Create a Data Reference Architecture to Enable Self-Service BI

Published: 5 December 2017 **ID:** G00333398

Analyst(s): Soyeb Barot

Self-service analytics can't be achieved solely with the right set of tools. It must be embedded in the underlying data management architecture. This document gives technical professionals a reference architecture for a multitool BI environment that enables decision making without dependence on IT.

Key Findings

- Most organizations implement multiple business intelligence (BI) tools because of varying capabilities and features desired by the user community.
- A centralized implementation model works best for managing data because it maintains integrity. That is not the case when it comes to BI because it takes agility from users and creates BI silos.
- It is difficult to effectively provide self-service analytics using a single BI tool. Hence, organizations need to embed capabilities for a self-service environment within the overall data management and BI architecture.
- In a multitool BI environment, you run the risk of creating BI silos because the interpretation and implementation of business logic within individual semantics layers can vary.

Recommendations

Technical professionals responsible for implementing and managing BI tools should:

- Evaluate the multiple BI tools implemented and identify capability gaps and overlapping features within the overall architecture. Align BI tools and technologies to business user groups based on their skill sets and functional roles.
- Use a federated implementation model for BI to deliver, manage and scale an effective BI solution by providing autonomy to business users and maintaining enterprise architecture standards.

- Support a self-service analytics platform by implementing the proposed reference architecture to integrate multiple BI tools in a secured and governed manner to enable users to discover and analyze data in real time using a logical data warehouse (LDW) that can store diverse datasets.
- Centralize the core business logic within a data mart or cube to help maintain portability across multiple BI tools and provide uniformity to the information being consumed by applications, reports and dashboards.

Table of Contents

List of Tables

List of Figures

Problem Statement

This document was revised on 8 December 2017. The document you are viewing is the corrected version. For more information, see the Corrections page on gartner.com.

Data is the lifeblood of today's competitive business environment. The goal of every organization is to effectively turn data into information that eventually presents insights to business users. This information helps executives make fact-based business decisions to plan for the future. Data helps them understand business operations to optimize efficiency and improve customer experience. Operating a company without BI derived from data is like flying a plane without instrumentation. Given the importance of data, it is no wonder that organizations are clamoring for new tools, reports and dashboards. "Data-driven" is the new catchphrase, and organizations aren't waiting for IT to deliver BI.

Unfortunately, this race for BI has created a challenge for technical professionals to effectively store, manage, transform and deliver data to end users via tools of their choice. Many organizations begin with rolling out an enterprise BI and reporting tool. As new use cases exceed the capabilities of these tools, companies implement new tools to support new capabilities. Based on Gartner

research, most midsize to large organizations have implemented multiple BI tools to address the reporting and BI needs of their users.

Traditionally, IT has played an important role in delivering analytics projects by providing tools and technologies that power BI. Business users and stakeholders come to IT with a set of requirements for their reporting needs. The IT teams then look at existing tools to deliver the analytics solutions and, in the absence of certain capabilities, buy and deploy infrastructure to support a new BI tool. Analytics projects are use-case-driven, and they come from multiple lines of business (LOBs), which requires the implementation of personalized end-to-end solutions. And to add more complexity to the architecture, most BI tools today are SaaS offerings, which makes it easier for business to implement solutions independent of IT and provides them with agility and faster go-to-market capabilities.

The result of all these factors is a multitool environment. This environment is often complex, difficult to maintain and expensive. Hence, organizations look for ways to simplify and better manage their multitool BI environments. Fundamentally, there are two approaches for tackling the issue:

1. Consolidate the number of BI tools
2. Manage the multitool environment better

The first approach to reducing the number of BI tools requires looking at overlapping features within existing tools. On getting the consolidated list, the tool with the most maturity within the user community and coverage for the most number of use cases across the enterprise is usually selected as the primary. In some cases, this kind of reconciliation exercise even paves the way for the introduction of a completely new BI tool to be considered for the transition. This is a relatively straightforward approach, but it comes with its own set of challenges — a longer migration path with significant impact on business processes and certain unhappy users for whom their preferred BI tool is no longer available.

If the first approach is not an option, then IT is stuck with managing a multitool BI environment. And when it comes to supporting a multitool BI environment, IT runs the risk of creating BI silos within the enterprise architecture. This happens due to lack of governance and a reference architecture to support multiple BI tools.

The primary goal of this document is to provide enterprise and solution architects with a guidance framework to design and build a reference architecture to support a multitool BI environment while addressing the following questions:

- How can we build, manage and scale an architecture comprising of multiple BI tools?
- How do we integrate multiple BI tools?
- How can we provide features and capabilities consistently to all users within a multitool BI environment?

The target audience for this document includes enterprise architects, data and BI solution architects, and analysts. This research can also be leveraged by chief data officers (CDOs), chief

analytics officers (CAOs) and project managers to better understand the feasibility and implementation of a multitool BI environment.

This document focuses on:

- Identifying BI tool capabilities desired by the various user groups within the organization
- Building the prerequisites to support a multitool BI environment
- Designing a federated reference architecture that supports the integration of multiple BI tools to derive maximum business value while reducing operational overhead and maintaining a scalable architecture

The Gartner Approach

We have observed that selecting the one ideal BI tool or a set of multiple tools does not address the problem of catering to the requirements of multiple users within various LOBs of an enterprise. Implementing multiple BI tools in a decentralized model is not the solution to the problem either. It's the architecture that helps integrate a set of BI tools, address business needs and provide a scalable BI platform for IT to manage. In most cases, organizations struggle with managing multiple BI tools due to lack of an implementation model, an integrated cohesive architecture and a well-defined analytics development workflow.

This guidance framework provides a systematic approach for evaluating users based on their expertise, aligning the capabilities they look for within the various BI tools, deciding on the implementation model, building the prerequisites and building the integrated multitool BI solution. The document also provides a walk-through of an iterative analytics and BI project development cycle that will provide autonomy to the users within a governed environment.

Gartner recommends that you follow the sequence laid out within the framework to identify capability gaps, build the required necessary prerequisites and integrate the right set of BI tools to build a self-service analytics platform.

The Guidance Framework

The implementation of the federated BI architecture follows Gartner's familiar guidance framework to plan, build and run the platform.

The Plan and Build stages involve significantly more time and effort than the Run stage. Hence, the first two stages have been stretched across multiple steps. The complete course of action is shown in Figure 1. The Prework step, along with Steps 1 through 3, define the Plan stage. Steps 4 through 6 cover the Build stage. The Run stage is covered by the Use the Analytics Project Development Life Cycle section.

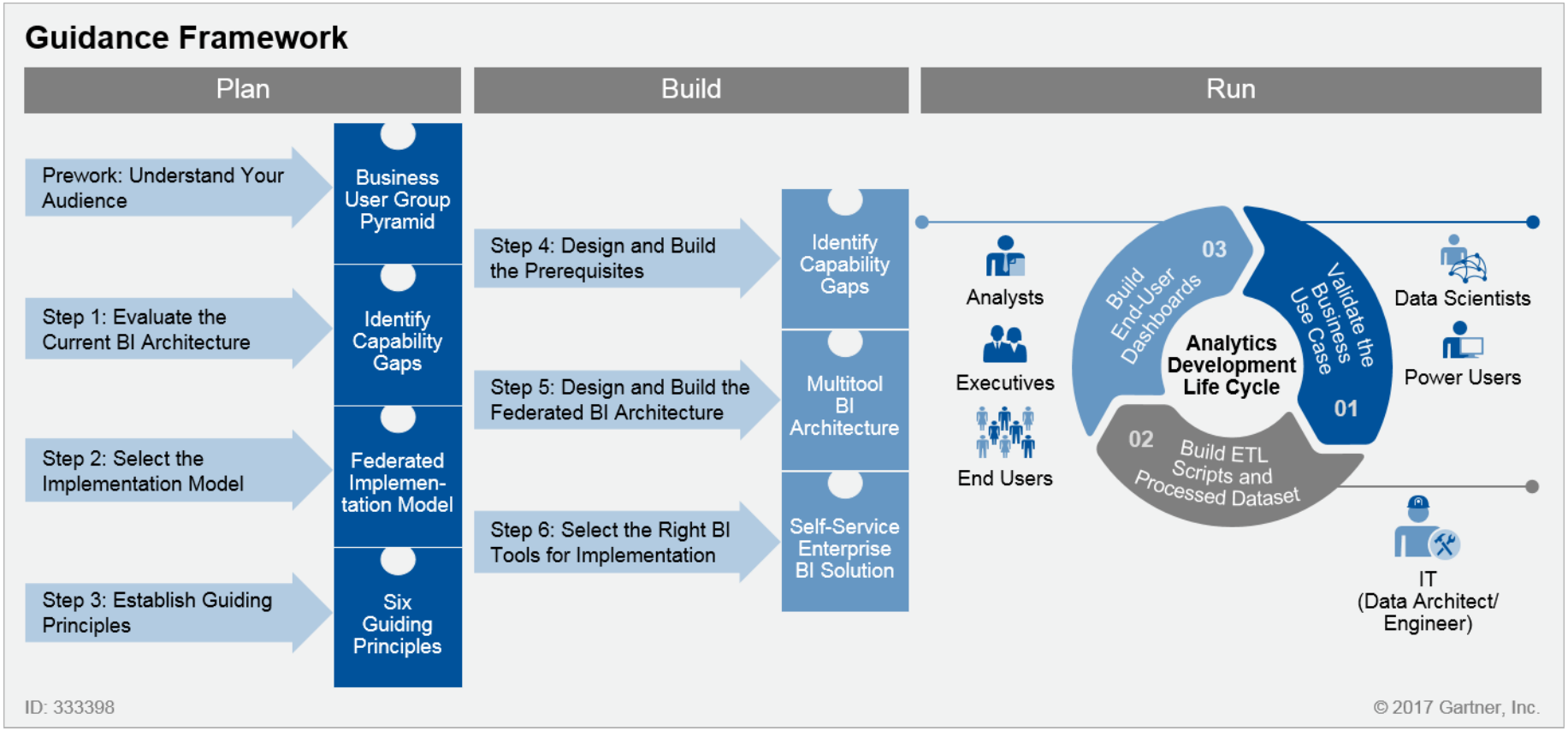
- Plan:

- Prework: Understand your audience and the capabilities they expect from a BI solution.
- Step 1: Evaluate the current BI architecture and identify capability gaps.
- Step 2: Select the right implementation model.
- Step 3: Establish guiding principles for the BI architecture.

- Build:
  - Step 4: Design and build the prerequisites.
  - Step 5: Design and implement the federated BI architecture.
  - Step 6: Select the right BI tools for implementation.

- Run:
  - Use the analytics project development life cycle.

Figure 1. Guidance Framework



Source: Gartner (December 2017)

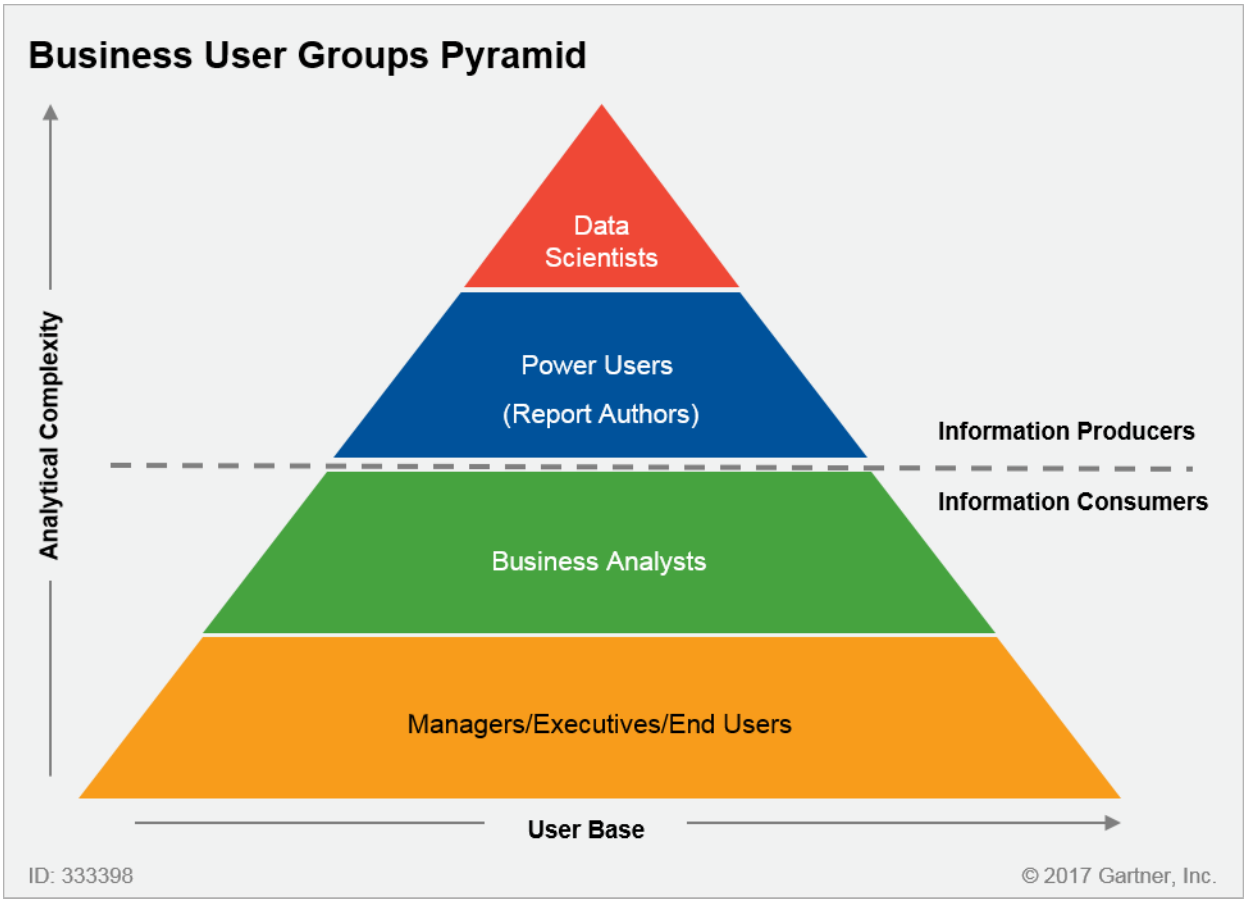
Prework

An architecture must meet the needs of business users and scale a BI solution in an effective and efficient manner. The best place to start before defining a reference architecture is to identify the audience. Furthermore, it is important to understand the function the audience performs within the data analytics life cycle and what its expectations are from a BI tool or solution.

Understand Your Audience

Figure 2 is a pictorial representation of the various business user groups identified by Gartner.

Figure 2. Business User Groups Pyramid



Source: Gartner (December 2017)

A typical organization consists of the following four business user groups, which are further classified into two major categories depending on the functions they perform:

- **Information producers:**
  - Data scientists

    - Power users (report authors)
- **Information consumers:**
    - Business analysts
    - Managers, executives and end users

Information Producers

This category includes data scientists and power users. The individuals within this category help derive valuable insights that can be leveraged by business users. They have a set of unique technical skills, like running queries, transforming multiple datasets, and building statistical models to support descriptive and predictive analysis combined with business acumen and understanding of the domain.

Data scientists or statisticians usually come from a computer science or mathematics background. They are well-versed with languages such as SQL, Python, R and Java and have an understanding of algorithms like linear regression, clustering and random forests. Most of them like to access raw data at its lowest level of granularity.

Power users are often referred to as citizen data scientists. They understand the business domain and data very well and usually come from a management background but are tech-savvy, even with the limited set of technical skills at their disposal. These individuals are very creative when it comes to using basic tools like Excel to transform the data to do their analysis. They are usually hired to focus on data-centric initiatives to help create retention models, investigate market trends and performance metrics, and provide root cause analysis. Along with the data scientists, they too prefer access to the most granular level of raw data.

Apart from full autonomy and access to the raw datasets, these information producers expect the ability to analyze, model, prep and integrate multiple datasets from a BI tool. Additionally, they expect support for building predictive and machine-learning models using advanced analytics libraries.

Information Consumers

Information consumers comprise business analysts, managers, executives and internal/external end users. The individuals within this category prefer prepackaged data for daily operations and standardized reporting for improving business processes. They have a significant understanding of the business domain and act upon the insights provided by the information producers via reports and interactive dashboards. In most cases, the executives and managers provide power users and data scientists with business use cases to help validate certain hypothesis in order to grow the business.

Analysts come from a variety of professional backgrounds with a deeper understanding of the business operations within the enterprise. They help define and create metrics, format the reports, pivot the data and occasionally modify existing reports and dashboards to present data differently.

They have some level of understanding of SQL and are well-versed with using reporting and BI tools at an operational level.

Data — transformed from its most granular level to the prepackaged format — is delivered to this group of managers, executives and end users. Information consumers have the final reports or access to dashboards where data is periodically refreshed and then have the ability to export the data or the graph to Excel, Word, PowerPoint or a PDF file.

In terms of capabilities, the expectations of the information consumers from the BI tool are the ability to filter, sort, pivot, visualize and export the data to a preferred format or file. They like to analyze the data to a certain depth and hence look for drill-down capabilities or the ability to augment the data, but they always prefer a refined dataset.

As highlighted in Figure 2, the analytical complexity is higher toward the top half of the pyramid than the bottom. But when it comes to a number of users, the majority of individuals fall within the bottom half of the pyramid. The information consumers make about 75% to 85% of the overall strength of the enterprise. Information producers represent the remaining 15% to 25% of the enterprise. These numbers can vary depending on the organization, corporate culture and the kind of industry the enterprise is associated with. Organizations within insurance, finance, healthcare, telecommunications and advertising media tend to have higher percentages of data scientists and power users as compared to analysts and managers.

Figure 3 summarizes the functional roles of the individual user groups and the capabilities they expect out of the BI solution.

Figure 3. Functional User Roles Mapped to Desired BI Tool Capabilities

Functional User Roles Mapped to Desired BI Tool Capabilities

Categories	Groups	Functional Roles (Skill Sets)	Desired Capabilities
Information Producers	Data Scientists Power Users	• Ad hoc queries • Data wrangling • Data modeling • Advanced and predictive analytics • Dashboard and report authors	• Online and offline data integration • Self-serviced data prep • Support for data modeling • Advanced analytics and statistical libraries • Report and dashboard authoring capabilities

Information Consumers	Business Analyst
	Executives, Managers, End Users
		• Active/passive users
		• Schedule refreshes and automated delivery
		• Require access to:
		  • Standard reports
		  • Interactive dashboards
			• Curated datasets (marts, cubes, warehouses)
			• Automated data refreshes and scheduling
			• Filtering/drill-down capabilities
			• Visualization and exporting results (PDF, Excel)

ID: 333398

Source: Gartner (December 2017)

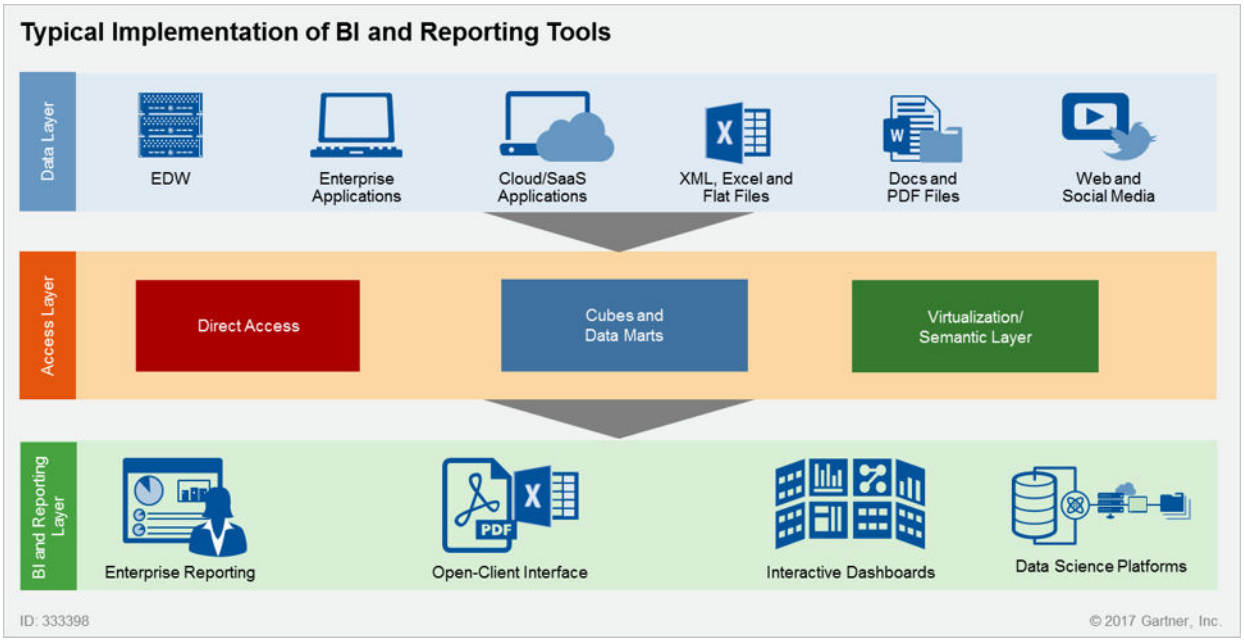
Having completed the prework, identified the users and the capabilities they look for with BI solution, we can now proceed with the steps laid out within the guidance framework.

Step 1: Evaluate the Current BI Architecture

As part of the first step, architects and business partners should evaluate the current BI architecture and identify potential capability gaps. This involves doing a thorough assessment of the data management system, performing integration of various BI tools and understanding the flow of data across various systems. We should look at data integration, metadata management, governance, security and — most importantly — data movement between systems. While evaluating the individual components or modules of the architecture, the focus should be on the capabilities rather than the specific features the tool or the technology provides. This will help create an end-to-end visual of the set of tools, capabilities and interaction points of various user groups within the entire data analytics life cycle. It also presents a unique opportunity to look at overlapping capabilities that some tools provide and can facilitate consolidation efforts within the existing architecture.

Figure 4 represents a typical implementation architecture of BI tools across most organizations.

Figure 4. Typical Implementation of BI and Reporting Tools



Source: Gartner (December 2017)

The architecture consists of three distinct layers with a unidirectional data flow.

Data Layer

The data layer comprises all of the sources and data stores that exist within an organization:

- **Enterprise data warehouse (EDW):** The EDW is a central repository of integrated datasets from one or more disparate sources. It stores both current and historical data in a single data store.
- **Application system data stores:** These are the operational data stores (ODSs) for the individual enterprise applications including third-party vendor apps and those built in-house.
- **Cloud and SaaS applications:** Data stored in cloud-based ERP or CRM solutions is a form of ODS that do not reside on-premises within the organizations' data center.
- **Data within Excel, CSV, document and PDF files:** This comprises data sitting in files stored on shared network drives or content management applications like SharePoint.
- **Web and social media:** These market data sources are made available in the form of CSV files that are delivered via FTP, Secure Shell FTP (SFTP) or application programming interfaces (APIs).

Access Layer

The access layer usually consists of a set of tools and technologies that provide an interface to query the underlying data defined within the data layer. Each organization has a different approach to providing access for users.

- **Direct access:** via reporting and BI tools or applications that use a set of Open Database Connectivity (ODBC), Java Database Connectivity (JDBC) or application programming interfaces (APIs) to query the data.
- **Data marts and cubes:** Data marts or cubes are subsets of a data warehouse that are oriented toward the needs of specific LOBs within the organization (e.g., finance, marketing and HR). These data marts and cubes are built on traditional relational data management systems (RDBMSs) like Oracle and Microsoft (MS) SQL Server or in-memory databases like SAP Hana, Teradata and MemSQL.
- **Virtualization/semantic layer:** A semantic or virtualization tool helps map complex data to familiar business terms like "product," "customer" and "revenue." This provides a unified, consolidated view of the data across the organization. These layers help to create a logical model that combines data from multiple sources on top of a disparate physical data schema. Examples are products like MicroStrategy, SAP Business Objects (BO) Universe and AtScale or virtualization tools like Composite, Red Hat's JBoss and Denodo.

Depending on the scale and the number of use cases an organization needs to support, access to data is provided to the users in one of the ways listed above. The guidance framework specifically focuses on this layer to address the risks of creating data silos within the implementation of multiple BI tools.

BI and Reporting Layer

Enterprise BI and reporting tools are implemented within this layer to provide end users with the capability to access the data via a graphical user interface (GUI). This is where all business users

query, analyze and consume the data through reports and interactive dashboards. Hence, it is referred to as a consumption layer. The individual tools provide the following capabilities to the end users:

- **Enterprise reporting platform:** This provides standardized and ad hoc reporting for regulatory, compliance and daily operational business processes. Typical implementations include SAP BO, OBIEE, IBM Cognos Analytics or MS SQL Server Analysis Services (SSAS).
- **Open-client interfaces:** Access to data is provided via Excel or an open-client connection using ODBC or JDBC drivers. Here the users can establish a connection to a specific data store or data mart, extract it within the tool, and then transform it for further analysis. They use inherent functions and graphics libraries with Excel to build reports.
- **BI tools and dashboards:** BI and visualization tools provide an interactive interface for users to analyze and consume the data in the form of tables, graphs and dashboards. Tools like Tableau Software, Qlik and MS Power BI provide these capabilities to the users.
- **Data science platforms and notebooks:** Products such as Jupyter and Apache Zeppelin allow data scientists to access the data within their machine learning algorithms written in languages such as Java, Scala, Python and R.

An implementation of this format, even though streamlined, presents business users and IT operations with a lot of challenges. The following section looks at some of these challenges, and later sections show how the new federated BI architecture helps mitigate the risks.

Identifying Challenges With the Typical Implementation

- Building and maintaining multiple data pipelines for the various tools within the consumption layer and eventually ending up with a spaghetti of extraction, transformation and loading (ETL) processes
- Designing and maintaining multiple data schemas for each LOB within the enterprise and at times even for individual use cases
- Dealing with the significant amount of data movement that occurs between source systems, EDW, BI and reporting platforms
- Avoiding the creation of data silos within individual BI solutions
- Avoiding delivering users with multiple versions of truth using the same underlying data
- Embedding business logic and rules across multiple BI and reporting tools that may lead to reporting inconsistencies and governing changes

All of these make it extremely difficult to secure and scale the analytics and BI architecture. It provides an inconsistent way for users to access the data and creates an operations nightmare for IT.

Having completed the assessment and identified the gaps in the current implementation, we can look at the newly proposed framework that will help address all of these challenges and provide

business users with a single, secured and governed analytics platform using their preferred tool of choice.

Step 2: Select the Implementation Model

When it comes to managing data across multiple systems and providing it to users for analysis, the centralization model has proven to be the most effective. Hence, we see the implementation of EDWs and data lakes or data hubs. Information from multiple source systems is brought together in one single location. Users are then provided access to the data via multiple tools and applications, including BI tools. This centralization approach helps improve efficiency in managing the data, improves data quality and helps govern it.

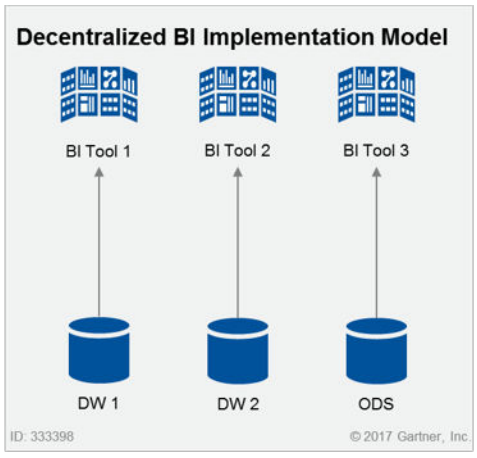
It makes complete sense to implement the centralized model for data management. But can this model be applied toward implementing a BI solution as well?

It's just not possible to scale a centralized analysis platform or architecture. It is impossible to keep hiring enough people with expertise within a specific BI tool and expanding the underlying infrastructure to support it. The next section discusses the challenges and risks associated with the centralized model.

In Step 2, we will look at the various implementation models and then select the best one for implementation of the multitool BI architecture.

Decentralized Model

Figure 5. Decentralized BI Implementation Model



DW = data warehouse; ODS = operational data store

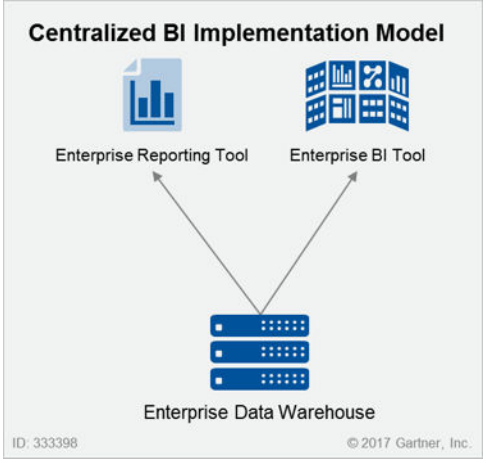
Source: Gartner (December 2017)

Figure 5 is a good representation of a decentralized implementation model. Each individual BI or reporting tool has its own end-to-end implementation with data flowing all the way from source to consumption via the access layer.

Most organizations facing challenges managing a multitool BI environment have a decentralized implementation model. Each LOB provides its own set of BI requirements, and IT — using the business users' tool of choice — helps design and implement the solution. These end-to-end deployments are quick, at times affordable and tailored for individual use cases. Because the implementations are business-user- and use-case-driven, over a period of time, the architecture will need to accommodate multiple BI tools and create redundant ETL processes. This eventually creates BI silos because the level of expertise and data consumed via individual end-to-end implementations varies. There is also a significant impact on the way data is interpreted, and this diminishes the quality and consistency of delivering data to the end users. The same data coming from the source systems gets processed multiple times — each time with a different set of business rules and logic — which generates conflicting outputs for users.

Centralized Model

Figure 6. Centralized BI Implementation Model



Source: Gartner (December 2017)

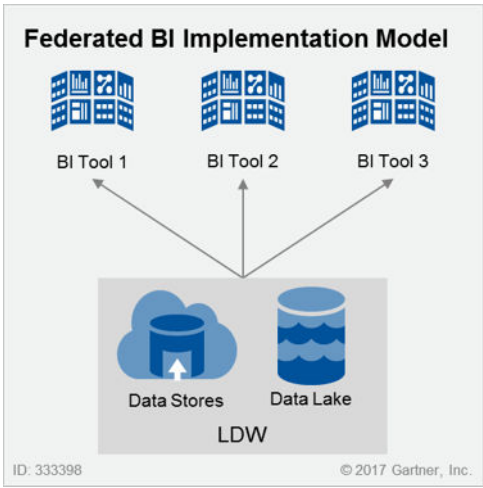
The centralized implementation model (see Figure 6) helps reduce redundancy of processes and tools, so enterprises tend to lean toward a centralized analytics architecture. It surely helps mitigate some of the risks associated with the decentralized model and helps create an enterprise view of the data.

One of the key components to implementing this model requires IT to build an EDW. It looks like the right approach because it helps create standardized data definitions, and managing and governing a centralized data hub is relatively easy. But when it comes to BI tools and analytics, it creates significant bottlenecks for the user community. The centralized model comes with its own set of challenges if an organization decides to reconcile its set of multiple BI and reporting tools to just one enterprise tool. The implementation requires a significant amount of restructuring of the existing BI architecture. IT teams need to rebuild the ETL scripts to support the migration of business rules and logic for populating the new centralized data schema. Data architects have to design and build an evolving physical and logical data model to accommodate new data sources and business

requirements in the future. And lastly, new dashboards that mimic the features and representation of the data as provided via the BI tools being replaced must be built.

Federated Model

Figure 7. Federated BI Implementation Model



Source: Gartner (December 2017)

The federated model (see Figure 7) provides the best of both worlds by addressing the weaknesses presented within the centralized and decentralized models when designing an analytics BI solution.

> A federated model is a pattern within the enterprise architecture that allows interoperability and information sharing between semiautonomous noncentrally organized LOBs, IT systems and applications.

Within a federated implementation model, the corporate BI team would open up its data warehousing environment to all of the divisions within the enterprise. Each division or LOB would have its own dedicated partition in the EDW or data lake to develop its own data marts, reports and dashboards. IT teams provide self-service data preparation tools and train business users on how to blend local data with corporate data inside their EDW or data lake partitions. For groups that have limited or no BI expertise, the corporate BI team would continue to build custom data marts, reports and dashboards as before.

The federated model provides a perfect blend of both the top-down and bottom-up approaches:

- Standardized scalable architecture
- Support for an extensible physical and logical data model

- The centralized data repository, with dedicated partitions and zones for individual LOBs
- Consistent data definitions
- Agility to deliver multiple end-user analytics products and services
- A stronger partnership between IT and business
- True support for self-service governed analytics platform

This model does come with its own unique set of challenges, but they are more process-dependent rather than being limited by the design of the architecture. Overcoming these challenges requires building a stronger communication line between the various business groups, users and IT. The success of the federated implementation model requires all constituents to adhere to some basic operating standards when using the enterprise solution. In order to succeed, enterprises need to establish a clear roadmap and robust business-led governance program. It would help to establish a BI center of excellence (COE) and for IT to relinquish control back to business when it comes to data analytics.

> The BI and analytics market is in the final stages of a multiyear shift from IT-led, system-of-record reporting to business-led, self-service analytics.[1]

Refer to "The 30 Capabilities That Your Analytics Center of Excellence May Be Lacking" for more details on establishing a BI COE.

Table 1 summarizes the strengths and weaknesses of the three implementation models.

Table 1. Summarizing Strengths and Weaknesses of Implementation Models

Model	Strengths	Weaknesses
Decentralized BI	■ Agile delivery ■ Greater flexibility ■ Customized views	■ Siloed BI solutions ■ Inconsistent BI architecture ■ Conflicting data definitions ■ Operational and licensing cost overhead ■ Disparate set of capabilities
Centralized BI	■ Eliminates redundant costs ■ Consistent data definitions ■ Presents an enterprise view of the data ■ Standardized, scalable architecture	■ Introduces bottlenecks ■ Single data model ■ Restricts exploratory analysis ■ Creates significant backlogs
Federated BI	■ Standardized, scalable architecture ■ Extensible physical and logical data models ■ Centralized data and metadata repository ■ Supports self-service across the data management architecture ■ Agility and modularity to grow the analytics practice in an organic fashion	■ Requires strict adherence to the process and guidelines laid out within the implementation model

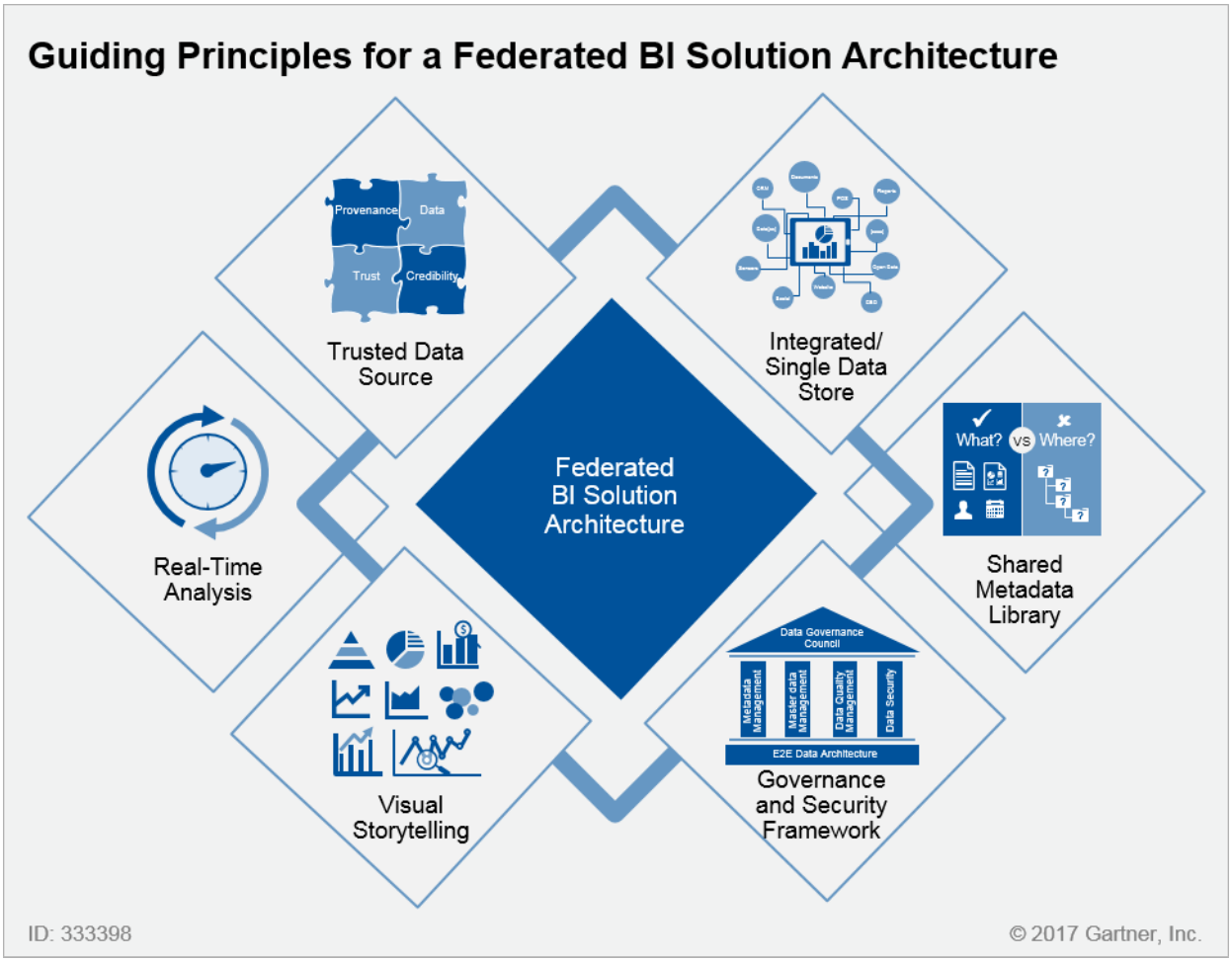
Source: Gartner (December 2017)

Step 3: Establish Guiding Principles

Before we start designing the federated BI architecture, we first need to establish a set of guiding principles to help us identify the set of core capabilities to be considered within the architecture.

Figure 8 highlights the six guiding principles identified within this framework to build an enterprise BI solution.

Figure 8. Guiding Principles for a Federated BI Solution Architecture



Source: Gartner (December 2017)

1. **Integrated single data store:** Users need the ability to start the data analysis from its most granular format and roll it up to a refined, curated dataset that provides actionable insight. An LDW would help provide a centralized data hub to store all the data, help identify new relationships between data elements and reduce data movement across multiple data stores and analytical systems. Hence, this is a core foundational architectural component of a federated BI solution.

2. **Shared metadata library:** Another key component of the architecture is building a centralized business glossary that helps standardize data definitions. Metadata, defined as the set of data that describes and provides context about the data elements can help support discovery and self-service analytics. Providing the ability to search and filter an enriched metadata library helps users locate relevant information to use for further analysis. This metadata library can be created via an automated process using profiling tools, or it can be created manually with the help of data stewards.

3. **Governance and security framework:** Data Governance is a four-way framework that helps ensure integrity, security, applicability and access to reliable data. The governance framework will provide a set of processes, implemented and used by stakeholders leveraging technology, to ensure that critical data is protected and well-managed. This becomes significantly more important when building an LDW that is built upon a data lake. If the data is governed properly at the physical layer, it reduces the complexity of managing access for the users from within the BI tools.

4. **Trusted data source:** The combination of the LDW with a metadata library that is both governed and secured presents a trusted data source for the users to do their analysis. This provides them with a unique set of capabilities and promotes self-service analytics across the enterprise with less reliance on IT. Users now have one single location to fetch pristine quality data. This supports ad hoc analysis and standardized operational reporting and even advanced analytics and machine learning in the future.

5. **Real-time analysis:** Information about consumers and sales received in real-time is more relevant than from last month or even last week. It is more useful to your business users because it gives them a true understanding of what is happening within their business operations — as it happens. Hence, access to data being generated in real time by systems has become an important consideration to maintain an enterprise's competitive edge. How else can a business provide an accurate picture of where it stands at a given point, both in terms of its internal decision making and to satisfy its regulatory requirements.

6. **Visual storytelling:** Once you have provided the capabilities to process the data and extract value, it becomes extremely important for users to communicate insights back to the decision makers. What better way to do it than by visually representing it in the form of graphs and charts. This is where the tools and architecture come into play. The architecture should support discovery, self-service data prep, and modeling alongside an interactive way to visualize and deliver the data to the end users.

Plan Stage Summary

We have now come to the end of the Plan stage of the guidance framework. Let's summarize what we have discovered so far from the outcomes of the Prework and Step 1 through Step 3.

- Each organization has a set of four business user groups classified within two categories as information producers and information consumers that have their own set of capability requirements from a BI solution.
- IT teams will need to support a multitool BI environment to serve the needs of the various business user groups.
- A federated implementation model is an ideal approach to supporting business analytics.
- There are six guiding principles to be considered when designing the reference architecture.

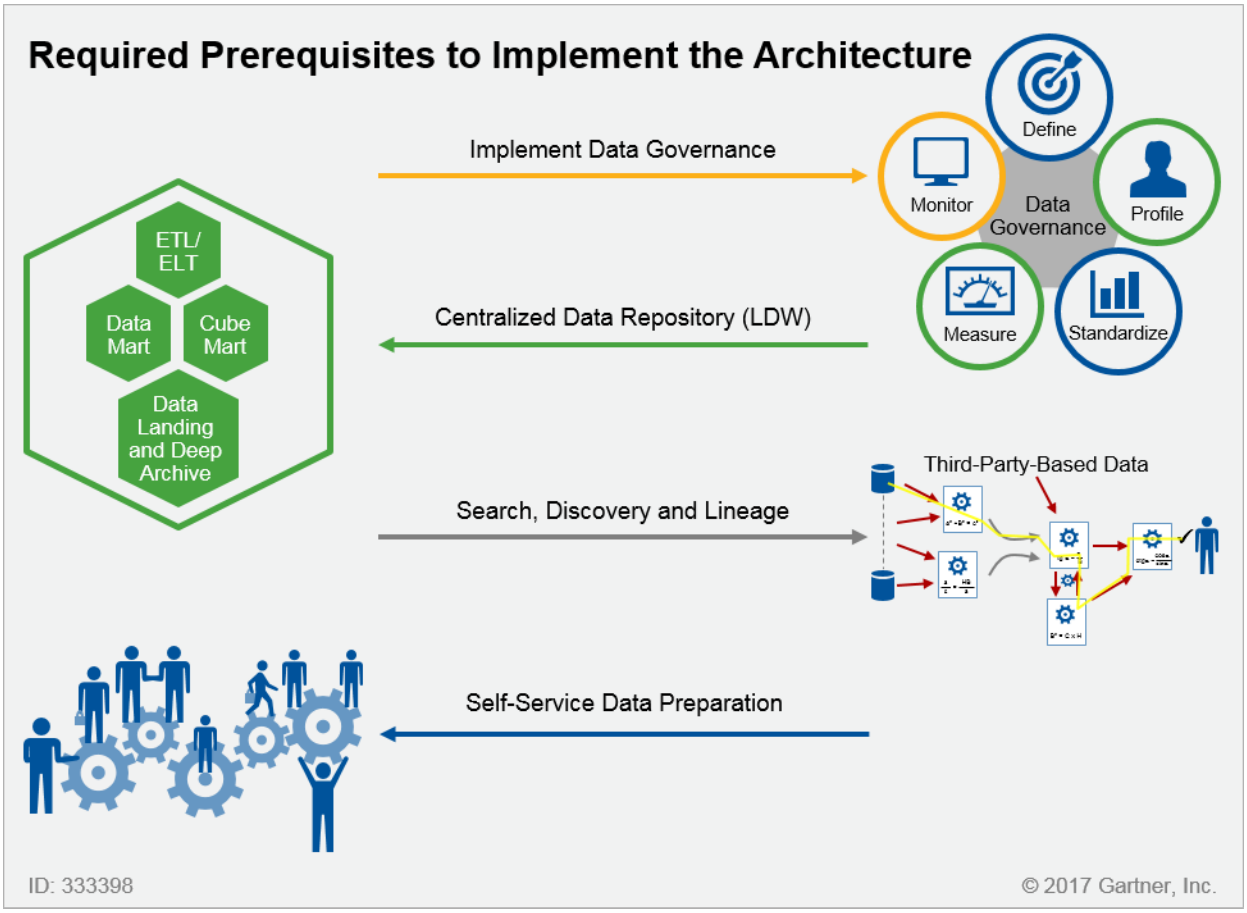
We will now enter the Build stage of the guidance framework. First, we will design and build the prerequisites in Step 4, then build the federated BI architecture in Step 5 and finally select the right BI tools in Step 6.

Step 4: Design and Build the Prerequisites

In order to implement the federated BI architecture, it is important to design and build the required components, based on the guiding principles defined in Step 3.

Figure 9 highlights the four prerequisites required to build the architecture.

Figure 9. Required Prerequisites to Implement the Architecture



Source: Gartner (December 2017)

1. **Implement data governance framework:** "*Data Governance is a system of decision rights and accountabilities for information-related processes, executed according to agreed-upon models which describe who can take what actions with what information, and when, under what circumstances, using what methods.*"[2] Implementation of a successful data governance practice within an enterprise is 70% process and 30% technology. A typical data governance includes:

   - **Aligning and defining:** This should include the business strategy and how to leverage data to solve the business problem.

- **Policy and processes:** This helps manage data effectively and maps processes to the roles and responsibilities of all the constituents (business and IT) within a data management life cycle.
- **Definition of roles and responsibilities:** These are key components as part of the stewardship to help monitor and control the flow of data within the enterprise. Hence, defining a clear set of roles and responsibilities helps implement a complex governance model that involves various business and IT user groups.
- **Tools and technologies:** Selection of the right set of tools and technologies helps operationalize a governance framework that includes:

    1. **Data security:** These tools can help locate sensitive data elements across multiple systems. They can also enforce regulatory, contractual and architectural compliance, support access management and detect security breaches.

        Refer to "Securing the Big Data and Advanced Analytics Pipeline" for further guidance.

        Table 2 shows a list of data security products in alphabetical order.

Table 2. Data Security Tools

Vendors	Products
BlueTalon	BlueTalon
Dataguise	DgSecure Protect
Micro Focus (Hewlett Packard Enterprise [HPE])	Voltage SecureData
Protegrity	Big Data Protector, Database Protector and Mainframe Protector
Thales e-Security	Vormetric Data Security Platform

Source: Gartner (December 2017)

2. **Data quality:** These tools support data quality and are an absolute prerequisite to ensure successful implementation of the governance framework. The tool should provide profiling, cleansing, merging and visualization capabilities and help to create a workflow to maintain data in its pristine state.

    Refer to "Agile Data Quality to Maximize Your Business Results" for further guidance on data quality tools.

    Table 3 lists some of the data quality tools to be considered for evaluation.

Table 3. Data Quality Tools

Vendors	Products
Ataccama	Data Quality Analyzer
IBM	InfoSphere Information Server
Informatica	IDQ
Melissa	Data Quality Suite
Microsoft	Data Quality Services (DQS)
Oracle	Enterprise Data Quality
SAP	Data Services
Syncsort	Trillium
Talend	Data Quality Management Tool

Source: Gartner (December 2017)

3. **Metadata management:** The definition of data governance implies effective management of data, which is impossible without managing its associated metadata. The selected tool should support manual and automated tagging of data to specific business domains alongside an ability to search and steward the underlying data.

   Table 4 lists some metadata management tools that can be considered for implementation.

Table 4. Metadata Management Tools

Vendors	Products
Alation	Alation
Collibra	Data Governance Center
Datum	Information Value Management
IBM	InfoSphere Metadata Workbench
Informatica	Metadata Management
Oracle	Enterprise Metadata Management
Smartlogic	Semaphore

Source: Gartner (December 2017)

You can also establish a data governance framework by selecting tools supporting the individual capabilities. However, there are tools available in the market that provide end-to-end data governance solutions. Listing the tools below in alphabetical order, the selection for the appropriate one should be done after conducting a comprehensive proof of concept (POC) based on your evaluation criteria. Table 5 lists vendor products that provide complete end-to-end solutions for implementing a data governance platform.

Table 5. Data Governance Platforms

Vendors	Products
Cask	Cask Data Application Platform (CDAP)
Cloudera	Navigator
Collibra	Data Governance Center
Infogix	Data3Sixty Govern
Informatica	Axon
Apache	Atlas (open source)
Podium	Data Marketplace
Talend	Data Fabric
UniFi	UniFi
Waterline Data	Data Catalog
Zaloni	Data Lake Management Platform

Source: Gartner (December 2017)

2. **LDW:** The LDW is a growing data management architecture for analytics that combines the strengths of traditional repository warehouses with alternative data management and access strategy — specifically federation and distributed processing. It constitutes establishing a centralized trusted data source for users to perform their analysis. The LDW will help store the active raw data, historical processed and raw datasets, semistructured and unstructured data, and curated datasets in the form of data marts and cubes. The key here is to build a data management system that reduces the amount of data movement across multiple systems.

   Refer to "Solution Path for Planning and Implementing the Logical Data Warehouse" for further guidance on implementing an LDW.

   The LDW, sometimes referred to as a data lake or a data hub, can be built using a set of relational, NoSQL, data virtualization and distributed storage and computing frameworks.

Depending on the scale of the implementation, the LDW can be built using a single persistent data store.

Refer to "Identifying and Selecting the Optimal Persistent Data Store for Big Data Initiatives" for further guidance on selecting the optimal data store to build the LDW.

3. **Search, discovery and lineage:** Once you have established the ETL pipeline to bring data from the various source system to the LDW, it is important for business users to be able to search the data, discover business context using metadata tags and confirm lineage before using the datasets for analysis. Having a tool or technology to support this not only helps build the federated BI architecture, but also promotes self-service analytics. Users don't have to rely on IT to provide them access to a table, view or file with the most updated snapshot of the dataset.

   Table 6 provides a list of stand-alone data cataloging tools providing search, discovery and lineage capabilities within the architecture.

Table 6. Data Cataloging Tools

Vendors	Products
Alation	Data Catalog
Attivio	Attivio
Waterline Data	Waterline Data Catalog
Zaloni	Data Lake Management Platform

Source: Gartner (December 2017)

4. **Self-service data prep:** The primary goal of implementing a modernized self-service analytics platform is to empower business users' capabilities to derive faster insights from the data. The biggest bottleneck for any analytics project is the data preparation phase, which accounts for more than 65% of the time required to deliver data in a consumable format. The key for a self-service data preparation platform is to provide users the ability to combine, cleanse and transform relevant data for further analysis (see Table 7).

Table 7. Data-Prep Tools

Vendors	Products
Alation	Alation
Alteryx	Designer
Datameer	Datameer
IBM	SPSS Data Preparation
Informatica	Real-Time Integration Products
Oracle	Data Prep Editor
Paxata	Self-Service Data Prep Application
Trifacta	Wrangler Enterprise

Source: Gartner (December 2017)

Step 5: Design and Build the Federated BI Architecture

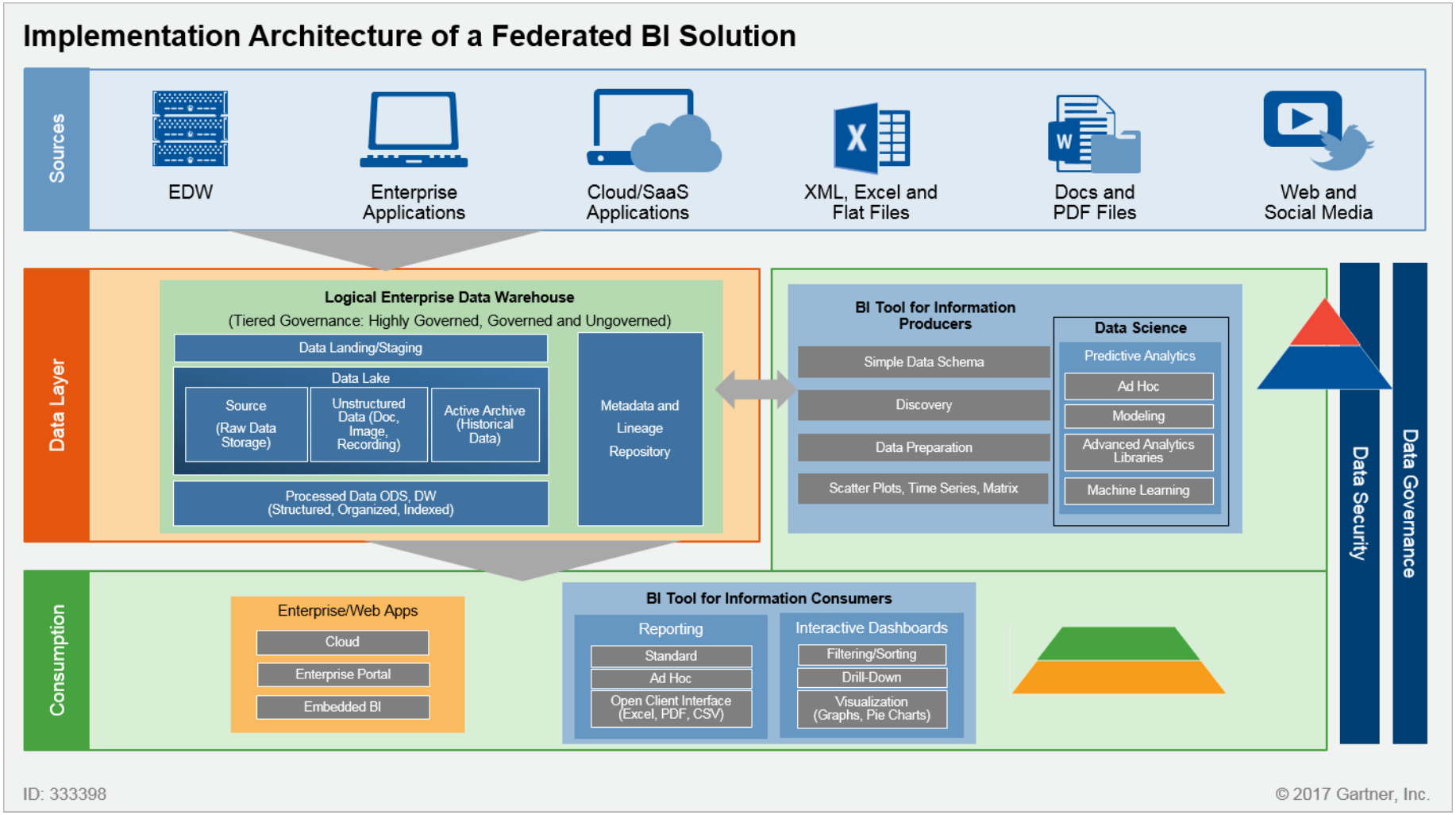
All of the analysis done until this point indicates that, in order to implement a scalable multitool BI architecture, we need to align the capabilities of the individual tools to the requirements of the user groups identified within the pyramid in Figure 2. Hence, the BI solution architecture should be built on a capability model.

> A capability model provides a comprehensive set of features an enterprise requires via the integration of individual tools or a single platform, in order to execute its business model. In terms of designing a BI solution, think of the capability model as a set of features and functionalities the users look for based on their level of expertise and technical skill sets within a given tool.

The architecture design should also factor in the federated implementation model and take into consideration the guiding principles identified as part of the initial analysis within the Plan stage.

In order to build the reference architecture, we are making the assumption that the organization has selected two BI tools for implementation. Each tool individually caters to the user groups within the two major categories: producers and consumers. However, this architecture is extensible and can support more than two BI tools. Implementation of more than two BI tools using the same reference architecture is covered in the Can We Support More Than Two BI Tools? section.

Figure 10. Implementation Architecture of a Federated BI Solution



Source: Gartner (December 2017)

Figure 10 shows the federated BI architecture supporting two BI tools.

The architecture consists of three main layers: source, data and consumption. Data from all sources needs to be brought into the LDW, which is implemented within the data layer. The LDW, often referred to as a data hub, can be implemented using a traditional RDBMS, NoSQL databases, Hadoop or a combination depending on the type of data (structured, semistructured or unstructured) that needs to be stored. This centralized hub should hold data in its most granular format in addition to the preprocessed data in the form of data marts and/or cubes.

The LDW should be divided into high-level zones:

- **Landing/staging:** This is where data from the source systems is pulled in and stored as is. This process can be implemented using the enterprise data integration tool. No transformation should be performed on the data during the process of moving it from source to destination, but you need to ensure that all of the metadata is collected and stored in the centralized metadata repository to provide lineage.
- **Data lake:** Gartner defines a data lake as a collection of storage instances of various data assets in addition to the originating systems. Data is stored in a near-exact format as its source. The data lake can be further partitioned to hold structured raw source data, semistructured and unstructured data, and a historical archive. Most organizations have to abide by compliance regulations set by the government or industry, so it is critical to secure the data. Hence, when moving data from landing to either of the partitions within the data lake, sensitive data elements should be either encrypted, masked or tokenized.
- **Processed zone:** This particular zone would hold the data marts for the individual LOBs, use cases and even golden copies of business domain reference data like the client and product master. Data stored in this zone can be highly indexed to support faster reads via enterprise applications and multiple reporting and BI tools.
- **Metadata repository:** The metadata repository can act as a business glossary or catalog for users to access relevant data within the data lake, so it should be stored within the LDW. Metadata should be collected from the acquisition of data from originating systems as it flows through the various zones or partitions and is transformed. This repository is a critical component within the architecture to help support self-service analytics due to the search, discovery and lineage capabilities it provides to the end users.

> The LDW will help reduce the amount of data movement between various systems for analysis by business user groups.

Now let's take a look at the integration of BI tools within this architecture. First, we begin with associating BI tools based on the features they provide to individuals within the respective categories as identified in Figure 2.

- **BI tool for information producers:** The BI tool for information producers should provide features and capabilities that the information producers look for, such as data discovery, ad hoc data modeling, self-service data prep and support for advanced statistical libraries. This BI tool

should be integrated into the data layer to facilitate access to the raw granular datasets for data scientists and power users. Individuals within this category would eventually help validate hypotheses and use cases for businesses and executives. Using their preferred BI tools of choice, they can provide valuable insights based on trends and patterns they see when combining data from various sources in the LDW. As part of their analyses, these individuals can help create the business and ETL logic for the resultant analytics report and dashboard.

The resultant data mart or cube should be built at the data layer with the help of data engineers within IT, rather than within the information producers' BI tool's semantic layer. This helps with portability and reusability of the data via another BI tool or even an enterprise application.

How this can be achieved is explained within the Use the Analytics Project Development Life Cycle section.

- **BI tool for information consumers:** The second BI Tool is for information consumers and should be integrated into the consumption layer within the architecture. It could be an enterprise BI or reporting platform that provides access to the processed data in the form of data marts for information consumers (business analysts, managers, executives and end users). The tool selected needs to provide individuals in this category with features and capabilities to slice and dice the data, pivot it and provide visualization capabilities in the form of charts and graphs. Data visualization is the crystallization of numeric algorithmic outputs into images. Images help provide clarity to the message being delivered by the underlying data. The preprocessed, curated datasets built by the information producers can be leveraged by enterprise web and mobile applications alongside the information consumers' BI tool and integrated within the consumption layer. The level of drill-down and access to granular data can be set within this BI tool to provide complete lineage if required by the end users. The ability to drill down and roll up for aggregation is possible within either of the BI tools implemented within the different layers because all of the data resides within the LDW.

The selection of the BI tools based on the requirements and criteria laid out for the individual user groups is covered in Step 6.

Can We Support More Than Two BI Tools?

It is possible to implement and support more than two BI tools using the same reference architecture. The number of BI tools to be implemented is usually dictated by the business requirements and scale of the organization. In order to accommodate more than two BI tools, we need to establish the ideal integration point within the above reference architecture. This can be achieved by aligning each BI tool to the user groups that will leverage its inherent capabilities for accessing the data.

Let's take an example. If the team of data scientists and power users prefer two different BI tools, then you should integrate the two BI tools at the data layer within the architecture. This way, you are supporting the individual user groups and letting the architecture scale in a federated manner. Each user group now has its own preferred BI tool for accessing the data.

Similarly, you can have multiple BI and reporting tools integrated into the consumption layer for managers, executives and internal/external end users. This federated BI architecture helps support a bring your own tool (BYOT) model and scale the enterprise analytics platform.

Step 6: Select the Right BI Tools for Implementation

When it comes to selecting the right BI tools for implementation at either the data production or the data consumption layer, we need to look at the critical capabilities for each of the tools from a technical architecture perspective.

The top 10 critical capabilities to consider are shown in Table 8.

Table 8. Critical Capabilities for BI Tools

Capabilities	Description
Admin, security and architecture	Capabilities that enable platform security, administering users, usage monitoring, auditing platform access and utilization, optimizing performance, and ensuring high availability and disaster recovery
Data source connectivity	Native support for a broad range of connections to query structured and unstructured data contained within various types of storage platforms, including personal data sources, relational, NoSQL and direct Hadoop Distributed File System (HDFS)
Metadata management	Enables users to leverage the same system-of-record semantic model or create a semantic model automatically
Interactive visual exploration	Enables exploration of data via an array of visualizations that go beyond basic pie, bar and line charts, to include trellis, heat and tree maps, and scatter plots
Ease of use and visual appeal	Ease of use to administer and deploy the platform, create content, and consume and interact with content — all with visual appeal
Self-contained ETL	Platform capabilities for accessing, integrating, transforming and loading data into a self-contained performance engine with the ability to index data and manage data loads and refresh scheduling of loaded data
Self-service data preparation	Drag-and-drop cleansing, modeling, and blending of multiple data sources and creation of analytic models
Smart data discovery	Enables exploration of data via visualizations, autogenerated voice or text narration, search, and natural language query technologies
Platform and workflow integration	Provides capabilities offering a single, seamless product with little work required for integration with other data management components
Embedded advanced analytics	Enables users to easily access advanced analytics capabilities that are self-contained within the platform, or are imported from and integrated with externally developed models

Source: Gartner (December 2017)

For criteria to consider when evaluating the BI tools, refer to "Evaluation Criteria for Business Intelligence and Analytics Platforms."

The critical capabilities will help support the requirements for scaling and integrating the BI tools with other data management components. The feature requirements of the users within the business groups can be addressed by taking into consideration the evaluation criteria of the BI tools.

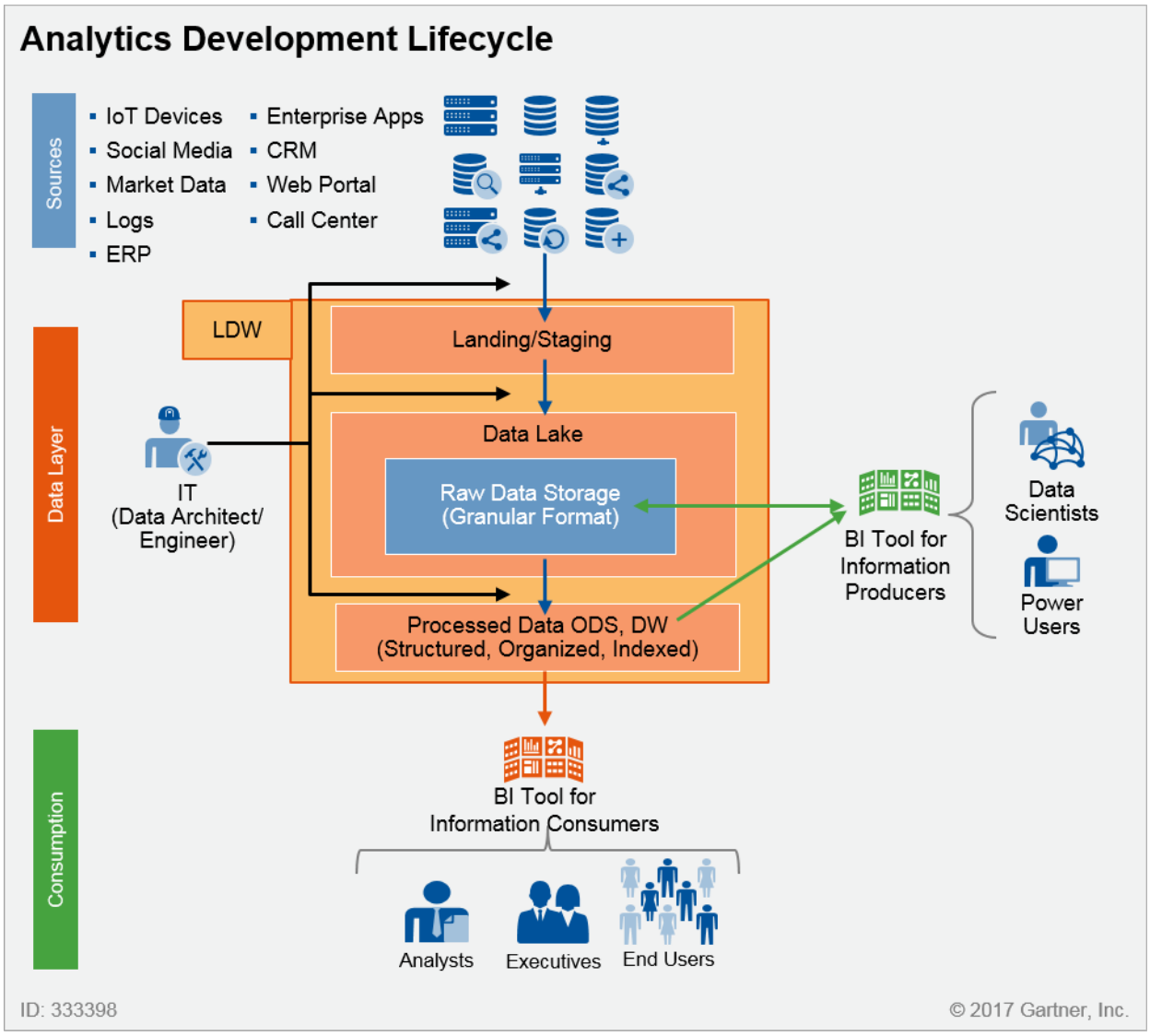
Most organizations already have a set of BI tools, licenses and infrastructure in place. Gartner suggests that IT teams can rearchitect the implementation of the current BI solution based on the newly proposed federated architecture. It is not required to bring in a set of new tools as long as the prevailing BI tools provide all of the capabilities that the business users are looking for.

With this, we come to the end of the Build stage of the guidance framework.

Use the Analytics Project Development Life Cycle

Having designed and built the multitool federated BI architecture, let's look at the recommended workflow for executing an analytics project using the newly federated BI architecture.

Figure 11. Analytics Development Life Cycle



Source: Gartner (December 2017)

Figure 11 depicts the flow of information through the various layers within an analytics development life cycle and its interaction points with the various users.

On the top, we have the list of sources. In the middle, we have the LDW as part of the data layers, and at the bottom is the consumption section. The blue arrows represent the flow of data from the source systems to landing/staging within the LDW and then into the data lake where it's stored in its raw format. Curated, preprocessed data that is structured and indexed is made available within the processed zone for downstream consumption. This flow is usually managed using a data integration platform or ETL scripts. The IT team is responsible for building and managing the data pipeline from source to processed zone, as highlighted by the black arrows in Figure 11.

The steps of an analytics project development life cycle are:

1. A business use case is proposed by an executive, manager, analyst or end user
2. The data scientists and power users use their preferred BI tool to access the raw data and transform it to validate the hypothesis by building custom reports or dashboards for their specific LOBs.
3. The curated data built as part of the hypothesis is shared with other teams across the enterprise.
4. The business logic embedded within the transformation is provided to the IT team as part of the specifications for building a new data mart.
5. Data architects and engineers help build the data mart using a data integration tool within the processed zone of the LDW.
6. Maintenance and periodic refreshes of the data within these data marts and cubes are managed by IT operations.
7. The corporate BI team educates power users to build custom dashboards for the executives and end users within its own constituency.
8. For divisions that have limited or no BI expertise, the corporate BI development team can support building the underlying data marts and dashboards.

The success of the newly designed analytics platform is fully dependent on adherence to the recommended workflow because it ensures:

- Data portability across multiple BI tools: The transformation logic resides within the data layer, rather than semantic layers of the individual tools. This prevents creating BI silos within the organization.
- A governed, secured self-service analytics platform
- New capabilities are introduced for users to analyze the data better and derive faster insights to help grow the business

Follow-Up

After the implementation of the federated BI solution, it is important to assess whether the underlying objectives for building a scalable analytical platform were achieved:

- Did the introduction of new capabilities for both business and IT help scale a multitool BI platform?
- Did BI and analytical practice transition to business from IT? IT should take on a supporting role of managing the infrastructure of the tools and data management systems alongside providing access to data.

- Is there a clear definition of roles and responsibilities of data stewards, architects (building data pipelines) and power users (report authors and data scientists)?
- Did you establishing a self-service analytical platform that caters to all user groups within the enterprise?
- Are you promoting a franchise model for execution of future analytical programs by providing business users with tools and technologies to give them a personalized, specialized and simplified solution?
- Have you reduced technical debt by reconciling process and tools, which are now much more aligned to the capabilities and use cases within the enterprise architecture?

Organizations benefit by the introduction of innovative tools and features rolled out by vendors, who are under constant pressure to deliver more value. This results in a market that has an abundance of tools and technologies that is constantly evolving. Hence, it is important to document all of the findings throughout the implementation process and outcomes of POCs conducted while establishing or rearchitecting existing/new BI tools. The key findings gathered along the Plan, Build and Run stages should be revisited to look for opportunities to further optimize the architecture.

Risks and Pitfalls

When starting with a clean slate or introducing a second BI tool, enterprises have limited risks. That is not the case when they already have an implementation of multiple BI tools that needs to be rearchitected. Following are a list of potential risks, along with mitigation strategies.

Risk: Not defining the business use cases. This means there is a lack of business use case and clear definition of the problem statement that users are facing with the current implementation.

How to mitigate: Identify the enterprise's roadmap and capabilities that will be required to achieve both short- and long-term business goals. Data, being a key component driving the modern-day digital revolution, needs to be acquired, organized, analyzed and delivered to provide value. Business should articulate a minimum of four to five strategic use cases to effectively design the future architecture.

Risk: Not having business on board throughout the implementation process. The absence of business partnership when designing the solution can result in standing up an analytics platform that does not meet business requirements.

How to mitigate: Use an agile implementation model and have someone from the business group be the product owner. This helps create transparency throughout the design and implementation stages. Business users should be actively involved in the tool selection process. This will mitigate any risk of selecting the wrong tool and inhibiting users from effectively working with their data.

Risk: Implementing an end-to-end data management solution using a single product. A number of tools provide multiple capabilities within the data management systems. Going with a

single integrated end-to-end solution creates a limitation on scalability, cost, innovation and integration with future technologies.

How to mitigate: Design and build a modular analytics platform. This provides you with the flexibility to scale a specific capability or introduce a new one in the future with minimal integration effort and impact on the overall architecture. You don't want your business and IT operations teams to be limited by a stand-alone solution that is not modular or extensible.

Risk: Rushing with the implementation. The speed to market is always tempting, and business loves it when IT can deliver faster tactical solutions. But in rushing, you run the risk of not delivering on the enterprise's long-term vision.

How to mitigate: Consider strategic business use cases to design the federated BI architecture. It is imperative to conduct a POC of the newly designed architecture to validate the data flow, required analytical capabilities and user interaction points before proceeding with the implementation. Organizations conduct POCs when selecting products, but fail to do the same with the architecture. It is recommended to do a POC focused on the user requirements and capabilities rather than just focusing on the underlying tools.

Risk: Decoupling existing BI tools can take significant time. When moving away from a BI tool that has been used for a significant amount of time, enterprises end up with a large number of reports and dashboards that need to be transitioned to the new platform.

How to mitigate: Perform the migration of existing reports and dashboards in phases. That way, users can start leveraging capabilities of the new BI tool and then slowly transition from the older one. This approach helps prevent any operational impact on critical business processes while providing a parallel test environment.

Risk: Lack of training. Introduction of new processes such as governance, data preparation and self-service can at times be overwhelming for end users. This can affect the success and adoption of the newly designed BI platform.

How to mitigate: Provide appropriate communication and adequate training. Build strategic partnerships within the business, and clearly communicate the roadmap and changes to the existing process. Educating and training users on new processes and tools will ensure successful onboarding and adoption of the new platform. This can be done by partnering with vendors and conducting workshops to train the users.

Risk: Nonadherence to the recommended workflow. If users from individual LOBs start building their own data pipelines and dashboards, enterprises run the risk of creating BI silos that prevent collaboration, portability of data and sharing insights and that lead to the inconsistent interpretation of data.

How to mitigate: Strictly enforce policies and processes to adhere to the recommended workflow. This can be achieved by a central BI or data governing council. Conduct periodic reviews of analytics project implementations to prevent any deviations from the recommended approach.

Gartner Recommended Reading

Some documents may not be available as part of your current Gartner subscription.

"Solution Path for Planning and Implementing a Data and Analytics Architecture"

"Solution Path for Planning and Implementing the Logical Data Warehouse"

"Magic Quadrant for Metadata Management Solutions"

"Identifying and Selecting the Optimal Persistent Data Store for Big Data Initiatives"

"Securing the Big Data and Advanced Analytics Pipeline"

"Agile Data Quality to Maximize Your Business Results"

"Critical Capabilities for Business Intelligence and Analytics Platforms"

"Magic Quadrant for Business Intelligence and Analytics Platforms"

Evidence

[1] "Gartner Says Worldwide Business Intelligence and Analytics Market to Reach $16.9 Billion in 2016."

[2] "Data Governance: The Basic Information," The Data Governance Institute.

GARTNER HEADQUARTERS

Corporate Headquarters
56 Top Gallant Road
Stamford, CT 06902-7700
USA
+1 203 964 0096

Regional Headquarters
AUSTRALIA
BRAZIL
JAPAN
UNITED KINGDOM

For a complete list of worldwide locations,
visit http://www.gartner.com/technology/about.jsp