A Guidance Framework for Deploying Data and Analytics in the Cloud

Refreshed 13 October 2022, Published 18 March 2021 - ID G00741083 - 62 min read

FOUNDATIONAL This research is reviewed periodically for accuracy.

By Analyst(s): Ramke Ramakrishnan

Initiatives: Analytics and Artificial Intelligence for Technical Professionals; Data Management Solutions for Technical Professionals; Evolve Technology and Process Capabilities to Support D&A

Organizations need to select the right deployment architecture, technologies and best practices for deploying a best-of-breed analytic platform. Data and analytics technical professionals can use this guidance framework to develop a strategy and approach for cloud analytics deployments.

Overview

Key Findings

- Organizations that are successful with data and analytics cloud deployments invest considerable time and effort to understand deployment approaches, entanglement and complexity for the end-to-end architecture. It requires evaluating each component, benefits of cost optimizations, compliance obligations and price-for-performance constraints before moving to the cloud.
- Establishing a preliminary roadmap decision path and architecture, and defining the migration, is critical to identify the right deployment model for your data and analytics platform.
- Cloud data and analytics deployments are complex and come with many potential risks and pitfalls. Technical professionals struggle to balance the need to deliver solutions quickly to provide customized, optimized and complete services for end-to-end migration.
- Successful data and analytics cloud migrations tend to take an iterative approach, starting small to deliver a core set of cloud-based services, and expanding over time toward a broader set of use cases. This approach allows technical professionals to manage change, build competency in managing cloud services and avoid security risks.

Recommendations

To create a data and analytics strategy, technical professionals responsible for Analytics and Artificial Intelligence should:

- **Identify** your data and analytics platform requirements and cloud deployment approach early on, as this will have ramifications to achieve desired business and technical outcomes for your deployment.
- **Make** the best use of the proposed framework to create a robust data and analytics platform based on the implementation approach by including a roadmap to address both the current and future use case scenarios.
- **Streamline** your deployments across the organization that will endure transparency and consistency to drive effective decisions and outcomes across the entire data and analytics platform with accessibility to data in a governed fashion.

- **Focus** on the process of deploying data and analytics, avoid retrofitting old practices, and take a radically new approach and continuous change management to integrate emerging technologies to help serve new business demands in the cloud.

Problem Statement

Organizations face increased demand for business growth and are looking for new ways to deploy analytics strategies and practices. Data and analytics professionals are to overcome several challenges to foster data-driven decisions to enable new capabilities and monetization on data. They need to modernize their data and analytics platform by developing new deployment models, identifying new technological capabilities, and creating and executing new strategies for migrating their analytical workloads to the cloud.

Traditional D&A platforms are constrained to meet the growing business demands. In addition to keeping up the request in tackling these new sets of data and analytics demands, the total cost of ownership of on-premises solutions continues to grow due to complexity, increased resources and maintenance of the environment. In contrast, cloud data and analytics offer more value and capabilities through new services, simplicity and agility to handle data modernization — and new types of analytics demands, such as streaming analytics, specialized data stores and more self-service-friendly tools to support end-to-end deployment.

While many organizations prefer to align with a single cloud service provider strategically, other challenges make it practically impossible due to several factors, such as:

- Existing investments in best-of-breed tools — Many enterprise D&A solutions have chosen different cloud providers for their SaaS offering, forcing additional multicloud diversity simply by having chosen D&A platform for particular requirements in the past.

- Business units are making independent decisions — D&A teams are often not in control of cloud purchasing decisions. The broader enterprise may decide that multicloud analytics makes better sense to deliver a broader set of capabilities to different stakeholders.

In addition to the complexity of deciding on the service provider in the cloud, the migration or deployment of your data and analytics in the cloud can be challenging, and choosing the best approach is an important step. Cloud deployment approaches also must account for the many components, such as data ingestion, data integration, data modeling, data optimization, data security, data quality, data governance, management reporting, data science or machine learning, etc. Also, for a step by step-by-step approach to create a governed, self-service, augmented data-driven analytic architecture, please see Solution Path for Modernizing Analytic Architectures.

The D&A cloud strategy also needs to understand and evaluate the existing workloads and application characteristics to determine the best path for migration. This includes portability of applications, the role of technologies — such as containers, PaaS and open-source technologies — and vendor dependencies for a hybrid or multicloud deployment.

The common cause of failure in migrations to the cloud is that technical professionals fail to take an orderly and structured approach to migration. Streamlining the migration by retrofitting these components and their functions from an existing on-premises analytics platform to the cloud can be challenging and time-consuming without adequate planning, guidance and support. This guidance framework provides a set of considerations that technical professionals can use to evaluate what is involved in the migration, the compatibility of D&A components with the desired strategy, the complexity of the target architecture and the proper practices for completing the migration to the cloud.

“How to create a strategy for deploying data and analytics components in the cloud, hybrid and multicloud platforms?”

The Gartner Approach

The cloud has become the most pervasive style of computing — essential for data and analytics innovation, and anything not cloud is considered legacy.

Organizations move to the cloud for different reasons and require justification — especially when migrating from on-premises to the cloud. Some organizations adopt cloud technologies to deliver improved products and services that drive competitive advantage. Other organizations strive to reduce the amount of technology they manage to lower TCO and improve efficiency. However, organizations moving to the cloud require significant investment in time and resources to overcome security and regulatory concerns, optimize performance, and realize efficiencies.

A well-thought-out approach of migrating the data and analytics platform to the cloud can maximize the cost savings and reduce the ability to take full advantage of the cloud services and service provider capabilities in the cloud analytics deployments. Migrating the data and analytics platform to the cloud requires evaluating each component and workload separately to determine the suitability, compliance obligations, benefits of cost optimizations and price-for-performance constraints in moving to the cloud.

Many of the on-premises data and analytics platforms have grown incrementally over the years, resulting in overlapping and redundant technologies and some inconsistency in how the business users consume information. They need a strategic approach and plan when migrating their data and analytics to the cloud as they become indispensable to help modernize their analytics platform. The broad deployment still needs to be approached with a greater level of scrutiny and caution.

Today, organizations are facing the challenge of dealing with conflicting priorities:

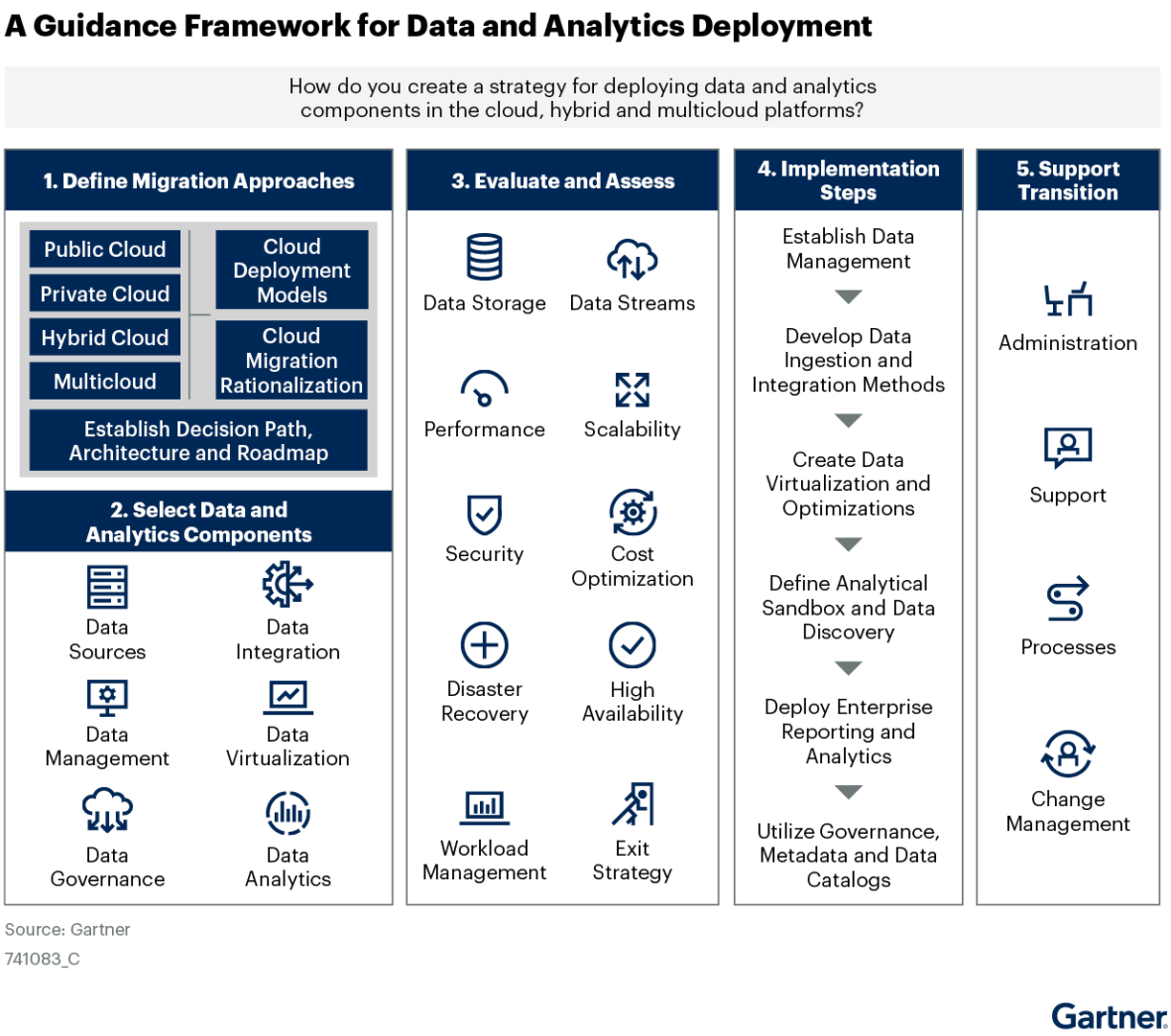
- **Continuous Change (Agility Yet Efficient):** Design and deliver analytical applications that respond to the business needs' desire and use the right technology to achieve differentiation and drive innovative processes.
- **Controlling Change (Scalability Yet Controlled):** Providing a scalable, reliable, secure and cost-optimal platform to support analytical decision-making and mission-critical business processes.

In order to balance these conflicting priorities, it is essential to know when and where to apply more agility and prioritize the controls and efficiencies. So, organizations must establish a renewed approach to deploying data and analytics platforms in the cloud. One of the keys to developing this strategy is understanding the business vision and creating a phased approach to migration. In addition to having a strategy for selecting, deploying and managing the data and analytics platform, it is essential to focus on understanding business capabilities that help combine resources, competencies, information, processes and platform technologies to deliver consistent value to the consumers.

The Guidance Framework

The guidance framework aims to provide a step-by-step process for defining your migration approach; establish a preliminary decision path, architecture and roadmap; and incorporate them into the implementation, deployment and supporting the transition activities (see Figure 1).

Figure 1: A Guidance Framework for Data and Analytics Deployment



Prework

Many organizations are either considering or migrating their new data and analytics platforms to a cloud-based system to support the growing business demand for analytics. The acceleration to the cloud is due to the increased capabilities on the data management platform and support for integrated self-service business intelligence augmented analytics, data science and machine learning capabilities readily available in the cloud.

In most organizations, D&A is not a greenfield or unilateral solution but a brownfield — it consists of multiple layers of technologies that need to work and operate independently. One of the most common mistakes is underestimating the complexity and the time frame planned for the migration, thereby delivering suboptimal results or extended timelines for deployment.

Plan Your Deployment of D&A Applications

To avoid these problems, organizations need to introduce a proactive approach. Without proper planning and execution, the new D&A platform will add more complexity to an already fragmented landscape that is currently in place. Unfortunately, migrating an existing D&A platform from on-premises to the cloud or establishing a hybrid cloud is not trivial — primarily when it already serves up the data and analytics to the business in critical business decisions.

Before deploying your data and analytics platform in the cloud, you need to have a purpose and goals for the migration. Cloud migration demands modernizing your data and analytics platform and is very different from the traditional way of managing and supporting on-premises analytics platforms.

An organization needs to focus more on the process of deploying data and analytics than retrofitting old practices. So, a radically new approach and continuous change management are required. It is also essential to review and evaluate each stage of the process and apply those guidelines to integrate emerging technologies to help serve new business demands in the cloud. It is also helpful to understand the significant shortcomings in the current implementation, and where it worked and delivered to meet or exceed stakeholder expectations. Understanding these can uncover critical information that applies to the migration project.

Also, the decision to migrate varies based on the individual components that constitute your migration. In some cases, you may continue maintaining the existing on-premises solutions or possibly pursue individual components' migrations opportunistically to the cloud.

Depending on the level of maturity of your current D&A deployment, there may be multiple permutations and combinations to define for the cloud migration. It is essential to understand and evaluate those factors and incorporate them into the overall planning for your D&A platform migration and deployment. **We will look into this in more detail in Step 1 of the guidance framework.**

Considerations for Migrating Your D&A Applications

The following are some of the foundation items to consider when migrating your analytical application to the cloud:

- Conduct a feasibility study to review your existing D&A platform. Additionally, determine the best path for migration as many times it may be easier to redeploy in the modern platform rather than retrofitting the existing design by setting up similar technology and deployment practices in the new platform.

- Ensure that both IT and business leaders fully understand your decision process for enabling the D&A platform. Define and document the success criteria, operating model and governance controls, establish a RASCI (responsible, accountable, supporting, consulted and informed) matrix to identify the specifics on the roles for the migration strategy.

- Initiate a program to identify and prioritize the D&A initiatives that can better support the business outcomes, identifying mission-critical priorities, implementing a data literacy program and focusing on operational areas to deliver success.

- Create a steering committee and governance structure to manage the scope, purpose and value. Have a process to measure the success and as the mechanism for achieving D&A outcomes and build a communication plan to support the governance process.

- Access multiple deployment architectures for your end-to-end D&A system. Facilitate a high degree of collaboration and prepare to address conflicts as and when they emerge, and have an arbitrage process as necessary.

- Consider agile, adaptive and incremental steps with your analytics migration or deployment to demonstrate the new platform's viability and long-term sustainability.

For example, a typical D&A deployment on-premises consists of multiple components. The most simplified D&A system, at the minimum, will have a data warehouse for data storage, ETL for data integration and BI tool for Analytics as candidates for migration. **The details on these components are discussed further in Step 2 of the guidance framework.**

Depending on the cloud deployment model chosen, you can keep some of these on-premises or migrate them to infrastructure as a service (IaaS) in the cloud that will require little or no redevelopment effort. However, if you are considering cloud-native solutions such as platform as a service (PaaS), the migration will require additional effort — including but not limited to the target cloud components and service provider selected for such migration.

Typically, the data store migration and analytical reporting tools migration for the same vendor solution to the same cloud solution is straightforward. Many of the service providers offer migration tools to do some essential migration. You still have to manually perform several post migration steps or develop your migration framework to identify the workarounds and alternatives to complete the migration successfully. However, if you consider migrating to a different vendor solution or the cloud service provider's tools, the task can become daunting. The migration steps will involve either third-party services or use the cloud service provider's schema conversion tool or other migration utilities. Even such migration will not be straightforward and require multiple workarounds and additional development steps to successful completion.

Additional details on the deployment steps and postmigration steps are discussed in Steps 4 and 5 of the guidance framework.

Migrating prepackaged analytical applications, such as SAP Business Warehouse (SAP BW) and Oracle Business Intelligence Applications (OBIA) to the cloud is not simple — other than rehosting them, as these applications are built with tighter integration across all layers. Architecting them to use the cloud-native services will also require a complete reengineering and rebuilding effort. Some of the ETL processes, metrics and calculations are predefined, developed and managed by vendors through application patches and upgrades. While there is a significant effort involved in redeveloping them, it also requires a considerable measure in maintaining them as source ERP systems go through significant upgrades and changes.

While we cannot develop a migration plan for every possible scenario, the proposed framework will help understand the steps, general guidance and important considerations for creating a migration strategy for your D&A deployments.

[Framework Component]

Step 1: Define Migration Approaches

Organizations evaluating cloud strategies for data and analytics applications must first identify whether to deploy their new system on-premises using the private cloud, have a single public cloud service provider or use multiple public cloud service providers. Depending on the plan, your migration approach can be either all cloud, hybrid or multicloud.

Understand the Different Cloud Deployment Models

One of the key factors for organizations to migrate to the cloud is to eliminate the need to manage the infrastructure and the software applications internally, as well as maximize the opportunities and enterprise capabilities that cloud computing can provide.

Organizations that consider deploying their D&A platform in the cloud have to go through multiple points of consideration and assessment before finalizing the approach for migration.

In simple terms, each cloud service or a set of services offered by the cloud service provider is called “cloud.” However, as we review the cloud migration, we usually consider several cloud deployments, and the most common cloud deployments are:

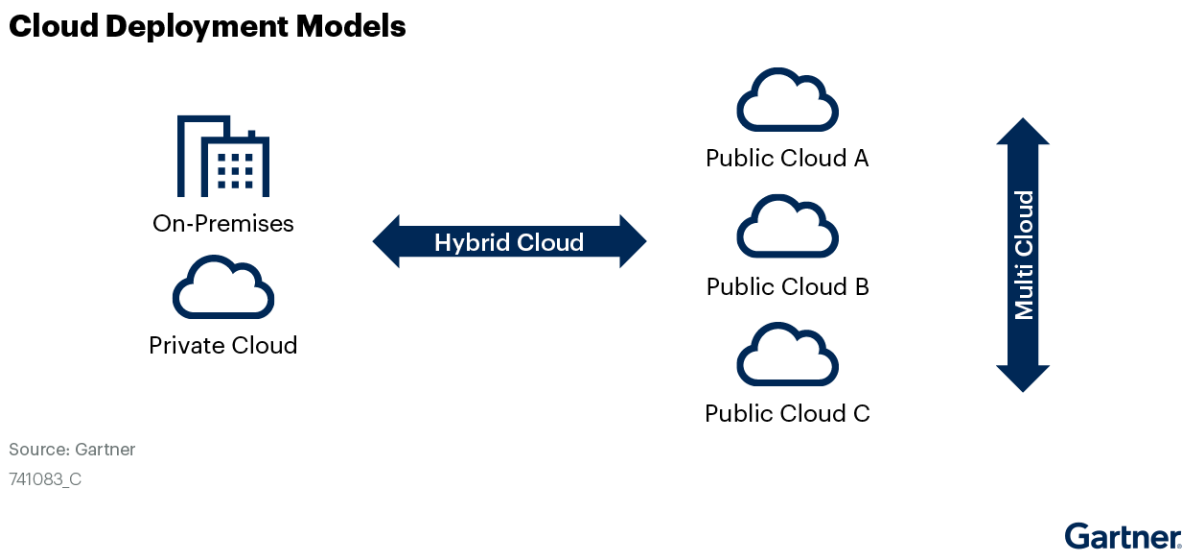
Private cloud: A private cloud is a wholly dedicated service to one organization and not shared with other organizations. Although private clouds traditionally ran on-premises, organizations now build private clouds on rented, vendor-owned data centers located off-premises. So, depending on who manages the private cloud environment and hosting location, a private cloud can be classified into a virtual private cloud, managed private cloud, hosted private cloud and on-premises private cloud. Private cloud usually costs more to deploy due to dedicated resources allotment as compared to public clouds.

The most prominent cloud service providers for deploying the D&A solutions in the private cloud are Hewlett Packard Enterprise (HPE), Dell EMC, Oracle, IBM-Red Hat, VMware, Amazon Web Services (AWS), Microsoft Azure and Apache OpenStack.

Public cloud: A public cloud is a service run by an external vendor that may include servers in one or multiple data centers. Unlike a private cloud, public clouds provide a shared infrastructure to various organizations. Individual servers may be split into virtual services and shared by different companies, a situation called “multitenancy,” because multiple tenants are renting server space within the same server.

Some of the most popular public cloud services available are Amazon Web Services (AWS), Google Cloud Platform (GCP), Microsoft Azure Cloud (Azure) and Oracle Cloud Infrastructure (OCI) (see Figure 2).

Figure 2: Cloud Deployment Models



Hybrid cloud: Hybrid cloud deployments combine one or many public and private clouds, and may even include on-premises systems. An organization may use its private cloud or on-premises for some services and its public cloud for others, or they may use the public cloud as backup or secondary for their on-premises or private cloud.

Using Tableau Online or installing a Tableau Server in the Amazon Elastic Compute Cloud (Amazon EC2) instance and connecting to Oracle or Teradata data warehouse databases on-premises is considered a hybrid cloud deployment.

Multicloud: Multicloud is a type of cloud deployment that involves using multiple public clouds, and the multicloud deployment uses services from several external service providers. Multicloud implementations can also be hybrid cloud deployments.

An example of a multicloud is connecting the Power BI services running on Azure Cloud to Google BigQuery cloud data warehouse on the Google Cloud Platform, (or) to Snowflake Cloud data warehouse running on AWS.

Google Anthos provides a combination of hybrid and multicloud deployment using an umbrella of services, including CaaS built on Google Kubernetes Engine (GKE).

Intercloud: Intercloud deployment is when the cloud's data integrates or exchanges between cloud service providers as part of logical application deployment.

For best optimization of the cloud deployment models, please review and assess the components defined in Step 3 of the guidance framework.

Assess Cloud Architecture to Rationalize Your Migration

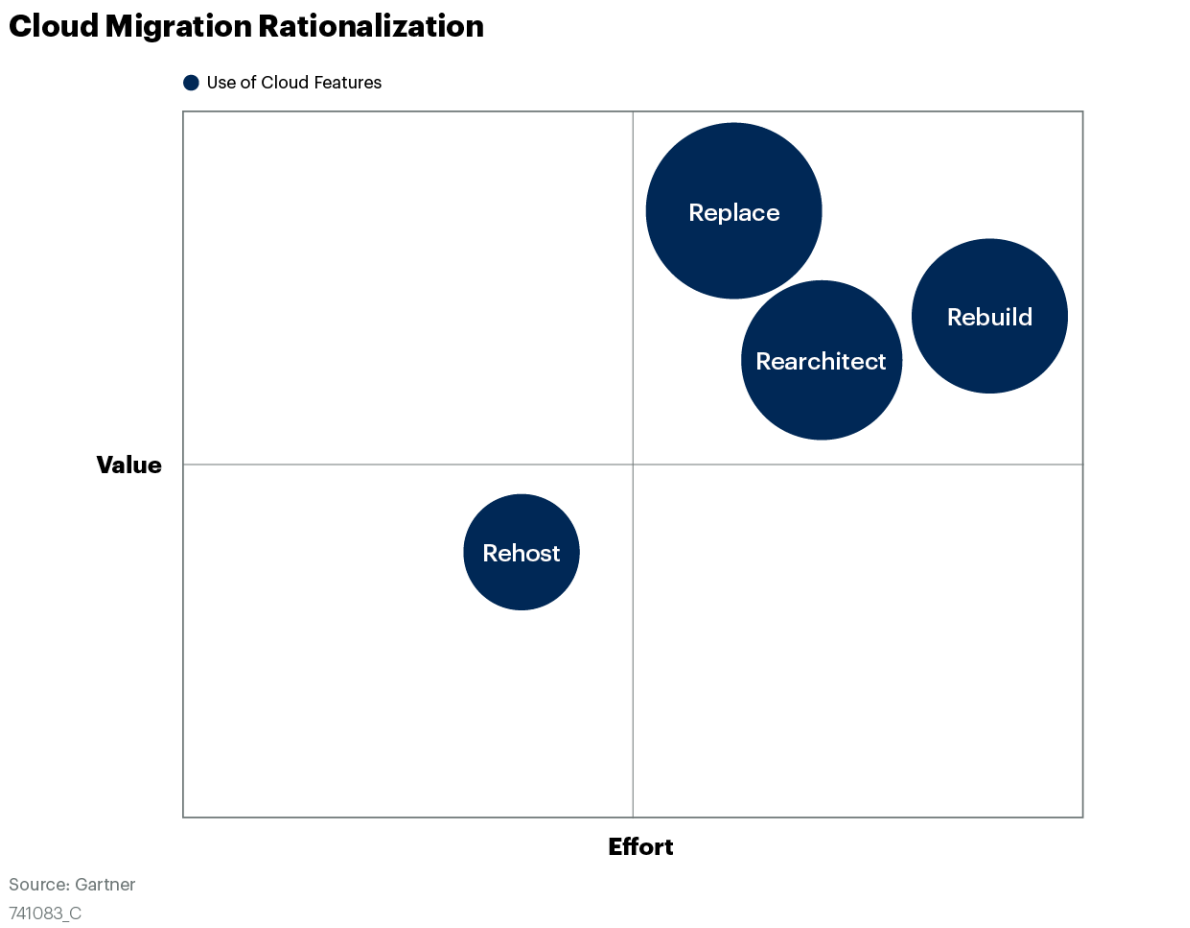
Migrating the D&A application to the cloud requires you to select the best fit. They usually intersect with the commonly available cloud platform architecture types, such as IaaS, PaaS, containers, low-code application platforms and SaaS. The strategy can vary for each of the analytical components, such as selecting data ingestion tools, data management platform, BI tools, AI/ML platform, etc., based on the scope, complexity, deployment architecture and level of effort involved.

There are several factors in play to prepare your D&A migration architecture. Many of these are dependent on the application deployment architecture that you select for your cloud migration.

- The deployment options offered by the cloud service provider to meet your D&A application requirements and architectural considerations of your analytical application.
- The level of controls and expertise you want to leverage from the cloud service provider services, based on internal capabilities.
- The analytics architecture components that you want to keep in on-premises or possibly distribute across more than one cloud service provider.

The following chart provides the effort to value on the different migration approaches. The bubble's size represents the level of usage and usefulness of the native cloud features to maximize the price for performance in the cloud (see Figure 3).

Figure 3: Cloud Migration Rationalization



The most common alternatives available for the D&A platform migration to the cloud are:

Rehost: This is the quickest way to move to the cloud by just doing a "lift and shift" of your existing components without any modifications other than adapting to the hosting environment.

The components are migrated from the on-premises infrastructure to the service provider cloud platform by installing in private, public cloud IaaS platform. One of the other advantages of this strategy is that you can identify individual components to migrate to the IaaS platform while the others remain on-premises. The other factors that drive this migration are the process dependencies, life span and sunken costs associated with the current software license and hardware costs — and the ability to rebuild or replace in the native cloud platform. However, this approach requires some analysis of the dependencies to determine any impact and implement additional controls or processes for proper migration and execution.

An example of rehosting is equivalent to migrating one or more of the following components running on physical servers or virtual machines to the cloud.

- Data Lake on Apache Hadoop Clusters
- Data Warehouse on SQL Server Data Warehouse Database
- Data Integration on SQL Server Integration Services (SSIS)
- Data Analytics in Tableau Server

The effort requires the lift and shift either manually or automatically from on-premises to cloud virtual servers running on Amazon EC2 instance, or Azure Virtual Machines or Google Cloud Compute Engine Instance based on the public cloud service's choice provider.

Rearchitect: This encompasses thoroughly looking into altering the application to fit the cloud-optimized architecture, leveraging some of the native cloud-native capabilities and using managed cloud services such as PaaS or container as a service (CaaS). These can reduce operational complexity, elasticity on-demand and optimal resource utilization. This alternative requires in-depth evaluation and assessment of the technology, as well as identifying ways to make them cloud compatible before migrating the application components to the cloud-native platform services. The rearchitect approach introduces cost and operational efficiencies than running them as-is by keeping them on-premises or rehosting in the cloud.

Some of the common examples of rearchitecting or replatform of a D&A solution are:

- Migrating your on-premises SQL server databases to Azure managed instances running on PaaS or porting the SQL databases to run as CaaS in the cloud.
- Deploying the streaming data architecture running on Apache Kafka in on-premises to a managed cloud alternative, such as Azure Stream Analytics, Amazon Kinesis or Amazon Managed Streaming for Apache Kafka (Amazon MSK).

Rebuild: This requires rebuilding the application entirely from scratch, prescribing to PaaS architecture as part of modernizing your D&A platform, allows the organization to adopt new technologies with better productivity and ease of use, and is potentially better suited to meet the current and future business requirements. This approach will help realign the existing business processes by translating into a complete cloud-native — thereby accelerating the innovation through reinvestment in the cloud.

A rebuild will involve a complete overhaul of your D&A deployment to leverage all the latest modernization and cloud-centric tools for your deployment. Some common examples for rebuild include:

- Migration of data to cloud object storage to optimize the storage costs.
- Integrating microservices, open-source data processing engines and serverless architecture by maximizing the best use of pay as you go, pricing models.
- Utilizing specialized data stores that are determined based on the use-case driven approach.

Replace: This alternative is about overhauling the entire application to align with modernizing your overall business strategy. Some of these decisions are driven based on the replacement or overhaul of your source systems, and make it necessary to put together your custom analytical application entirely — or opting for a commodity SaaS analytical approach. The drivers supporting this approach are transforming your D&A deployment to make it more business process-driven and aligning to standard industry practices as part of the transformational effort.

Replacement is a sweeping change to the product components in the cloud deployment, compared to the existing on-premised D&A platform. In this example, you may choose to implement everything brand new using the cloud-native tools and components in the cloud, such as

- Substituting your existing on-premises Oracle data warehouse with Amazon Redshift or Snowflake.
- Replacing your existing business objects reporting with Microsoft's Power BI, Amazon QuickSight or Google's Looker in the Cloud, etc.

As part of the replacement, you may choose to bring some historical data, business logic and metadata from your legacy systems. Most of the service providers offer tools for such migration, even though it may not be complete.

Replacement also means moving to a commercial off-the-shelf (COTS) such as Salesforce, Microsoft Dynamics 365 or software as a service (SaaS) analytics solution such as Adobe Analytics, Google Analytics, etc., deployed in conjunction with the migration of your transactional sources or other business transformation initiatives, such as ERP, HCM or CRM implementations.

The data and analytics platform consists of multiple components. Each of these components will require its own set of assessments and rationalization to determine the best approach to meet both your short-term and long-term goals. The Decision Point for the migration is driven based on your understanding of these components, characteristics and architecture that best fit the development and operational skills required to build and deploy the system.

Establish a Preliminary Architecture and Roadmap

Defining an approach is an important decision that will drive creating the strategy needed to deploy the D&A platform and the associated applications in the cloud. This decision will also determine how much effort, complexity and benefit that you get from moving an existing D&A platform, or deploying them as a brand new D&A system into the cloud.

Some of the key deliverables and outputs that you should consider producing at the end of the first step are:

- Identifying the cloud deployment model that is best suited for your organization and the factors that led to making the deployment decisions.
- Assessing the cloud migration approaches and architectures.
- Evaluating the decision paths for the various migration alternatives.
- Creating a deployment plan and a roadmap for your D&A deployment.

As you review the different approaches, architecture and models, you need to derive a preliminary strategy for your D&A migration, understanding that it is still fluid and subjected to change as you iterate through the next set of steps.

Step 2: Select Data and Analytics Components

Data and analytics platform is not a single thing, and it includes multiple pieces of architecture and technology elements. Building a complete end-to-end data analytics platform requires identifying and assembling various technological components. To further understand how to build a holistic data analytics platform, see Solution Path for Building a Holistic Data Management and Analytics Architecture.

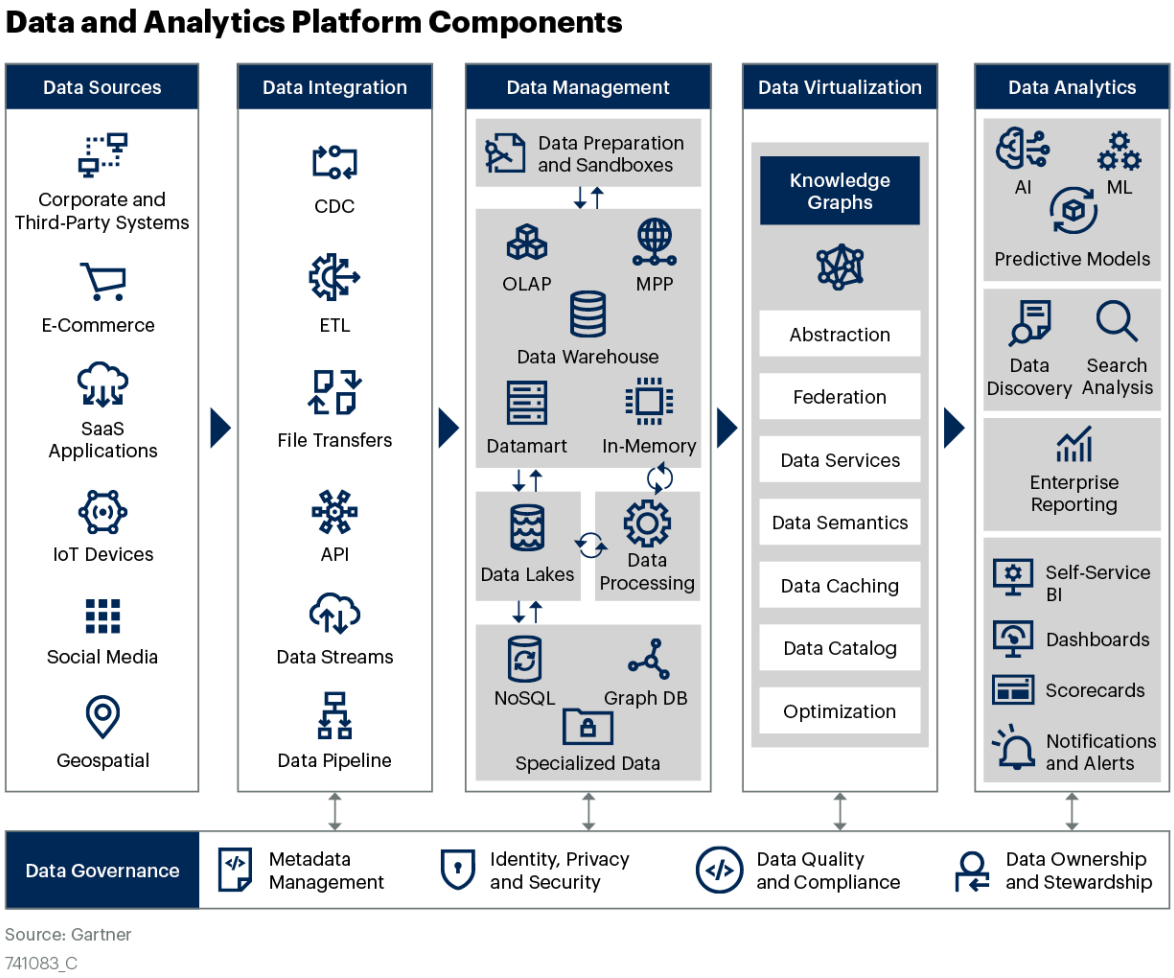
Every organization considering D&A deployments — regardless of whether it is an on-premises migration to the cloud or establishing a brand-new analytics system in the cloud — has some common challenges to deal with, such as:

- Focusing on the technical aspects and less on solving business problems.
- Lack of understanding of the business use cases and outcomes.
- Ability to support new use cases and unknown data growth and workloads.
- Building silo investment and in an ad hoc fashion without a good vision for your future-state analytics deployment.

It is essential to assess the data and analytics components individually to determine the impact on the new deployment challenges and a decision path needed for each component.

The following diagram represents the end-to-end D&A architecture and the various components that are integral parts of the data and analytics platform (see Figure 4).

Figure 4: Data and Analytics Platform Components



Data Sources

Data sources represent data from various structured data sources: transactional, operational, external references, etc. The data and analytics platform intended to integrate data from these data sources — either in batch or in near real time — through Change Data Capture (CDC), streaming technologies bringing in structures, unstructured and semistructured content from different types of sources. These include social media, sensors, machine learning log data, weblog files and data generated through the Internet of Things (IoT). Additionally, all sorts of data are available in various business systems, on-premises applications and cloud applications, public domain and other complex data sources introduced every day at much faster rates.

While a D&A platform deployment to the cloud has any impact or changes to the source systems, the data sources play an essential role in your data and analytics strategy. In most cases, a change in the data sources, a replacement of the data sources or adding new capabilities to an existing data source or new data sources can potentially drive significant changes to the D&A platform deployment.

In many organizations, a change in the data sources as part of an ERP transformation, replatform or migration of the source system to different technologies drives your D&A strategy shift. Other factors include possibly migrating the source system from one service provider to another provider can significantly impact the sources and the complete data and analytics platform. Introducing a new set of data sources — such as streaming through sensors, IoT devices or blending of data from multiple sources to broaden your analytics scope — can advance changes to your analytics deployment. Each component requires evaluation, assessment and entirely overhauled in the new platform deployment.

Data Integration

Data integration should consist of the characteristics that can support relational, big data, cloud, application, IoT data and other complex data integration through batch and stream processing near real time. The integration also includes embedded processes to address data quality, data cleansing and master data management principles to support overall data integrity, validity and governance.

Some of the most common integration methods include batch processing through ETL/ELT technologies into data lakes, data warehouses, advanced Integration for loading data into specialized data stores, real-time enterprise messaging and lightweight integration services using structured data formats. These include JSON, XML, etc. and cloud-based integration through Restful API and integration services through IaaS and PaaS.

Since cloud data integration handles various systems and data types, there is no single service or tool that can provide full-fledged data integration capabilities. It is also essential to have the integration platform tightly aligned to the metadata management solution.

Defining a centralized hub for data integration through a data pipeline platform and subscribing to data streams can help manage data ingestion and scalability and can provide additional functions, such as cleansing and enrichment, before loading the data into data lakes and analytics targets.

Data integration is linked tightly with your data sources and data management. The data integration effort relies heavily on the source system and selecting the cloud's target data management platform.

Data Integration in the cloud also provides new ways to modernize your data ingestion and data processing requirements. These data integration tools need to support hybrid data environments to unify on-premises and cloud-based data sources and targets. They are also designed for scalability and performance and offer critical optimization capabilities — including push-down optimization so you can process the data efficiently.

Data integration tools typically support a combination of the following main data delivery styles.

Bulk/batch data movement: Bulk or batch data extraction and delivery approaches — such as support for extract, transform, load (ETL)/extract, load, transform (ELT)/extract, transform, load, transform (ETLT) — to consolidate data from distributed databases and formats. This capability draws on data from across systems and organizational boundaries and can play a role in multiple use cases.

Message-oriented data movement: This capability allows data integration tools to encapsulate data in messages that various applications can read so they can exchange data in real time. This is often via message queues and data as a service (DaaS), or even data services orchestration mechanisms.

Data replication/synchronization: This capability allows data integration tools to provide a copy of data by physically moving it in near real time from one location to another, always in a physical data store. This capability can be a basis for all data integration styles but, specifically, does not change the form, structure or content of the data it moves.

Stream data integration: The ability to address data integration requirements through interoperability with streams/events, including provisioning data in-stream for enabling downstream consumption or analysis.

It is typical for organizations in the on-premises deployment to have a single tool to consolidate multiple integration styles. As you get ready to migrate these into the cloud, it is optimal to leverage various services for different integration styles. For more details, see Assessing Cloud Data Integration Options. Also see Deploying Effective iPaaS Solutions for Data Integration.

Data Management

Data management includes multiple technologies to store and process data that can support the broad enterprise requirements. Cloud data stores have gained rapid adoption due to low-cost storage and the ability to scale on on-demand storage and processing using commodity hardware systems, innovation contributions through open source technologies and ease of portability to the cloud.

The concept of data lakes includes distributed file systems, object stores and specialized data stores. Cloud data management also has nonrelational and special-purpose data stores such as graphs, time series and in-memory databases. Enterprise data warehouse and data marts also form an integral part of the cloud's modern data architecture to help deliver conformed, structured and governed data (see Demystifying Cloud Data Warehouse Characteristics).

The rapid evolution of data warehouses to support optimized workloads through massive parallel processing, in-memory, columnar storage, native data transformation through ELT commands and routines deployable in the single, cluster or cloud computing make it more relevant in the cloud data management architecture. The data warehouse concept in today's enterprise has moved from physical consolidation and data management to more logical and data virtualization paradigms to support historical data analysis, batch processing, and transactional workloads and reporting. It is typical for organizations to introduce new data management layers in the cloud as part of the data modernization effort and use efficiency and cost management by selecting the data management layer.

Data management in the cloud addresses scalability, handles high concurrency workloads with low latency and faster throughputs, and uses it on a pay-per-use basis. Additionally, the cloud-based data stores are accessed through multiple methods using traditional query languages and APIs.

The deployment models include delivering databases through PaaS and microservices architecture orchestrated by containers. In addition to conventional relational and nonrelational databases for analytics, the cloud offers hybrid transactional/analytical processing (HTAP) data stores to support faster ACID transactions, as well as real-time and near-real-time analytical capabilities (see Solution Criteria for Cloud Analytical Data Stores).

To further learn on how some of the cloud service providers met the criteria and their corresponding scores, please see:

Solution Scorecard for AWS Cloud Analytical Data Stores

Solution Scorecard for Microsoft Azure Synapse Analytical Data Stores

Solution Scorecard for Google Cloud Analytical Data Stores

Data Governance

Data governance becomes the integral part that ensures only the highest-quality data is gathered, processed and funneled to establish a single source for all the enterprise data needs.

Once the broad governance guidelines are in place, the focus turns to catalog metadata — mapping fields and defining how to parse and publish data collected and managed from all the sources. Other challenges include implementing cloud governance practices to efficiently share, access and leverage data in the cloud to the analytics consumers while remaining compliant.

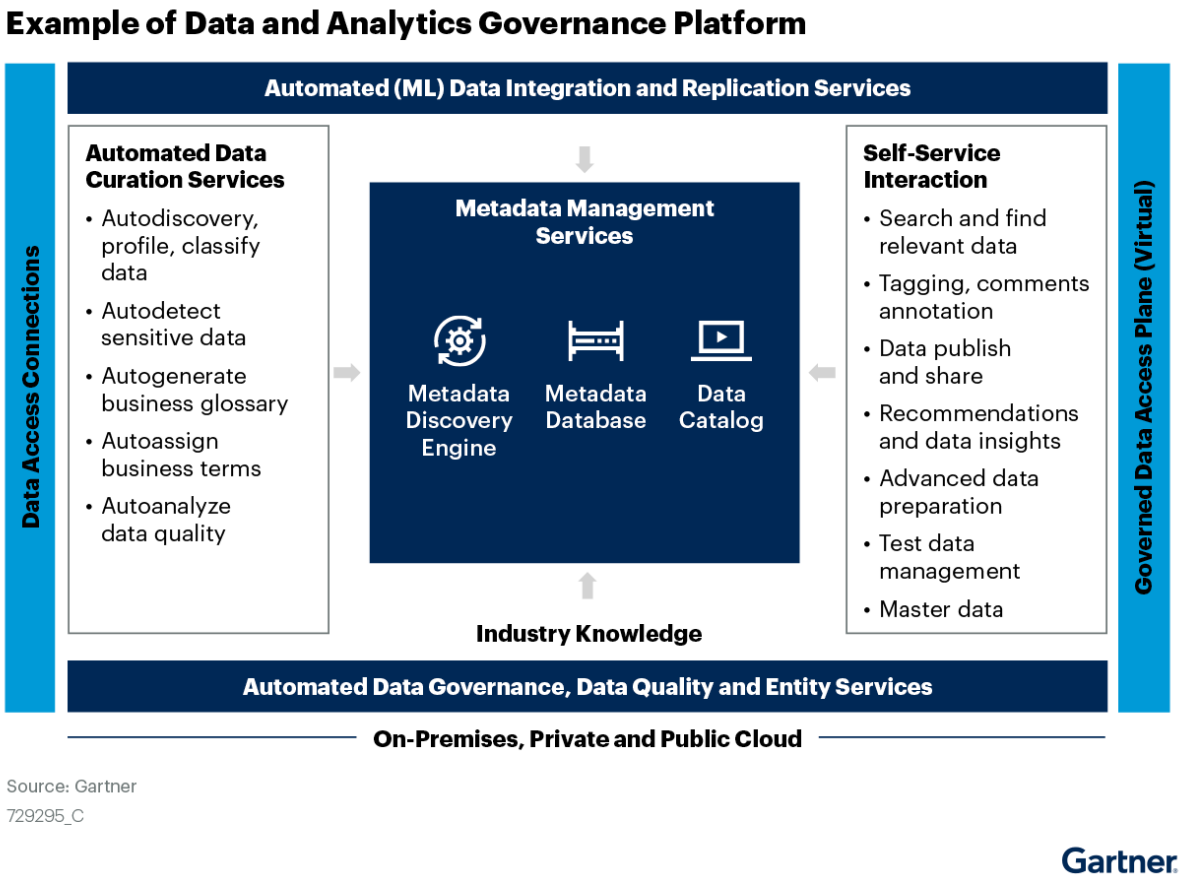
Today, with increasing numbers of data creators and consumers, governance must shift to being about enablement and appropriate levels of more granular control.

Business value driven by data and analytics is created within business functions and through formalized data and analytics teams. Because governance practices in business functions are localized, fragmented and inconsistent, they limit the value realized and scalability of investment (see Implement Your Data and Analytics Governance Through 5 Pragmatic Steps).

Data governance covers a broad range of activities in a D&A deployment from data access, automated data discovery and curation, data management, data catalog, and automated data integration using ML and self-service analytics. Organizations need a scalable and disciplined approach to apply governance across the entire data and analytics ecosystem (see Building a Comprehensive Data Governance Program).

Deploying D&A in the cloud introduces other aspects of data governance, such as regulatory compliance, security, availability, usage, data privacy and data use sensitivity (see Figure 5).

Figure 5: Data and Analytics Governance Platform



Source: Gartner
729295_C

Gartner

Data Virtualization

Data virtualization has become an emerging trend based on the execution of distributed data management processing, primarily for queries, against multiple heterogeneous data sources and federation of query results into virtual views. The data delivered through these virtual layers will be made available to analytical applications, reporting tools and other data delivery tools.

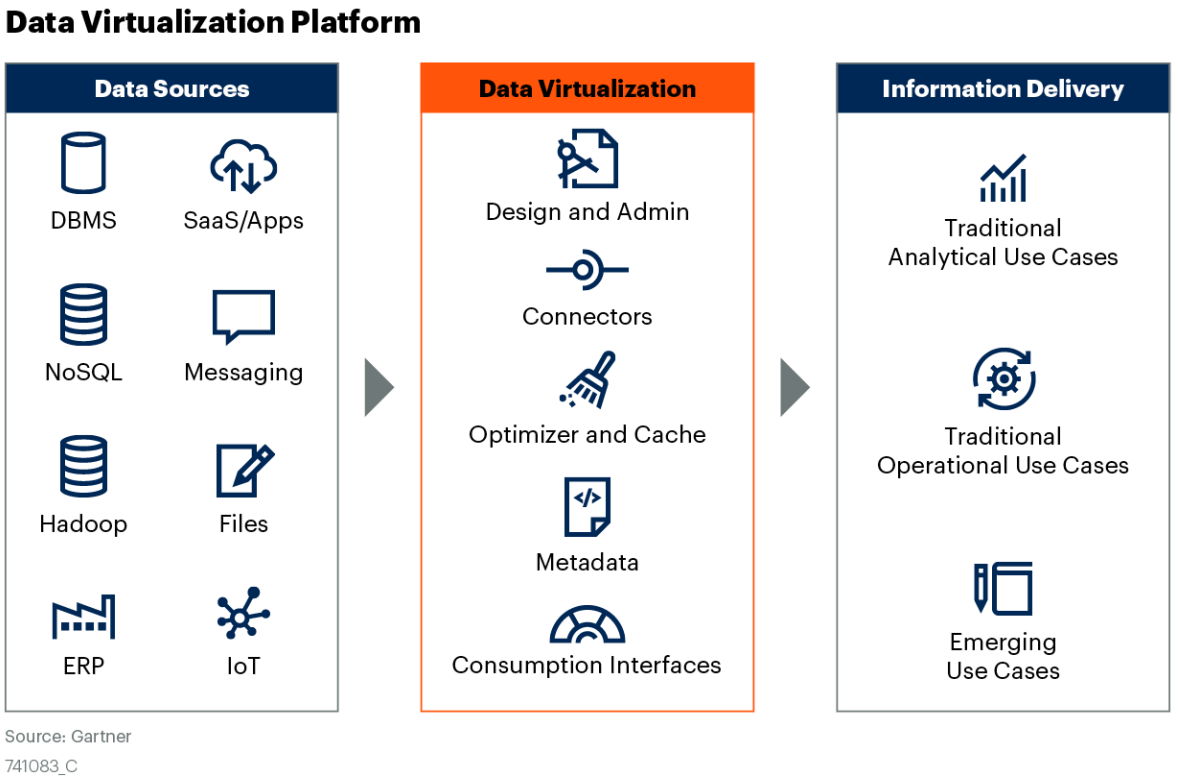
Data virtualization provides a semantic layer that simply stores the metadata and integration logic for viewing the data through a single point of access via centralized integration and management. Thereby, they enable the organization to support enterprise data governance and master data management initiatives.

The more-modern data virtualization tools have native connectors to some of the most popular data sources and applications. These sources need to be integrated, including relational DBMSs, flat files, XML data, JSON data, message queues, streaming/event data, IoT data, packaged applications and cloud data stores. They also include SaaS applications, big data stores (e.g., Hadoop data stores and NoSQL DBMSs), in-memory DBMSs, social media data and unstructured data sources to deliver an enhanced data and analytics platform.

Data virtualization plays a pivotal role in the migration of analytical workloads to the cloud. Using the virtualization concept, organizations can incrementally and iteratively bring their data into the cloud while keeping the core data in on-premises or other private clouds. In case of the need to keep the data in the cloud for performance, data replication of your on-premises data to the cloud can be an option.

Most data virtualization tools available today provide on-premises and cloud, as well as hybrid deployment options. Some vendors support hosted versions of their on-premises offerings on private clouds and popular public cloud service providers. Organizations look for the right PaaS offerings that provide them with the flexibility and scalability of cloud-based data integration tool options. Also, there is still a gap in the market for tools that can support multicloud options, allowing workloads to be stored/processed across clusters in different cloud service providers (see Figure 6).

Figure 6: Data Virtualization Platform



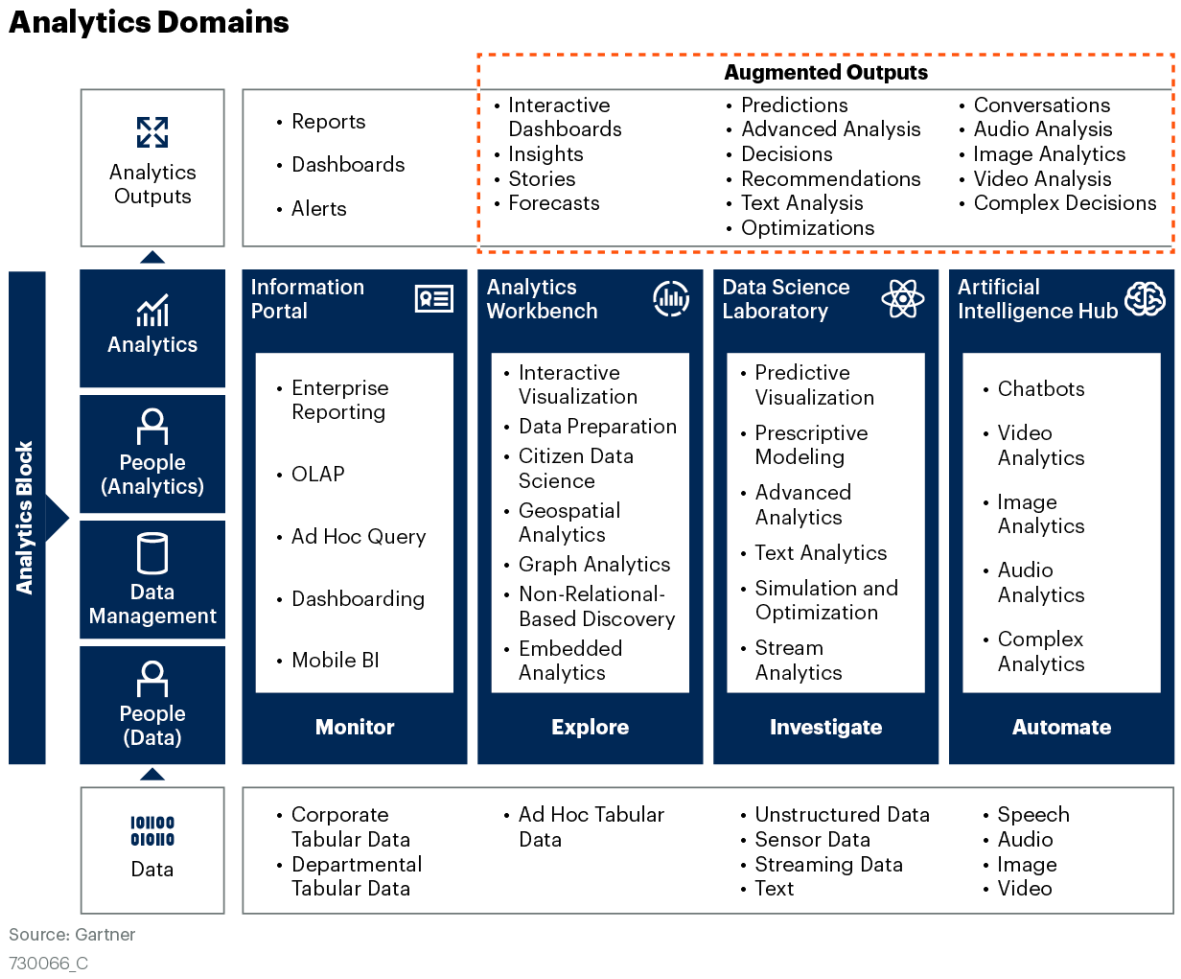
Data Analytics

Data analytics supports a wide variety of features through business intelligence tools to enable businesses to drive critical data-driven decisions. The most common features and capabilities include delivering data visualizations, dashboards, ad hoc query and analysis, custom reports, data discovery search analysis, predictions, alerts, and workflows.

Business users will have all the self-service tools to help correlate their business requirements through user stories and map visualization by developing storyboards in powerful BI and analytics tools. Analytics centered on embedding artificial intelligence (AI), machine learning (ML) and predictive analytics within the modern data architecture. It is a collection that includes model training and predictive models derived on new events based on a specific learned pattern, and used to continuously improve and redeploy the analytics models overtime on a production scale. Predications are formulated and executed within an AI application or microservice. With the complete breadth of data available from the enterprise data lakes and data warehouses, AI systems will significantly benefit through models and insights.

Augmented data management refers to applying AI and ML for optimization and improved operations of data management tasks. AI and ML are used — based on the existing usage data — to tune processes and optimize configuration, security and performance (see Figure 7).

Figure 7: Data and Analytics Domain



Source: Gartner
730066_C

Gartner

To further evaluate the analytics, BI, data science and machine learning platforms, Gartner has defined a set of solution criteria. For more details, see:

Solution Criteria for Analytics and Business Intelligence Platforms

Solution Criteria for Data Science and Machine Learning Platforms

To further learn on how some of the major vendors met the criteria and their corresponding scores, please see below and other analytics Solution Comparison and scores available through cloud decisions. [1]

Solution Scorecard for ThoughtSpot

Solution Scorecard for Tableau

Solution Scorecard for Qlik Sense

Solution Scorecard for Amazon SageMaker

Solution Scorecard for Microsoft Azure ML

Solution Scorecard for the Google Cloud Data Science and Machine Learning Platform, February 2020

Step 3: Evaluate and Assess Cloud Service Characteristics

As we review these steps, there are several ways to approach and implement your D&A platform in the cloud and, in most cases, they are ideally implemented and deployed in concert. In addition to the steps, it is vital to evaluate and assess the cloud's multiple service characteristics — including but not limited to event streams, data storage, performance, scalability, services and pricing models, governance, compliance, and others.

One of the biggest drivers for data and analytics applications to migrate to the cloud is to leverage some of the unique characteristics that the cloud has to offer. D&A applications are likely to run more efficiently if we use the cloud-based performance optimization tools and architectures designed with proven methodologies. The use of cost optimization and monitoring tools can help organizations navigate through the complex vendor pricing models.

The overall decisions of selecting the best deployment must lead to those priorities and objectives, as there is no single approach.

Data Storage

Organizations moving to the cloud have a comprehensive set of data storage and management options.

Organizations are increasingly migrating their current data lakes to use object storage as the primary data storage for data lake platforms and alternative or complementary to the on-premises Hadoop-based data lake storage. Object storages have improved drastically to offer high availability and scalability to deliver better economies of scale for data-intensive applications. Object store also eliminates the need to manage and maintain a complicated Hadoop-based data lake storage in the cloud by transferring some of these workloads, which supports direct access to the data through database interfaces and various analytical tools.

Data storage in the cloud offers multiple services and types that include replicating data across multiple clusters. The replication of data provides maximum reliability and data storage in various locations for redundancy, recovery and improved access to the end user.

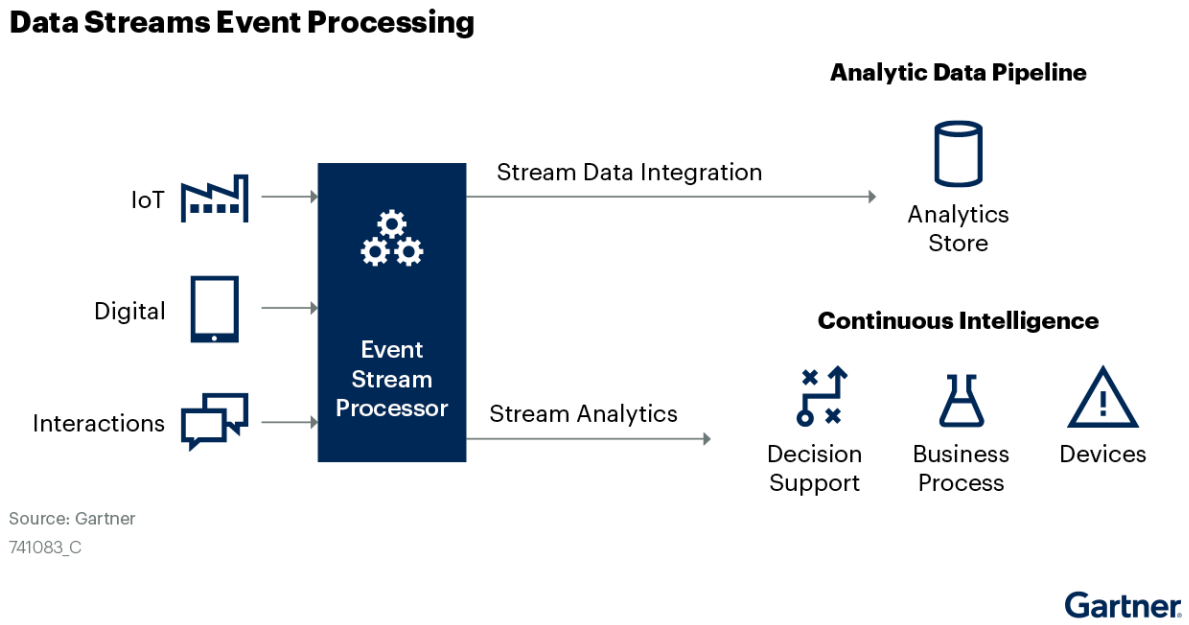
A single data store for all data is not practically feasible — especially with the growth and variety of systems that deliver analytics data. Cloud data management offers several other types of data stores, such as relational data stores, columnar data stores, event stores, graph stores, in-memory data stores and other distributed storage options.

Cloud also offers hybrid access to multiple storage to minimize data copying between the different storage components for various consumption types. The support for structured and semistructured data types such as Apache Avro, Optimized Row Columnar (ORC), Apache Parquet, JavaScript Object Notation (JSON), XML and other data types is increasing among the cloud service providers. Some are integrated into the data warehouse's core feature, whereas others do the same job in conjunction with the storage cloud. This approach provides additional flexibility and eliminates redundancy for data analysis, data integration and data virtualization scenarios.

Event Streams

Organizations are doing more event processing through data and event streams due to continuous intelligence, better situational awareness and faster, more precise business decisions. There is a growing demand for the adoption of streaming platforms because of all the cloud benefits. However, some considerations are required in the hybrid cloud model on data transfer — especially for low-latency applications. The rise of ML and advanced analytics workloads as part of the cloud migration strategy makes it more critical to use event stream processing and distribution of multiple streams for various types of integrations (see Figure 8).

Figure 8. Data Streams Event Processing



Performance

As individual D&A components migrate to the cloud, it is essential to ensure that the deployed solution delivers optimal performance and availability. D&A developers have to think about performance management as they build or migrate upfront as part of their cloud deployment process.

Deployment of D&A applications, especially on a hybrid or multicloud environment, has to be approached with greater scrutiny as it can introduce latency and other performance problems.

Managing application performance in the cloud needs to define and communicate the business and technical requirements correctly, and would include performance standards around availability, reliability, connectivity, number of users, performance, such as response times, latency, etc.

Performance becomes a critical factor if the suitability and architecture are not correct — especially the issues related to network latency on the hybrid cloud deployment. The tier dependencies do not get set up locally between and across the cloud environments.

It is essential to monitor the underlying infrastructure resources that the D&A components support and continuously measure their cost for performance to prepare and predict for managing future performance in the cloud.

The other factors that affect performance are the placement of the data to minimize the movement across the D&A platform, selection of instance types and microservices to match the workload characteristics, and the best storage layer for better performance and throughput. Other dependent factors include allocating resources to run a specific workload or query, the assignment of compute nodes, the level of concurrency and the volume of data processed by these workloads.

While there is no single method to achieve high performance, the following are the options to consider to improve your D&A application's overall performance.

- Selection of different Instance types based on the characteristics of the workloads.
- Implementation of autoscaling with the ability to set thresholds and readjustment of resources based on demand.
- Identifying the most optimal storage, data store option and caching methods to improve latency and faster execution of workloads.
- Implement query optimization techniques and reorganize your data storage to minimize query joins where applicable.
- Use serverless computing, event-driven architectures and microservices architecture to execute your workloads in a modular manner and optimize the resource consumption in the cloud deployment.

The service provider automatically allocates, deallocates and redistributes the data across nodes. The provider also allows the pausing of clusters based on the demand and usage of the data. Although this approach can save costs, it may sometimes conflict with providing consistent performance for unplanned workloads or user queries.

Scalability

On-demand scalability is unique to the cloud that provides a planned level of increase or decrease in your application components' capacity as needed.

It is crucial to evaluate the individual D&A components that are already horizontally scalable as a better fit for migration onto infrastructure that offers the ability to scale resources in response to demand elastically. Determining the degree to which the components are autonomous and stateless, and finding the most critical processing bottlenecks, are required prework for this decision.

Cloud eliminates the barriers by allowing the capacity to start small and support the growth based on demand. With each D&A application component needing its own sets of scaling principles, the cloud allows configuring them independently to support horizontal scaling by adding more capacity and vertical scaling by scaling out more resources to meet the demands.

Some of the other best practices to meet the cloud's scalability requirements are implementing automatic scaling, balancing your workload across multiple compute resources, and using containers to create scalable and portable workloads. However, implementing or lifting and shifting a nonscalable application to an elastic and pay-as-you-go environment won't work and becomes expensive. Several cloud D&A components have native capabilities to support autoscaling, both up and down and balancing workloads.

Most of the cloud's analytical data stores are built on Massively Parallel Processing (MPP) principles to deliver the desired throughput and performance. However, several factors can affect performance — including the volume of data and the other overheads manifested by data distribution on a shared-nothing architecture.

Other factors to consider for managing efficiencies to scalability include:

- Use of fine-grained control of preconfigured computing and storage resources to independently scale up and scale down and avoid overpaying for expensive, unused, fixed blocks of cloud resources.
- Flexible scaling and autoallocation of resources for continuously changing workloads.
- Use of serverless approach to scaling dynamically for selective workloads.

Cost Optimization

Cost optimization is the process of reducing your overall cloud spend by identifying mismanaged resources, eliminating waste, reserving capacity for higher discounts and rightsizing computing services to scale.

Unlike on-premises, organizations are continuously billed for consumption in the cloud, and if not managed with appropriate controls and optimization strategies, it can quickly go out of control. Cloud also poses difficulty in creating or forecasting accurate cost estimates due to complex and multifaceted pricing, ease of provision of resources and change of cloud offerings, multiple paths to deployment, and lack of standardization between and across the cloud platforms.

Some of the factors that are critical for optimizing the cloud cost are:

- Selecting the cloud service provider to best suit your organization's overall cloud strategy and D&A deployment. Cost provides multiple ways to deploy and run your application, and selecting such options drives the overall costs. The goal is to find the right balance of identifying the on-premises, cloud and multicloud architectures that can serve your organization's enterprise needs and overall operations costs.

- Capacity planning and sizing your D&A deployment is to identify the correct instance types and storage types for migration. The workloads meet the service-level requirements and autoscale those resources as the demand increases over time.

- Cloud computing brings innovation to your applications through redundancy. Based on the deployment option selected for the individual D&A component, the redundancy is achieved either by distributing the workloads within a region, or across multiple regions. The selection of different distribution options reduces the overall cost.

- Data access and movement is an important part, and establishing a distributed data management for your D&A platform impacts latency and performance, and the egress costs for actively moving the data across the different cloud and on-premises systems.

- Develop a process to manage, monitor effectively and secure the D&A platform that spans across cloud, on-premises and multicloud environments. Sometimes, you may have to supplement the native tools offered by cloud providers with other third-party tools.

- Use of automation to proactively manage and monitor the resources and control the overall cost leakages.

For cost optimization, AWS provides Cost Explorer and Trusted Advisor, and Azure offers similar capabilities through Azure Advisor. Though Google Cloud Platform does not render one, it does offer some cost optimization functionality distributed across other tools. For example, Google Cloud Platform provides the Cloud Scheduler service, which automatically starts and stops infrastructure resources on a user-defined schedule.

Some of the other third-party vendors that offer cost optimization on these cloud service provider platforms are:

- Apptio (Cloud Business Management and Cloudability)
- CloudCheckr
- Flexera (Flexera Optima)
- Turbonomic (Turbonomic Platform and ParkMyCloud)
- VMware (CloudHealth by VMware and vRealize Operations)

Security

Cloud D&A deployment requires a comprehensive view of the security ranging from data breaches to data access, data encryption and protection in various forms. In general, the cloud service providers have purpose-built technologies and continuously deliver new offerings and innovations in securing data than traditional on-premises systems. Data and analytics platforms are subjected to the same security challenges and threats as any other application deployed in the cloud. Still, the impact of not having appropriate controls on security measures for D&A deployment is severe, due to the data-centric nature of D&A applications.

Most of the common challenges that organizations need to tackle with D&A systems are data breaches and data access attempts through the applications hosted in a public cloud. Others include database access, data integration process intrusions, safeguarding deletion of records through data purging, and ingestion of malware functions across database and application networks.

Although these challenges are familiar, even to traditional on-premises data warehouse systems, they are more prevalent in cloud-based systems. Also, in many cases, it is not the cloud system but the lack of knowledge, controls and governance that are in place to secure the D&A platform.

The storage and management of data lakes and data warehouses are managed through network isolation policies and virtual private cloud (VPC). Some of the other areas of security that are of importance to the cloud data warehouse systems are:

- Maintaining control over data access and governance through access controls — authentication, authorization, data storage (encryption, protection, and restrictions), key management, monitoring and alert mechanisms.
- Automatic encryption of data both at rest in storage and transit during data transformation, integration and access. Data encryptions using hardware-accelerated AE-128, Secure Sockets Layer (SSL) connections and protected access using a virtual networking environment to prevent external access outside of the organization's corporate network.
- Identity and access management (IAM) to manage resource access, two-factor authentication, federation and single sign-on authentication. Support for integration with external authenticators and data, and application security using role-based access controls.
- Supporting numerous encryption mechanisms, including SSL/Transport Layer Security (TLS), IPsec and Advanced Encryption Standard (AES).
- Undertaking and addressing some of the formal third-party security evaluations. These include the International Organization for Standardization (ISO) 27001 and Service Organization Control (SOC) 2, as well as the Federal Risk and Authorization Management Program (FedRAMP), Payment Card Industry Data Security Standard (PCI-DSS) and encryption keys (provider and customer-based).
- Cross-region replication of access control list (ACL) overwrite, cross-region replication with KMS and detailed inventory report.
- General Data Protection Regulation (GDPR) presents a joint opportunity for cloud service providers and customers to implement new strategies and data protection methods. This opportunity helps meet the legal requirements of the responsible authority.

Although there are security models and controls available across the cloud service providers, there are differences in how they approach security. In many cases, the tools they apply to control these security measures are entirely incompatible.

Making the most public cloud benefits means carefully governing your organizational practices for the cloud to avoid introducing security or regulatory exposures.

Although end-user organizations benefit enormously in the public cloud, large service providers leverage sophisticated systems and vulnerability management that are not typical in on-premises systems. Their economies of scale make it possible to deliver around-the-clock safety monitoring and response.

Snowflake data warehouses focus more on enhancing features and security of applications, and less on protecting the cloud data center or the underlying cloud services. Those processes are generally managed and governed by the cloud service providers.

Similarly, cloud data warehouse vendors like Oracle address security and vulnerabilities through autonomous, ML and self-security capabilities.

Most cloud service providers offer enhanced and integrated D&A security platforms to keep the customers within the provider platform to help serve and address broader security needs.

Although the cloud provides a secure environment for the D&A platform, the customer's controls are of the utmost importance. The safe use of all forms of public clouds involves new skills, procedures, activities and policies. Cloud access security broker (CASB) tools are becoming increasingly popular. These bridge cloud service provider users and cloud applications, enforce policies, and monitor usage. No technologies can ever be considered secure or reliable when users and technical professionals are provided a new capability with no guidance on its use or management.

High Availability and Disaster Recovery

The primary focus on high availability (HA) and disaster recovery (DR) is to keep the system operational all the time. However, the main difference is that HA tackles the problem as the system runs, whereas the DR comes in after it fails. Regardless of how highly available a system is, any production application needs to have some sort of disaster recovery plans, as they are not necessarily mutually exclusive. By making them work in concert, the organization achieves a higher level of fault tolerance with no disruption to the end users.

With cloud deployment, organizations automatically back up their applications — including the cloud storage data as an alternative to a secondary DR site — and have them available instantaneously for recovery purposes. You can store your backed-up data across multiple geographical locations, thus eliminating a single point of failure. Some modern cloud data warehouse platforms offer flashback and time travel functionality to restore data-related objects that were accidentally removed or deleted from the system. You can always have a backup copy, even if one of the cloud data centers fails. Cloud providers' access to cloud storage is generally a cost-effective and long-lasting solution to data protection and disaster recovery.

Most public cloud service providers offer 99.95% or above on the overall availability of the services. Under exceptional circumstances, if such occasional outages are beyond the control and extends the SLA thresholds, the service provider usually offers credit — which is not likely to equate to the outage's actual cost. Most importantly, they don't offer terms that can guarantee or protect against customer data loss. You can consider hosting and replicating the data to multiple providers and locations in case of extreme availability requirements. However, it is also dependent on the technology and services you have chosen for your D&A deployment. It may not be feasible to bring them up quickly in other service provider platforms.

Workload Management

Managing and executing your analytical workload must be a critical part of your deployment. As any typical D&A platform consists of multiple components and services in the cloud, it can be challenging to determine these workloads' ideal placement. Some of the factors that can typically influence the workload placement are:

- **Deployment model:** The use of a single service provider versus using a hybrid or multicloud deployment can potentially introduce technical dependencies, network latency and other operational thresholds.
- **Manageability:** Managing deployment across multicloud and using external services such as cloud service technical support, consulting services and third-party managed services introduce challenges to the operational processes, governance and risk management.
- **Features and functionalities:** Understanding the differences between the niche provider's capabilities versus cloud service providers for specific core analytical components. As each product is unique with differentiated capabilities, it presumes that some providers may have more advanced, mature and proven offerings that may best suit your deployment needs.

- **Data gravity:** The degree of data concentration in a limited number of cloud providers. The greater the number of providers, the greater the intercloud data exchange and data integration challenges.

- **TCO:** The ability to maximize the cost-efficiency of a given workload. The costs involve not just the service provider's pricing model, but include other factors such as volume, location, characteristics that determine the total cost and overall manageability of the workloads.

Exit Strategy

Cloud providers are not entirely interchangeable. You can select technologies and design solutions with portability being the highest priority, but there will continue to be significant differences between cloud providers.

Every solution you build in a provider's cloud will have specific dependencies on that provider's APIs, services and tooling. You must consider the migration of all of these dependencies when moving to an alternative cloud provider.

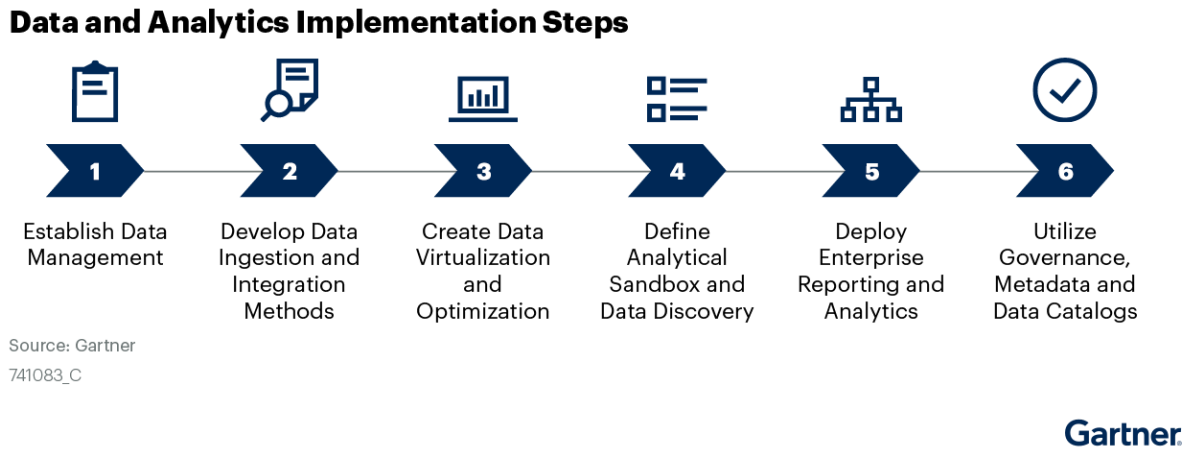
As part of solution planning, estimate what a migration to a new provider involves, and include that as part of your exit strategy.

Step 4: Identify Implementation Steps

As we review the various data and analytics components and other factors critical for the migration of your D&A platform, it is vital to be selective in identifying the steps that affect the migration as migrations can be costly to do everything.

The decision in scope is an issue of migration and one of optimizing your D&A deployment (see Figure 9).

Figure 9: Data and Analytics Implementation Steps



Establish Data Management

Data management is the core foundation of the D&A platform. There is an increased need for organizations moving to the cloud to support both on-premises and cloud data management platforms.

Hybrid and multicloud deployments and the use of database services, run in containers-based architectures, have become more relevant.

Traditional databases in the on-premises can be migrated to PaaS in the cloud delivering serverless and autonomous capabilities. Cloud also provides multiple storage options to drive agility, optimization and maximize the price for performance.

Cloud and other third-party services offer multiple ways for data migration — from bulk data migration to cloud object storage to data transfer using database and schema migration services for heterogeneous database transfers.

Cloud data warehouses support denormalization as complex joins across distributed data, and columnar storage in the cloud offers better performance and scalability. So, some data models and design changes are to be considered part of the cloud migration.

As part of establishing your data management in the cloud, it is critical to review the following:

- Explore the various methods of migrating your existing data into the new data management platform.

- Cloud offers multiple methods and options to store and organize data.
- Data lakes and data warehouse storage provides new, modern and overlapping capabilities.
- Diverse storage options in the cloud accelerate user case development and user adoption.
- Migration of your data store from one engine to another engine will require conversion of schema, data types, data mapping rules and code conversion.

Develop Data Ingestion and Integration Methods

Developing the data ingestion is the foundation for data extraction from source data systems and orchestration of different data management integration methods.

Cloud provides multiple data delivery methods, such as batch and bulk data movement, stream data delivery, change data capture, data replication, synchronization, message-oriented data movement, etc. As part of the cloud migration, it is essential to assess the different methods and determine your deployment's most suitable option.

The selection of the ingestion framework and data ingestion tools will determine the sources and different types of ingestion methods supported to meet your service-level agreement. Most organizations typically adopt multiple ingestion methods.

Cloud data ingestion and data integration tools provide a low-code and no-code method for automating data flow and open-source options with batch processing built on the premise of ELT than ETL. Also, cloud data integration works much differently than on-premises data ingestion tools, with the bulk of the transformation occuring at the data layer. So, the migration development will be a complete overhaul and architecting the entire pipelines instead of a lift-and-shift migration approach.

Developing data ingestion and data integration requires a framework and implementation of best practices and includes:

- Develop future-proof integration using cloud PaaS tools to add new sources and data formats, and adapt to changing enterprise.
- Integrate with the data catalog to support data lineage, data discovery and governance.
- Detect and handle data drift and schema drift.

- Provide complete automation and orchestration of data ingestion workflows and scheduling of workloads.
- Scale on demand to meet the increased data movement and provide better traceability on the data and integration routines.

Create Data Virtualization and Optimization

Data virtualization provides new ways to harness the data across multiple data sources and cloud, hybrid, and multicloud environments. It is also a great alternative and yet another option to establish a unified semantic layer by pushing down to the data management layer.

Using data virtualization, the users get the flexibility to access a single view of data across the enterprise. By embracing data virtualization, organizations don't have to take a big-bang approach with their cloud data migration and can incrementally build their new data management platform in the cloud.

Creating data virtualization also allows businesses to tap into new data sources and existing on-premises data warehouses for historical data, or other data usually brought in subsequent migration phases.

While data virtualization accelerates the data delivery to the business users, they should not be considered a replacement for the data lakes and data warehouses. Data virtualization is an excellent substitute for a hybrid data lake and data warehouse approach. It offers performance optimization using the data query and data persistence engines, and various connectors to deliver faster analytics to the enterprise.

Some of the factors that can influence the creation of data virtualization include:

- The ability to support integrated data analytics with data dispersed across multiple data stores, or between on-premises and cloud data platforms.
- Drive agility by delivering faster access to the data by connecting to different sources than collecting data into data lakes and data warehouses.
- Provide the ability to do data discovery upfront, and determine the data that is genuinely required for analytical purposes eliminating data swamps and unwanted data in the analytical data stores.

Also, new trends are emerging on data virtualization and optimization that support augmented data management and integration capabilities using integrated data pipelines, active metadata, knowledge graphs, and the use of artificial intelligence (AI) and machine learning (ML) capabilities using data fabric (see Demystifying the Data Fabric).

Define Analytical Sandbox and Data Discovery

Cloud D&A deployment breaks away from traditional ways of implementing analytics to adapt to the changing business, eliminating the poor practice of building solutions and supporting faster ways to deliver incremental value over time. It is about giving the business users the opportunities to be more involved and take few things into their own hands to contribute to successful deployment in the cloud.

While it is critical to managing controls to a centralized D&A process, it is also equally crucial for specific organizations to drive agile analytics by empowering the business users to access through an analytical sandbox quickly. The analytical sandbox is built contextually for particular business users to make data discovery and develop prototyping solutions that can serve as an input to the centralized IT team for deploying into production to help broader audiences.

The use of analytical sandbox for data discovery:

- Provides a framework for agile development and modular data architecture with higher interoperability.
- Enables data stewards and analysts to explore and experiment with new use cases development.
- Empowers business users to bring their data into addition to tapping into existing data assets in the enterprise for experimentation and study data patterns and anomalies.
- Saves time and cost overall in production deployment as the business has the opportunity to vet through the data to determine the value in the sandbox before onboarding them into the centralized data platform.

Deploy Enterprise Reporting and Augmented Analytics

Delivering analytics has become an interdisciplinary approach that uses the concept of traditional reporting, data analysis, statistics, machine learning and other related technologies that can provide better insights with predictive and prescriptive outcomes.

Analytics and business intelligence (A&BI) is evolving and extended beyond the standard reporting, data visualization and interactive dashboards. Traditionally, in an on-premises environment, organizations involve IT to set up an enterprise standard analytics tool to deliver to their requirements. As more capabilities add to providing analytics with data and reporting distributed across the enterprise, the use of multiple tools across the different business domains or between on-premises and cloud deployments has become more prevalent. The democratization of data in the cloud has evolved these A&BI platforms to deliver more self-service, augmented and advanced analytics capabilities.

As you formalize and consolidate your use cases as part of the cloud D&A deployment, it is essential to access, rationalize and combine them. You may choose different tools, as no single analytical tool or vendor platform delivers a comprehensive analytical solution for all use cases.

To learn more about the different approaches and how to deploy your A&BI tools best, see Reference Architecture to Enable Real-Time Self-Service Analytics.

Organizations migrating their D&A platform to the cloud have to look beyond the traditional reporting platform and evolve into augmented and advanced analytics capabilities. These include self-service data preparation, advanced visualizations, augmented data discovery and use of natural language query (NLQ), natural language processing (NLP) with advanced search capabilities, and the use of artificial intelligence (AI) and machine learning (ML) integrated capabilities within the analytics platform.

To further understand the different analytics domains and a detailed assessment of their capabilities, review the following:

Self-Service Analytics Success With Metrics, Data Literacy and Organizational Models

Demystifying the Analytics and BI Space

Demystifying the Data Science Lab and AI Hub for Business Analytics

Utilize Governance, Metadata and Data Catalogs

As organizations migrate their data and analytics platform to the cloud, it is imperative to establish a structure to govern their data assets and the process that can promote trust and improve transparency and operational efficiencies.

As technical professionals move their data and analytics platforms to the cloud and how they operate these systems change drastically with the cloud, it is crucial to have a robust process to track, measure, monitor and report to ensure that the system provides the value for their investments. With the ability to leverage multiple services spontaneously in the cloud, more robust governance and controls are necessary (see Three Approaches to Support Data and Analytics Governance).

It is even more critical for organizations migrating their data and analytics platforms to the cloud to have concrete governance at the program and at the domain levels to collect and manage their data assets, measure the analytics usage, and map to their stakeholder's value.

Data catalogs have become increasingly important for analytical use cases, and for supporting a robust data stewardship and governance in the organization. They offer greater transparency and a shared understanding of data, especially when your implementation spans across hybrid and multicloud data management scenarios. The data catalog provides more significant insights on everyone's data availability — from business users to data analysts, data engineers and data scientists to help deliver informed decisions faster. Additionally, it helps with the organization of distributed datasets. It maintains an inventory of data assets for ease of understanding and access to extract business insights. Cloud technologies allow modernization with augmented AI and ML tools to automate various tedious tasks involved in data cataloging — including metadata discovery, ingestion, translation, enrichment and the creation of semantic relationships between metadata.

Enterprise metadata management (EMM) is about delivering shared governance of metadata between use cases and implementations. EMM process tightly integrates into multiple data and analytics practices, such as data integration, metadata ingestion, defining and building relationships between metadata, business glossaries, and documenting standards.

Therefore, these next-generation data catalogs can propel EMM projects by allowing business users to participate in understanding, enriching and using metadata to inform and further their data and analytics initiatives. With proper metadata management and automation, and interoperability capabilities, EMM plays a crucial role in collecting, analyzing and sharing metadata from the overall data management landscape and turning this metadata into actions.

Step 5: Support Post Launch

One of the final steps in the D&A deployment to the cloud is understanding the overall management process of maintaining oversight and administrative control of various products and services components. These will be implemented as part of your analytics platform — whether in public, private or hybrid cloud environments.

While cloud deployment offers excellent potential for eliminating the similar IT responsibilities that you typically manage in the on-premises environment, it is still necessary to manage and control the cloud resources and services for efficiencies and optimize your D&A deployment.

Some common themes that you can adapt to manage your cloud D&A deployment effectively include:

- Use of self-service to access cloud resources, create, monitor usage and cost, and adjust resource allocations as necessary.
- Automate your workflow from the platform level, such as managing cloud instances, migration of software code, workloads and other artifacts at each component level to minimize or eliminate human interventions.
- Able to monitor and track cloud resources, workloads, service-level engagement, performance and overall user experiences.

Administration and Support

In a traditional on-premises environment, IT is responsible for managing the underlying infrastructure and management of tools. They must have built several processes, procedures and tools to administer and support the D&A platform. While many of these will continue to be useful based on the deployment architecture selected, many of these have to remain on-premises and may not be compatible with the cloud solutions. Also, cloud deployments are a shared responsibility between the customer and the cloud service provider, and the level of commitment determines the deployment model chosen.

Some of the common administrative steps involved in cloud deployment include:

- Migration planning and execution.
- Provisioning and automating the various technical and cost management controls.
- Monitoring and approving capacity.
- Implementing cost optimization.

- Pushing approved changes on different technology layers to your multiple D&A deployment environments.

The service components for each of the D&A products provide an application-level administration console based on their core characteristics for data storage provisioning, cluster management, audit logging, controlling object privileges, user authentication and policy management, query optimization, monitoring, alerts, and overall usage of the services.

Organizational Structure and Processes

Migrating your D&A to the cloud brings changes to your organization's structure and processes. While the fundamental goals don't change, they provide a unique opportunity to integrate flexibility and understand how the cloud technology is delivered, managed and deployed, rather than about the technology itself.

Every organization has a set of goals and objectives defined as part of its broader analytics program. Each set of goals has a different set of priorities that may lead to various deployment models. Analytics deployments are so unique and built around the centralized IT strategy with a strong emphasis on governance and consistency across the organization. It also needs to support different department-level analytics across business units and geographies that require their own set of deployment models. The idea is to deliver consistency, consensus and shared practices with flexibility, autonomy, innovation and an organizational model that provides both sets of values.

Organizations must evaluate multiple approaches (see Create a Hybrid Centralized and Decentralized Data and Analytics Organizational Model) and define their most critical goals and analytics needs, as no organization set structures and processes can be perfect.

Change Management

Organizations accustomed to traditional waterfall methodologies and long release cycles can have challenges or have to pivot their implementation approach when migrating to the cloud. One of the main drivers to migrate to the cloud is to change the way that the applications manage an agile delivery model, automation and autoprovisioning of services and resources, based on the demand.

A reliable change management process enables the delivery of business value while balancing any associated risks. It should maximize productivity and reduce cost and effort for all participants in the process. Automation, integration and deployment tools in the cloud allow businesses to make small, frequent changes that reduce risk and deliver business value at an increased rate.

A change process is about developing and deploying new services, configuration changes, patching, and software updates and implementation as automated.

Some of the common process for the change management in the cloud deployment involves:

- Allowing preapproved changes for autoscaling, network updates, storage, etc.
- Automatic configuration and custom setting to the services, tracking changes and recording them across the analytics deployment.
- Establishing security groups, firewalls and monitoring the services for any unauthorized changes and intrusions.

Other standard practices are to take advantage of configuration management, infrastructure as code, automated testing and validation, and continuous integration and delivery. They can implement lightweight approval processes tightly integrated into the overall development process.

The change management goal does not intend to minimize the business risk; the process should ensure that overall business risk is optimized.

Evidence

[1] Gartner Cloud Decisions

Recommended by the Author

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

Solution Path for Modernizing Analytic Architectures

Solution Path for Building a Holistic Data Management and Analytics Architecture

Reference Architecture to Enable Real-Time Self-Service Analytics

Self-Service Analytics Success With Metrics, Data Literacy and Organizational Models

Evolving the Capabilities of Analytics and Business Intelligence Platforms

Assessing Semantic Layers to Achieve Successful Self-Service Analytics

Data and Analytics Strategies Primer for 2021

A Guidance Framework for Selecting Cloud Management Platforms and Tools

---