VIEW KEY INSIGHTS (HTTPS://WWW.GARTNER.COM/DOCUMENT/3738069/KEY-INSIGHT?REF=DDISP&REFVAL=203899904&QID=A49D9BAC64D1FB7EC477B655E14C0E3B)

Solution Path for Planning and Implementing a Data and Analytics Architecture

Published: 16 June 2017 **ID:** G00324344

Analyst(s): *Jason Lewis*

Summary

Data-driven organizations need a flexible, end-to-end architecture for
integrating and analyzing diverse data sources at scale. This research provides a step-by-step methodology to help technical professionals envision, architect and implement a comprehensive data management and analytics strategy.

Overview

Key Findings

The range of new analytics use cases is driving the need for diverse architectural capabilities, forcing organizations to adopt new technologies, proficiencies and processes. Cloud-based services can simplify this transition.

A modern and agile data management and analytics architecture requires support for real-time data pipelines, self-service data preparation and machine learning.

Digital transformation demands a comprehensive, modern data architecture capable of supporting real-time decision making, process improvement and automated actions.

Recommendations

For technical professionals focused on modernizing their data and analytics infrastructure:

Align data projects with your enterprise information management strategy. Enterprise information management (EIM) underpins the architecture and should be a first step in developing a modern data and analytics strategy.

Ensure your data and analytics team has skills to address gaps in critical areas of the architecture, especially in rapidly evolving categories like streaming analytics and machine learning.

Develop a transition plan for analytics using cloud-based services. Cloud-enabled data and analytics achieve superior scalability and offer the high-performance computing needed by machine learning.

Modernize and transform your architectural capabilities by prioritizing opportunities to expand advanced analytics, high-frequency decision making and automated actions.

Problem Statement

How can I build a holistic data management architecture to support business intelligence and advanced analytics?

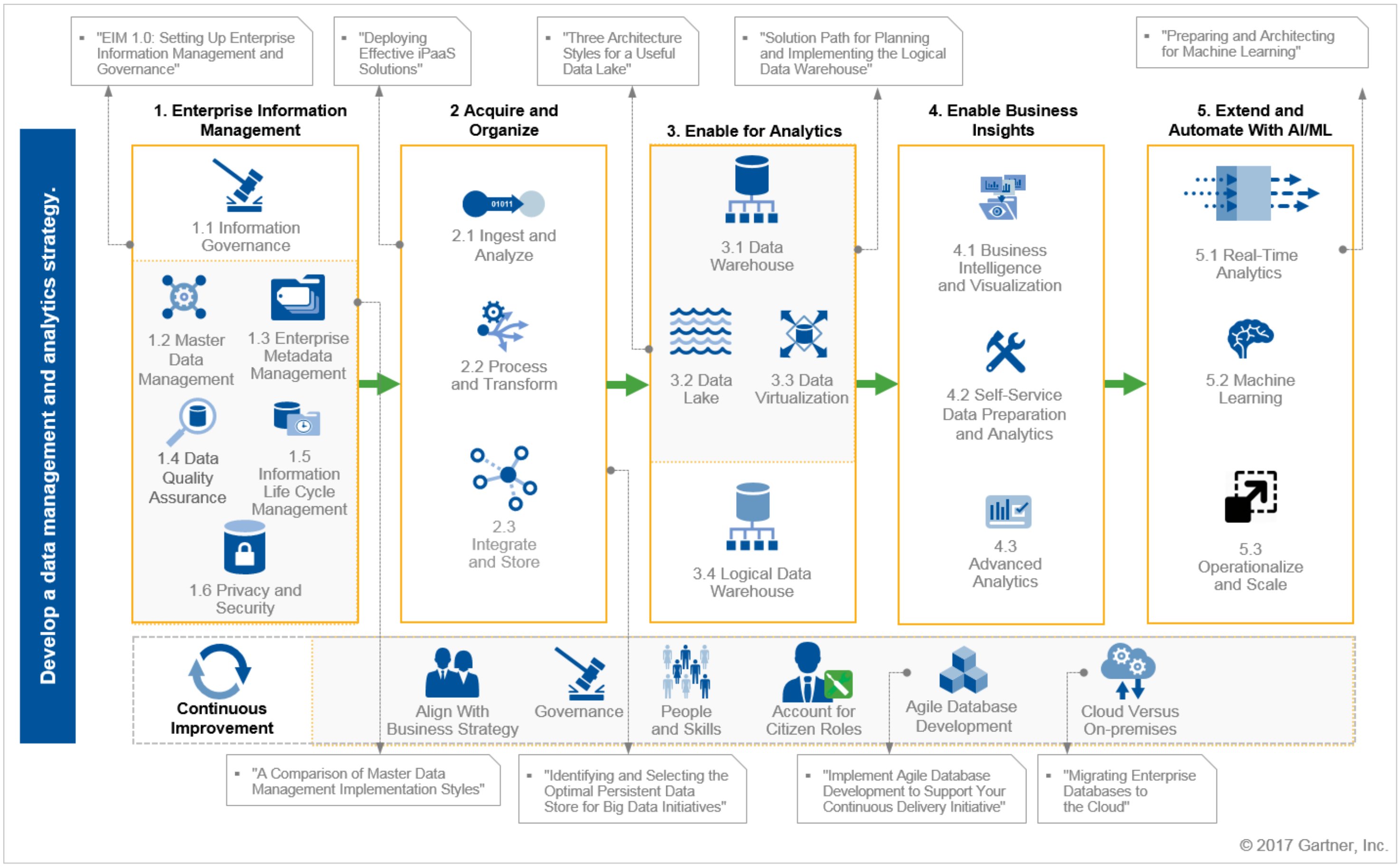
Solution Path Diagram

Many organizations struggle with how to implement a data management and analytics strategy, and ask Gartner about holistic data management architecture.

Gartner recommends that organizations adopt a build-for-change mindset to establish a highly modular data architecture. Each new capability should be implemented using an iterative execution model based on a framework that ensures consistency across key technical considerations.

This Solution Path outlines the steps in the framework (see Figure 1) and provides links to related Gartner research.

Figure 1. How Can I Build a Holistic Data Management Architecture to Support Business Intelligence and Advanced Analytics?


"EIM 1.0: Setting Up Enterprise Information Management and Governance"
"Deploying Effective iPaaS Solutions"
"Three Architecture Styles for a Useful Data Lake"
"Solution Path for Planning and Implementing the Logical Data Warehouse"
"Preparing and Architecting for Machine Learning"
Develop a data management and analytics strategy.
1. Enterprise Information Management
1.1 Information Governance
1.2 Master Data Management
1.3 Enterprise Metadata Management
1.4 Data Quality Assurance
1.5 Information Life Cycle Management
1.6 Privacy and Security
2 Acquire and Organize
2.1 Ingest and Analyze
2.2 Process and Transform
2.3 Integrate and Store
3. Enable for Analytics
3.1 Data Warehouse
3.2 Data Lake
3.3 Data Virtualization
3.4 Logical Data Warehouse
4. Enable Business Insights
4.1 Business Intelligence and Visualization
4.2 Self-Service Data Preparation and Analytics
4.3 Advanced Analytics
5. Extend and Automate With AI/ML
5.1 Real-Time Analytics
5.2 Machine Learning
5.3 Operationalize and Scale
Continuous Improvement
Align With Business Strategy
Governance
People and Skills
Account for Citizen Roles
Agile Database Development
Cloud Versus On-premises
"A Comparison of Master Data Management Implementation Styles"
"Identifying and Selecting the Optimal Persistent Data Store for Big Data Initiatives"
"Implement Agile Database Development to Support Your Continuous Delivery Initiative"
"Migrating Enterprise Databases to the Cloud"
© 2017 Gartner, Inc.

Source: Gartner (June 2017)

Solution Path

Digital business transformation demands a modern data and analytics architecture, but organizations are still struggling to maintain existing processes and coping with outdated technologies and skills. The old methods are no longer working.

Organizations need an iterative framework (as shown in Figure 1) that builds on the knowledge acquired at each stage of the process and integrates new and emerging technologies to address changing business demands. Moreover, organizations must constantly evolve their data and analytics architectures to support automated decision making and action based on an ever-changing flow of new information.

Using a framework also allows businesses to more quickly add new functionality and scale while ensuring that they retain governance and control.

At each stage of the framework, technical professionals must revisit common elements within the framework that ensure the consistent application of key decisions. These elements are shown across the bottom of the framework in Figure 1 and represent areas for continuous improvement.

This Solution Path for data and analytics will help technical professionals plan their strategy by addressing key questions, such as:

How can we ensure that information is properly classified and protected according to business requirements without sacrificing agility?

How do we manage the acquisition and integration of data from disparate sources, including cloud and on-premises?

What patterns and technologies are needed to organize and store data in support of operational and analytical projects?

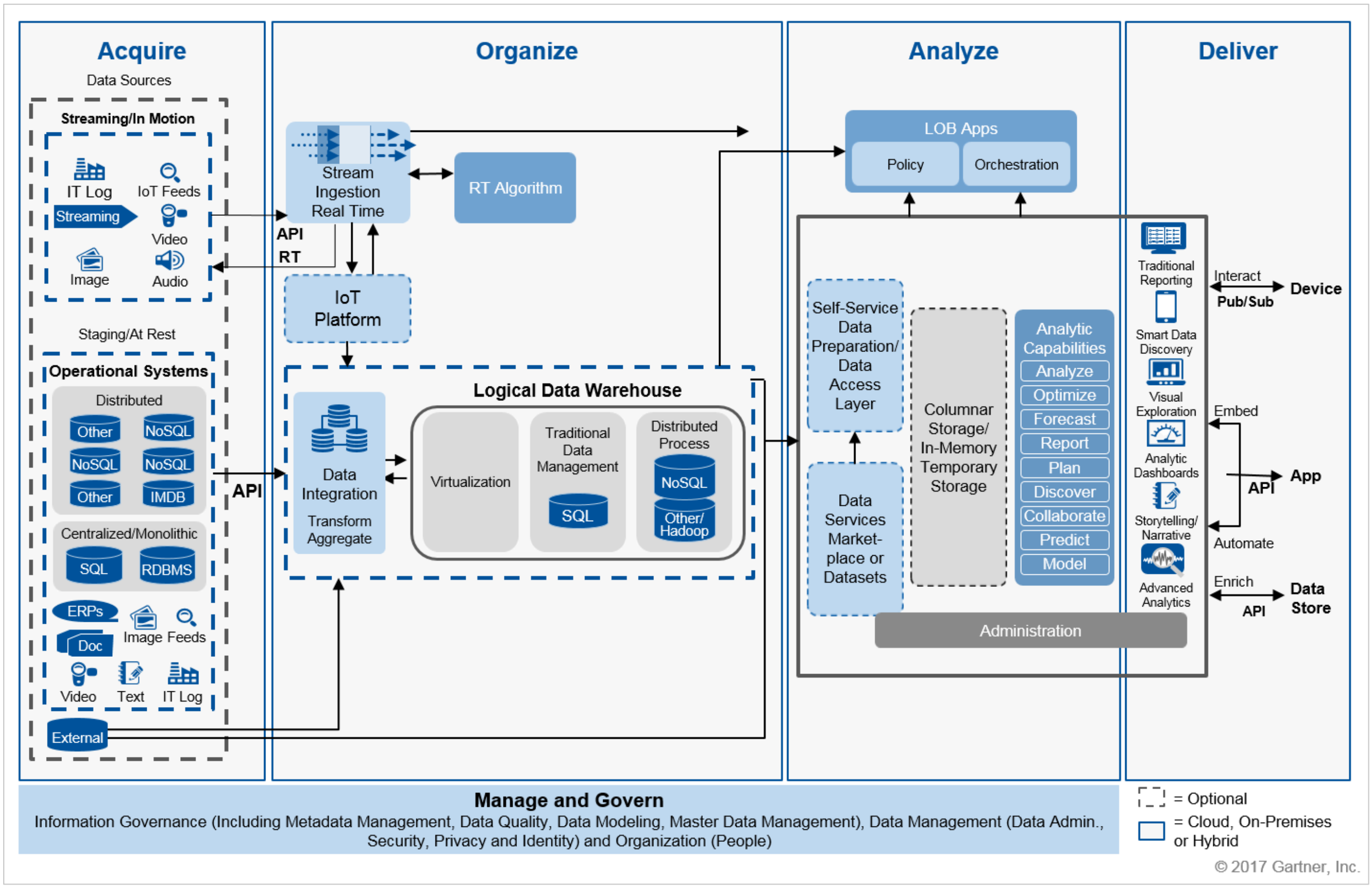
How do we extract value from data to support ongoing business decisions and create competitive advantage?

How can we adopt an agile approach to analytics, delivering insights earlier and more frequently while learning from each initiative and evolving our approach?

How can advances in machine learning and artificial intelligence help us automate actions based on these insights?

The program described by this Solution Path helps data management professionals structure the business and technical projects that will lead to an efficient and agile data management strategy with advanced architectural capabilities supporting data through creation, capture, distribution and consumption. The comprehensive, end-to-end data and analytics architecture (as shown in Figure 2) supports the full spectrum of projects outlined in this research.

Figure 2. A Comprehensive, End-to-End Data and Analytics Architecture



Source: Gartner (June 2017)

Step 1: Enterprise Information Management

The first planning step in this Solution Path, enterprise information management (EIM), builds on a business-focused foundation. EIM creates alignment with business at every stage of the architecture by promoting enterprisewide information governance, extending the business data dictionary, and encouraging data quality and context.

> Review your information governance, metadata and business data dictionary. These underpin the architecture and should be a first step in developing a modern data and analytics strategy.

Gartner defines enterprise information management as an integrative discipline for structuring, describing and governing information assets, regardless of organizational and technological boundaries, and to improve operational efficiency, promote transparency and enable business insight.

The EIM program is typically owned and championed by the CIO, or in the CIO's place, the chief data officer (CDO). It is the program that helps coordinate and organize all information initiatives to achieve alignment with business objectives. Although EIM is owned and led by the CDO, the technology components supporting the EIM program are normally managed by IT with support from and collaboration with business stakeholders.

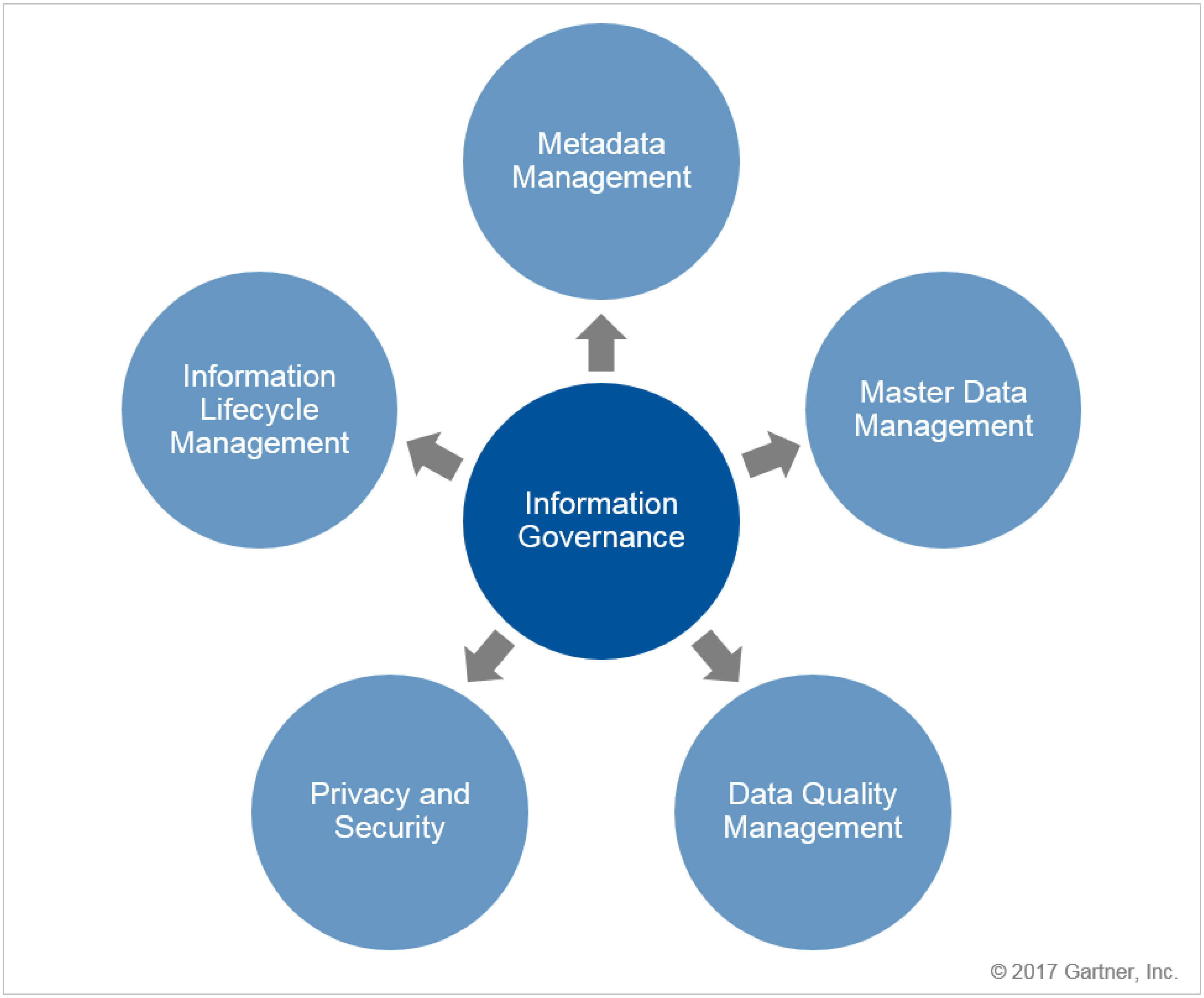
EIM synchronizes decisions between strategic, operational and technical stakeholders, coordinating efforts to improve the organization's information capabilities.

A data management and analytics strategy benefits from an EIM program with technology, processes and tools supporting the key areas surrounding information governance:

- Metadata management
- Master data management
- Data-quality management
- Information life cycle management
- Privacy and security

Each of these areas promotes the goals of information governance, as shown in Figure 3.

Figure 3. EIM Components and Their Relationship to Information Governance



Source: Gartner (June 2017)

Each category is covered in more detail in subsequent substeps in the Solution Path. However, it's important to note that many EIM programs fail to achieve their full potential when individual projects are managed separately without coordination and recognition of their interdependence within an EIM framework. Technical professionals must adopt a cross-functional view of data and its use as the foundation for information and knowledge across the organization.

EIM transcends technical problem solving and helps technical professionals think beyond functional job responsibilities and project-specific objectives.

During this step, the data management professional should be asking:

What information assets do we have, and where are they?

What information should be managed?

What information and processes should be governed?

Can we agree on the interpretation of data in each business context?

Who is responsible for the data we store?

Who should be accessing the data?

Where did the data originate and where is it being used?

How can we avoid disruptive events caused by bad data?

Gartner has published a decision framework to help organizations through this process (see the Gartner research "EIM 1.0: Setting Up Enterprise Information Management and Governance" (https://www.gartner.com/document/code/294451?ref=grbody&refval=3738069&latest=true)). Gartner's method for designing and implementing EIM and data governance helps information management professionals to create EIM programs that are easy to understand and adaptable to change.

1.1 Information Governance

Information governance is a decision-making framework for assigning rights, responsibilities and authorities to ensure that an enterprise, its regulators and its shareholders receive reliable, authentic, accurate and timely information.

Although information governance is presented as an early step in the Solution Path, information governance is not a project, but an ongoing program that must be integral to each new data management initiative. Therefore, organizations should consider this step as both a foundation and an ongoing program with regular checkpoints and considerations for new information sources (see the Iterate for Continuous Improvement section below).

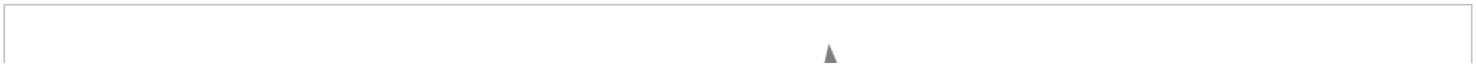
The foundational information governance program will establish processes and controls that will ensure that information is accurate and nonredundant. This requires the introduction of formal management controls in the form of systems, processes and data stewards who serve as custodians of the data.

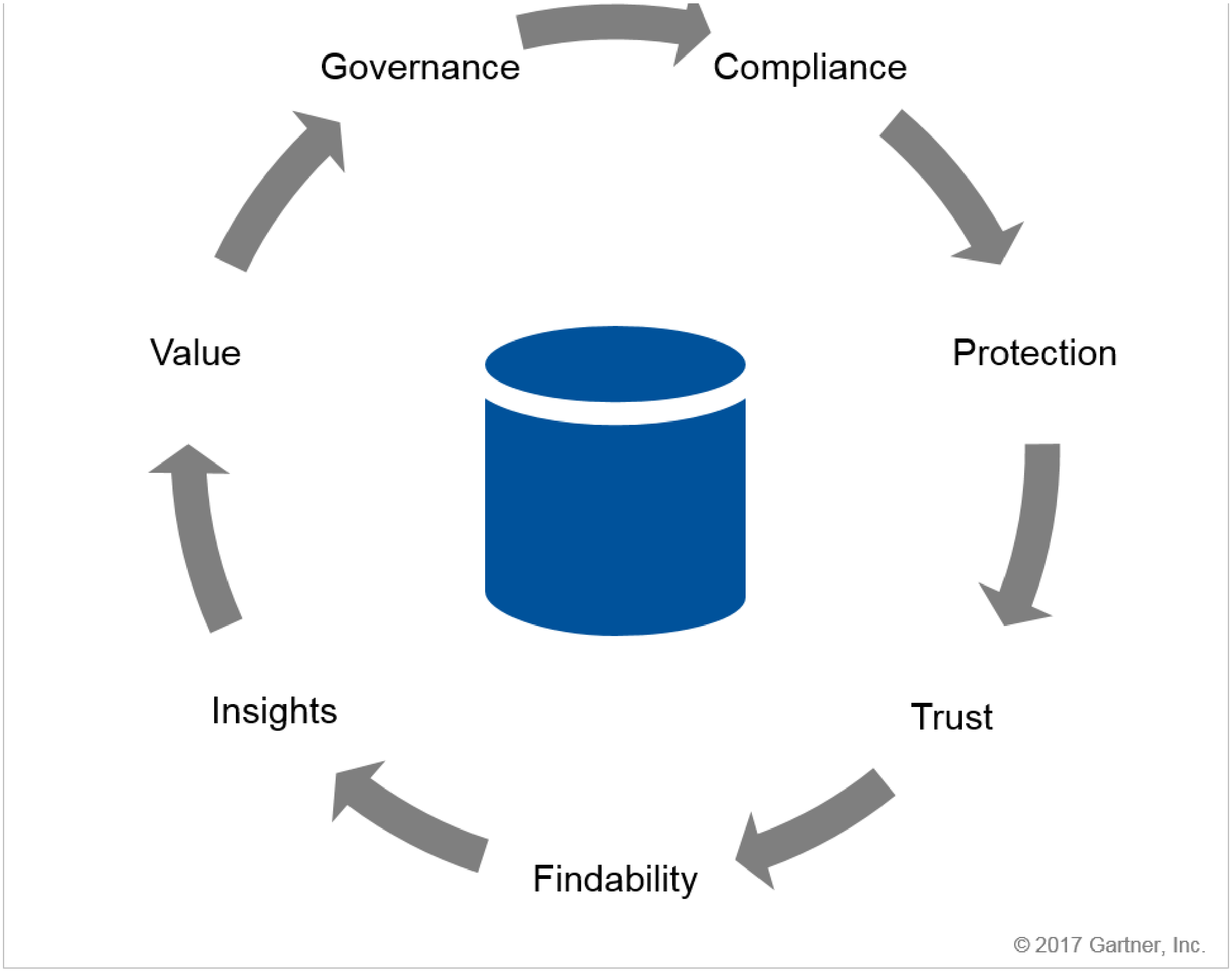
Each new information management initiative must operate within the constraints of the information governance program to create ongoing consistency and quality. This means that each new data source must undergo data cleansing to prevent ingestion of corrupted, inaccurate, duplicated or nonessential data (see Step 1.4, Data Quality Assurance).

Information management and governance is more than just protecting sensitive data by restricting access and enforcing compliance (see Step 1.6, Privacy and Security). Successful information governance protects information while simultaneously promoting awareness and understanding of the business's data assets to help drive desired business outcomes.

This interrelationship between EIM components creates a virtuous cycle, as shown in Figure 4.

Figure 4. EIM and Information Governance Creates a Virtuous Cycle





Source: Gartner (June 2017)

1.2 Master Data Management

The proliferation of enterprise applications that store information about customers, products and other information assets has made it difficult to create, maintain and enable a single, trusted, shareable version of master data across business domains. Organizational changes such as mergers and acquisitions introduce additional challenges as new information assets must be reconciled to support business processes and decision making.

The contention between the mastering of shared data objects leads to inconsistencies in data quality and classification, requiring reconciliation through complex and error-prone data transformation processes. Data inconsistencies result in low confidence in data and inhibit business decisions that rely upon that data. Without enterprisewide agreement on commonly reused master data domains, entities and attributes, organizations cannot be totally effective or efficient in the execution of many business and IT programs. To address these inconsistencies, Gartner recommends master data management (MDM) to help data practitioners create a single version of truth for specific information assets.

MDM is a technology-enabled business discipline in which business and IT work together to ensure the uniformity, accuracy, stewardship, governance, semantic consistency and accountability of an enterprise's official shared master data assets. Master data is the consistent and uniform set of identifiers and extended attributes that describe the core entities of the enterprise.

Each organization should assess the cause of data inconsistencies to decide where they will focus their MDM efforts. There are six characteristics that will drive the MDM implementation's focus:

- Where the master data is authored
- Where the master data is verified

The latency of master data movement

The degree to which a physical "golden record" is instantiated

The usage of the master data

Search complexity

Gartner has published a guide to four MDM implementation styles to help organizations choose a strategy aligned with their unique requirements (see the "A Comparison of Master Data Management Implementation Styles" (https://www.gartner.com/document/code/276842?ref=grbody&refval=3738069&latest=true)).

1.3 Enterprise Metadata Management

The next step is to establish enterprise metadata management (EMM). Because of the growing variety and volume of data, data lakes are becoming an essential part of the architecture. The data lake pattern employs technology that supports data from a variety of different sources, such as files, clickstreams, Internet of Things (IoT) sensors and social networks. The unconstrained storage model of the data lake offers extreme flexibility, but lacks the assurances and context enforced within an enterprise data warehouse. EMM offers an effective solution to this problem.

An effective metadata management approach enables the following information capabilities:

Describe: To collect knowledge about information assets

Organize: To align and structure information assets so they can be readily found and easily consumed by other capabilities of the platform

Share: To make data available to consumption points

Govern: To provide for control, levels of consistency, protection, quality assurance, risk assessment and compliance

To understand the benefits of EMM, it's useful to review four types of metadata:

Technical metadata: Describes the technical attributes including the form, type and structure of each dataset.

Operational metadata: Captures the lineage and provenance of the data, along with audit details about the success or failure of job runs and update frequency.

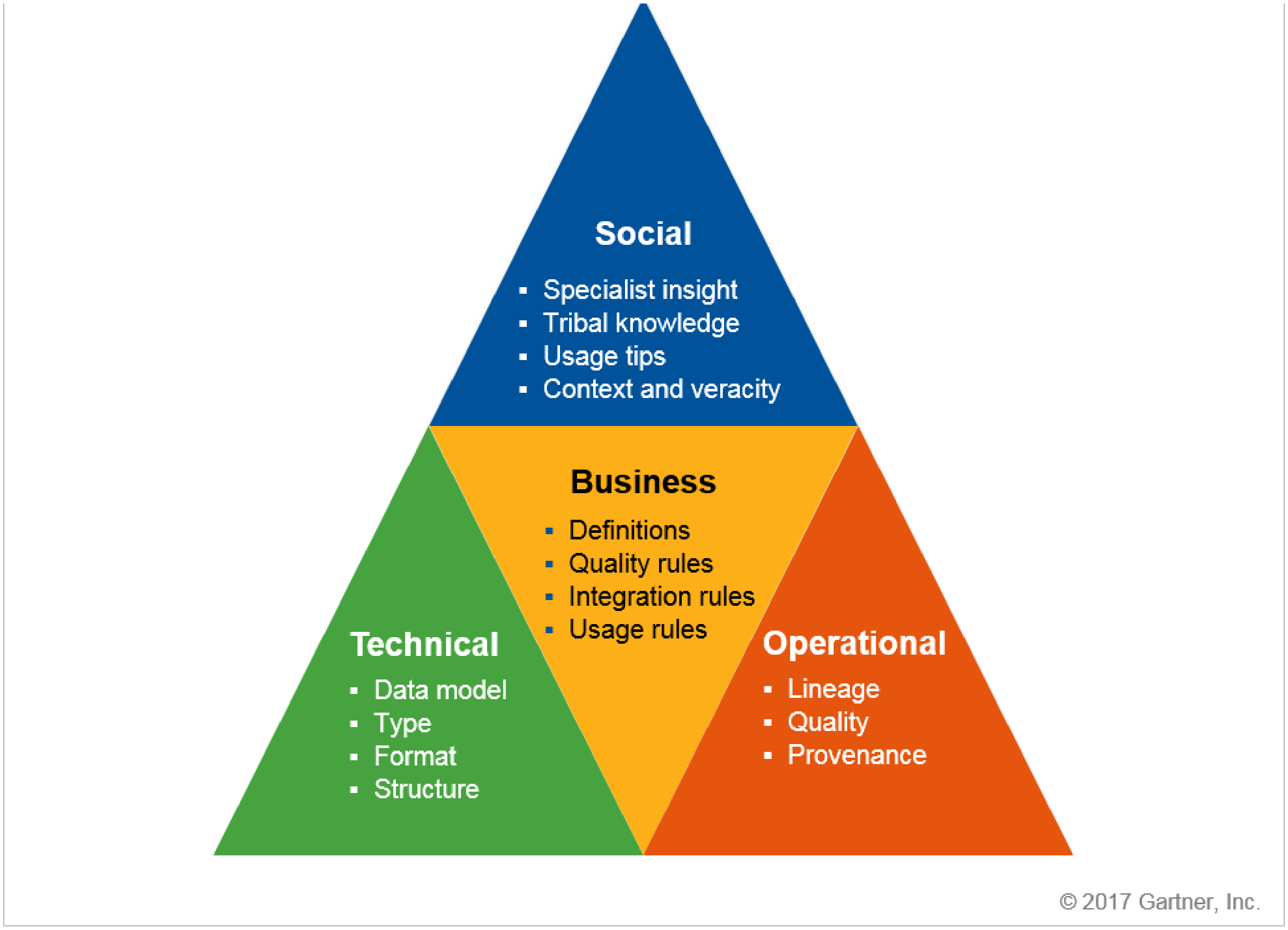
Business metadata: Captures specific business attributes to help promote context, findability and meaning. Business metadata gives end users a way to find and understand information assets based on business semantics.

Social metadata: Gives end users the ability to tag and describe data with information meaningful within their specific business context. This crowdsourced metadata can offer tremendous collaborative benefits to citizen data scientists working with new and exotic data sources.

> Social metadata promotes discovery, but more importantly, provides meaningful insights from specialist and citizen data scientists that evolve new insights about the data on each new project.

Social collaboration plays a critical new role in metadata strategy by providing context from users that work closely with the data — see Figure 5.

Figure 5. Social Metadata — The Tip of the Metadata Spear



Source: Gartner (June 2017)

Organizations can publish a centrally managed EMM data catalog/repository and business glossary to provide a consistent set of definitions and descriptions for each data element. Such a strategy adds coherence to business metadata and ensures reliability in data semantics.

By combining different types of metadata, the data catalog/repository gives insight into:

- The movement of data within jobs or between systems
- What data is available within the organization and how it can be used
- Impact analysis to determine the dependencies between information assets
- Common business vocabulary and accountability for its terms and definitions
- Audit trail for compliance

EMM technologies offer automated metadata discovery to inspect data elements for semantics and take specific actions based on workflow rules. For example, inspecting datasets for operational metadata combined with business metadata to determine the information classification of incoming data. This could be used to ensure policy/regulatory compliance or for the dynamic application of data masking or obfuscation techniques.

As previously mentioned, EMM is an important capability for information governance. To ensure consistency with information governance, metadata management processes must support the following:

- Metadata needs to be produced through reliable, sustainable, governed processes based on defined standards.

Storage in a common model that enforces standards and is managed in an integrated repository.

Data consumers need to be able to access the data from one central place. They must be able to provide feedback about the metadata to improve it over time.

1.4 Data Quality Assurance

As mentioned earlier, information governance promotes awareness and understanding of the business's data assets to help drive desired business outcomes. In this step, a data quality assurance strategy is adopted or adapted to improve reliability and usability of information by ensuring that data is fit for purpose in downstream business processes.

These processes range from those used in core operations to those required by analytics and for decision making, regulatory compliance, and engagement and interaction with external entities. As a discipline, data quality assurance covers much more than technology. It also includes roles and organizational structures; processes for monitoring, measuring, reporting and remediating data quality issues; and links to broader information governance activities via data-quality-specific policies.

Successful data quality assurance depends on a strong partnership between IT and business stakeholders, along with clear roles and responsibilities. For example, organizations often rely on information stewardship for the enforcement of information governance policies and rules. Data steward is a role that is established within line of business — not IT. Data stewards work on behalf of the business stakeholders, enacting the policies created and working to ensure data conforms to expectations. They actively monitor the quality of the data (via data quality metrics and visualization techniques) and take corrective action when data in certified sources does not align with policy. In addition, data stewards are instrumental in influencing the people around them to rely on certified information sources in their reporting and analysis.

Technical professionals support the techniques and technologies used by data stewards for discovering and investigating data quality issues, such as duplication, lack of consistency, and lack of accuracy and completeness. This is accomplished by analyzing one or multiple data sources and collecting metadata that shows the condition of the data and enabling the further investigation into the origin of data errors.

> Data quality assurance should be applied at all stages of the data and analytics architecture. However, efforts to improve data quality early in the data life cycle will reduce friction and improve efficiency.

Organizations should apply data quality assurance across the enterprise and data life cycle:

Adopt comprehensive data quality assurance processes to ensure consistency and avoid duplication of efforts.

If data quality is not recognized as a priority, enlist support from business sponsors to create a business case for data quality investments.

Codify expectations of data usability by cataloging requirements for data completeness, accessibility, format, presentation, timeliness, relevance and accuracy.

Use metadata to assess and classify data quality characteristics.

Implement data profiling during ingestion and integration, but offer self-service data preparation to extend quality assurance capabilities to end users who are closest to the data.

1.5 Information Life Cycle Management

Information life cycle management (ILM) is an approach to data and storage management that recognizes that the value of information changes over time and that it must be managed accordingly. ILM seeks to classify data according to its business value and establish policies to migrate and store data on the appropriate storage tier and, ultimately, remove it altogether. ILM has evolved to include upfront initiatives like master data management and compliance.

Data management professionals should consider ILM classification strategies, such as:

- Categorizing data according to business life cycle rules
- Identifying frequently and infrequently accessed data and classifying as hot, warm, cold
- Identifying data by type for different storage tiers
- Considering data retention, compliance and security requirements mandated by regulations

Business continuity strategies strive to keep businesses operating in the event of a disaster. Data protection is a key component of business continuity.

Particularly with historical or analytical data use cases, organizations can take advantage of tiered storage models to move data from nodes with faster storage (and higher compute capability) to nodes with lower-cost storage models. This strategy allows for cost-effective and scalable storage platforms for data that would otherwise require higher-cost resources. Cloud service providers (CSPs) offer tiered storage to take advantage of different ILM classifications.

Information governance may alter or preclude ILM handling procedures. You should:

- Require encrypted storage for certain classifications
- Prevent storage or replication across regional boundaries
- Require archival or purging of data based on information retention policies

1.6 Privacy and Security

Data privacy and security is another aspect of information governance that requires the orchestration of data security policies across disparate data stores. Data discovery products can help identify unprotected sensitive data and trigger appropriate data protection measures to:

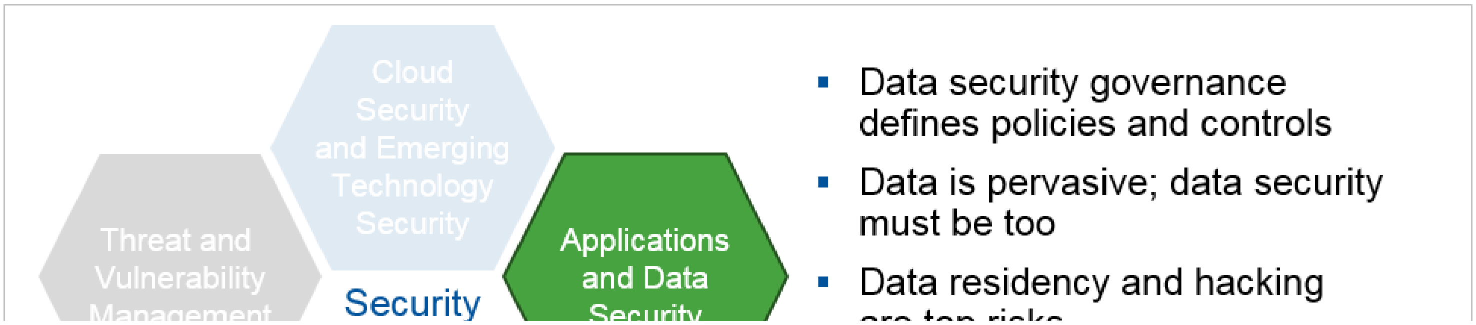
- Apply metadata or labeling
- Apply data tokenization, masking and/or redaction

Technical professionals should implement technologies to enforce information governance policies, such as:

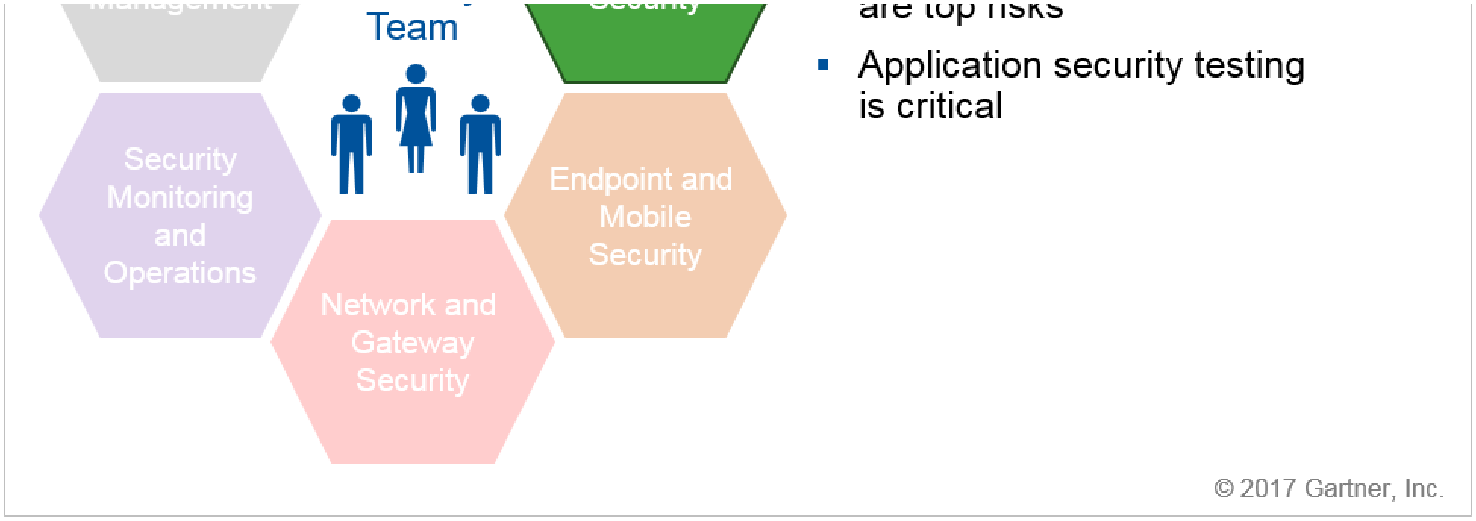
- Requiring encryption at the application or data layer, including storage and at-rest
- Implementing role-based access controls to protect data elements from unauthorized consumption or alteration
- Segmenting data for different audiences based on information classifications

Data privacy and security are part of a much broader information security agenda (see Figure 6). Data management professionals should work closely with information security professionals to ensure appropriate security measures are applied.

Figure 6. Application and Data Security Overview


Cloud Security and Emerging Technology Security
Threat and Vulnerability Management
Applications and Data Security
Security

- Data security governance defines policies and controls
- Data is pervasive; data security must be too
- Data residency and hacking are top risks


Management
Team
Security
Security Monitoring and Operations
Endpoint and Mobile Security
Network and Gateway Security
© 2017 Gartner, Inc.

- are top risks
- Application security testing is critical

Source: Gartner (June 2017)

For more information on data privacy and security, see the following Gartner research:

"Securing the Big Data and Advanced Analytics Pipeline" (https://www.gartner.com/document/code/313004?ref=grbody&refval=3738069&latest=true)

"Four Steps to Secure Modern Databases" (https://www.gartner.com/document/code/290961?ref=grbody&refval=3738069&latest=true)

"Protecting Big Data in Hadoop" (https://www.gartner.com/document/code/271209?ref=grbody&refval=3738069&latest=true)

What to Consider Before Moving to the Next Step

Gartner recommends that most organizations should consider the following before moving on to the next step:

Consolidate EIM initiatives to ensure functional collaboration, and reduce overlap and inconsistencies arising from fragmented projects.

Establish a framework for applying data governance to new data management and analytics programs.

Step 2: Acquire and Organize

The next step of the Solution Path, Acquire and Organize, prepares the architecture to collect and assimilate information, regardless of frequency, structure or origin. An effective data and analytics strategy allows for discovery, ingestion and integration of all available data from the source, as fast as it is produced in any format and quality.

This stage involves three related subtasks:

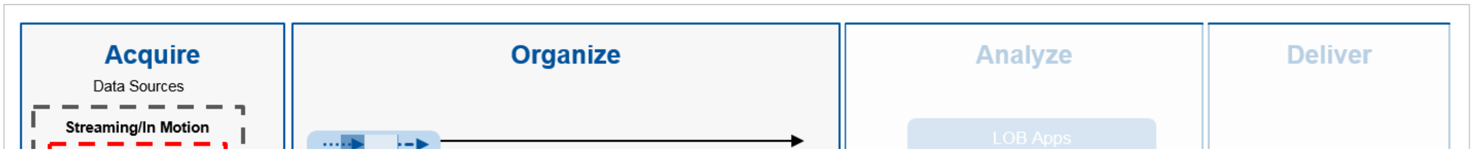
Ingest: Ingestion describes the process of acquiring, importing, transferring or loading data for transformation, processing or storage.

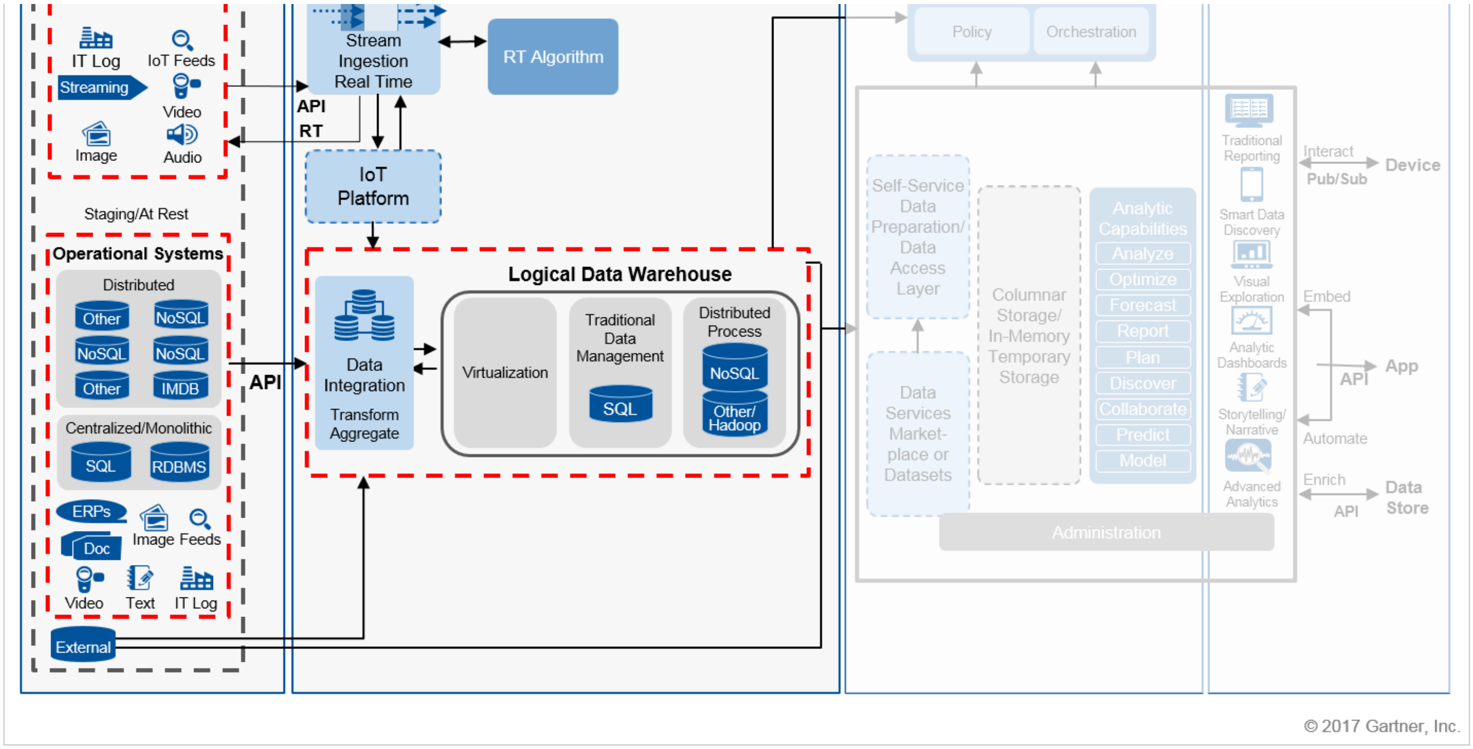
Process and transform: This is the stage where data is prepared, cleaned, modified or enriched before use or storage.

Integrate: Integration describes the process by which data is retrieved from disparate sources and combined to provide a unified view of the data.

The architectural components that support ingestion, transformation and integration are closely related and are commonly used in concert throughout this stage — see Figure 7.

Figure 7. Data and Analytics Architecture — Acquire and Organize





Source: Gartner (June 2017)

Data practitioners should conceptualize this part of the architecture as a managed data pipeline composed of a series of stages through which the data flows.

The following patterns support different use cases for ingesting, transforming and integrating data:

> **Data replication and synchronization** is the process of copying data from one data store to another. This is a well-established integration pattern and is used in traditional ETL processes to manage the acquisition and integration of information from heterogeneous or compatible systems.
>
> **Batch processing** is an asynchronous method for ingesting and storing an immutable and constantly growing dataset in bounded intervals — that is, the incoming data has a beginning and an end. This method is extremely effective for processing high volumes of data for specified intervals and allows for the computation of arbitrary functions on the dataset, which in turn, are output in batches.
>
> **Stream processing** is an asynchronous method of ingesting an infinite, constantly growing data stream. Such streams often include unbounded, unordered datasets produced at high velocity. Examples include web logs, mobile events and sensor data, but can represent any unbounded dataset. Discrete portions of the stream, known as windows, can be captured through some characteristic of the data, such as event timestamps.
>
> **Data virtualization** is a form of data integration based on data abstraction and provides a consistent interface to data distributed across multiple, disparate data sources and repositories. Modern data virtualization tools provide both read and write access to a host of popular data types and sources (such as relational, Hadoop, NoSQL, flat files and cloud data stores).
>
> **Message-oriented** patterns are often associated with service-oriented architecture (SOA), messaging captures, transforms and delivers data through message brokers. Data is often delivered in near real time.

These patterns are not mutually exclusive and are often used together to provide a broad range of options within the architecture. When implementing each subtask, revisit these patterns to select the most appropriate style for each use case.

2.1 Ingest and Analyze

The ingest step ensures that the data and analytics architecture is capable of collecting, replicating and storing data from many different types of data sources with support for high volumes and low latency.

The acquisition of data requires different strategies depending on the data source, the velocity of information and the type of information to be consumed. Organizations should adopt an architecture capable of supporting multiple styles of data ingestion:

- Batch data ingestion and ingestion of data at rest
- Real-time data stream ingestion with low latency
- Reliable and durable message ingestion
- Scalability, to support high throughput of incoming data
- Decoupling of data sources from data subscribers
- Support for some transformation before handing over the output stream to consumers
- Ability to act as both input source and output sink to other services

Build a scalable ingestion process with a repeatable data pipeline:

- Capture and add metadata during ingestion to understand where data originated (lineage) and identify where it was processed or stored
- Implement ingestion workflow to halt processing, reroute data or take actions based on metadata, quality or other characteristics of the data
- Implement support for a variety of endpoints and data types with low impact to data sources
- Avoid custom code and automate ingestion with managed, automated and repeatable pipeline definitions
- Implement change data capture to optimize ingestion workloads

2.2 Process and Transform

Once data is created or ingested, it usually requires some processing and transformation before it can be used. In this step, the data is prepared, cleaned, modified or enriched so that it will be fit for the business use case.

A traditional approach used an ETL process that involved extracting data from the source, transforming it to fit operational needs and loading it into staging tables or a data warehouse.

Another approach relies on an ELT process where data is extracted from a source, then loaded directly into a staging table or data lake in its raw, unmodified state. Transformation occurs in the staging area before being loaded into a target database or data warehouse.

As with data ingestion, data transformation can benefit from a managed data pipeline.

- Monitor the quality of data as it flows through the pipeline and implement predefined processing rules to apply transformation, enrichment or structure when needed.
- Apply data privacy and security policies by segmenting, masking or tokenizing data before it gets published for consumption.
- Automate scheduling and orchestration of data movement between heterogeneous storage environments.
- Apply data life cycle policies to move data across tiered storage or cloud and on-premises based on hot, warm, cold classification.

2.3 Integrate and Store

DATA INTEGRATION

Data integration comprises the practices, architectural techniques and tools for collecting data from disparate sources and combining that data into a unified view to meet the data consumption requirements of all applications and business processes.

Unfortunately, the process of integrating data has become more complicated due to several factors:

- Data is no longer centrally stored and managed in corporate data centers. Integration strategies must account for data sourced from diverse locations, cloud infrastructure and external parties.

- Continuous data streams demand different patterns and technologies with strategies for real-time integration and analytics.

- Today's integration techniques can't rely on predictable data structure and schema, and must be capable of dealing with unstructured data in a variety of formats.

- Business users are demanding self-service integration capabilities — (see the Self-Service Data Preparation and Analytics section below).

These challenges necessitate an agile and flexible data integration strategy that can support different styles of data integration, each at the appropriate stage in the architecture. However, care should be taken to avoid data integration strategies that rely on hard-coded, unmanaged interfaces. Without proper governance, point-to-point integrations can become unmanageable and result in redundant work and higher maintenance costs.

Data management professionals should be familiar with the following integration styles:

- **Embedded integration:** Typically delivered through commercial applications, embedded integration is a component of broader IT solutions. It includes three subtypes: embedded in databases, embedded in operational software and embedded in analytics software.

- **Stand-alone integration:** Independent of databases and applications, stand-alone integration is often delivered through separate integration middleware. Stand-alone integration is business-friendly and includes three subtypes. The first subtype is integration platform as a service (iPaaS). The second subtype is integration software as a service (iSaaS). The third subtype is data federation/virtualization, which is mentioned in more detail later in the next section.

- **Data preparation:** Data preparation provides business-friendly and data-centric capabilities, such as data access, data discovery, data cleansing, data transformation, data enrichment and data collaboration. Data preparation functionality can be offered either as stand-alone tools or as embedded capabilities in an analytics platform. More detail on data preparation can be found later in this document.

For more information about the pros and cons of data integration styles, see "Use Data Integration Patterns to Build Optimal Architecture." (https://www.gartner.com/document/code/270543?ref=grbody&refval=3738069&latest=true)

For additional information, please refer to the following research:

- "Comparing Three Self-Service Integration Architectures" (https://www.gartner.com/document/code/297311?ref=grbody&refval=3738069&latest=true)

- "Comparing Four iPaaS-Based Architectures for Data and App Integration in Public Cloud" (https://www.gartner.com/document/code/289852?ref=grbody&refval=3738069&latest=true)

- "Deploying Effective iPaaS Solutions for Data Integration" (https://www.gartner.com/document/code/324279?ref=grbody&refval=3738069&latest=true)

DATA STORES

Data stores range in variety from the simple, such as text files, to the complex, such as distributed processing and storage frameworks like Hadoop. There are architectural choices, such as persistent disk storage versus in-memory processing and infrastructure considerations, such as cloud versus on-premises. Technical professionals have a multitude of options when it comes to data stores and database management systems (DBMSs), and often ask Gartner, "What data store is most appropriate to address my particular use case?"

Answering this question depends on several factors and characteristics of the data and the consumption model, including:

- Transactional guarantees

- Consistency

- Availability

- Schema flexibility
- Scalability
- Cost performance
- Manageability
- Recoverability

There are also architectural and business considerations, such as:

- General versus use-specific
- Community support versus vendor support
- Technical maturity
- On-premises versus cloud
- Managed versus unmanaged
- Single vendor versus multiple vendors

Choosing the components that comprise the data storage and management architecture is a critical step and will have long-lasting implications for the maintainability, performance, availability and efficacy of the overall architecture.

Data architects will need to evaluate the requirements and constraints for each type of DBMS to decide which option is most appropriate for a particular use case. To assist with this process, refer to the following research:

- "Decision Point for Selecting a DBMS Architecture" (https://www.gartner.com/document/code/274013?ref=grbody&refval=3738069&latest=true)
- "Identifying and Selecting the Optimal Persistent Data Store for Big Data Initiatives" (https://www.gartner.com/document/code/322578?ref=grbody&refval=3738069&latest=true)

Today's cloud-based data services represent a strong alternative to on-premises deployments. Cloud-based operational and analytical database services offer deployment speed, flexibility and scalability that are difficult, or even impossible, for enterprises to achieve with on-premises systems.

For a guide to evaluating cloud-based database services, see — Evaluating Microsoft Azure's Cloud Database Services (https://www.gartner.com/document/code/311029?ref=grbody&refval=3738069&latest=true) and Evaluating the Cloud Databases From Amazon Web Services. (https://www.gartner.com/document/code/303836?ref=grbody&refval=3738069&latest=true)

For a comparison of cloud-based managed big data services, see "Assessing Cloud-Based Big Data Services: Amazon EMR vs. Microsoft Azure HDInsight." (https://www.gartner.com/document/code/317100?ref=grbody&refval=3738069&latest=true)

What to Consider Before Moving to the Next Step

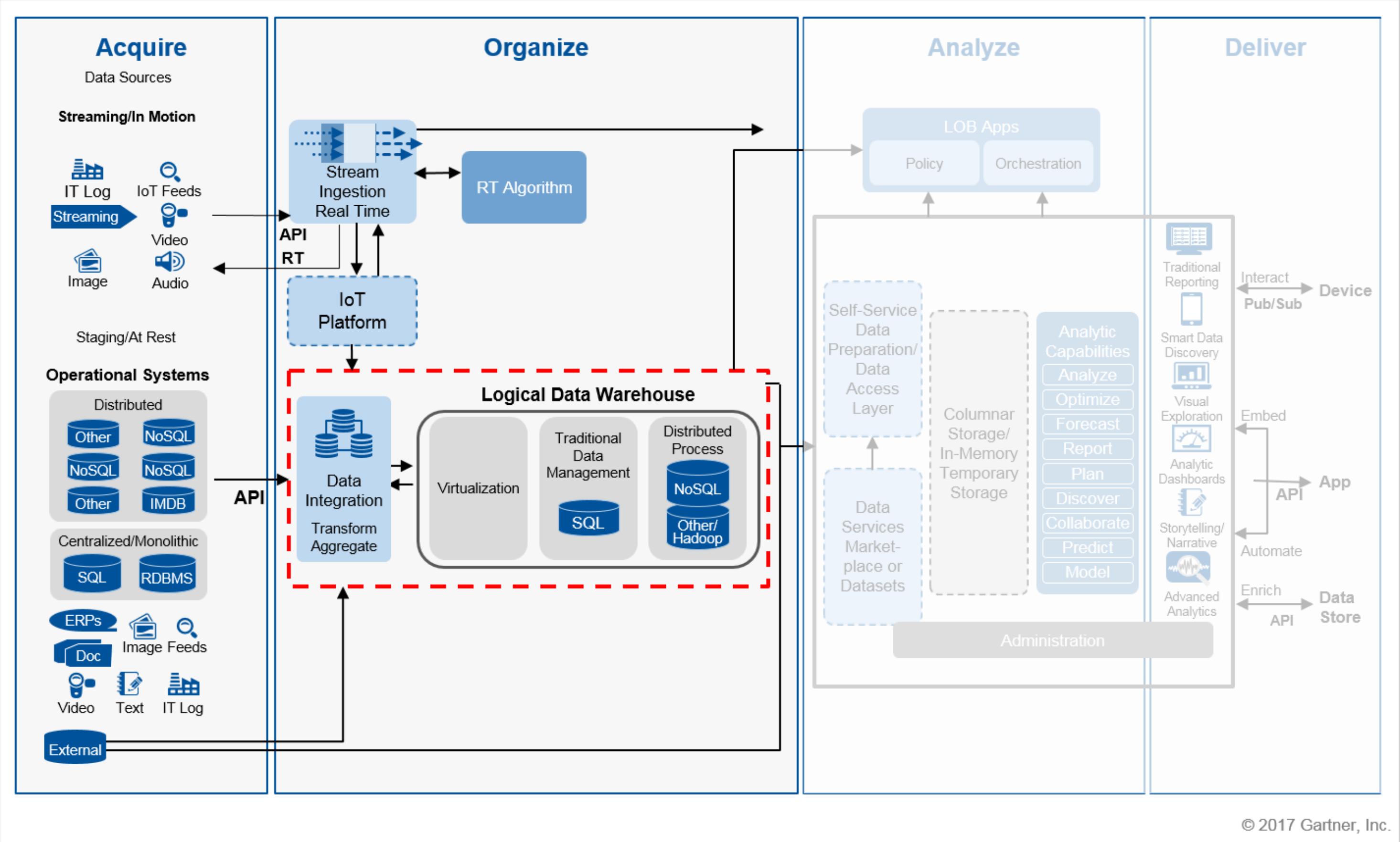
Gartner recommends that most organizations should consider the following before moving on to the next step:

- Review known and anticipated requirements necessitating new architectural capabilities, such as streaming ingestion, real-time data streams and integration between heterogeneous systems.
- Evaluate cloud services and hybrid cloud strategies to support data migration requirements.

Step 3: Enable Data for Analytics

In this step, the data is prepared for consumption by end users, applications and business processes in support of analytics. The data could be stored in memory, on disk, and/or distributed across physical or logical data stores — see Figure 8.

Figure 8. Data and Analytics Architecture — Store and Enable



Source: Gartner (June 2017)

3.1 Data Warehouse

A data warehouse (DW) is a storage architecture designed to hold detailed data extracted from transactional systems, operational data stores and external sources. The DW then combines that data in an aggregate, summary form suitable for enterprisewide data analysis and reporting based on predefined business needs. The DW also supports ad hoc queries on the detailed data to enable analysis beyond predetermined use cases.

The DW architecture depends on a "schema on write" model where data conforms to a known and expected structure and format. This means data can't be loaded into the data warehouse without considerable planning and effort to analyze data sources and the business requirements for the individual data elements.

The DW construct is ideal for business users that analyze operational metrics, since it is purpose-built to answer predefined and ad hoc questions about established business metrics. It's this design goal that makes the data warehouse an important component in most organizations.

However, the DW model has certain shortcomings in modern digital business. Due to the highly structured and predefined data requirements, the DW can't accommodate nontraditional data sources such as web server logs, sensor data and social network activity. Because of the planning and expense of adding new information sources, if data doesn't answer specific questions, it will usually be excluded from the data warehouse. This makes finding new and unexpected insights impossible using a DW alone.

Implement a data warehouse when you need:

A system that can aggregate and analyze data from transactional and operational systems

A prestructured data model designed to support enterprise reporting

- An efficient analytical system optimized for reporting and analysis of common business subjects with high concurrency and low latency

Gartner recommends that every data warehouse, whether new or old, be built on or evolving toward a logical data warehouse (LDW) model to effectively support business intelligence and analytics. The LDW is covered in more detail in a subsequent section.

For more information on data warehouse strategies and platforms, refer to the following research:

- "Comparing Cloud Data Warehouses: Amazon Redshift and Microsoft Azure SQL Data Warehouse" (https://www.gartner.com/document/code/309107?ref=grbody&refval=3738069&latest=true)

- "Solution Path for Planning and Implementing the Logical Data Warehouse" (https://www.gartner.com/document/code/320563?ref=grbody&refval=3738069&latest=true)

3.2 Data Lake

A data lake represents a collection of storage instances of various data assets, which are often stored in a near-exact, or even exact, copy of the source format. Unlike the data warehouse, it is up to the users of the lake to interpret the data and to determine the best data applicability for the identified use cases.

Data lakes are not replacements for primary systems of record or data warehouses. Instead, they complement existing efforts and support the discovery of new questions. A data lake contains unrefined data where the data structure may not be known in advance, or when organizations want to increase analytics and operational agility by complementing their systems of record with systems of insight.

Use a data lake for:

- Consolidating data in its raw unrefined state from a variety of different data sources

- A general-purpose staging area

- Collecting all data and attributes with the idea that new insights can be derived from analyzing larger volumes of data or new types of data

- An active archive of historical data

The data lake is a useful construct for managing data in different zones or layers, with each layer optimized for different styles of consumption. These patterns are optional and can easily coexist within a lake, each layer serving its own use cases:

1. **Transient layer:** This acts as a landing area supporting the ingestion of many disparate data sources. This layer holds all raw, unrestricted data including sensitive information (personally identifiable information [PII], protected health information [PHI]) in its original, unaltered form. The layer is deemed transient because it serves merely as a staging area to fill other layers.

2. **Discovery layer:** This holds all raw data to promote exploration and discovery. However, sensitive data is tokenized or masked to prevent unauthorized consumption. This might be to conform to compliance or regulatory requirements to protect sensitive information. This layer can be thought of as an unrefined and unaltered sandbox for exploration and advanced analytics.

3. **Refined layer:** In this layer, enrichments and transformations are processed to create new datasets, which are made available to downstream applications and processes. Data is integrated into a common format with data validation and cleansing techniques applied.

4. **Trusted layer:** The trusted layer is a step beyond the refined layer, with reference data reconciled to ensure consistency with master data policies.

For a deeper analysis of data lake architectural styles, see "Three Architecture Styles for a Useful Data Lake." (https://www.gartner.com/document/code/303817?ref=grbody&refval=3738069&latest=true)

Along with the data warehouse, the data lake is an important component of the LDW.

3.3 Data Virtualization

When combined with the data warehouse and data lake, data virtualization (DV) becomes an integral component of the LDW (see next section). DV can provide a uniform interface to multiple data stores, allowing users easy access to all the organization's data (subject to security controls). DV can also used by business intelligence (BI) and reporting tools to enable analytics on data residing in different repositories.

To understand the concept of data virtualization, it's useful to look at the different styles/types:

> **Embedded virtualization:** The virtualization technology is functionally embedded in a BI tool, allowing the BI software to make multiple calls to back-end databases to provide a consolidated view for analytics or reporting.
>
> **Physical virtualization:** Data is retrieved from disparate data sources and consolidated into new physical data structures for consumption from a unified source.
>
> **Dynamic virtualization:** In this model, a virtualization engine acts as a query orchestration manager, accepting queries and decomposing those queries into subqueries to be run against multiple data sources. The virtualization engine is capable of delegating subqueries to the back-end source system for independent processing or caching data from multiple sources so that it can perform processing tasks on its own.

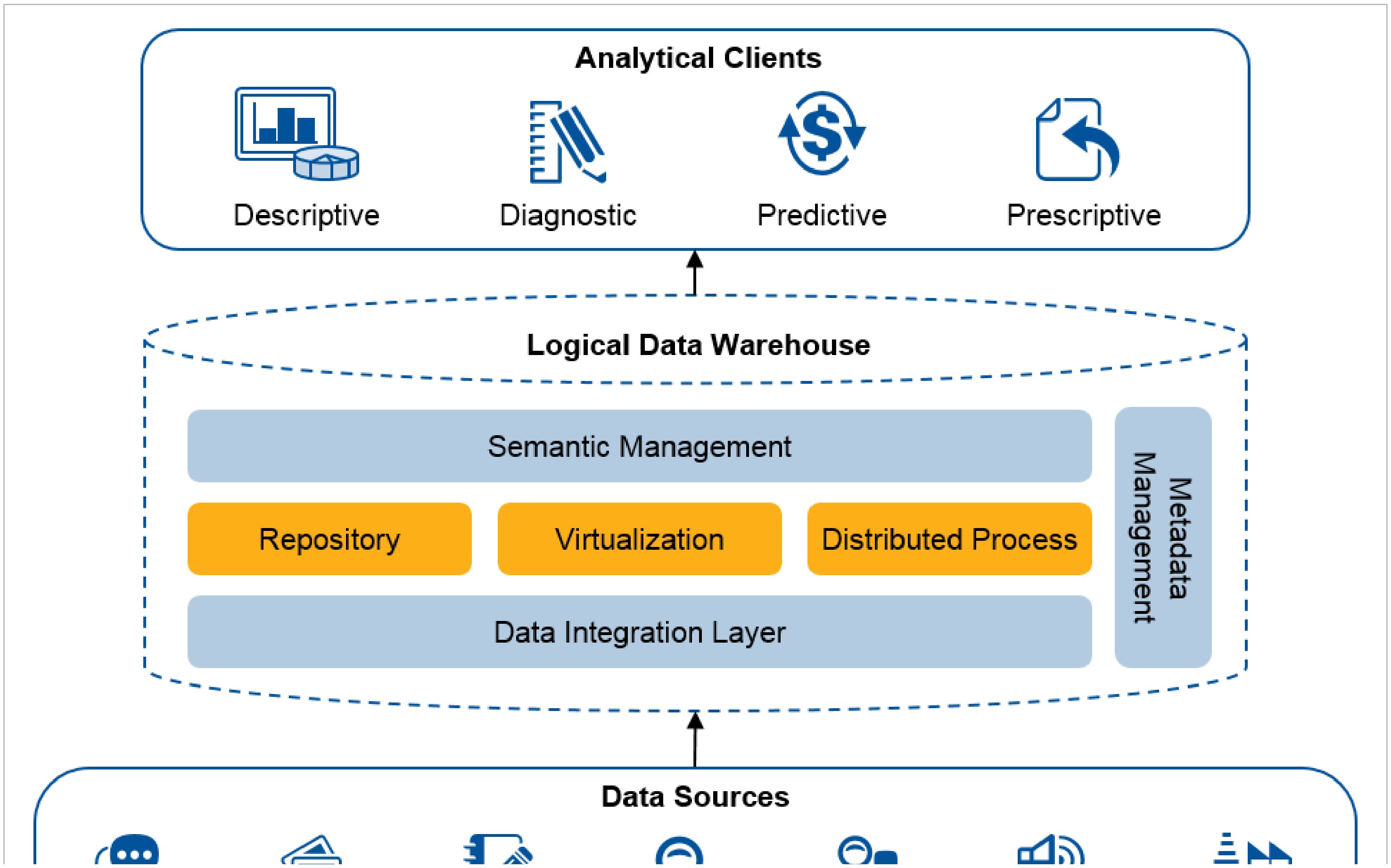
For a detailed analysis of data virtualization, see: " Solution Path for Planning and Implementing the Logical Data Warehouse." (https://www.gartner.com/document/code/320563?ref=grbody&refval=3738069&latest=true)

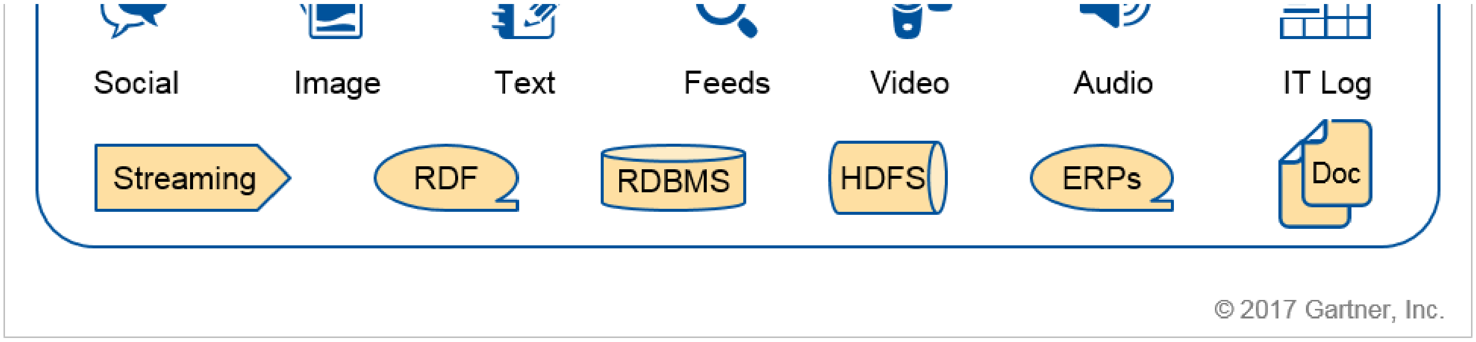
3.4 Logical Data Warehouse

The LDW is Gartner's recommended data management architecture for analytics, combining the strengths of traditional data warehouses with alternative data management and access strategies such as data lakes.

Figure 9 depicts the LDW as a conceptual layer that unifies a collection of architectural components into a connected logical view. This layer provides the logical definitions, processes and repositories that integrate the storage and persistence architecture underneath where source data resides.

Figure 9. The LDW Conceptual Architectural Diagram





Source: Gartner (June 2017)

Data architects can implement and evolve the LDW using three complementary architectural approaches:

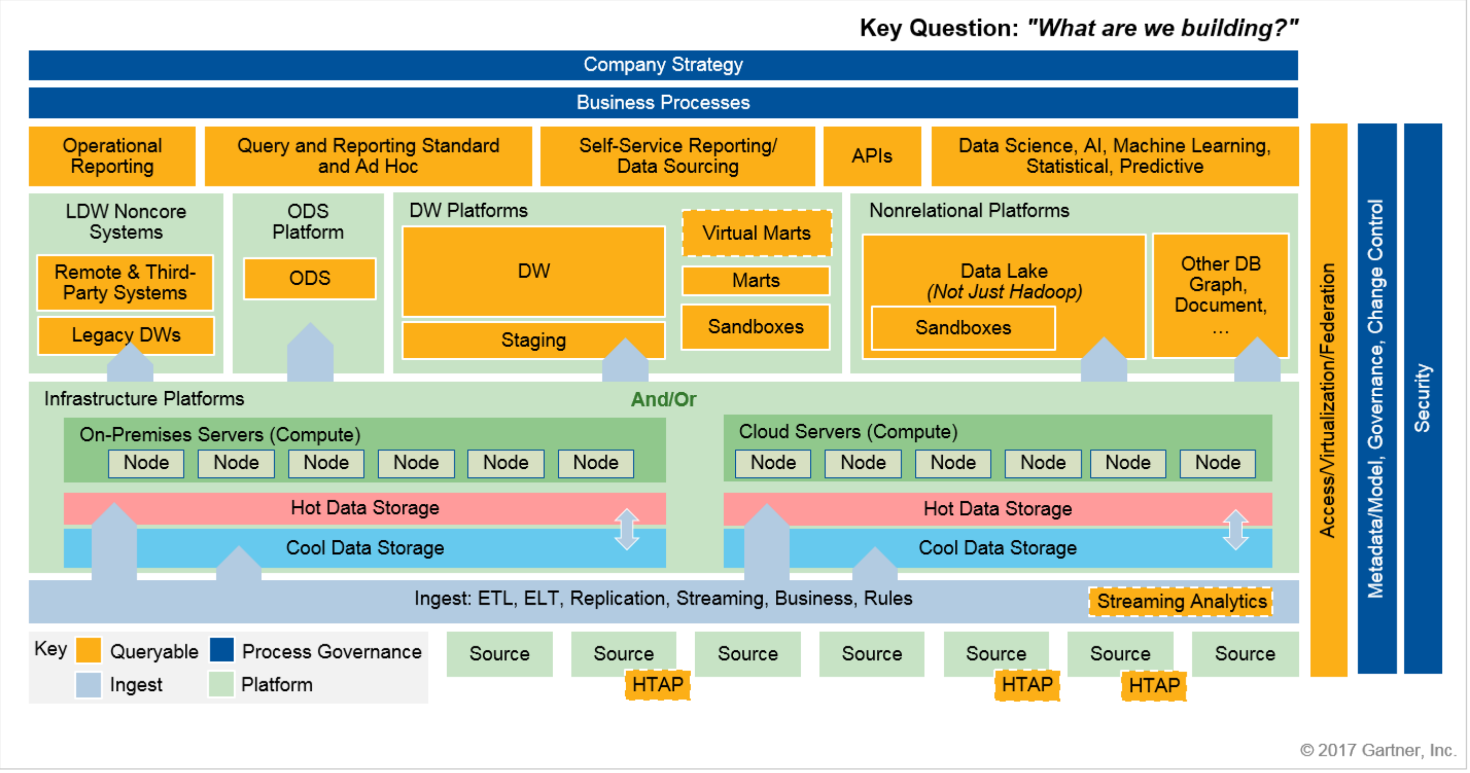
Classic data warehouse: Build the traditional data warehouse component to deliver high performance, predictable, prequalified and (possibly) summarized data.

Data lake: Enable working with very large scale and/or unstructured data with source data as near to native format and structure as possible.

Virtualization: Combine data virtualization and physical and virtual data marts to add new ways of using data.

The LDW is a concept based on the combination of architectural components and is not a commercial-off-the-shelf solution. The components comprising the LDW can be implemented by a range of software from different vendors and custom-built components. In this way, the LDW strategy offers a separation of concerns, with different parts of the architecture servicing different needs based on required capabilities (see Figure 10).

Figure 10. The LDW Architecture



Source: Gartner (June 2017)

For example, relational DBMSs can be used for the DW and operational data store (ODS), while Hadoop Distributed File System (HDFS) can be used for the data lake. Other database technologies, such as graph and document databases, can also be used to service special requirements.

Over time, the convergence of analytic and operational platforms will become more common. In-memory computing (IMC) provides performance that removes the need for competing operational and analytical strategies for accessing data. An emerging technology supporting such convergence is hybrid transaction/analytical processing (HTAP). By utilizing IMC, HTAP systems are capable of performing both online transaction processing (OLTP) and online analytical processing (OLAP) for real-time processing and analytics. Adoption of IMC technology is somewhat limited by the lack of commonly agreed upon standards, skills scarcity and high cost, as well as technology and market fragmentation. However, IMC is having a notable impact on software vendors' strategies. A growing number of vendors provide IMC-enabling application infrastructure products.

For assistance with LDW implementations and decisions, see:

- "Solution Path for Planning and Implementing the Logical Data Warehouse," (https://www.gartner.com/document/code/320563?ref=grbody&refval=3738069&latest=true) which provides a step-by-step guide to planning and implementing the LDW, along with details about LDW architectures and styles.

- "Embrace Sound Design Principles to Architect a Successful Logical Data Warehouse," (https://www.gartner.com/document/code/268536?ref=grbody&refval=3738069&latest=true) which reviews the LDW's strengths and weaknesses, along with a case study of a successful LDW.

What to Consider Before Moving to the Next Step

Gartner recommends that most organizations should consider the following before moving on to the next step.

- Its flexibility makes the LDW a critical capability for modern data management and analytics. Data architects should align their user SLA requirements with use cases and then determine how to architect their LDW.

- Evaluate managed cloud-based services, including iPaaS and database platform as a service (dbPaaS) for "greenfield" deployments or where IT skills or resources are lacking.

Step 4: Enable Business Insights

Today's competitive business environment demands effective analytics and business intelligence to enable fast, reliable insights and decisions. Achieving this goal depends on the tasks outlined in this step.

During this stage, data practitioners should be asking the following questions:

- What are the business objectives, and what analytics capabilities are needed throughout the architecture to deliver business value?

- What roles, skills and tools will enable technical professionals to do their own jobs while empowering business users to perform their own data preparation and analysis when appropriate?

The objective should be to deliver analytics capabilities to the decision maker at the point where the decision will have the greatest impact. To accomplish this, the organization needs an architecture nimble enough to respond quickly to changing demands driven by new data sources (internal and external), and new types of data with increasing volume and velocity. In many cases, timely access to insights means pushing data services outside traditional IT boundaries in the form of self-service functionality.

4.1 Business Intelligence and Visualization

Many organizations rely on IT-centric, enterprise-reporting-based platforms for large-scale systems of record reporting. While these platforms serve an essential function, they fail to meet business expectations due to a continued focus on IT-centric analytic strategies. Such traditional BI platforms typically emphasize reports and dashboards based exclusively on OLAP technologies, managed as an isolated solution. In these environments, end users submit requests for reports or data and wait for IT to deliver what they need. In some cases, users have ad hoc access to data through web interfaces or SQL for limited data mining use cases.

However, the BI and analytics market has moved from traditional, descriptive tools purchased and managed by IT to predictive, prescriptive, self-service tools and applications bought and used by lines of business. While enterprise reporting remains an important requirement for organizations, technical professionals need to anticipate the shift to self-service BI and analytics and plan accordingly.

Technical professionals should consider the following trends:

- Platform buying decisions have shifted more heavily to the business with a de-emphasis on IT's role.
- The need for analytic agility and business user autonomy outweighs the requirement for centrally provisioned, highly governed and scalable system-of-record reporting.
- Today's BI and analytics platforms must provide self-service capabilities, interactive visual exploration, analytic dashboards, and sharing and collaboration.

Basic BI, reporting and visualization capabilities are essential to any organization, and the implementation of these tools is a necessary stage in the data and analytics strategy. However, business modernization depends on empowering business users with the latest tools and the necessary data to explore new analytic opportunities. When evaluating BI platforms and tools, technical professionals should consider the BI platform's ability to support self-service exploration.

Refer to the following research for more information about BI:

- "Evaluation Criteria for Business Intelligence and Analytics Platforms" (https://www.gartner.com/document/code/297598?ref=grbody&refval=3738069&latest=true)
- "How to Build Data Visualization Capabilities as Part of a Modern Business Intelligence Platform" (https://www.gartner.com/document/code/297545?ref=grbody&refval=3738069&latest=true)

4.2 Self-Service Data Preparation and Analytics

Modern business demands faster and deeper insights from a wider range of data sources than ever before. While IT oversees the architecture and technologies supporting data management, analytics is happening throughout the enterprise.

As self-service and advanced analytics become more pervasive, technical professionals have an opportunity to shift and expand their role from being BI content creators and data access controllers to user and data enablers. This step is a critical stage of growth in the evolution of an organization's data and analytics strategy.

> Technical professionals can enable business analytics by conceptualizing IT's role as providing "data as a service" and enabling analytics throughout the enterprise.

This strategy does not require a lessening of data governance, security and privacy requirements. In fact, data classification and handling becomes even more critical. Leverage the architectural elements mentioned earlier to support data governance, but embrace self-service capabilities to enable business transformation.

At this stage, technical professionals should endeavor to:

- Empower business users to prototype and model new strategies and concepts.
- Establish a culture of innovation with a cross-functional business analytics foundation.
- Allow small, opportunistic projects with limited life spans that serve the needs of one person or a few people to bypass formal architecture disciplines, even for the underlying information model and software technology layers.

SPECIALIST AND CITIZEN USERS

The data scientist now plays an indispensable role in many organizations. Unfortunately, the proliferation of data science use cases creates anxiety among veteran data management professionals, who ask:

- Who are these users?
- Why do they need access to so much data?
- How can we safeguard the data once it's out of our control?

Compounding the problem, business users and self-styled analysts are demanding unprecedented access to data and analytical tools, often under the banner of data science. These "citizen" users may have legitimate use cases, but technical professionals are dubious.

SUPPORT SELF-SERVICE WORKFLOW

Self-service capabilities accelerate time to insight by giving business users a way to find, access, clean and prepare data for analytics. Technical professionals can support self-service workflow in the following ways:

Self-Service Data Preparation

Self-service data preparation technologies can be stand-alone or embedded in modern BI and advanced analytics platforms. They give users the tools to combine, prepare and manipulate data as well as collaborate with others by providing descriptive information about the data (metadata). Self-service data preparation tools can access and combine data from on-premises and cloud repositories and enable blending of enterprise data with data acquired from partners and third parties, such as data management platforms.

Provide Data Services

In lieu of or along with self-service data preparation, IT can help speed the creation of curated, trusted data for a range of distributed analytics content authors. This function is being addressed by the emerging role of data engineer: a centralized data specialist within IT.

Data engineers have the technical expertise and familiarity with IT processes, technologies and requirements to expedite curated datasets on behalf of business and specialist users. They typically possess software development experience and are capable of writing complex queries. They work closely with data scientists and analysts to understand requirements and design systems and processes that support their users.

4.3 Advanced Analytics

Once the traditional analytical tools that comprise basic BI and reporting are in place, organizations should turn their focus to advanced analytics capabilities that assist business decision making. Advanced analytics empower business stakeholders to conduct what-if analysis to envision the outcome of events based on today's decisions. For example, manufacturers can analyze buying patterns, forecast future trends and optimize inventory accordingly.

FROM EXPLANATORY TO EXPLORATORY

While traditional data reporting and analysis tend to look in the rearview mirror, advanced analytics peer into the future.

Predictive analytics: Answers the question of "What is likely to happen?" This category relies on techniques such as predictive modeling, regression analysis, forecasting, multivariate statistics and pattern matching.

Prescriptive analytics: Addresses the question of "What should be done?" or "What can we do to make "X" happen?" This category relies on techniques such as graph analysis, simulation, complex-event processing, recommendation engines, heuristics and, increasingly, neural networks and machine learning.

> Technical professionals should evaluate advanced analytics through two different lenses: IT-related and business-related.

IT can benefit from advanced analytics by looking at its own internal processes, systems and metrics to identify opportunities to improve operational effectiveness.

In addition to using advanced analytics to improve IT, technical professionals play a critical role in supporting business analytics. Technical professionals provide the architecture, tools and in some cases, prepare the data supporting business-related analytics. This requires planning for new data types, data sources and real-time use cases.

What to Consider Before Moving to the Next Step

Gartner recommends that most organizations should assess capabilities before moving on to the next step. Data architects should be asking themselves the following questions:

Do we have the foundational BI and reporting capabilities?

Can we support self-service capabilities for our BI platforms?

Can we support specialist and citizen users with the data they need to accomplish their goals?

Can we provide this data while applying appropriate safeguards within established data governance policies?

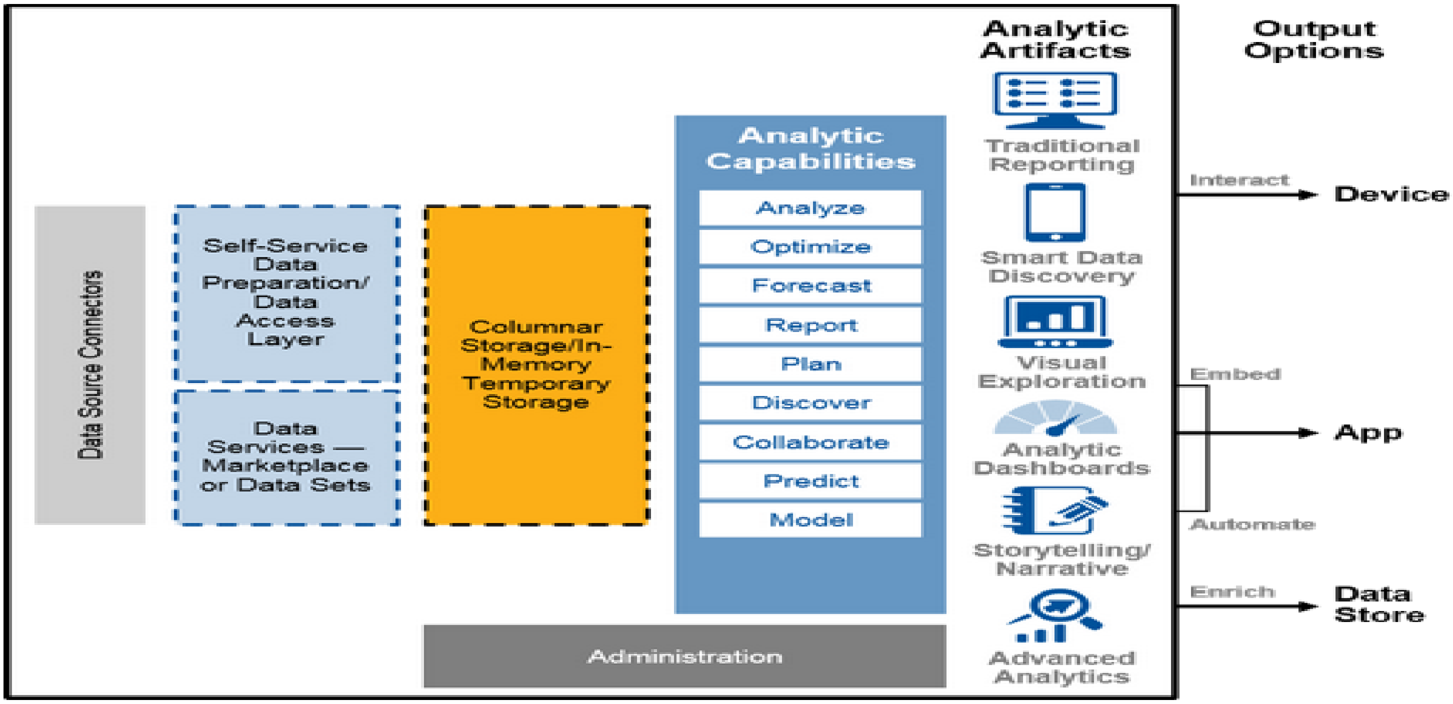
Step 5: Extend and Automate With Artificial Intelligence and Machine Learning

This step expands the strategy to enable the full spectrum of analytic capabilities by automating processes based on real-time data analysis and adopting machine learning to add intelligence to automation.

5.1 Real-Time Analytics

Supporting the full range of analytic capabilities requires an architecture that can support ingestion of streaming data, perform real-time analysis, take automated actions and deliver business insights. Figure 11 provides a reference architecture for an overall model of the key components and analytic capabilities.

Figure 11. Business Analytic Reference Architecture — From Data to Insight to Action



Source: Gartner (June 2017)

As the volume of data collected from sensors, devices, appliances and other endpoints grows, the potential business value that can be extracted from this data is growing exponentially. Technical professionals must plan for new and varied data sources such as the Internet of Things. The velocity associated with these new data sources requires careful planning and new architectural patterns. For more details, see "A Guide to Deploying IoT Analytics, From Edge to Enterprise." (https://www.gartner.com/document/code/317014?ref=grbody&refval=3738069&latest=true)

For additional details about building an analytics program, see the following research:

"Solution Path for Evolving Your Business Analytics Program" (https://www.gartner.com/document/code/292685?ref=grbody&refval=3738069&latest=true)

"Solution Path: Implementing Big Data for Analytics" (https://www.gartner.com/document/code/294453?ref=grbody&refval=3738069&latest=true)

"Hyperscaling Analytics: Comparing Streaming Analytics in the Cloud With AWS, Microsoft Azure and IBM" (https://www.gartner.com/document/code/310019?ref=grbody&refval=3738069&latest=true)

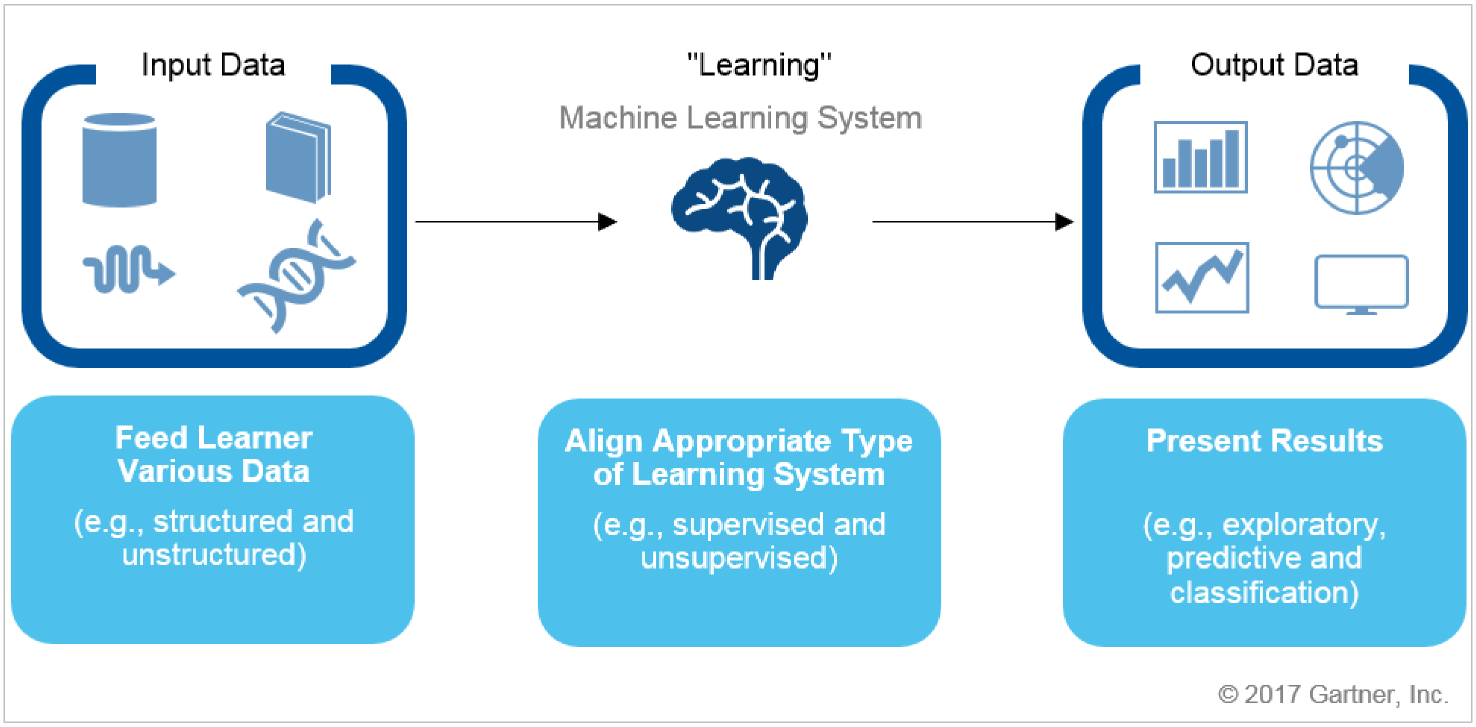
5.2 Machine Learning

Although data mining and manual analysis can be powerful, many organizations don't have the resources to apply data science in a broad fashion. Machine learning is an attractive option for evolving and automating analysis.

Machine learning (ML), a subset of artificial intelligence (AI), is more than a technique for analyzing data. It's a system that is fueled by data, with the ability to learn and improve by using algorithms that provide new insights without being explicitly programmed to do so. ML enables advanced systems that appear to understand, learn, predict, adapt and even operate autonomously, rather than being programmed only for a finite set of prescribed actions.

Using learning algorithms that simulate human learning, ML can leverage data and perform computations to learn and improve. Using input data, as shown in Figure 12, ML recognizes patterns that can be used to make a prediction or classify an object.

Figure 12. The Basics of Machine Learning Technology



Input Data
"Learning"
Machine Learning System
Output Data
Feed Learner Various Data
(e.g., structured and unstructured)
Align Appropriate Type of Learning System
(e.g., supervised and unsupervised)
Present Results
(e.g., exploratory, predictive and classification)
© 2017 Gartner, Inc.

Source: Gartner (June 2017)

Architect for Machine Learning

ML demands a flexible architecture capable of supporting elastic learning patterns, and consuming large and varying volumes of data. These design requirements often necessitate considerable processing power and storage versatility. Cloud infrastructure is an excellent proving ground for new ML initiatives with its elastic capabilities and scaling algorithms. Figure 13 shows Gartner's suggested reference architecture for ML covering the functional areas required for the ML process:

Data acquisition: Where data is collected, prepared and forwarded for processing.

Data processing: Where steps such as preprocessing, sample selection and the training of datasets take place, in preparation for execution of the ML routines:

Feature analysis or feature engineering (a subset of the data processing component), where features that describe the structures inherent in your data are analyzed and selected

Data modeling or model engineering: Includes the data model designs and machine algorithms used in ML data processing (including clustering and training algorithms):

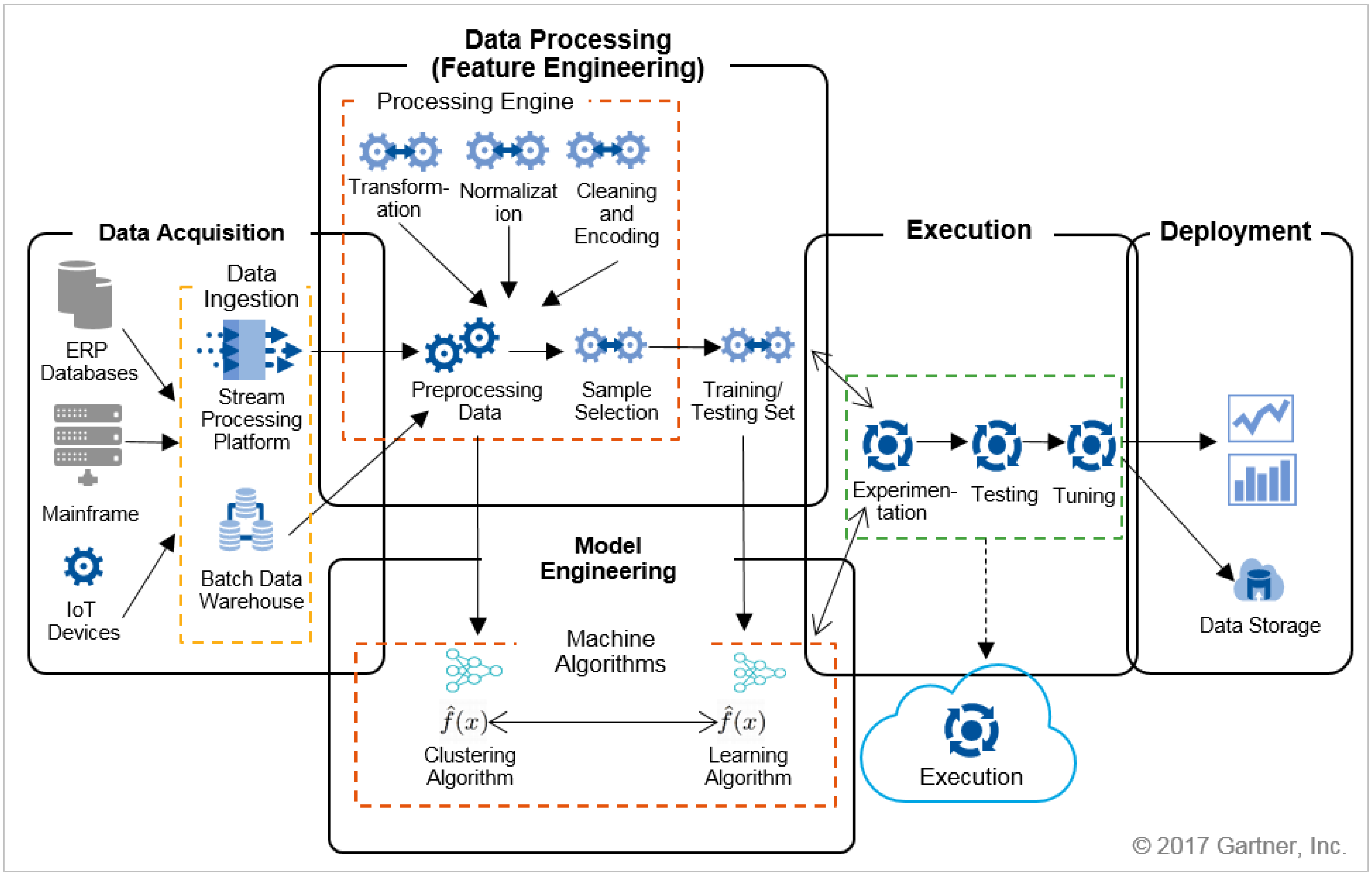
Model fitting, where a set of training data is assigned to a model in order to make reliable predictions on new or untrained data

Model evaluation, where models are evaluated based on performance and efficacy

Execution, the environment where the processed and trained data is forwarded for use in the execution of ML routines (such as experimentation, testing and tuning).

Deployment, where business-usable results of the ML process — such as models or insights — are deployed to enterprise applications, systems or data stores (for example, for reporting).

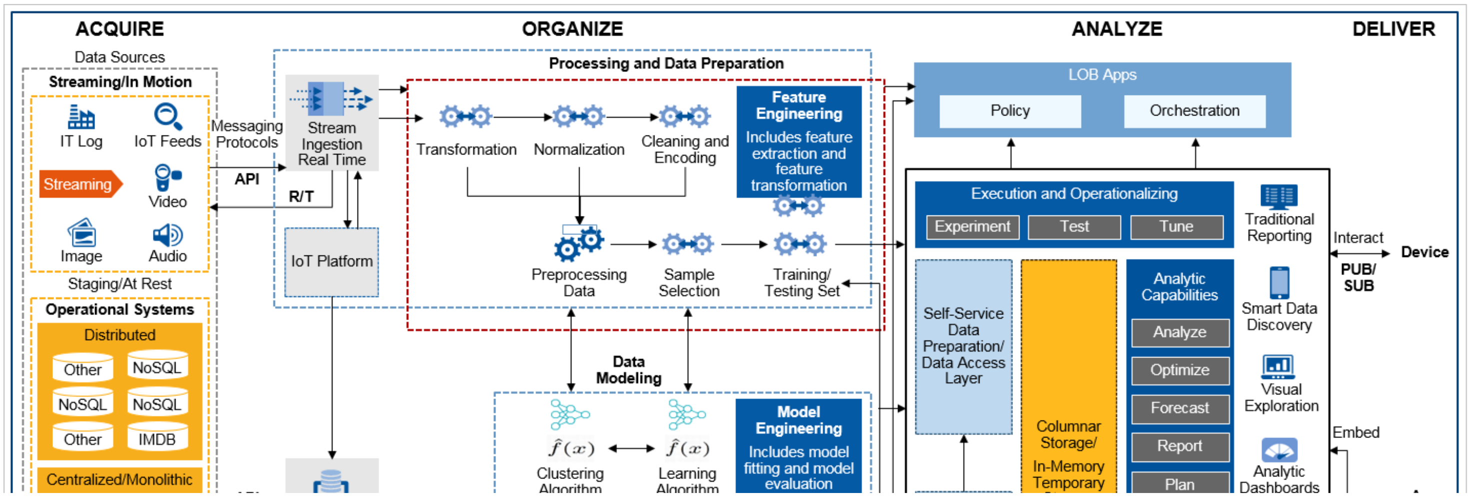
Figure 13. Machine Learning Architecture

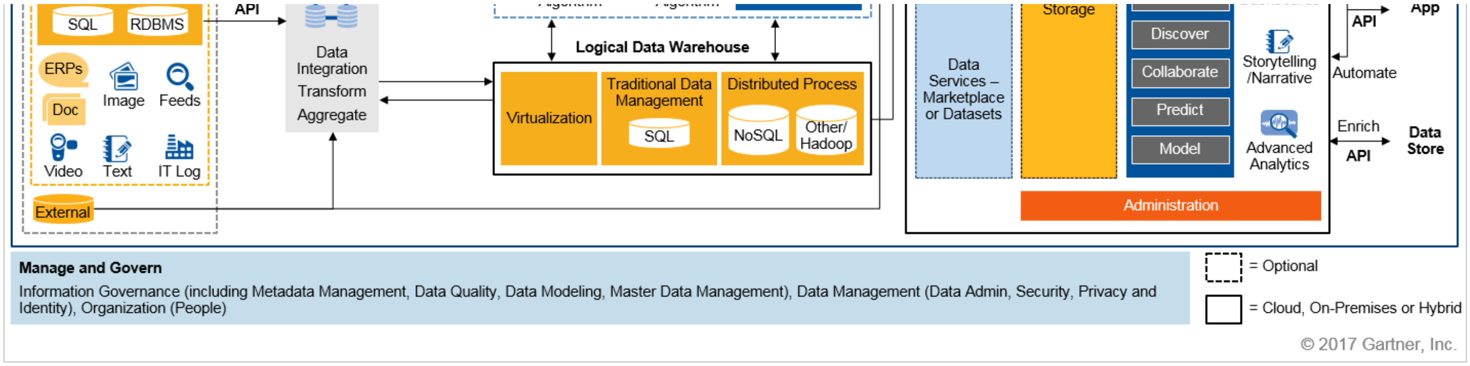


Source: Gartner (June 2017)

The comprehensive, end-to-end architecture referred to earlier is updated in Figure 14 to incorporate the diverse data, models and algorithms supporting the stages of ML.

Figure 14. End-to-End ML and Analytics Architecture





Source: Gartner (June 2017)

Steps to get started with machine learning:

- Learn about and experiment with ML concepts and technology
- Work closely with data science teams and business users to identify a use case
- Build a use case in the cloud
- Iteratively expand your ML platform and services over time

For more details, see the following Gartner research:

- "Preparing and Architecting for Machine Learning" (https://www.gartner.com/document/code/317328?ref=grbody&refval=3738069&latest=true)
- "How to Create a Data Strategy for Machine Learning-Powered Artificial Intelligence" (https://www.gartner.com/document/code/324342?ref=grbody&refval=3738069&latest=true)

Gartner recommends that technical professionals adopt ML techniques as part of their personal "tradecraft," which will improve their ability to support digital business efforts — as well as tackle data management and operations challenges that arise within IT — see "Top Skills for IT's Future: Cloud, Analytics, Mobility and Security." (https://www.gartner.com/document/code/297698?ref=grbody&refval=3738069&latest=true)

5.3 Operationalize and Scale

This step focuses on supporting the architectural capabilities outlined in the previous steps with process automation and scalability.

DATA SCIENCE AT SCALE

Analysts and data scientists leverage specialist, stand-alone tooling when data integration requirements exceed the capabilities embedded in analytics platforms. Many data scientists also supplement embedded capabilities with multiple open-source tools, such as Python, R and Scala, or tools that come with Hadoop distributions for different parts of the data pipeline process. Using these stand-alone tools, they extract small amounts of data from the data lake and perform analysis on their local systems.

There are several problems with this stand-alone strategy:

- Users may base models on incomplete data or datasets that are too small
- Storing data on personal laptops leads to data sprawl
- Data scientists spend time acquiring and preparing data rather than on high-value tasks
- Processing occurs on stand-alone systems instead of distributed, scale-out platforms
- Models and algorithms may require refactoring for production rollout

By implementing self-service data preparation tools combined with support for data science utilities and platforms, technical professionals can accelerate and streamline the entire process resulting in more efficient and productive data science teams.

This approach also requires a process and guidelines for how business analysts and data scientists can "promote" or operationalize findings and models into trusted, recurring analysis, either themselves or via formal data integration practices.

DECISION SUPPORT

Many organizations claim that their business decisions are data-driven. But they often use the term "data driven" to mean reporting key performance metrics based on historical data and using analysis of these metrics to support and justify business decisions that will, hopefully, lead to desired business outcomes. While this a good start, it is no longer enough. ML and predictive analytics can take decision support to new levels.

Using ML to Support Decisions

Forecasting is nothing new, and organizations have long relied upon a combination of intuition, expertise and past results to anticipate future state. For decades, organizations have applied data analysis based on statistical models to distinguish previously unknown relationships and patterns within data. However, the confluence of ML techniques and advances in computing and processing capabilities have increased the speed and efficiency of predictive analytics — especially in cases where designing and programming explicit algorithms is either impractical or unfeasible.

ML is now being used in a variety of business and IT contexts to support decision making, as shown in Tables 1 and 2.

Table 1. ML Decision Support Examples — Business Context

Business Category	Decision Context
Supply Chain	Analyze buyer behavior and purchasing trends to identify changes in the marketplace and forecast demand.
Sales	Analyze CRM data to identify actions and events conducive to desirable sales outcomes.
Manufacturing	Utilize sensor data to anticipate failure based on events, temperatures, usage, etc.
Customer Service	Analyze churn probability to prioritize service and sales efforts to maximize retention.
Marketing	Identifying negative brand sentiment though mentions in social media

Source: Gartner (June 2017)

Table 2. ML Decision Support Examples — IT Context

IT Category	Decision Context
IT Operations Management	Anticipating equipment failure and implementing proactive maintenance to prevent disruption.
Information Security	Identifying and resolving IT and security incidents by automatically detecting anomalies and patterns in data

Source: Gartner (June 2017)

AUTOMATED ACTIONS

In this step, data and analytics become the brain of the enterprise — becoming proactive as well as reactive, and coordinating a host of decisions, interactions and processes in support of business and IT outcomes.

Infuse predictive intelligence directly into your systems, to combine ML and advanced analytics with workflows and automated processes.

Shape and mold external and internal customer experiences, based on predicted preferences for how each individual and group wants to interact with the organization.

Drive business processes, not only by recommending the next best action but also by triggering those actions automatically.

As organizations become more adept at using data to drive decisions, the next logical stage is using data to take action. The ability to automate decisions and actions based on information is a powerful differentiator in modern business. Technical professionals can expand on the examples listed earlier to take automated actions (see Tables 3 and 4):

Table 3. ML Automation Support Examples — Business Context

Business Category	Automated Actions
Supply Chain	Trigger automatic inventory replenishment, implement discounts and promotions to reduce inventory.
Sales	Classify, prioritize and channel leads based on scoring models. Optimize models as data evolves and route leads automatically.
Recommendations	Make real-time product recommendations, ad placement, content filtering, and ranking.
Fraud Detection	Block transactions or suspend accounts when fraudulent activity is detected.
Failure Prediction	Ship replacement parts and schedule service.

Source: Gartner (June 2017)

Table 4. ML Automation Support Examples — IT Context

IT Category	Description
Security	Security actions can be based on data thresholds, patterns or trends. For example, security anomalies can trigger account lockouts or temporarily suspend transactions.
IT Operations	Engineers can have new nodes added automatically to a compute cluster when metrics exceed preset thresholds.
Service and Support	Algorithmic chatbots using machine learning can become more effective as they learn from each interaction and draw on contextual information about their users.

Source: Gartner (June 2017)

See "2017 Planning Guide for Data and Analytics." (https://www.gartner.com/document/code/311517?ref=grbody&refval=3738069&latest=true)

Iterate for Continuous Improvement

Technical professionals should revisit the following steps when adjusting the architecture, adding new data stores, or implementing new technologies. Reviewing these steps ensures alignment with business goals, promotes comprehensive data governance and creates technical consistency in the selection of tools and services. Organizations should use the steps in this research to iteratively improve their data management and analytics architecture as they expand into new use cases.

Align With Business Strategy

At the beginning of each project, senior managers and business stakeholders should explain the vision and strategy of the company. This can occur through initial meetings, interviews and company documentation, such as the annual report. This step ensures that the team understands how information needs relate to high-priority items in company strategy. It also allows that prioritization to balance the longer-term strategic goals that may have very high value, and which are essential to company growth and survival, alongside other immediate and urgent requirements.

Review business drivers and business outcomes. Start with the problem you are trying to solve. Examples:

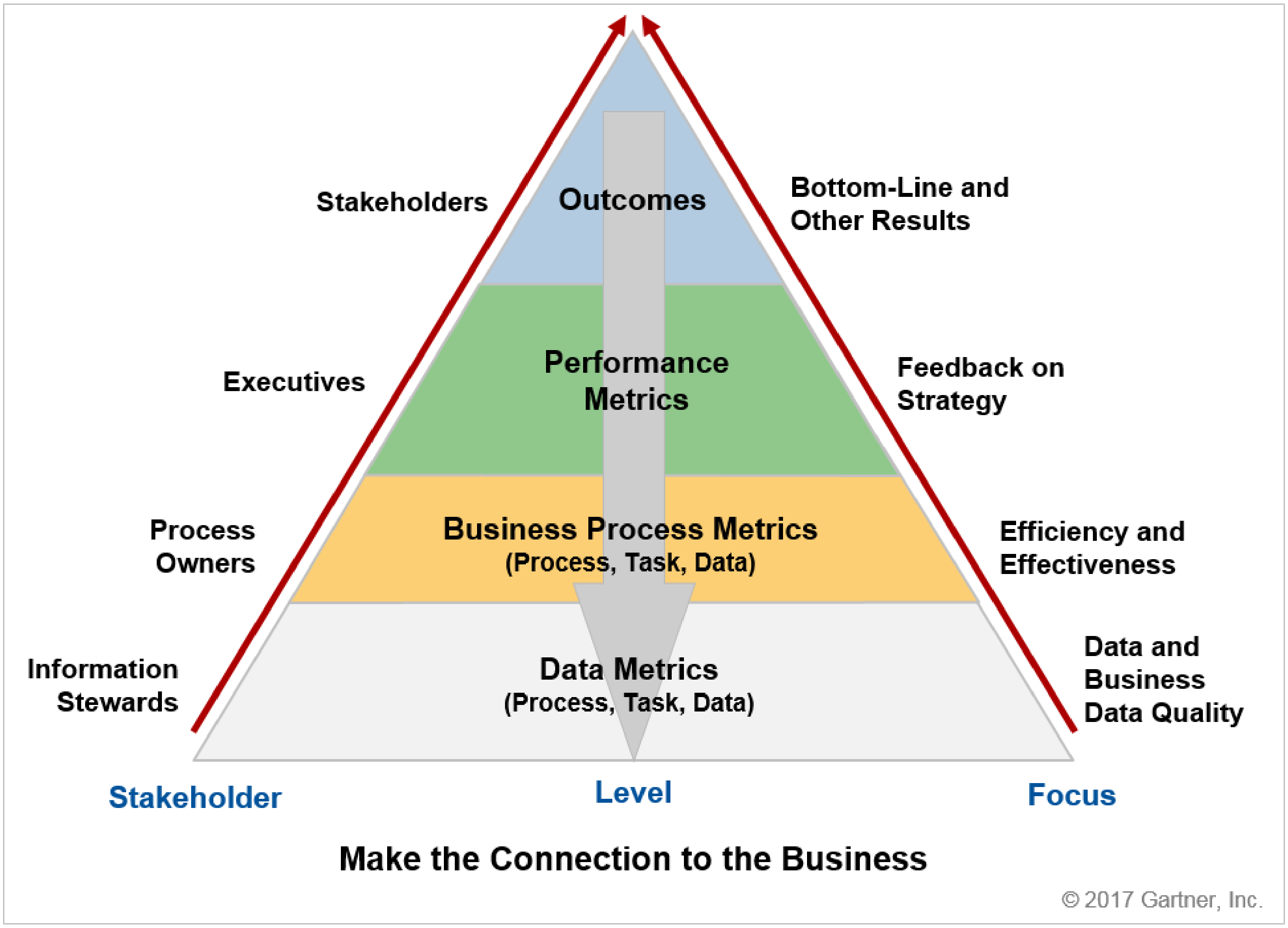
- How can we identify and reduce/prevent fraud?
- How can we optimize purchasing across divisions?

If possible, reframe business challenges within the context of analytical questions to be answered.

Identify the datasets needed to solve those problems: What are the use cases, and what changes are needed in the architecture?

Figure 15 from "Key Recommendations for Implementing Enterprise Metadata Management Across the Organization" (https://www.gartner.com/document/code/315412?ref=grbody&refval=3738069) shows the relationship between business and technical metrics.

Figure 15. Aligning Business Strategy With Development


Stakeholders
Outcomes
Bottom-Line and Other Results
Executives
Performance Metrics
Feedback on Strategy
Process Owners
Business Process Metrics (Process, Task, Data)
Efficiency and Effectiveness
Information Stewards
Data Metrics (Process, Task, Data)
Data and Business Data Quality
Stakeholder
Level
Focus
Make the Connection to the Business
© 2017 Gartner, Inc.

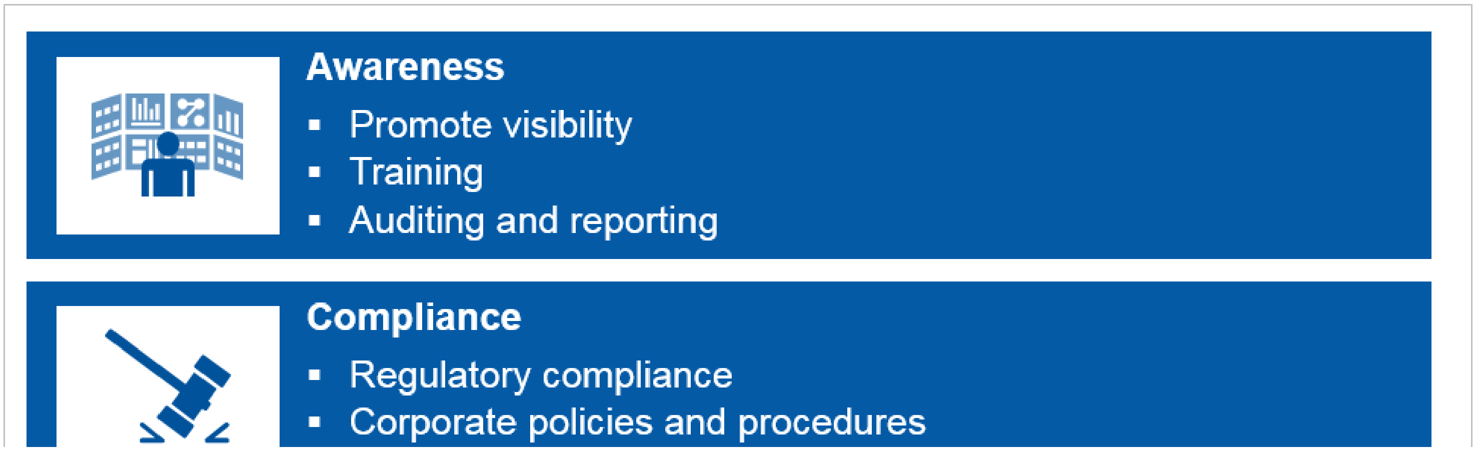
Source: Gartner (June 2017)

Governance

Revisit the components of information governance outlined earlier. Take steps to ensure consistent application with each new project or additional technology.

It can be useful to adopt information governance stages to ensure that new initiatives are properly socialized within the organization, that there is agreement on policies and procedures, and that various aspects of information management are considered during the project (see Figure 16).

Figure 16. Information Governance — Stages

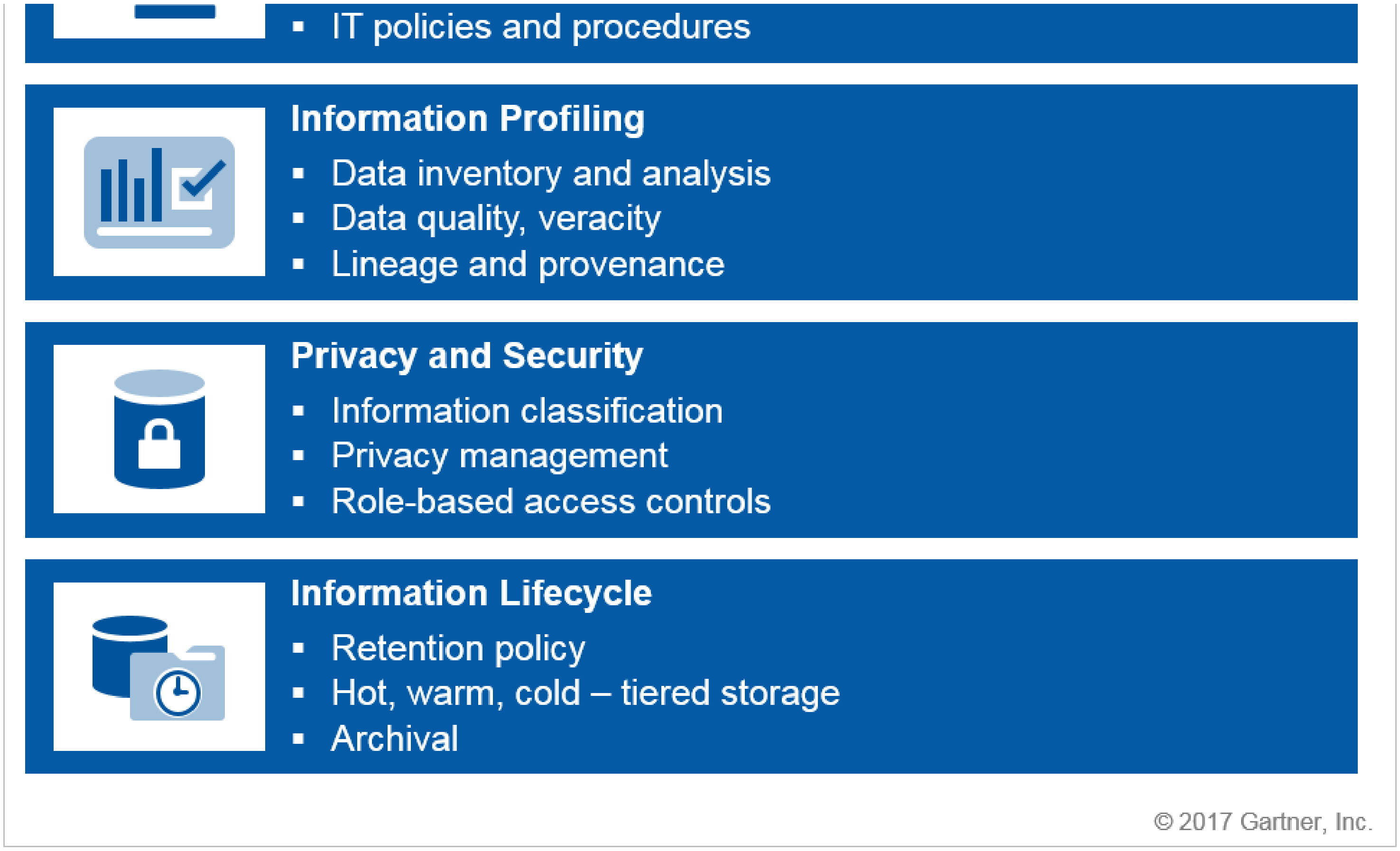


Awareness

- Promote visibility
- Training
- Auditing and reporting

Compliance

- Regulatory compliance
- Corporate policies and procedures



- IT policies and procedures

Information Profiling

- Data inventory and analysis
- Data quality, veracity
- Lineage and provenance

Privacy and Security

- Information classification
- Privacy management
- Role-based access controls

Information Lifecycle

- Retention policy
- Hot, warm, cold – tiered storage
- Archival

Source: Gartner (June 2017)

People and Skills

Identify roles and responsibilities along with skills needed for each new initiative. The following roles can be helpful in managing a modern data and analytics architecture.

Business user: Someone who understands the domain area and usually benefits from the results. This person can consult and advise the project team on the context of the project, the value of the results and how the outputs will be operationalized. Usually, a business analyst, line manager or deep subject matter expert in the project domain fulfills this role.

Project sponsor: Responsible for the genesis of the project. Provides the impetus and requirements for the project and defines the core business problem. Generally provides the funding and gauges the degree of value from the final outputs of the working team. This person sets the priorities for the project and clarifies the desired outputs.

Project manager: Ensures that key milestones and objectives are met on time and at the expected quality.

Business intelligence analyst: Provides business domain expertise based on a deep understanding of the data, key performance indicators (KPIs), key metrics and business intelligence from a reporting perspective. Business intelligence analysts generally create dashboards and reports, and have knowledge of the data feeds and sources.

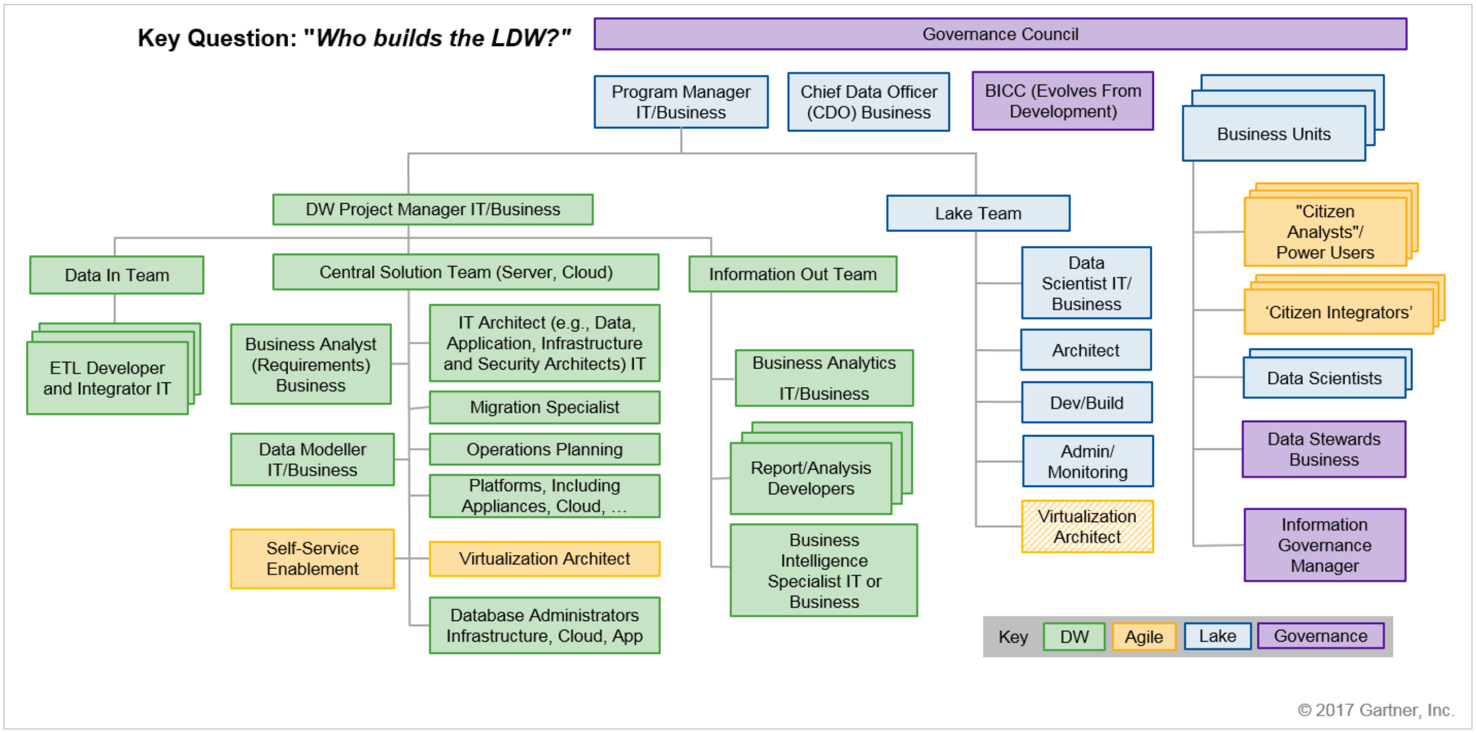
Database administrator (DBA): Provisions and configures the database environment to support the analytics needs of the working team. These responsibilities may include providing access to key databases or tables and ensuring the appropriate security levels are in place related to the data repositories.

Data engineer: Leverages deep technical skills to assist with tuning SQL queries for data management and data extraction, and provides support for data ingestion into the analytic sandbox. Whereas the DBA sets up and configures the databases to be used, the data engineer executes the actual data extractions and performs substantial data manipulation to facilitate the analytics. The data engineer works closely with the data scientist to help shape data in the right ways for analyses.

Data scientist: Provides subject matter expertise for analytical techniques, data modeling and applying valid analytical techniques to given business problems. Ensures overall analytics objectives are met. Designs and executes analytical methods and approaches with the data available to the project.

Figure 17 shows a sample team structure for building an LDW. Documenting the roles and responsibilities in this way can be a useful reference when designing a data and analytics strategy.

Figure 17. Example Team Structure for Building an LDW


Key Question: "*Who builds the LDW?*"
Governance Council
Program Manager IT/Business
Chief Data Officer (CDO) Business
BICC (Evolves From Development)
Business Units
DW Project Manager IT/Business
Lake Team
Data In Team
Central Solution Team (Server, Cloud)
Information Out Team
ETL Developer and Integrator IT
Business Analyst (Requirements) Business
IT Architect (e.g., Data, Application, Infrastructure and Security Architects) IT
Migration Specialist
Operations Planning
Platforms, Including Appliances, Cloud, …
Data Modeller IT/Business
Self-Service Enablement
Virtualization Architect
Database Administrators Infrastructure, Cloud, App
Business Analytics IT/Business
Report/Analysis Developers
Business Intelligence Specialist IT or Business
Data Scientist IT/ Business
Architect
Dev/Build
Admin/ Monitoring
Virtualization Architect
"Citizen Analysts"/ Power Users
'Citizen Integrators'
Data Scientists
Data Stewards Business
Information Governance Manager
Key
DW
Agile
Lake
Governance
© 2017 Gartner, Inc.

Source: Gartner (June 2017)

Account for Citizen Roles

Account for citizen roles performed by business users. These new, hybrid roles often span functions and departments, blending business and IT functions and blurring traditional organizational boundaries. Organizations must plan for the emergence of these hybrid roles and determine how the information management and analytics strategy will support these new functions.

With each new initiative, it's important to consider cross-functional boundaries and identify where IT's responsibilities begin and end. Citizen roles can't be ignored, so IT should provide tools, processes and support to enable a strong partnership with business stakeholders.

Agile Database Development

Agile and DevOps require application and data developers to adopt iterative and incremental design and implementation processes. To achieve continuous delivery, technical professionals focused on agile development must apply these same processes to their application database changes.

To effectively support agile development, the methods used to design, implement and evolve databases must follow the same incremental, evolutionary and automated approaches used to develop the application code. Agile developers, under the guidance and support of the DBA, can remove or overcome constraints on velocity by developing the knowledge and skills to:

Create, maintain and incrementally evolve the data model as the project advances.

Write, review, modify, optimize and validate changes to database interface code, SQL scripts, database schemas and functional objects, using agile technical practices.

Manage, maintain and version the test data used for validation.

Implement automation as part of the continuous integration and DevOps pipelines to minimize the amount of manual effort involved in supporting the tasks listed above.

Gartner has published guidance to help organizations through this process (see "Implement Agile Database Development to Support Your Continuous Delivery Initiative" (https://www.gartner.com/document/code/310316?ref=grbody&refval=3738069&latest=true)).

Cloud Versus On-Premises

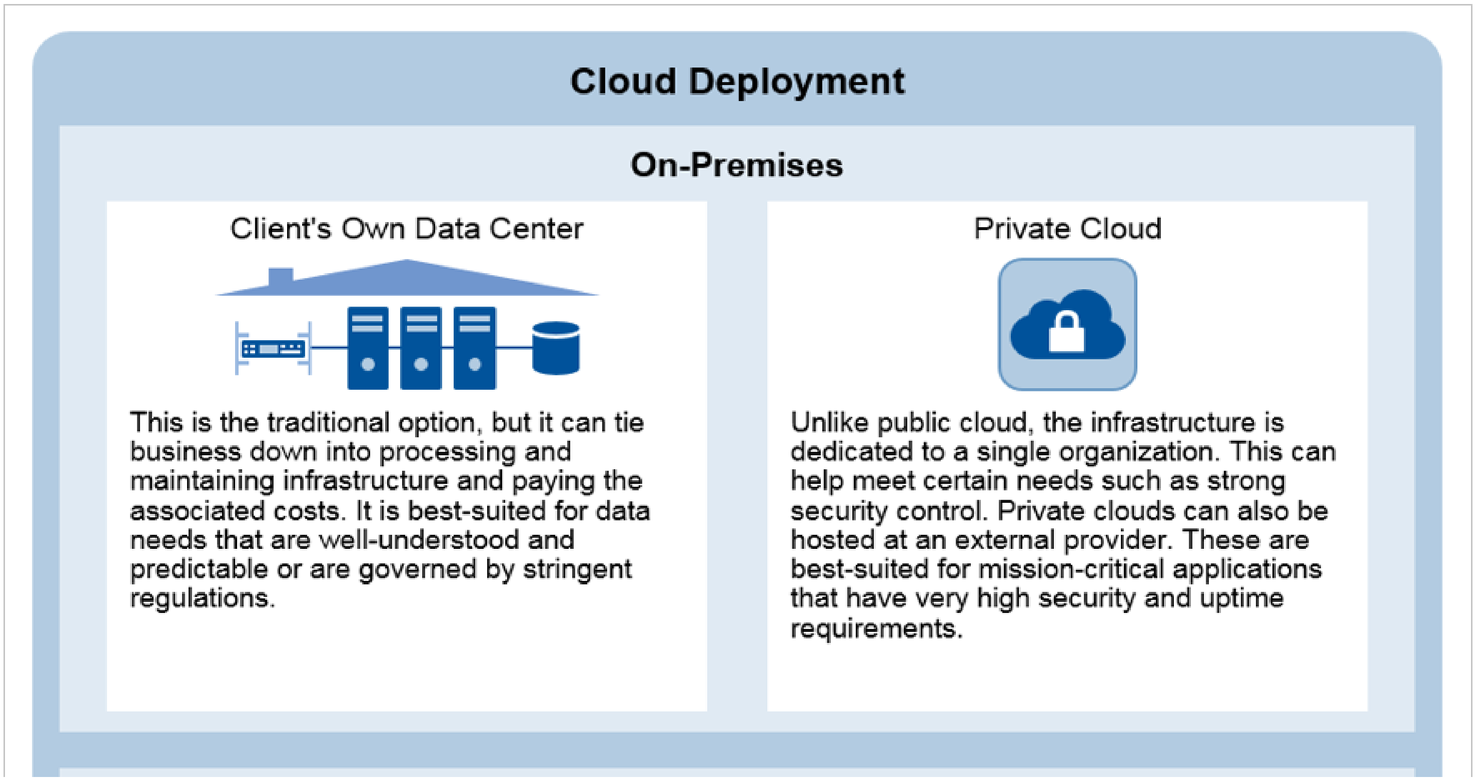
With each new initiative, technical professionals should consider important technology placement and deployment considerations, such as:

- On-premises versus cloud
- Cloud service providers
- Multicloud and hybrid strategies
- PaaS and IaaS

Key factors in the decision might include:

- **Reducing or controlling costs:** To optimize business and reduce costs (such as costs associated with capacity upgrades or infrastructure operations), organizations opt for external IT sourcing alternatives, such as colocation, managed hosting and cloud.
- **Improving speed of delivery:** Organizations leverage technologies, such as virtualization and orchestration, to improve the speed of delivery and to adapt more easily to the changing business demands.
- **Focusing on core competencies:** Organizations want to refocus on core competencies or capabilities that provide competitive differentiation to the business, and leave more commoditized services to external providers.
- **Overcoming skills shortages:** Building and operating a data center requires a complete range of skills — including people, processes and technology — that many organizations find difficult to maintain. External hosting models require a subset of the skills required by the data center; for example, facilities management skills are no longer required. In the case of public cloud, required skills include vendor management and business strategy.
- **Improving service levels:** Many corporate-owned data centers are overcrowded and lack availability features, such as backup generators or multiple internet connections. Moving applications from these data centers to more robust facilities can improve service levels and availability (see Figure 18).

Figure 18. Cloud Deployment Options for Data



Cloud Deployment

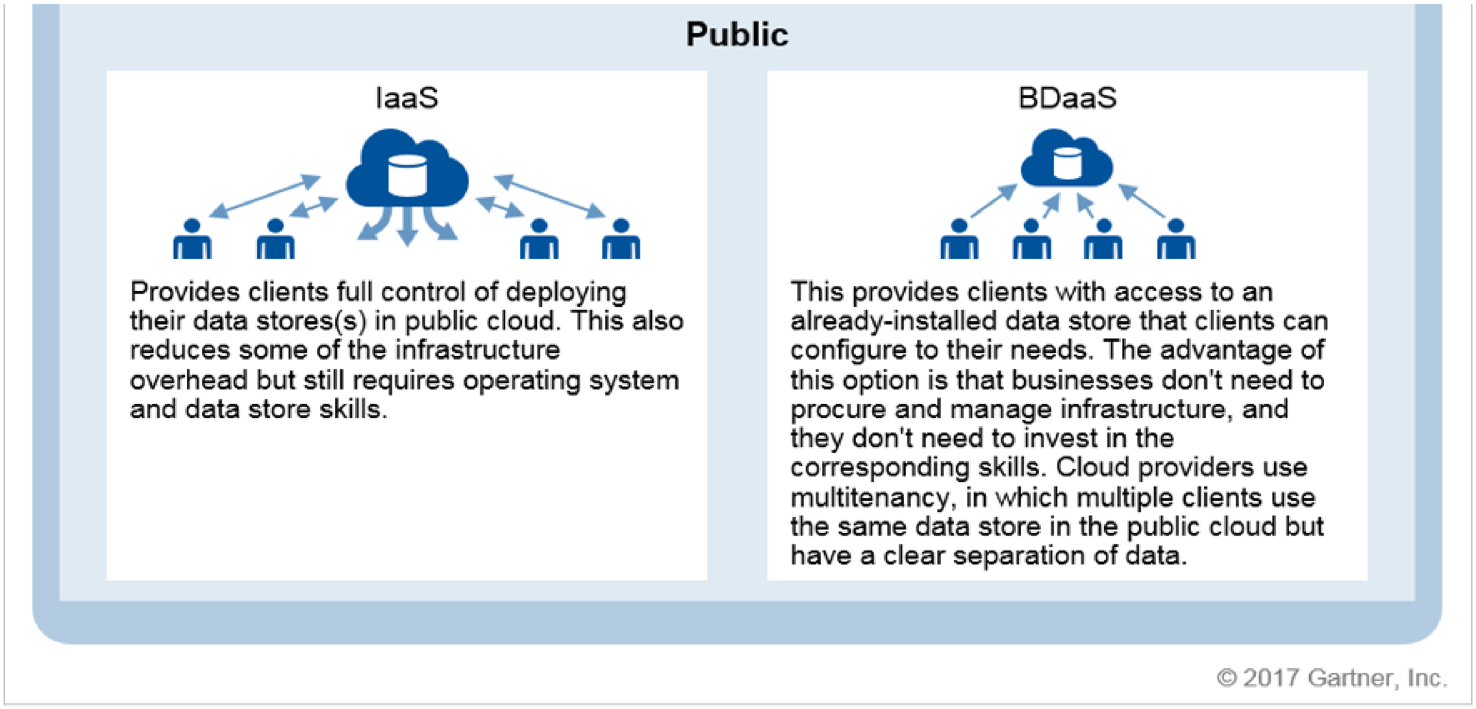
On-Premises

Client's Own Data Center

This is the traditional option, but it can tie business down into processing and maintaining infrastructure and paying the associated costs. It is best-suited for data needs that are well-understood and predictable or are governed by stringent regulations.

Private Cloud

Unlike public cloud, the infrastructure is dedicated to a single organization. This can help meet certain needs such as strong security control. Private clouds can also be hosted at an external provider. These are best-suited for mission-critical applications that have very high security and uptime requirements.



Provides clients full control of deploying their data stores(s) in public cloud. This also reduces some of the infrastructure overhead but still requires operating system and data store skills.

This provides clients with access to an already-installed data store that clients can configure to their needs. The advantage of this option is that businesses don't need to procure and manage infrastructure, and they don't need to invest in the corresponding skills. Cloud providers use multitenancy, in which multiple clients use the same data store in the public cloud but have a clear separation of data.

Source: Gartner (June 2017)

When it comes to deployment, clients have two basic choices: cloud or on-premises. Technical professionals may also deploy a hybrid model consisting of the data store in their own data center or in a private cloud that bursts into public cloud when it needs to meet loads that are beyond its design.

As technical professionals migrate on-premises systems to the public cloud, they need to know which databases to migrate, which tools and techniques to use and how to run a successful migration project.

Gartner has published guidance to help organizations through this process (see "Migrating Enterprise Databases and Data to the Cloud" (https://www.gartner.com/document/code/317167?ref=grbody&refval=3738069&latest=true)).

Gartner Recommended Reading

The analyst(s) have suggested further reading that is not available under your Gartner subscription.

"Data Consistency Flaws Can Destroy the Value of Your Data" (https://www.gartner.com/document/code/304174?ref=ggrec&refval=3738069)

"Applying Gartner's Pace Layer Model to Business Analytics" (https://www.gartner.com/document/code/226265?ref=ggrec&refval=3738069)

"Organizing for Effective Data Management in Digital Business" (https://www.gartner.com/document/code/315253?ref=ggrec&refval=3738069)