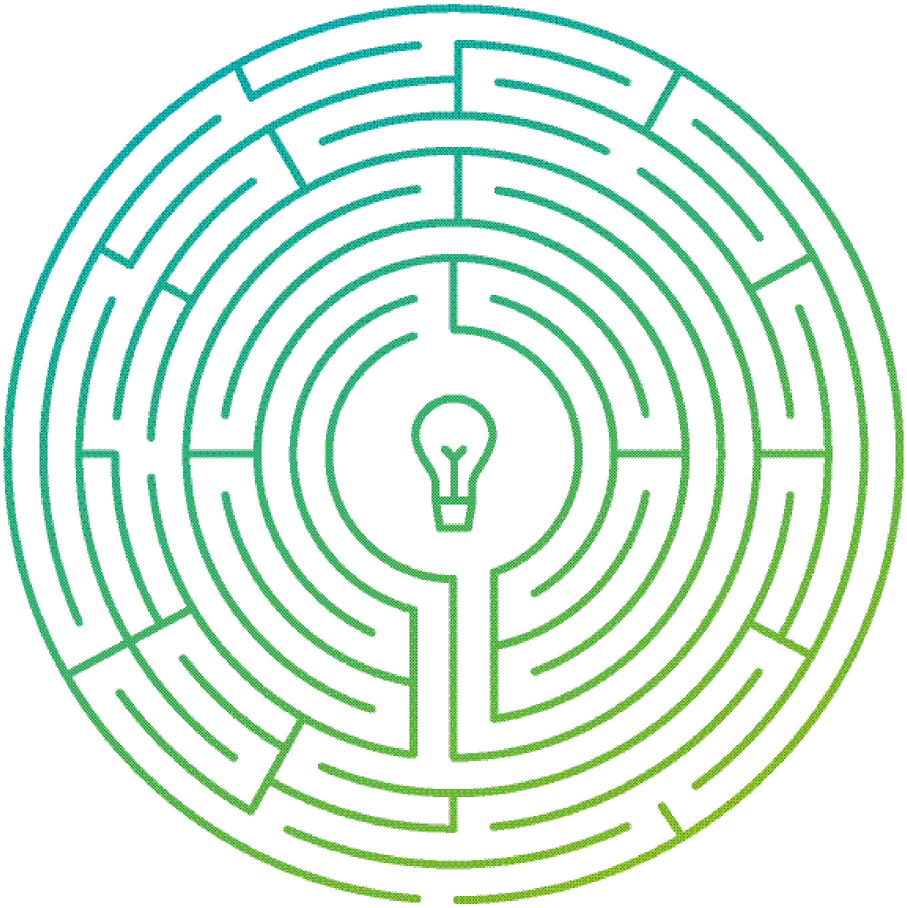
Deloitte.



Modern Business Intelligence

The Path to Big Data Analytics

April 2018

Introduction

In a world where the amount of data produced grows exponentially, federal agencies and IT departments face ever-increasing demand to tap into the value of enterprise data. With the potential to increase business value and overall mission effectiveness, many agencies are seeking new and innovative ways to turn organizational data into valuable insights. Making sense of the technologies, tools, and, techniques required to derive insights from such vast amounts of data can seem overwhelming. However, a modern enterprise analytics solution often doesn't require a complete reboot of previous investments. By investing in a modern business intelligence (BI) platform that complements existing business intelligence systems, businesses can expand their range of insight-driven capabilities. With this investment comes a shift in data ownership from IT to business groups, giving more users the power to answer any question, with any data, at any time. By implementing a modern BI platform, federal agencies can use analytics to more effectively achieve mission objectives such as protecting and maintaining the health of the American people, keeping the country safe and secure from foreign and domestic threats, and preventing waste, fraud, and abuse of government resources.

In this paper, we discuss the challenges of traditional business intelligence and reporting, the need for solutions that answer today's toughest data challenges, and the accompanying people, processes, and technology that support the shift to modern enterprise analytics.

Traditional Business Intelligence Platforms

The traditional Business Intelligence platforms of the past two decades have chiefly succeeded in providing users comprehensive historical reporting and user-friendly ad-hoc analysis tools. The availability of this functionality is largely due to the underlying data architecture, which consists of a centralized data storage solution such as an Enterprise Data Warehouse (EDW). EDWs form the backbone of traditional data platforms and often connect an immense web of source systems into a central data repository. Data is then standardized, cleansed, and transformed in the EDW before being pulled into various reports and dashboards to display historical business information, such as quarterly sales or weekly revenue metrics. While traditional BI offers a basis for these types of dashboards and ad-hoc reporting, this IT-developed solution has presented its own unique challenges.

While users have been able to gain tremendous value from traditional platforms for historical reporting capabilities, more users now require data analyses techniques that require direct access to data without relying on IT specialists. The following challenges associated with traditional BI solutions have been highlighted by federal agencies in the analytics space:

- **Lack of On-Demand Analysis Capabilities** – Today's advanced BI users don't want to wait to get answers to their most complex business problems. More users require self-service capabilities in to relate and analyze specific data sets based on their own understanding, at any time, for any purpose.
- **Need for Predictive Analyses** – Historical reporting capabilities only provide one piece of the puzzle: insight into what happened in the past. To truly become data driven and forward thinking, businesses are looking to predictive analytics – or insight into the future. With predictive models, businesses can use patterns and forecasting to gain actionable next steps based on their data.
- **Analysis of Mixed Data Types** – Traditional BI Platforms have largely been focused on structured data, but today, users need the ability to also view and analyze semi-structured, unstructured data, and third party data. The sheer amount of information produced has skyrocketed in recent years, in part due to the Internet of Things (IoT), new data mining

> techniques, and the proliferation of sensors and other automated data collection tools. Data scientists and advanced BI users now require access to untapped data in various formats, where they have the ability to create their own algorithms and blend data types, and where insights are available on demand for rapid and accurate decision-making.

Many organizations that lack the people, processes, and technology necessary to expand their data analysis capabilities to the next level become discouraged. These challenges require an analytics platform and strategy that goes beyond the breadth of traditional BI platforms, as seen in **Figure 1**.

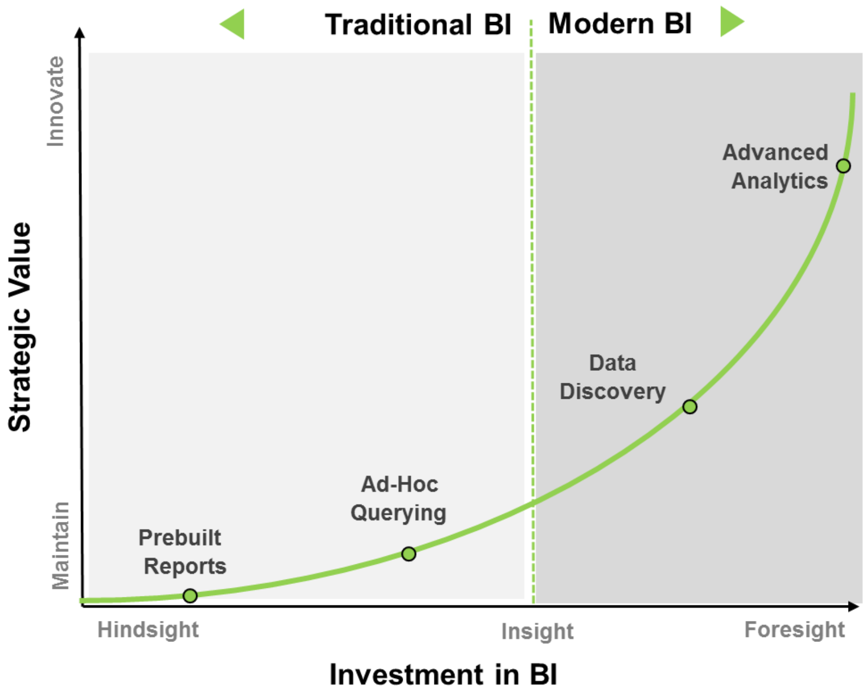


Figure 1: As organization investment in data modernization increases, value grows exponentially and changes from hindsight to insights to foresight.

This document demystifies modern business intelligence technologies and helps explain how the modern platform can co-exist in a traditional BI reporting environment to expand the capabilities of the businesses in the realm of data exploration and advanced analytics.

What is a Modern Business Intelligence Platform?

While traditional BI platforms often provide analyses that answer the question “What happened?” in a historical perspective, modern platforms have the ability to answer the question of “What is happening, what will happen, and why?”, offering the ability to not only obtain and monitor a continuous pulse of the organization through rapid analytics, but to accomplish mission objectives through predictive analytics.

Integrating Traditional and Modern BI Platforms

Data platform changes are necessary to shape the foundation for an enterprise-wide data transformation and organizations are rightfully wary of scrapping their entire IT architecture and starting fresh. Data warehouses continue to play a key role in existing data platforms, providing the thoroughly cleansed, organized, and governed data needed for most businesses. The data warehouse allows business executives and others without deep technical knowledge to gain insights from historical data with relative ease. This data, sourced from the data warehouse, is highly accurate due to IT scrubbing, rigorous testing, and thorough knowledge of layers of the data by IT specialists. However, the challenges associated with traditional BI are creating demand for augmenting an EDW with another form of architecture optimized for quick access to ever-changing data: the Hadoop Data Lake.

Organizations looking to modernize their analytics platforms have started to adopt the concept of data lakes. Data lakes store information in its raw and unfiltered form, be it structured, semi-structured, or unstructured. As opposed to the stand-alone EDW, data lakes themselves perform very little automated cleansing and transformation of data, allowing data to therefore be ingested with greater efficiency, but transferring the larger responsibility of data preparation and analysis to business users.

Using Hadoop’s distributed file system (HDFS), data lakes offer a low-cost solution for efficiently storing and analyzing many types of data in its native form. A data lake solution coupled with a data warehouse defines the next generation of BI and offers an optimal foundation for data analysis, as shown in **Figure 2**.

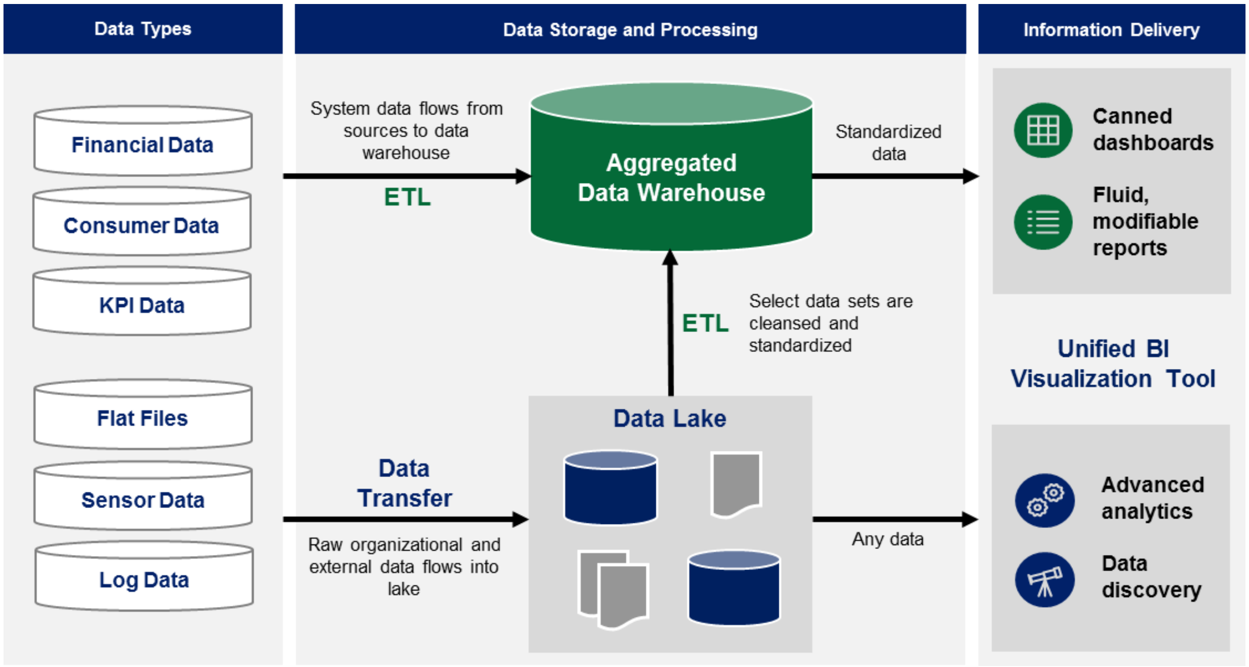

Data Types
Data Storage and Processing
Information Delivery
Financial Data
Consumer Data
KPI Data
Flat Files
Sensor Data
Log Data
System data flows from sources to data warehouse
ETL
Aggregated Data Warehouse
Standardized data
ETL
Select data sets are cleansed and standardized
Data Lake
Data Transfer
Raw organizational and external data flows into lake
Any data
Canned dashboards
Fluid, modifiable reports
Unified BI Visualization Tool
Advanced analytics
Data discovery

Figure 2: Data begins in source systems on the left. The data warehouse receives data in large batches for BI reporting, while the data lake collects raw organizational data used for advanced analytics and data discovery.

In the system displayed in Figure 2, the EDW receives system data from various sources through an ETL (Extract, Transform, and Load) process. After being cleansed, standardized, and transformed, the data is ready for analysis by a wide variety of users via reports and dashboards. Meanwhile, the data lake collects raw data from one, many, or all of the source systems, and data is ingested and immediately ready for discovery or analysis. The result: a wider user base exploring and creating relationships between enormous amounts of diverse data for individual analyses, on demand.

Understanding Your Data within a Modern BI Environment

While the data lake can quickly ingest and store organizational data, it does not provide a one-size-fits all solution for every data type. As seen in **Figure 3** below, the higher the complexity and veracity (required precision) of the data, the greater the need to cleanse, transform, and organize the data. Data lakes offer the ability to do all three, but may not always be the most effective solution. Given this, there is a tradeoff of having quick access to raw, unfiltered data, or spending the time to cleanse and prepare data based on business requirements.

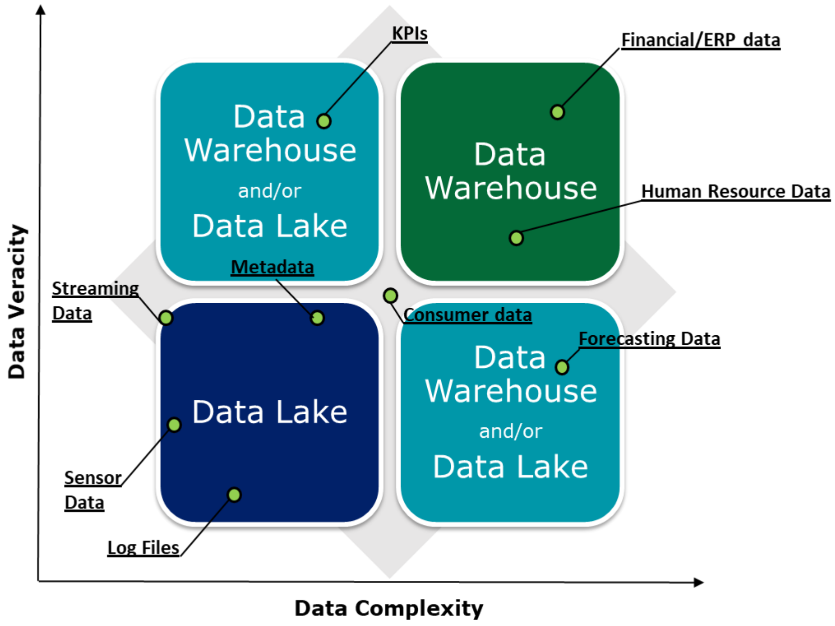


Figure 3: The higher the veracity and complexity of the data set at hand, the more cleansing, transforming, and organizing the data set will require.

For example, a monthly costing report which requires auditor precision may be better suited for development in a traditional data warehousing model where financial subject areas can be established, calculations are defined, thorough testing completed, and predefined reports built. On the other side of the business, machine and log data present a classic use case for the data lake. Log, sensor, and other streaming data are great candidates for data lakes given their semi-structured and flat, less-complex nature. Often, it does not make sense to spend time modeling, loading, and converting log data into a reporting structure within the data warehouse. This process can be arduous, and can be further complicated when an analysis involves additional data, requiring additional data modeling and IT processing. Conversely, loading data into the data lake can be done with relative ease due to the limited amount of conversions and transformations initially required. With the data lake, open-ended data discovery and analysis allows any questions to be asked, and data structures or sets to be determined in support of those questions on-demand.

Beyond the Foundation

A modern BI platform allows a wider base of employees to leverage huge amounts of data for rapid, insight-driven decision-making. However, the platform is only the foundation for advanced analytics. The people, processes, and technologies that support the platform ultimately drive the impact of the system's ability to derive insights and achieve mission objectives. In the following section, we will discuss topics of consideration when managing a modern business intelligence solution.

Rapid Insights at a Federal Manufacturing Facility

A federal manufacturing facility needed quicker access to large volumes of data in its native format in order to scale and adapt to the changing needs of the business. The Deloitte team implemented a Hadoop Data Lake to complement the client's existing data warehouse in order to support self-service and open-ended data discovery. By using the data lake, users are be able to perform advanced analytics of sensor and log data and analyze various file types on-demand.

Modern Business Intelligence Management

A BI Platform without data management is a data swamp – a place where data goes in, but is unable to be retrieved or provide the desired value. Modern business intelligence data management focuses on increasing the value, and thus impact, of the modern business intelligence investment.

It is important to discuss how data lakes can, and should be, divided into three zones. These zones aid in the process of data loading, defining user access and security, and creating a more user-friendly environment. The zones are depicted and described in more detail in **Figure 4**.

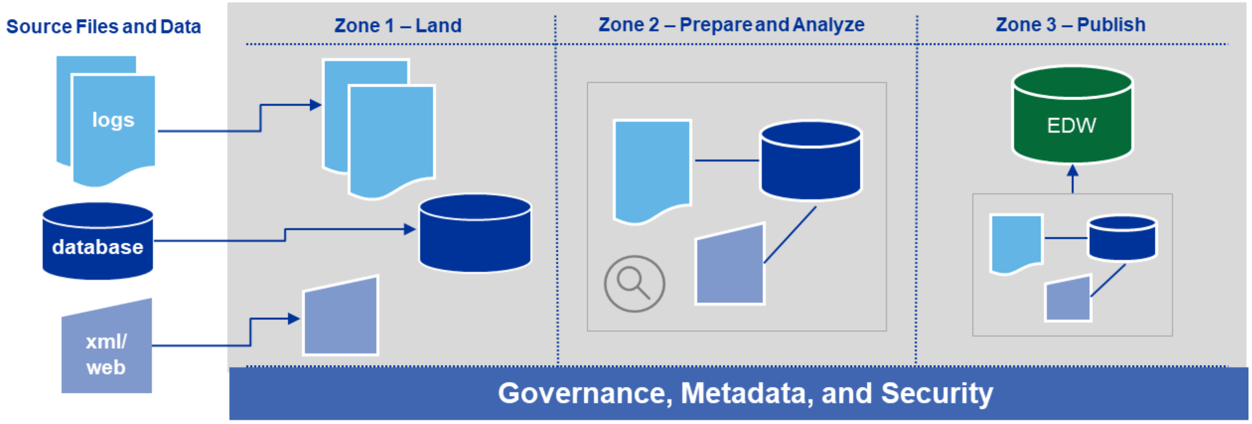


Figure 4: Data becomes increasingly cleansed and standardized as it moves from Zone 1 to Zone 3.

- **Zone 1**, the landing region, consists of raw, untransformed data gathered directly from the source systems. Here, data is often automatically ingested and maintained by IT, with very little room for manipulation.
- **Zone 2**, the data sandbox, where data is lightly processed, cleansed, and combined for exploration and analysis. Each user may have a private region alongside a collaborative, shared region, and security control is typically less strict than the landing and publishing zones.
- **Zone 3** consists of refined data, stored in its optimal form for reporting or treated as trusted data. Often called the production zone, Zone 3 has the strictest governance controls. Data stored here can be considered a published, trusted data set, and can be used or manipulated in the data warehouse or by other users for analysis.

To provide an analogy, Zone 1 is like finding a diamond in the rough. It is raw and uncut, and may not look like it has much value without performing additional processing. Zone 2 is where the diamond is cut and polished. Just like preparing data, diamond cutting requires specialized skills, tools, and techniques. It is then able to be examined for cut, clarity, color and other measures. If the diamond is found to be of value, it would then be sent to Zone 3, or made available to others. The true value is now known and the diamond can be sold or placed in a setting – similar to using analyses developed within the data lake on a predictive analysis or dashboard in the data warehouse.

In this section we provide the three keys to keeping a modern BI platform from turning into a swamp: effective governance and data enhanced with metadata, selecting and leveraging the right software products, and adopting organizational change. We will discuss these three keys in more detail below.

Governance, Metadata, and Security

Any analytics platform is only as useful as its data. While governance may not seem imperative or valuable for small data sets, proper standards are crucial as the data platform scales to accommodate an entire agency or several agencies. Governance is typically defined as an internal body that helps organizations oversee changes to analytics solutions and processes, resolve analytics/data issues, and facilitate decision making amongst agency stakeholders. As part of its regular duties, the governance body helps prioritize data sets to be ingested into the data lake, defines best practices for performing analyses and creating efficient self-service data sets, and sets the criteria for publishing data sets for other users.

As higher volumes of data are ingested into the data lake, the risk of misinformation and incomplete or undefined data grows, reducing the overall usefulness of the data stored, and ultimately the quality of any downstream analyses produced. This is where Metadata Management comes in to play. Metadata management can best be illustrated by considering a library. The books represent various pieces of data; as the library grows, it is important to catalog, index, and describe each book in the context of larger categorizations, such as genre, publication date, and author name. In any large library, it would be impossible to locate books without a sorting method. In the same way, designing a metadata process from the beginning enables efficient data organization and trust throughout the pipeline, preventing the data lake from degrading. Effective metadata management not only builds trust through clearly identified data, but also enables shared knowledge of how data is defined and related, expediting future analyses.

Security also plays a key role in the development and proper use of a data lake solution. Comprehensive identify management and authentication systems are key to controlling access to content stored in the data lake. Role-based access and security groups offer a way to regulate which users have the ability to access and interact with the data lake, minimizing the risk of non-cleared users accessing potentially sensitive or confidential data. Through these processes, agencies can increase the consistency with which users can locate and trust data, increasing user adoption and trust.

Analytics Software

Today's modern analytics software provides the ability to power both agency decision making and comprehensive growth. In order to support modern analytics capabilities, today's analytic software must power the following components of data analysis:

- **Data Ingestion** describes the tools and software that collect and store the various types of data in Zone 1 and making them available for analytics. Logs and streaming data require different ingestion mechanisms than data residing in a database. Various open source and commercial software can ease the data ingestion process with flows and visual representations of the process from various data sources.
- **Data Preparation,** or data wrangling, is the cleansing, consolidation, and standardization of data prior to data analysis that is typically performed in Zone 2. With the responsibility of data preparation now falling into the hands of the business user, software is emerging to aid in the heavy lifting. With well-documented metadata, users can input the expectations and rules for how data should be processed, resulting in a user-prepared and tailored data set.
- **Data Discovery** is used for analyzing patterns and relationships through summary statistics, what-if analysis, and visualizations and is also performed in Zone 2. Many visualization software products are able to connect and combine data from both data warehouse and data lake platforms, yielding results not previously possible given the nature of structured and unstructured data. There are two reasons it is important to select software that leverages the data platform when performing analysis. First, results should be processed in the platform, not

users' desktops, which yields much more efficient results due to scaled architecture. Second, once data sets have been created, it's important for those results to be easily shared with proper security.

- **Advanced Analytics** consists of a collection of data analysis techniques that expand beyond historical reporting and trend analysis to gain deeper insights, actionable intelligence, and next steps from diverse sets of data. The data lake platform supports software and languages that power the tools necessary for enabling advanced analytics. Methods such as machine learning, artificial intelligence, data and text mining, network/cluster analysis, sentiment analysis, and random forest regressions are changing and shaping the future of modern analytics. For example, through predictive regression analysis of population data, a data scientist can map how average temperature, population density, and proximity to standing water relates to the spread of disease. From there, agencies could identify target locations that are most at risk and potential solutions for mitigating the risk of an outbreak.

Advanced Analytics at CFPB

The Consumer Financial Protection Bureau (CFPB) was receiving large volumes of unstructured data in the form of 40,000 consumer complaints each month. Deloitte implemented an analytics solution based on machine learning, advanced data mining, and algorithms to find new insights and automatically classify consumer complaints. The solution developed by Deloitte is now processing over 40,000 complaints a month, with accuracy exceeding that of humans by 30%.

Enabling Insight-Driven Organizations

Accompanying any technological shift is a change in tools, processes, and equally as important, people and behavior. Even with the most cutting-edge technology and well-documented processes, users need to feel empowered to adopt modern BI solutions, as they are the ones who will drive insight-driven decision making. For that to happen, the organization must have the skills and knowledge as it relates to data. The shift to a modern business intelligence solution requires support to the user in the following forms, to name a few:

- **Culture of Ownership** – With the shift to self-service technologies at the forefront of the analytics space, it is important to recognize that the power of data is moving to the hands of the user. Helping users understand the ways in which they can use the modern BI platform – and use it correctly - will support employees to feel empowered and incentivized to operate independently in the new technical environment.
- **End User Training** – Standing up a new business intelligence platform without organizing proper training and available change management support almost guarantees a lack of commitment from the target users. User training is key to aspects of the transition – from platform training (such as the nuances of operating Hadoop) to training within the visualization tool. Hands-on training sessions, deep dives, and ongoing support should be available to potential users.
- **Community of Engagement** – While users may have received the technical training required to physically use the tool, keeping user interest from waning is key to a platform's adoptability. By fostering interest through interactive workshops, office-specific demos, and clear means of communication, effective data management can keep users active and interested, spread awareness, and preserve the momentum once the technical transition itself is complete.

Investing in people upfront simplifies the shift from a traditional platform to a modern BI solution, preparing users from the start to get the most out of this new investment.

Conclusion

Today's BI landscape is rapidly changing and doesn't show signs of slowing down. The challenges associated with traditional business intelligence platforms have driven business leaders to look for modern, forward-looking, flexible solutions. Modern business intelligence platforms help organizations take advantage of mass amounts of existing and new information in vastly different ways than were previously possible, allowing users to ask and find answers to any question, with any data, at any time. To do this, organizations don't have to replace their existing data platforms but rather leverage existing investments and expand capabilities by augmenting with modern tools and technologies.

As self-service analytics technology continues to put more power and responsibility into the hands of the business user, the need for proper management of modern solutions becomes critical. With proper tools and software, established processes, and comprehensive change management, organizations can save time, resources, and better achieve mission effectiveness by harnessing the power of data to become modern, data-driven, insight-driven agencies.

Contact Us

Paul Needleman is a Manager with Deloitte Consulting LLP and based in the Rosslyn office. He has more than 11 years of experience in delivering enterprise data solutions to the Federal Government and currently serves as the Enterprise Data Architect at a federal manufacturing facility. Paul Needleman can be reached at pneedleman@deloitte.com.

Mary Kate Sternitzke is a Consultant with Deloitte Consulting LLP and based in the Rosslyn office. She has experience in delivering analytic solutions through the full SDLC and is currently serving as a functional analytics lead at her federal client. Mary Kate Sternitzke can be reached at msternitzke@deloitte.com.