A Reference Architecture for Self-Service Analytics

Balancing Agility and Governance

By Wayne Eckerson with Barry Devlin

September 2016

About the Author

Wayne Eckerson has been a thought leader in the business intelligence and analytics field since the early 1990s. He is a sought-after consultant, noted speaker, and expert educator who thinks critically, writes clearly, and presents persuasively about complex topics. Eckerson has conducted many groundbreaking research studies, chaired numerous conferences, and written two widely read books on performance dashboards and analytics. Eckerson is the founder and principal consultant of Eckerson Group, a research and consulting firm that helps business and analytics leaders use data and technology to drive better insights and actions.

About Eckerson Group

Eckerson Group is a research and consulting firm that helps business and analytics leaders use data and technology to drive better insights and actions. Through its reports and advisory services, the firm helps companies maximize their investment in data and analytics. Its researchers and consultants each have more than 20 years of experience in the field and are uniquely qualified to help business and technical leaders succeed with business intelligence, analytics, data management, data governance, performance management, and data science.

Report Sponsors

This report would not be possible without the generous support of our sponsors: Dundas Data Visualization, Information Builders, Looker Data Sciences, TimeXtender, ThoughtSpot.

Infographic

This report comes with a companion infographic titled **"Self-Service Analytics: Six Steps to Success".**

Executive Summary

Despite its promise to liberate users from reliance on the IT department, self-service analytics is not easy to achieve. Many companies that have deployed self-service analytics have become inundated by a tsunami of conflicting reports, spreadmarts, renegade reporting systems, and other data silos. These companies have learned that the goal of self-service is not unfettered liberation from IT, but rather a partnership that balances freedom and control, flexibility and standards, governance and self-service.

To succeed with self-service analytics, organizations need a reference architecture that maps business users, technology, and developers to an information supply chain designed to turn data and insights into action. The architecture stitches self-service capabilities into the supply chain so that designated business users can source their own data and create or modify reports to answer immediate questions without waiting for IT to create custom data sets or reports.

Although the current self-service revolution wouldn't be possible without technology, the keys to self-service success are organizational. In addition to a governed self-service architecture, companies need to establish data governance teams and gateways, create federated organizations with co-located BI developers, and provide continuous education, training, and support.

The Age of Self Service: Promise and Peril

GOVERNED SELF SERVICE

The Promise

We live in a self-service age. We are now expected to pump our own gas, book our own travel, and register for everything online. New technology has made it easier than ever to publish our own books, rent out our homes, prepare our taxes, and convert our cars into an income stream, among other things. We've replaced human intermediaries with technological ones, giving us greater control and convenience at lower cost.

The self-service zeitgeist has also pervaded business intelligence (BI). Since the 1990s, BI leaders have sought to empower business users with self-service tools so they could create their own reports and dashboards. Until recently, many of these initiatives fell short. The tools weren't easy enough for most business people to use without assistance. And the information technology (IT) department was reluctant to cede control to business users, fearing they would create data chaos.

New Technologies. Today, that tide is turning. New self-service technologies introduced in the past five years now enable business users to generate insights without assistance.

For example, *visual discovery* tools make it easier than ever for individuals without knowledge of SQL or coding to connect to databases, applications, and files and develop rich, interactive dashboards and analytic applications. *Data preparation and catalog* tools enable business users to find, clean, format, and merge data without having to rely on the IT department. Finally, *open source data platforms* such as Hadoop, Spark, and NoSQL databases, along with new *data pipelining tools*, enable departments and business units to quickly create data repositories to support local analytics requirements.

These new technologies help organizations make good on the promise of self-service analytics: the business gets what it wants, when it wants it, how it wants it, and the IT team gets to offload tedious custom report development tasks. This win-win proposition promises to speed time to insight by cutting out the middleman—the IT department—and give the business direct access to data and report creation functionality.

Business-IT Collaboration. The promise of self-service analytics is not to eliminate IT from the equation, but instead foster greater collaboration between the business and IT. Rather than submit requirements in writing, the business uses self-service tools to build ad hoc reports and data sets, some of which can serve as working prototypes for production applications that serve a broad-based audience. In essence, self-service analytics instantiates business requirements, accelerating delivery and optimizing business value.

With self-service analytics, IT's role shifts from application developer to data curator. Because IT knows data better than the business, it focuses on establishing an information supply chain that the business taps into to build custom applications. Just as an oil refinery takes a raw material—crude oil—and processes it into many different products, an information supply chain refines raw data from a variety of systems and turns it into many consumable data products.

The promise of self-service analytics is not to eliminate IT from the equation, but instead foster greater collaboration between business and IT.

At the same time, the role of business shifts from passive to active participant in the development process. Because the business knows its needs better than IT, it takes the lead in developing analytic applications. With self-service tools, the business quickly iterates through options, refining requirements as it goes until it finds something that works. IT supports the ad hoc process by providing curated data, which the business supplements with other internal or external data that it sources directly. If desired, the business can recruit IT to convert its ad hoc reports into production applications built on a secure, scalable, and reliable data and systems infrastructure. (See figure 2 on page 11.)

The Perils

Even with this new generation of technology, self-service analytics is not a slam dunk. Too often, self-service analytics makes the data environment worse, not better. **So what can possibly go wrong?**

Tower of Babel. Sometimes, when an organization deploys self-service tools, a small group of business users latch on to the tools with gusto and create a profusion of reports and dashboards with conflicting results. As a result, other business users can't find the reports they need, and they get confused when similar reports show different results.

Worse, report authors don't trust anyone's data but their own. They retreat to their own data bunker and refuse to use other data, whether for political reasons or to protect their turf, reputation, or career. This diffusion of conflicting reports creates a Tower of Babel where everyone talks, but no one communicates.

This diffusion of conflicting reports creates a Tower of Babel where everyone talks, but no one communicates.

In the resulting data chaos, executives can't get straight answers to simple questions, such as "How many customers do we have?" Such questions can ignite a firestorm when every business unit defines metrics differently and uses different data to answer the same questions. Executive review meetings devolve into arguments about whose data is right. Rather than use their time productively to make decisions, executives find themselves arbitrating data disputes to the detriment of the company.

Reporting Environments. Finally, many business units become stranded in dysfunctional self-service reporting environments that can't keep up with growing numbers of users, data, and application complexity. When key developers depart, the business unit may be left without the resources to fix, manage, or enhance the reporting environment. The business units eventually ask the IT department to step in and manage the environment before it implodes.

Why It's Hard

One Size Does Not Fit All. There are many reasons why it's difficult to achieve the promise of self-service analytics. One of the biggest is that self-service analytics is not a homogeneous thing that can be universally deployed. Self-service analytics means different things to different people.

For example, an executive might define self-service as the ability to view an online dashboard during an operational review meeting. A manager may see it as the ability to drill, sort, and filter dashboards and reports. A marketing analyst might think it is the ability to create custom data sets for a target marketing campaign that blends corporate and external demographic data.

In order to succeed with self-service analytics, organizations have to tailor analytic output to every individual in the organization. This is a tall order. It requires both deep knowledge of business users and their information requirements, as well as a sophisticated analytic platform that enables administrators to define and manage permissions for accessing data and analytic functionality.

To succeed with self-service analytics, organizations have to tailor analytic output to every individual in the organization.

Organizational Dynamics. In addition, self-service analytics pits intrinsic organizational forces against each other. On one hand, corporate executives seek business alignment, economies of scale, and a single face to the customer; on the other, business unit managers want greater autonomy to meet local needs quickly, efficiently, and inexpensively. One side wants greater centralization and standardization; the other wants decentralization and customization. (See our 2016 report **Governed Data Discovery: Balancing Flexibility and Standards**.)

This top-down versus bottom-up conflict wreaks havoc on BI implementations. Business leaders need to recognize this conflict and negotiate a truce. They need to redesign organizational models to federate responsibility for analytics between corporate and business units. And they need to implement standard architecture and data analytic platforms that make it possible to implement self-service analytics without sacrificing data consistency and alignment.

Success Factors

To successfully deliver a governed, self-service analytics environment, organizations need to implement:

1. A Reference Architecture. Organizations must first architect a self-service data environment. Simply distributing self-service tools willy-nilly to any user who wants them, or allowing vendors to sell directly to business units without including IT in discussions, are recipes for disaster. The ideal architecture maps users, technology, and developers to an information supply chain.

2. A Federated Organization. Companies also need an organizational model that balances top-down (centralized) and bottom-up (decentralized) forces. A federated organization with matrixed reporting assignments, co-located BI and data specialists, and business engagement specialists is the best way to balance freedom and control, agility and governance, self-service and standards. Creating this balance is critical to self-service success.

3. A Standard Data and Analytics Platform. The irony of self-service analytics is that it requires standardization. In addition to a reference architecture, organizations should seek to implement a standard data and analytics platform for all users. A unified environment makes it easier for administrators to tailor analytics to individual requirements and for business users and IT developers to collaborate around requirements and development, speeding the delivery of reports and dashboards.

The irony of self-service analytics is that it requires standardization.

4. Governance Processes. Ultimately, self-service analytics is not a technology or product; it's a process. Figuring out who has the right to access which data and functionality and share it (or not) with which people requires stewardship and oversight. The business (not IT) must form a committee to govern itself so it can optimize the use of data and protect against abuses. It must apply watermarks or stamps to differentiate between governed and ungoverned data artifacts so business users understand the context of the data they consume. (See "**How Watermarks Can Transform Your BI Organization**")

5. Continuous Training. Finally, organizations need to double down on training, as counterintuitive as that seems. Kevin Sonsky, senior BI director at Citrix Systems, says that "Self-service BI requires a lot of hand-holding." The fact is that even sophisticated power users won't be productive or constructive with self-service tools unless they receive heavy doses of formal and informal, peer-based training and support.

"Self-service BI requires a lot of hand-holding." —Kevin Sonsky, senior BI director, Citrix Systems

This report focuses on the first success factor above—a reference architecture—and touches on the remaining points as they relate to the architecture.

Reference Architecture

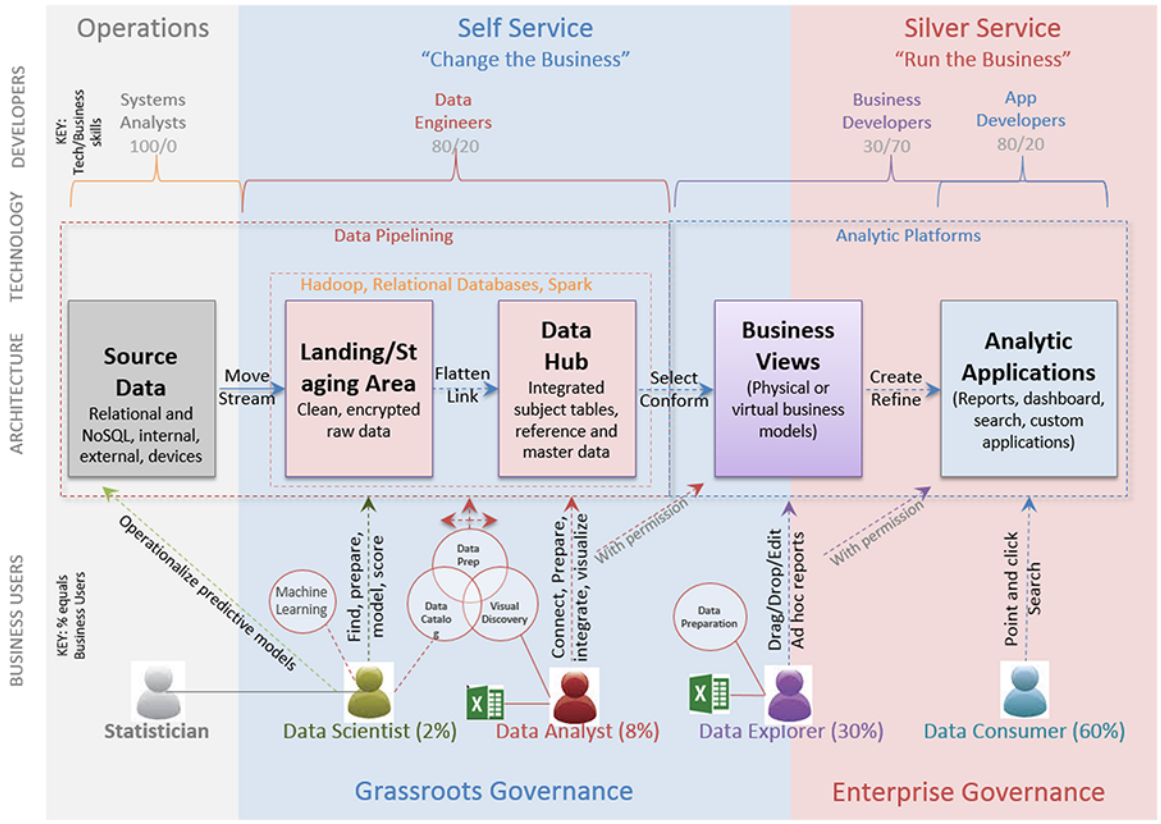
The terms "architecture" and "self-service" are contradictions in terms. For business users, architecture implies slowness, bureaucracy, and red tape. Most users want self-service tools to help them escape the tyranny of architects and architecture. Most fear that their projects will never see the light of day if enterprise architects get involved.

But without architecture, self-service analytics goes awry. Without standards, there is no center, and the organization descends into data chaos. Architecture provides a governance foundation upon which self-service analytics can flourish. It tailors self-service to individuals so they get everything they want and nothing they don't.

[Architecture] tailors self-service to individuals so they get everything they want and nothing they don't.

A good architecture defines the relationship between components in a clear, concise way. That is the goal of our reference architecture for self-service analytics. (See figure 1.) It is a conceptual architecture that describes the relationship between different types of business users, developers, and technologies, and offers an information supply chain designed to deliver data and reports to decision makers.

Figure 1. Eckerson Group's Reference Architecture for Self-Service Analytics



Ideal versus Real. Note that this reference architecture depicts an ideal environment; it doesn't model the reality that exists in most organizations. For instance, most organizations have at least two supply chains: a data warehouse for operational data, and Hadoop for multi-structured data. And in fact, many have dozens of supply chains crisscrossing their organizations in a tangled web. The chaotic reality of most data environments causes many IT managers to lie awake at night trying to figure out how to reconcile duplicate supply chains and eliminate overhead, inconsistent data, and synchronization conflicts.

Nevertheless, we're confident that most organizations will see their reflections in this model and benefit from examining how it maps to their environments. In fact, we hope this reference architecture will become a standard for how organizations think about deploying analytical resources and a playbook for how they balance enterprise governance with self-service requirements.

We will now deconstruct the reference architecture layer by layer, drilling into the details with text and supplemental charts, starting with a discussion of business users.

Business Users and Developers

Business Users

The point of any architecture is to meet the needs of business users in an efficient, effective manner. Consequently, the best place to begin a reference architecture discussion is to define different classes of business users who use an analytics environment. Once we understand types or classes of business users, we can map them to the information supply chain and other parts of the architecture.

Casual versus Power Users. Business users can be divided into two major classes: casual users and power users. Casual users use information to do their jobs. They make up 90% of knowledge workers in an organization—typically, executives, managers, and front-line workers. In contrast, power users are hired to collect and analyze information on a daily basis. They have working knowledge of databases, query techniques, statistics, and machine-learning tools and techniques. They make up approximately 10% of knowledge workers in an organization.

Business Roles. There are also four subclasses of business users based on role. Data consumers and data explorers are casual users, while data analysts and data scientists are power users. Data consumers make up roughly 60% of knowledge workers in an organization, while data explorers make up 30%. In contrast, data analysts comprise about 8% of users and data scientists just 2%.

Of course, these percentages vary widely by organization. Organizations in information-centric industries such as financial services, insurance, healthcare, and high-tech may have a much higher percentage of power users than those producing tangible goods, such as manufacturing, oil and gas, and retail. However, as the digital economy gathers steam, even traditional manufacturing companies such as General Electric and the Big Three automakers are pivoting their focus to data and information services.

Types of Casual Users

There are two major subclasses of casual users

- **Data consumers** simply want to consume reports and dashboards created for them. Some only view the content, while others interact with it, searching, drilling, sorting, pivoting, and creating snapshots for later viewing.

- **Data explorers** are data consumers who occasionally want to edit a report or dashboard or create one from scratch without coding. Using a BI or discovery tool, they drag and drop metrics, dimensions, controls, and predefined charts from an object library onto a report canvas to create ad hoc reports and dashboards. They may also create metrics using a point-and-click calculation engine and merge local and external data using integrated data preparation functions. (See "The Conundrum of the Data Explorer," below.)

Types of Power Users

There are three major subclasses of power users

- **Data analysts** are business-savvy data experts who are proficient with Excel and have a basic grasp of SQL and statistics. They are hired by department heads (e.g., finance, sales, marketing) to work on data-centric issues. For example, data analysts might create pricing plans, performance metrics, budget estimates, demand plans, or retention models or investigate performance anomalies via root-cause analysis. Data analysts need far-ranging access to data and use a triumvirate of analytical tools to do their work—specifically, data preparation, data catalogs, and data visualization tools.

- **Data scientists** are data analysts with a computer science background who know how to code using languages such as SQL, Java, Python, Hive, and Pig. The best are also conversant with statistics and data mining tools and can create predictive and machine-learning models. Most data scientists want to access raw data at the lowest level of granularity.

- **Statisticians** write statistical and machine-learning models but do not have a computer science background like data scientists. Many come from diverse fields, such as mathematics, econometrics, operations research, and social sciences. Like data scientists, statisticians want to use the most granular data possible to create predictive models.

Developers

Supporting business users are IT analysts and developers. There are numerous types of IT professionals who play key roles in a self-service analytics environment:

- **Systems analysts** are database administrators and systems analysts who manage an organization's operational systems. Any changes they make to operational systems can break downstream analytic processes if they don't coordinate their tasks with analytic professionals.

- **Data engineers** do the heavy lifting required to create and manage the information supply chain. We used to call these individuals ETL developers and data architects. They identify source data, map data flows, model databases, define and monitor data transformation jobs, and work with database administrators to create, manage, and tune databases and optimize performance. Some also design business views for business users, especially if they are built within a database.

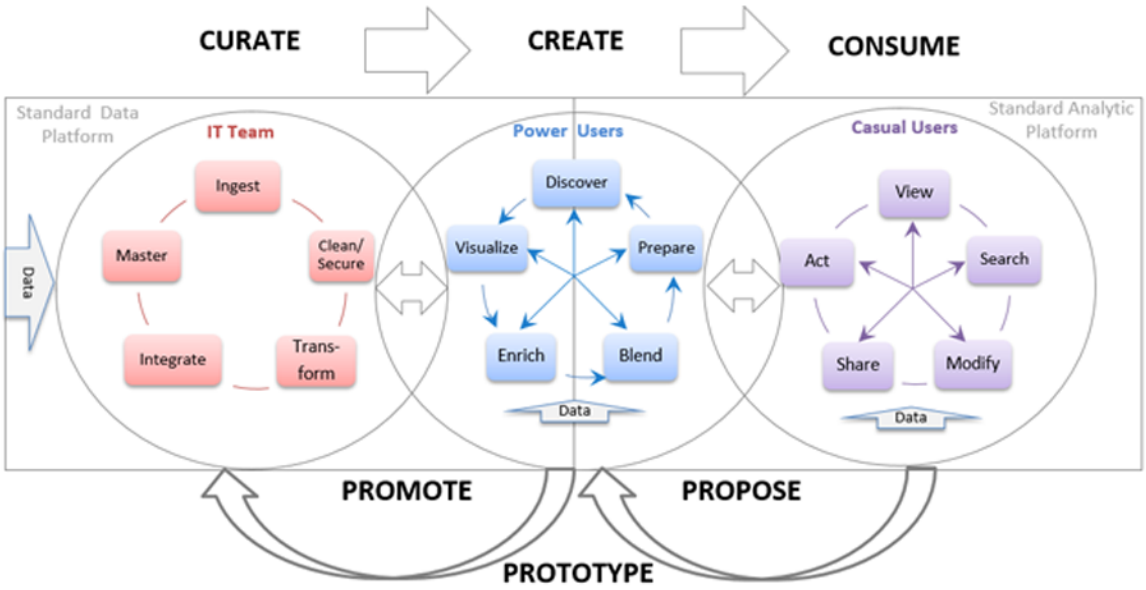
- **Business developers** build reports and dashboards for business consumption. Traditionally, they are corporate BI developers, or perhaps business-savvy data engineers. But increasingly, they are tech-savvy business users in each business unit. Ideally, they are BI developers co-located in the business unit as part of a federated organizational model. But business developers can also be data analysts who have time to build business views and reports for colleagues. Increasingly, they are tech-savvy BI analysts who can not only gather and consolidate requirements from business users, but also design the report or dashboard interface using point-and-click development tools or light scripting.

- **Application developers** build custom analytic applications using APIs, software development kits (SDKs), and various programming languages. Many also use their coding skills to embed analytic components and environments into other applications. Developers are increasingly being recruited to build analytic applications as organizations seek to use information to create a strategic advantage.

Self-Service Workflows

Collectively, business users and developers create a self-service analytics environment that refines data for consumption. The environment consists of bidirectional and iterative workflows that enable business users to refine and enrich curated data to meet their needs quickly, while promoting prototypes and requirements to iteratively expand the boundaries of the curated data environment. These workflows create a vibrant, self-reinforcing data environment that accelerates time to insight as well as user productivity. (See figure 2.)

Figure 2. Bi-Directional Self-Service Workflows



Left-to-Right Workflow. In most organizations, data flows from source to target, getting more refined and curated the closer it gets to business users. Traditionally, IT departments manage the curation process for both data and the creation process for reports. Consequently, IT often becomes a bottleneck that stands between business users and data.

With self-service analytic tools, however, power users can participate in the curation process. Using data preparation tools, they can transform, blend, and enrich enterprise and local data in an iterative manner. With visual discovery tools, they can visualize and analyze data and share their findings with departmental colleagues and managers. These casual users then view, search, modify, and discuss the power user reports and act on the results.

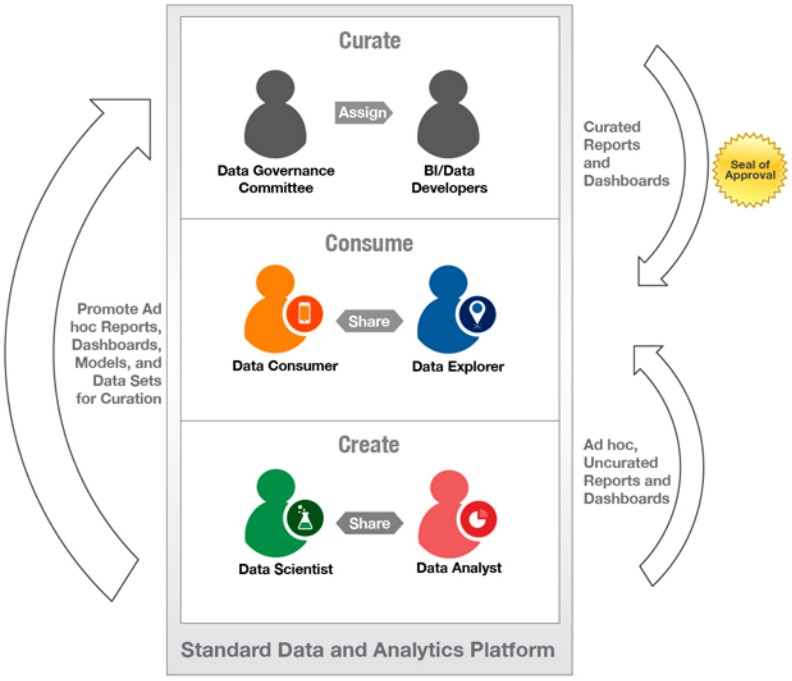
With self-service analytic tools, power users participate in the curation process.

Right-to-Left Workflow. Equally important is a reverse workflow where business users feed requirements

back into the curation and creation processes. After analyzing data in reports and dashboards, consumers better understand their requirements and can propose them to power users in their department. In turn, power users can build those requirements into an ad hoc report that casual users can review and refine.

If the report is popular, power users can submit the report to a governance team to review. The report serves as a prototype for a new production report. The committee inspects the report's metrics, data definitions, and GUI layout to ensure they align with corporate standards. And the IT department determines how to "productionize" the ad hoc report with adequate scalability, security, and reliability. (See figure 3.)

Figure 3. Report Workflow



Watermark. If the proposal is approved, the report then carries a seal or stamp of approval. This mark indicates to business users that the report is "safe" to use for decision making. The watermark helps users distinguish between curated and non-curated data. Before long, business users start refusing to use reports that don't carry the seal of approval. This creates a culture of governance from the ground up. Rather than work independently, power users recognize the value of working through formal channels to publish reports for broad-based consumption.

The watermark helps users distinguish between curated and non-curated data.

By understanding data and report workflows and the roles of different business users, organizations can accelerate the delivery of information to business users and eliminate BI bottlenecks. The result is a governed self-service environment that balances the business's need for speed, agility, and freedom with an organization's need for standards, control, and governance.

Self Service versus Silver Service

The reference architecture is vertically divided into three main sections that define different approaches for managing and packaging data: operations, self-service, and silver service.

Operations

Operations refers to applications, systems, and services that feed data into the analytic environment. In most companies, systems analysts and database administrators manage operational or transactional applications such as those for enterprise resource planning (ERP) and customer relationship management (CRM). But increasingly, these individuals oversee the acquisition of data from nontraditional systems, including social media environments, machines, and the sensor networks that form the basis of the Internet of Things.

Self Service

Lightly Governed. The self-service environment supports power users who need true ad hoc capabilities. Although power users benefit from a well-governed data warehouse or data lake, they inevitably hit the boundaries of these IT-generated data environments. Unless they can access any data inside or outside the organization, mash it together, and analyze it, they can't truly perform their jobs. Consequently, these power users need a lightly governed data analytics environment that gives them maximum freedom with minimal constraints.

Tools. Power users also need self-service tools that offer high degrees of flexibility and analytical horsepower that enable them (without IT assistance) to:

1. Locate any data they need (data catalog)
2. Access, profile, clean, transform, and merge data sets (data preparation)
3. Visualize, explore, and analyze the data and share it with others (visual discovery)
4. Prepare, model, score, and operationalize analytic models (machine learning)
5. Create reusable business logic and rules

Silver Service

Fully Governed. The silver service environment is geared to casual users. Unlike power users, casual users want and need a fully governed data environment that tailors data and information to their roles and gives them relevant, actionable content in an easy-to-digest format that accelerates time to insight. They also demand high-quality data that is fit for use (i.e., clean and accurate enough to make valid decisions). In other words, they need data delivered on a silver platter, hence the term "silver service."

Dashboards. Most of the time, these business users simply want to monitor the business processes for which they are accountable. They need information to monitor performance, track trends, and analyze anomalies. To "drive the business" on a day-to-day basis, these users are best served by a subject-specific, interactive dashboard that contains a dozen key performance indicators (KPIs) and 20 dimensions or filters.

Search and Ad Hoc Reports. Sometimes these casual users need to go beyond the confines of a predefined dashboard and conduct ad hoc analysis like power users. In the past, casual users would call the IT department or an analyst to answer their questions, but today, they can use search-based analytics tools or ad hoc (or editable) dashboards to get the information they need.

The Conundrum of the Data Explorer

One type of casual user—the data explorer—sits at the intersection between self-service and silver service. Many people argue that data explorers belong entirely in the "self-service" camp. We disagree. Data explorers, by definition, are business users who should be doing other things rather than collecting and prepping data. That's the role of the data analyst, who is hired full-time to do this kind of work. Data explorers want to consume more than create.

Data explorers want to consume more than create.

For data explorers to succeed, someone in IT (typically) needs to create a business view that explorers can use as a launching pad to occasionally (not constantly) create ad hoc reports. Without this business view and a graphical BI authoring environment, data explorers are lost. Data analysts, in contrast, often find business views too limiting and typically create reports and dashboards from scratch.

Today's Reality. Of course, the reality today is that most organizations have not built an adequate silver service environment to support the data explorer. That forces the data explorer to spend nights and weekends mashing together data with whatever tools they can find to create a departmental report. They would rather not do this.

An organization can improve user productivity significantly by enhancing its silver-service environment.

Ideally, a data explorer wants an extensible BI tool that lets them customize and enrich existing reports using a built-in semantic layer and calculation engine. They want a BI tool that comes with an integrated data preparation tool for the few occasions when they want to add data to an existing report. They also want an integrated data marketplace so they can drag and drop external data sets into the report without extensive data manipulation.

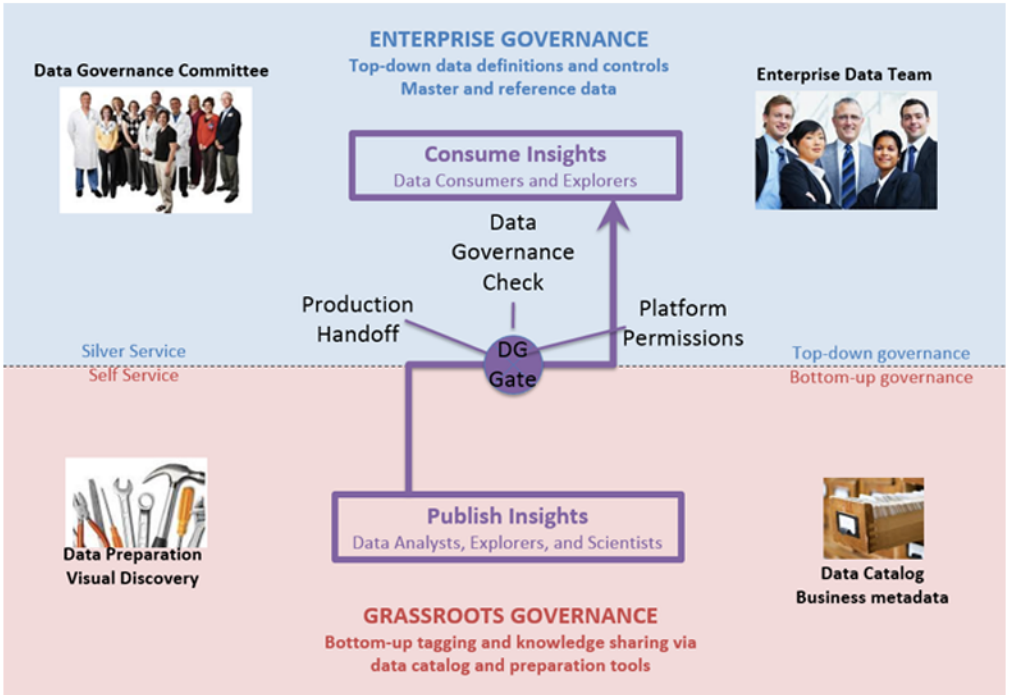
Really, they just want one tool to view, analyze, and augment existing reports, not a suite of tools, like a data analyst.

These distinctions come down to semantics. What is a data explorer? What is a data analyst? Our belief is that data explorers lose productivity if they spend too much time mashing data and creating reports, which is not their job. Organizations hire analysts to do that work. And they hire IT professionals to build rich, robust silver service environments that make it easy for casual users, including data explorers, to get the information they need quickly.

Enterprise versus Grassroots Governance

Finding Balance. As mentioned earlier, one of the biggest challenges with self-service analytics is balancing governance and agility. Applying the right amount of governance at different places in the architecture is key to success. Too much governance creates a request backlog that inhibits access to data and the free flow of insights and information. Too little governance creates chaos, with data shadow systems and conflicting reports and dashboards.

Figure 4. Bridging Two Worlds



So how do you empower business users without data hell breaking loose? Ultimately, the business must take responsibility for governing data and balance the twin forces of freedom and control.

Figure 4 shows how organizations can bridge (1) the top-down world of enterprise governance shaped by a central data team and a data governance committee, and (2) the bottom-up world of power users who use self-service tools to collaborate around data and create reports and dashboards.

Grassroots Governance

It's important to note that governance is not just a top-down exercise executed by centralized data teams and data governance committees, even though that is how most organizations practice it.

Tribal Knowledge. A parallel governance environment exists in the bottom-up world. It's always been there in the shadows, unseen by corporate planners and designers. It usually consists of one or two data analysts who know corporate databases inside and out. They know where data exists and how to get it. They hold all the tribal knowledge about data in their heads. Every analyst and business user who wants or needs data goes to these individuals for help.

New self-service tools—specifically, data catalogs and data preparation tools—make it easy for organizations to start capturing tribal knowledge.

New self-service tools—specifically, data catalogs and data preparation tools—make it easy for organizations to start capturing this tribal knowledge and augmenting it with other information supplied by data analysts and data scientists as they scour corporate databases. Data catalogs crawl databases, profile data, and extract samples for review. Data preparation tools create a portfolio of data products along with an audit trail of where each product originated (data lineage) and how it was created, by whom, and when (metadata).

Breadcrumbs. Using social media techniques such as tags, comments, ratings, and follows, self-service data integration tools effectively leave breadcrumbs for other analysts to follow when hunting down pertinent data to use for analysis. In particular, this helps new analysts learn their way around an organization's data environment and get up to speed quickly.

Symbiosis. But more importantly, the grassroots governance process fills the gaps in top-down processes, which move slowly and methodically and generally address only enterprise data elements and qualifiers. Both the enterprise data team and data governance office can borrow heavily from grassroots efforts when designing data models and data standards.

Data Governance Gateway

To make self-service analytics work, organizations need a "data governance gateway" to manage the flow of information from the bottom-up world of self-service to the top-down world of silver service. This gateway consists of three mechanisms designed to ensure data consistency and eliminate spreadmarts and data silos.

1. Data Governance Check

The first step in the gateway is the data governance check to review self-service reports that a business user

or business unit wants to promote to the entire organization. The check is performed by a cross-functional team at corporate or in the business unit where the request originated.

The committee evaluates the new report against its existing inventory of standard reports and checks for overlaps or duplications. It also cross-references all data elements and metrics against defined standards to check for alignment and conflicts. This check can be infuriatingly slow for a business person who wants to put a report into production quickly, but it's essential to avoid data conflicts and confusion among users.

2. Production Handoff

"If the data governance check gives the green light, then the report gets a watermark or seal of approval that signifies that the report's core data, metrics, and GUI layout have been vetted and it is safe to use for decision making. In some cases, the report may need to be hosted on a new BI platform or new data elements added to the data warehouse. In these cases, the report is turned over to the IT department to convert from an ad hoc report into a production one with adequate scalability, security, and reliability."

Corporate BI Developers. Analysts can hand their ad hoc reports to BI developers on the corporate BI team or a co-located one in the business unit. Corporate BI developers usually have deep expertise in the tool and work hand-in-hand with the data architects and data engineers who manage the corporate data repository. However, it might take a long time for the corporate BI developer to fully operationalize the report, depending on the corporate backlog of tasks.

Co-located BI Developers. A co-located BI developer is a good choice because they probably already know the analyst, the content of his or her report, and the data behind it. Thus, the local BI developer can service the request quicker and potentially better. However, the local developer might not have deep knowledge of the BI tool or corporate standards for production reports. The key to creating a federated BI program and center of excellence is developing a cadre of co-located BI developers in each business unit who are trained and supported by a corporate group.

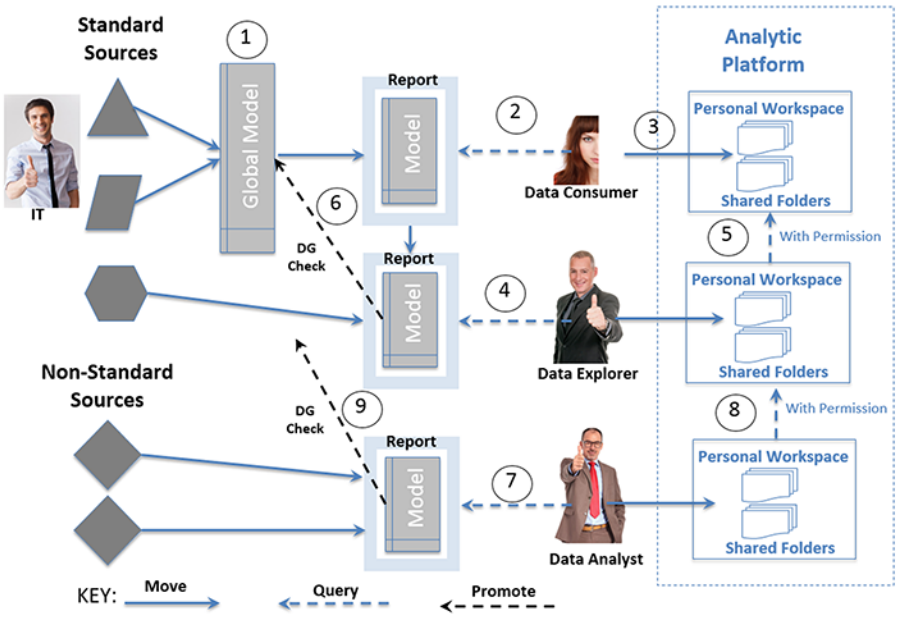
3. Platform Permissions

Most analytics vendors now recognize the need to balance governance and self-service. Consequently, they offer granular permissions based on a micro-service architecture that leverages existing security management frameworks, such as Active Directory and LDAP. This gives administrators the ability to define who can access what data and reports and who can share them with which colleagues and groups.

Extensible Views. Figure 5 shows a governed self-service workflow supported by an analytic platform with granular permissions. The platform gives different types of users different levels of access to data and reports. For example, data consumers can snapshot report views and save them to a BI portal; data explorers can edit existing reports and add governed data if desired; and data analysts can create new reports by adding ungoverned data. If permitted, users can share these custom reports with colleagues using an

analytic or collaborative platform. They can also submit the custom reports to a data governance committee to promote to a global report repository

Figure 5. Governed Self-Service Workflow



STEPS: (1) IT creates a global data model from multiple data sources and builds a report using elements of that model. (2) A data consumer views the report and creates a snapshot (3) which she saves in a personal workspace in the analytic platform. (4) A data explorer then edits the underlying report model, creating new metrics and dimensions from base fields and, if desired, adding data from a vetted, integrated corporate data source not in the report model. (5) Once he publishes the custom report to his personal workspace, the data explorer shares it with colleagues if he has permission. (6) He then submits the report to the data governance committee to review and promote it as a new standard report. (7) A data analyst creates a new report from scratch using unvetted, non-integrated data (8) that he saves to a personal workspace and shares only with permission. (9) The analyst submits the report to the data governance committee for review, and if approved, hands it off to a BI developer to turn into a production report.

A technology handoff works best if all business users—casual and power—use the same analytics platform. It doesn't work if they are using disparate tools. Fortunately, many BI vendors now offer a complete integrated stack of analytics functionality that can be used by both power and casual users. (More on analytic platforms below.)

If developed properly, governed self-service environments give casual users the right amount of analytic functionality without over-burdening them with complex features. And they give power users the right amount of data access and publishing rights so they can generate and share insights quickly without creating spreadmarts and data shadow systems.

Collaboration and Reuse. In addition, self-service environments foster greater collaboration and reuse, improving analytic productivity. With a shared analytic platform, business users can reuse the reports and analyses others have created and extend them to meet their own needs. This also gives the IT team visibility into what analysts are doing in the trenches. If they see an analysis done repeatedly by multiple users, they can build that functionality into the data warehouse, greatly increasing user productivity.

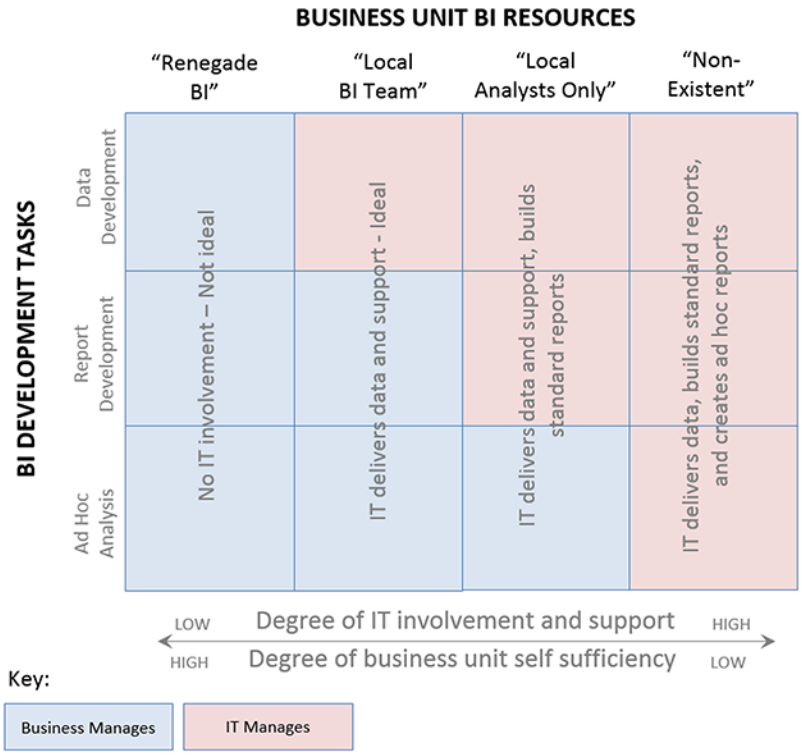
Tailored Governance

In the early days of BI, the IT department (i.e., the BI and data warehousing teams) not only built the data warehouse but also developed the data sets and applications for both power and casual users. That's because extracting, integrating, and transforming data and creating reports and applications required technical skills, including knowledge of SQL, programming, and in-depth tool training. Today, with the advent of powerful, low-cost self-service tools that enable business users to prepare data sets and create reports and dashboards, that is no longer the case.

One of IT's challenges is tailoring the degree of support it provides to business units, which vary greatly in their BI resources, skills, and interest in self-service analytics.

One of IT's challenges is tailoring the degree of support it provides to business units, which vary greatly in BI resources, skills, and interest in self-service analytics. For example, finance, sales, and marketing departments are typically more proactive in developing BI capabilities than human resources and legal departments. Some departments go "rogue" and build their own BI/DW environments, while others have their own BI teams but rely on IT to supply the data. Some only have analysts who do ad hoc reporting, while others don't even have analysts.

Figure 6. Tailored Governance Framework



In other words, the degree of IT support required is inversely proportional to a business unit's analytic self-sufficiency. This dynamic makes creating a business analytics center of excellence challenging. IT must support multiple models of engaging with the business rather than a single model. On one extreme, it needs to cajole renegade groups, to embrace the corporate information supply chain, and on the other, it needs to help other groups do everything from building standard reports and dashboards to creating ad hoc reports. (See figure 6.)

The ideal scenario is where the business unit has a BI team that is actively supported by the corporate BI team and relies heavily on the corporate data warehouse for shared data but also its own local data (column two in figure 6). But even here, the corporate BI team will continue to develop cross-functional or complex applications.

Information Supply Chain

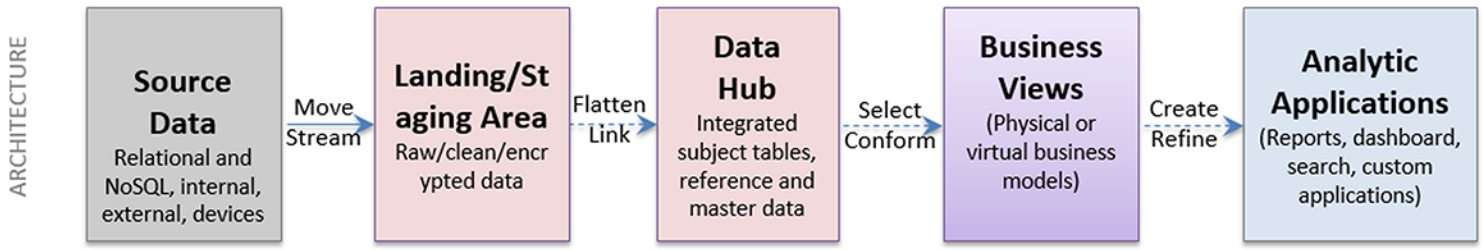
The backbone of the reference architecture is the information supply chain—the flow of data from source systems to business users. (See figure 7.) An information supply chain is critical to accelerate the flow of governed data and minimize self-service activity. A well-designed and agile information supply chain makes business users more productive by putting the data they need at their fingertips.

A well-designed and agile information supply chain makes business users more productive by putting the data they need at their fingertips.

This top-down, governed supply chain is managed by the IT department, which oversees the data warehouse and enterprise Hadoop implementations. Organizations can implement the information supply chain using any number of technologies, but Hadoop and relational databases are the preferred data platforms today, while Spark, NoSQL, and cloud-based platforms are emerging contenders. (See our 2015 report, **Selecting a Big Data Platform: Building a Foundation for the Future**.)

Organizations can use the supply chain to implement or feed an enterprise data warehouse (EDW). Conversely, an EDW can feed an information supply chain that is implemented in Hadoop. In essence, the information supply chain is a set of cascading data structures that can support multiple data processing and analytic requirements.

Figure 7. A Governed Information Supply Chain



A comprehensive, well-designed information supply chain minimizes the need for business users to secure their own data outside of governed channels. However, in reality, power users in particular need more data than the IT team can put into an enterprise data repository. Therefore, most organizations—whether intentionally or not—augment the information supply chain with self-service workflows. (See figure 7 above.)

A comprehensive, well-designed information supply chain minimizes the need for business users to secure their own data outside of governed channels.

New Thinking

Hadoop and its distributed file system (HDFS) as well as new visual discovery tools have reshaped the way corporate data architects design information supply chains.

Impact of Hadoop. Hadoop has dramatically reduced the time and cost of acquiring and loading diverse sets of data. Companies can load data quickly because they don't have to model data to fit a relational schema; they just dump the data into the Hadoop file system as is. And because Hadoop is open source, they can now store large volumes of detail data at low cost. So they don't have to dispose of data or aggregate it to keep costs down. Consequently, many organizations now use Hadoop to support a landing area to collect data and a staging area to clean it and lightly integrate it.

Impact of Visual Discovery. At the same time, popular visual discovery tools are reshaping the way data architects model enterprise data. Visual discovery tools were initially designed to consume files, particularly spreadsheets. Consequently, these tools work best when they query wide, flat tables containing denormalized data. For more complex data, they generally need wide, flat tables that share a common key and reside within the same database.

Consequently, many architects now design business-facing corporate repositories as a series of linked tables, each representing a major subject area in the business. This defines the shape of the data hub. Each table represents a major subject area such as customers, orders, or inventory. The tables are generic in nature, not designed or tuned with any application in mind. This enables the data hub to support a host of downstream applications and users. It is literally a hub that feeds other systems, including data marts, operational data systems, and BI tools.

Perhaps the most significant design change is that IT no longer tries to model the business at the enterprise level. That modeling is now pushed down to business units.

Minimal Data Modeling. Perhaps the most significant design change is that IT no longer tries to model the business at the enterprise level. That modeling is now pushed down to business units, which pull information from the data hub and create their own business views based on local requirements. This dramatically accelerates the data delivery cycle compared to traditional data warehousing development, where IT architects spend months learning the business and creating complex, global data models. Once built, the global models are difficult to change, further delaying the delivery of data to increasingly impatient business users.

Components

Here is a brief description of each of the components in the information supply chain.

- **Source Data.** This consists of transactional, structured, semi-structured, or unstructured machine- or human-generated data. Generally, organizations move structured, operational data in batch (or near-real-time microbatches) with changed data capture into a landing zone. They stream voluminous social media and machine-generated data into the landing zone using streaming tools such as Kafka, Flume, Storm, and Spark Streaming.

- **Landing Zone.** Batch or streaming processes deposit raw data of any type in a landing zone where it is persistently stored in its original form until it's no longer needed. Most data is encrypted when it is landed, and highly sensitive data might also be masked or tokenized. Hadoop has become a popular landing zone thanks to its file-based structure and low cost. Companies can also use a relational database as a landing zone, but this only makes sense if their data is structured and not terribly voluminous. Only IT administrators touch data in the landing zone.

- **Staging Area.** IT administrators perform data quality checks and cleansing processes to fix errors before moving data from the landing zone into the staging area. They also standardize naming conventions, creating base objects—or building blocks—at the lowest level of granularity possible. They apply masking or tokenization to protect sensitive fields, if not already done. Data scientists often access this layer, because they prefer using raw data before it's heavily transformed or aggregated.

- **Data Hub.** In the data hub, administrators prepare data for generic downstream usage. They summarize and join data, create relevant metrics, and merge base objects into linked subject area tables. For example, they might merge all of a customer's account data into one view stored within a wide, flat customer table. A data hub can feed a data warehouse or discovery tool as well as operational data stores or applications that are more transactional in nature.

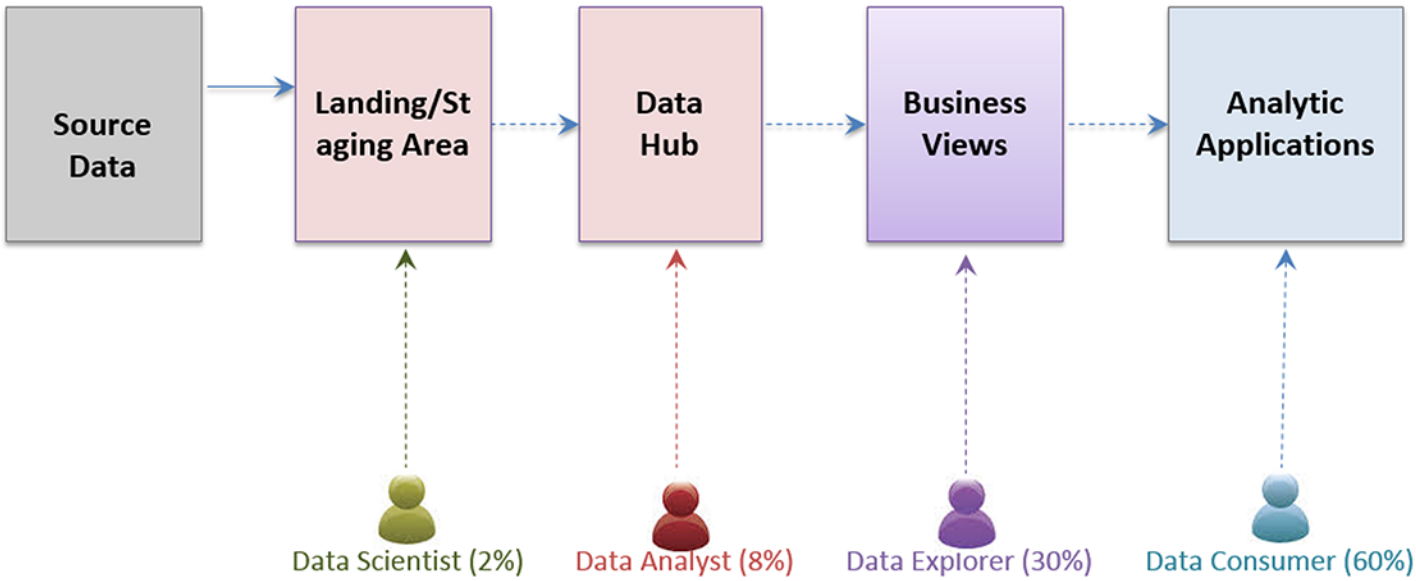
- **Business Views.** Business views transform data from the hub into data objects meaningful to a particular group of users. Created by business analysts or business-savvy BI developers, these views are either physically instantiated in a database or virtualized in an abstraction or semantic layer that displays data in business-friendly terms. Developers can create business views using a host of technologies: BI tools, data virtualization tools, OLAP cubes, and database views. Business views that exist outside of individual tools are often called "data services." Many financial services firms use data services to give users and applications seamless access to multiple back-end systems and external sources that would be too expensive or impossible to consolidate into a data warehouse. These views or services give IT administrators greater flexibility to change back-end data platforms and flows without impacting downstream users or applications.

- **Analytic Applications.** Finally, BI developers use the business view to create reports, dashboards, and custom analytic applications. These applications often use a subset of the business view, which is exposed to data explorers who want to modify reports or dashboards by adding new metrics or dimensions pulled or derived from base fields in the model. Many BI tools now offer rich application programming interfaces (APIs) that allow developers with knowledge of SQL and JavaScript to create highly customized graphical interfaces and workflows

Mapping Users to the Supply Chain

The most powerful element of this self-service reference architecture is that it maps types of business users to the information supply chain. (See figure 8.) In the past, the IT team locked down much of the information supply chain, only allowing users to access BI tools or data marts. Few individuals were granted direct access to the data hub (or data warehouse), and fewer still, if any, could query data in the landing zone or staging area. With a well-defined information supply chain, IT administrators should feel more confident about giving different classes of business users access to data.

Figure 8. Types of Business Users



Our reference architecture gives users access to the information supply chain based on their roles and requirements. To gain access to the landing/staging area and data hub, individuals in some organizations need to take an examination to prove that they have the skills and knowledge to leverage that data correctly.

1. **Data Scientists → Staging Area.** Data scientists (and statisticians) need raw, non-aggregated data to create predictive models. Rather than pull data from operational systems, they should be given access to raw data in the staging area, which at least has been cleaned and encrypted. Organizations should provide them with data catalogs to facilitate this exploration as well as data preparation tools to format, transform,

and blend data, and visual discovery tools to display and analyze the data.

2. **Data Analysts → Data Hub.** A true data analyst will want to pull data from the wide, subject-specific tables in the data hub to populate a visual discovery tool. Increasingly, organizations are prepopulating discovery tools with this data, saving data analysts (and data explorers) the task of connecting, exploring, and ingesting data. But data analysts often prefer to explore the data behind the tools. Like data scientists, they should be given data catalogs, data preparation, and visual discovery tools to assist their analysis.

3. **Data Explorers → Business Views.** Data explorers use business views to modify and extend existing reports and create simple ones from scratch. Many BI tools now embed or integrate with lightweight data preparation tools that enable data explorers to mash report data with other data sets not in the business view. And some are integrating data markets to help data explorers (and others) incorporate other internal and external data into their reports.

4. **Data Consumer → Snapshots.** Data consumers explore and analyze data within a report or dashboard. They never source new data, but they often want to save a snapshot of data from a report or dashboard so they can view it later with fresh data. This is an easy way to create a "custom" report or dashboard.

Technology

This report has mentioned numerous technologies required to build and maintain a self-service analytics environment. This section describes these core technologies in more detail and the roles they play in supporting a governed self-service environment. Eckerson Group's next report (to be published in December 2016) will explore many of these technologies in greater depth.

Data Pipelining

Data pipelining tools are a class of data management products that govern and manage the flow of data from source to target in a big data environment. They ingest, validate, clean, transform, merge, secure, and format data for analytical purposes and manage the processes required to operate a production environment. Most support both batch, near real time (i.e., microbatches), and real-time (i.e., streaming) updates as well as changed data capture (CDC) to load just deltas and minimize data volumes. Most importantly, they collect metadata every step of the way, so both technical and business users can track data lineage, understand how and when data was transformed, and evaluate the impact of any changes on downstream applications.

There are several types of data pipelining tools. Some are geared to the IT department and some are geared to power users.

IT-Oriented Pipelining Tools

- **Extract, Transform, and Load.** ETL tools are data warehousing workhorses, enabling developers to map source and target data and execute transformations. Originally designed as external execution engines, many now leverage underlying data processing systems (either relational databases or Hadoop) to support alternative extract, load, and then transform (ELT) processes, making them more like data lake management and data warehouse automation tools.

- **Streaming Engines.** These tools specialize in capturing real-time data off message buses, transforming data in flight, and landing data in real time with auditability and queue management.

- **Data Warehouse Automation.** These metadata-driven tools automate the processes to build, populate, operate, and maintain relational data warehouses and data marts. They facilitate agile development and now work directly with data in Hadoop and NoSQL databases.

- **Data Lake Management.** These products typically run natively on Hadoop and are geared to manage the bidirectional flow of large volumes of data in and out of Hadoop. Many contain discovery environments that can be used or accessed by business users.

- **Data Virtualization.** These tools generate a virtual, business-friendly view of back-end data resources. Also called a semantic layer or business metadata, these views are generated by BI tools, analytic platforms, and independent data virtualization vendors. Like data lake management software, these tools bridge business and IT.

Business-Driven Pipelining Products

- **Data Preparation.** These self-service tools enable power users to create data sets for analysis. The tools let users clean, format, transform, and merge data from multiple sources and publish them to select users.

- **Data Catalogs.** These tools crawl any database and create data profiles that users can augment with tags, comments, and ratings. Data catalogs help business users find data sets quickly and build a grassroots knowledge base of available data sources.

Analytic Tools

There are many categories of analytic tools and hundreds of products. Here is an overview of the major categories.

- **Search.** Search-based BI tools make it easy for casual users to submit ad hoc queries by typing words into a search box. The tools dynamically parse the words and suggest query or filter options.

- **BI Tools.** These are purpose-built tools designed to support a single mode of analytics such as reporting, dashboards, OLAP, or scorecards. Each uses a distinct architecture with unique charting, querying, security, administration, and other capabilities.

- **Analytic Platforms.** In contrast to BI tools, analytic platforms support every mode of BI on a single, integrated architecture with a common set of microservices, a granular security model, and a rich set of APIs to support custom development.

- **Ad Hoc Reports.** Most BI tools and analytic platforms offer ad hoc functionality that enables authorized users to edit reports and dashboards by dragging and dropping fields from a semantic layer or business model onto a canvas.

- **OLAP.** Online analytic processing is making a comeback thanks to new OLAP on Hadoop products that provide interactive dimensional analysis on big data sets stored in Hadoop. OLAP uses a dimensional model that serves as a business view.

- **Visual Discovery.** Designed initially for data analysts, these business-friendly tools let users connect, blend, and visualize data and create and publish highly interactive dashboards.

- **Machine Learning.** Designed for data scientists and statisticians, these tools make it easy to prepare data, design, test, and manage predictive models, visualize results, and generate scoring code

One of the keys to achieving self-service success is mapping business users to tools. Figure 9 shows how to map classes of business users to both analytic tools and data manipulation tools.

Figure 9. Mapping Users to Tools

	Casual Users		Power Users		
ANALYTIC TOOLS	**Data Consumer**	**Data Explorer**	**Data Analyst**	**Data Scientist**	DATA TOOLS
Dashboards					Snapshots
BI authoring					Embedded data prep
Visual discovery					Data preparation
Machine learning tools					SQL, Pig, Python, etc.
	"Silver Service"		"Self Service"		

The premise here is that casual users only use BI tools to analyze data, while power users use a variety of tools. As discussed above, data explorers sit between silver service and self-service. Ideally, they use only one tool—a BI tool or analytic platform—to both consume data and create ad hoc reports. They use a BI tool's semantic layer, authoring environment, and embedded data preparation capabilities to edit existing reports or create simple ad hoc reports from scratch.

Tools Triumvirate. In contrast, true data analysts and data scientists use a triumvirate of tools to consume and create reports: data catalogs, data preparation tools, and visual discovery tools. It's important to note that not every BI and reporting tool is a visual discovery tool. Although the market is standardizing on visual discovery tools, traditional BI tools (i.e., report and dashboard development platforms) as well as custom analytic applications should give data explorers the ability to create ad hoc content.

Putting It All Together

Although visually complex, Eckerson Group's reference architecture for self-service analytics maps business users, developers, and technologies to an information supply chain that serves as the foundation for analytic processing.

The reference architecture supports iterative data workflows in which IT, power users, and casual users collaborate to create information, reports, and dashboards that meet business needs quickly. It also defines self-service and "silver service" zones and the role of IT, developers, and technology in each. And it shows how to merge top-down and bottom-up governance methods and reporting workflows to ensure business users get the information they need quickly and can differentiate between governed and ungoverned content.

It is not easy to succeed with self-service analytics. Besides a governed self-service architecture, it requires well-designed governance processes, a standard analytics and data platform, a federated organizational structure with co-located BI developers, and continuous training and support. This report lays the foundation for success by defining a reference architecture to support self-service analytics.