IoT (roadside sensing, communications, and control), Big Data (to ingest data streams in traffic), and Machine Learning to perform new predictions of customer usage and road network performance. It is likely that all of this will be built on the back of IaaS hardware with some basic change in management of data and its security and privacy.

External --- Customers

The major drivers for customers are accuracy, speed, safety, and relevance.

Value of Emerging Technology and Connected Data

There are multiple emerging technologies that can support customer drivers including:

- Nowcasting - Crowdsourcing and streaming data from platforms and devices provide the basis for nowcasting – to shorten the time required to make accurate predictions about future conditions or opportunities. *For instance, by leveraging real-time, vehicle-source data for predictive nowcasting, traffic managers might be able to provide 'look ahead' routing information to better manage traffic flow during events like incident clearing. This could help customers and their devices solve lane positioning and exit questions created by an incident and improve traffic flow.*

- Inferential Statistics (through space time pattern mining) – Data collected with space and time associated with it allows for the computation of inferential statistics to determine if emergent patterns are significant. *For instance, as the customer travels through the road network, certain patterns are historical while others deserve more significant consideration for near-term decision making.*

- The Semantic Web – The semantic web refers to a web of relationships. Knowing those relationships can help traffic managers provided tailored advice to specific drivers. *For instance, as traffic patterns emerge, the known driving pattern could be used to leverage targets of opportunity, like errand stops or tasks, to get the most out of imperfect traffic patterns.*

Impact of Emerging Technology and Connected Data

The primary impact of both large data sets and the technologies that use them is to provide tailored responses within a framework of maximizing the utilization of travel time.

If a traveler requires service along the roadway system, tailored service responses, for instances invoking preferences related to vendors, can be applied to the overall decision-making pattern.

Enabling optimized decision making.

If a trip involves multiple waypoints in a specific order to complete a set of tasks then that trip can be made more flexible by technologies that, as conditions change, can find alternative routes that still accomplish a task-oriented framework for decision-making.

Future Use of Emerging Technology and Connected Data

Ultimately, customers will be looking for better situational awareness – to avoid delays or missing tasks to accomplish – and the ability to take advantage of opportunities as they arise.

Future enhancements to situational awareness include the ability to perceive the situation faster via big data, comprehend the impact of changes to the plan through AI-based planning, and to project future states with greater accuracy via Nowcasting and data analytics.

Customers are always looking to optimize their trips either by taking advantage of new opportunities or being presented with alternatives that allow them to exercise their tradeoffs and preferences. Even in the face of what might seem like a delay, if future systems understand enough about the priorities and habits of the customers, then relevant opportunities, for instance to stop and get a meal early at your favorite eatery where you have a e-coupon, can increase customer satisfaction with the overall trip experience.

External --- Government/Regional Transit and Local Partners

Major drivers for governmental agencies, regional transit and local partners are:

- Accuracy, flexibility, and scale of analytics
- Accurate application of regulatory compliance
- Optimization of capital expenditures

Predictive analytics allow agencies to investigate uses of their systems and identify areas of fraud, waste, and abuse simply by investigating outlying or emergent data. With big data analytics, agencies are past the point of trying to use static reporting to understand complex data patterns.

Value of Emerging Technology and Connected Data

Several important technical capabilities deserve further consideration by related agencies including:

- Nowcasting – By utilizing streaming data sources and applying data analytics to the stream, agencies open the possibility of nowcasting trends to support their operations. *For example, to control budgets, agencies can determine earlier in complex projects whether they are on budget or how far limited resources will take them.* Additionally, data analytics can be used across all the budgets to realize benefits that can offset predicted budget overruns and accomplish the goals of their project base.
- Inferential Statistics – Often, decisions based on historical data might be skewed because of a long history of data that biases understanding away from current trends. Inferential statistics applied to large data sets at scale open the possibility of finding significant emerging trends in data and determining its direction over time. *For instance, if determining accessibility is important, then current trends considering recently completed infrastructure projects would be important to discern.*
- Data analysis of regulatory and compliance data – Regulatory compliance and the creation of business policies are often driven by data analytics. By performing a more accurate analysis on larger data sets where emerging trends and outliers can be

identified, business policy and compliance regulations can be written to deliver a desired effect.

Impact of Emerging Technology and Connected Data

The impact of data-driven policies on the performance of related agencies should not be underestimated. Rather than attempting to glean information out of standardized reports of descriptive data and trends, big data analytics can deliver key insights into data that will allow both compliance regulations and business policies to be measured and refined using predictive models, clustering, and decision tree analysis. This enables better decisions to be made faster which saves tax dollars.

What occurs when loopholes in compliance regulations are exploited? Variance between measured and expected outcomes begin to increase. Those changes can be detected and examined quickly using big data analytics by using a more flexible and responsive analytics approach. It is possible that these methods will allow regulators to reduce fraud, waste, and abuse by pinpointing the source of variances quickly.

Not only can regulations be crafted that are flexible in the face of changing data, business policies can likewise be developed and then modified either to correct shortfalls or exploit new opportunities. All of this relies on being able to process large volumes of data quickly and accurately to provide insights that can support data-directed decision making. For example, if a local partner's tolling system shows a significant drop in revenue after an infrastructure change, big data can be quickly analyzed to find out what has changed in a related system that may have inadvertently caused the side effect. To mitigate impacts to partners, big data can also be exploited to possibly craft business policies that offset an unanticipated negative consequence to maintain strong partnerships.

Future Use of Emerging Technology and Connected Data

The speed with which big data is going to be collected and analyzed could revolutionize the ability of agencies to optimize their businesses. If, as anticipated, big data processing moves directly into the data stream, methods of successive approximation, like gradient descent search, will become the norm rather than the exception for learning how to control governments and businesses. Rather than waiting for aggregation, storage, and retrieval for processing by lengthy and monolithic jobs, it is possible that virtually all business processes will become 'agile' using machine learning models to classify observations and make control decisions. Continuous monitoring of these models for re-training could deliver a highly flexible system capable of immediately reacting to fundamental changes in the business landscape.

Data privacy is a major consideration for government organizations and the shift toward preserving private data rights is clear. How governmental bodies address this requirement has impacts that could both drive consumers toward or away from providing data and increase or decrease legitimate business development for private organizations. Data privacy concerns could be addressed by the adoption of personalized, computable data policies associated with private data. This approach, coupled with the issuance of data certificates that support

provenance tracking and demonstrate ownership, rights purchase, or rental might form a flexible and personalized basis for leveraging private data.

Appendix B. New Business Models for Big Data

Why?

The Business Cases and Drivers to Implement Data Governance

Data is becoming the core corporate asset that will determine the success of your business. Digital transformation is on the agenda everywhere. You can only exploit your data assets and do a successful digital transformation if you can govern your data. This means that it is imperative to deploy a data governance framework that fits your organization and your future business objectives and business models. That framework must control the data standards needed for this journey and delegate the required roles and responsibilities within your organization and with the business ecosystem where your company operates.

Business	IT
• Increasing organizational efficiency • Compliance with financial regulations, audit requirements, etc. • Increased data volumes and complexity • More efficient technology implementation • Reduction of risk through greater transparency • Increasing revenue • Customer optimization	• Poor data quality • Failed implementation • Infrastructure optimization and consolidation • Major application roll-out/upgrade • Technical challenges associated with BI/DW (environment) • Growth of unstructured content

The Data Governance Value Proposition

Data governance means better, leaner, cleaner data, which means better analytics, which means better business decisions, which means better business results, market positioning, mindshare in your space, reputation, and profit margin (everybody likes this one).

Barriers	Benefits
•Data ownership and other territorial issues •Lack of cross-business unit coordination •Lack of data governance understanding •Poor state of data management infrastructure •Resistance to change or transformation •Lack of executive sponsorship and buy-in •Resistance to accountability •Lack of business justification •Inexperience with cross-functional initiatives	• Quality of data • Consistent data definitions • Data as an enterprise asset • Appropriate use of data • Collaborations among teams, business units, etc. • Accountability for data use • Quality of master and meta data • Sharing of data • Visibility into the enterprise via data • Change management processes for data use and management • Data security • Data Lineage

What?

Governance is a critical enabler to address some of the most common pain points felt by the business, including:

- Drive process improvement.
- Increasing customer demands, new regulations.
- Streamlines and unifies the approach to managing data.
- Ensures the right people are involved in determining standards, usage, and integration of data across projects, subject areas, and lines of business.
- Balances silo-ed short-term project delivery focus.
- Traditional projects don't give enough focus to data management.
- Systems are becoming more challenging to manage.
- Data quality issues are persistent.

When?

Establishing key milestones in the implementation of your big data project is critical. These milestones will:

- Set the organizations expectations for the implementation team.
- Measure progress.

Who?

Deciding what operating model your organization will adopt is part of the initial steps in setting up your data governance program. The operating model selected should:

- Outline how your program will operate
- Set the expectations of escalation and decision making as well as program oversight
- Provide the infrastructure for ownership and decision making

User Role Definition

Data governance will involve the whole organization in a greater or lesser degree, but let's break down the most commonly involved stakeholders:

DATA OWNERS: First, you will need to appoint data owners (or data sponsors) in the business. This must be people that can make decisions and enforce these decisions throughout the organization. Data owners can be appointed at the entity level (e.g., customer, product, or employee records, etc.) and supplementary on attribute level (e.g., customer address, customer status, product name, product classification, etc.). Data owners are ultimately accountable for the state of the data as an asset.

DATA STEWARDS: Next, you will need data stewards (or data champions) who are the people making sure that the data policies and data standards are adhered to in daily business. These people will often be the subject matter experts for a data entity and a set of data attributes. Data stewards are either the ones responsible for taking care of the data as an asset or the ones consulted on how to do that.

DATA CUSTODIANS: Furthermore, you may use data custodians (or data operators) to make the business and technical onboarding, maintenance, and end-of-life updates to your data assets.

DATA GOVERNANCE COMMITTEE: Typically, a data governance committee will be established as the main forum for approving data policies and standards and handle escalated issues. Depending on the size and structure of your organization, there may be sub fora for each data domain (e.g., customer, vendor, product, employee).

The roles highlighted above should optionally be supported by a Data Governance Office with a Data Governance Team. In a typical enterprise, here are some examples of who might make up a Data Governance Team:

- **Manager, Master Data Governance:** Leads the design, implementation, and continued maintenance of Master Data Control and governance across the corporation.
- **Solution and Data Governance Architect**: Provides oversight for solution designs and implementations.
- **Data Analyst:** Uses analytics to determine trends and review information.
- **Data Strategist:** Develops and executes trend-pattern analytics plans.

- **Compliance specialist:** Ensure adherence to required standards (legal, defense, medical, privacy).

One of the most important aspects of assigning and fulfilling the roles is having a well-documented description of the roles, expectations, and how the roles interact. This will typically be outlined in a Responsible, Accountable, Consulted, and Informed (RACI) matrix describing who is responsible, accountable, to be consulted, and to be informed within certain enforcement—a processor for a certain artifact as a policy or standard.

Operating Model

Three models to consider for your environment are: centralized, decentralized, hybrid/federated; each with their pros and cons outlined below.

CENTRALIZED OPERATING MODEL

Similar to a top-down project management model, a centralized operating model relies on a single individual to make decisions and provide direction for the data governance program. There can be many different titles reflecting this position, such as Chief Data Officer, Chief Information Officer, Chief Data Steward, Data Governance Director, Data Stewardship Director, and so forth.

Pros

- Dedicated Data Governance Lead
- More efficient decision making
- Easier to focus on policy, guidelines
- Easier to control costs
- Reporting structure clearly defined based on the org chart

Cons

- Incompatible for a more matured data governance program
- Increased bureaucracy due to the linear structure
- Operation rigidity
- More time required to accomplish data governance operations
- Potential loss of oversight over unique and detailed business considerations
- Mostly concerned with enterprise priorities

DECENTRALIZED OPERATING MODEL

Almost the exact opposite, there is no single Data Governance owner as everything is committee-based.

Pros

- Relatively flat structure
- All-encompassing representation from the business
- Relatively easy to establish

Cons
- Reaching consensus tends to take longer
- Difficult to coordinate and commit the needed resources from participants
- The committee's direction can heavily be influenced by those stronger willed

HYBRID OPERATING MODEL

This is meant to be the best of both worlds. There is still a centralized structure which oversees the enterprise data level for which it has bottom-up input wide participation from the business units. The centralized structure provides a framework, tools, and best practices for the business units to follow, but in theory, it also provides the units with enough autonomy to manage business unit specific data and offers channels of influence to gather input for data sets impacting enterprise data or the other way around.

Pros
- Top-down decision-making regarding enterprise data with bottom-up inputs
- Centralized enterprise strategy with a decentralized execution and implementation
- Ownership is given to the application owners for the data and metadata
- Broad membership for working groups
- Provides the ability to focus on specific data sets at the business unit level and their relationship with the enterprise data
- Full autonomy to develop standards, policies, procedures for the business level
- Issue resolution at a bottom-up approach

Cons
- A highly skilled Data Governance lead position is required full-time – not an easy find
- Can get very political at the working group level
- Decisions made at the group level will be pushed up to the upper levels for approval
- Difficult to find the balance between enterprise priorities and those of the individual business units
- Oversight over the autonomy of the business units can be challenging and relies heavily on self-reporting
- Business unit's efficiency depends on localized skills
- Metadata management not simple to address as it can differ widely from one unit to another

Deciding on an operational model while you are initiating your data governance program is important, but it can also be adjusted at a later time. Small organizations typically benefit from a centralized structure because the data governance lead would have the capacity to not only wear multiple hats but be able to learn enough about the business, its environment, and challenges to address these issues. A decentralized model can work well for an organization which has dispersed its operations to several remote locations. As an organization expands, it is usually advised to look into a federated operating model to better support the data governance needs of the organization.

How?

How the organization chooses to structure the capture, control, and presentation of the data has been addressed throughout this white paper. In this section we will focus on the policies, process and training needed for successful big data implementation.

POLICIES AND RULES

Policies and rules establish the basic requirements for the general structure, format, identity, ownership, usage, and access for all information within the agency. These documented set of guidelines ensure the proper management and usage of information and are aligned with elements such as Data Security, Data Transformation, Data Catalogs, and Definitions as well as intended usage and ownership of data.

PROCESS

Big Data brings insights into existing business processes across the enterprise through data discovery and analytics, resulting in incremental improvement of existing business processes and the development of new business processes. The insight from this data discovery and process focus will improve cost, quality, and time resulting in continuous process improvement. Metrics and Key Process Indicators (KPIs) should be developed for each organization and process.

TRAINING

With Big Data comes the need to develop Analytics Literacy within/across the Organization – both for the Business and IT. This training can start with bringing a small core group of people up to speed on the use of the selected front end tool(s) who can provide needed reports and dashboards to the organization. However, over time basic report building capabilities need to be taught to a broader cross section of the organization to facilitate rapid analysis of issues and the creation of a data driven culture.

Where?

Where the organization decides to house the capture and processing of the data is dependent upon multiple factors that have been discussed in other sections of this white paper. Some of the factors to consider include:

- Existing IT capabilities (both infrastructure and personnel)
- Cost considerations (See Section: Cost to Build, Operate, and Maintain Big Data Capabilities)
- Agency policies
- Type of BI implementation