Looking Before You Leap Into the Cloud

A PLANNED APPROACH TO CLOUD IMPLEMENTATIONS PREVENTS PAIN, REWORK, AND FRUSTRATION

ENTERPRISE MANAGEMENT ASSOCIATES® (EMA™) Infobrief
Written by John Myers

March 2018

EMA™

IT & DATA MANAGEMENT RESEARCH • INDUSTRY ANALYSIS • CONSULTING

PREPARED FOR: cloudera® Microsoft

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In the era of data-driven cultures, the transformational impact of analytics initiatives grows more each year. Traditional analytics environments, such as dashboarding and reporting, expand from business analysts to include frontline employees in operations, the warehouse, and the point of sale. Advanced predictive and prescriptive analytics initiatives spread from marketing, to finance and customer care, to IoT real-world environments. Streaming data spurs the development and deployment of real-time analytics applications leveraging machine learning (ML) and artificial intelligence (AI) initiatives. These advancements in data and analytics are helping large enterprises reinvent themselves.

In each of these transformational areas, analytical initiative deployments benefit from cloud-based deployments. Speed of provisioning and pay-as-you-go (operational expense: OPEX) models make cloud solutions very attractive. Some organizations will jump into a "swipe and go" cloud implementation that focuses more on short-term convenience and speed of deployment. This is a short-sighted approach.

Modern data-driven organizations require more from their implementations. They need corporate-wide data management and visibility. They need speed and flexibility in their options for the configuration, processing, and deployment of complex, multi-disciplinary analytics use cases. They need solutions that benefit everyone and facilitate safe self-service and collaboration. A coordinated data environment strategy that considers private and public cloud resources and provides a standardized approach across all of a company's data assets and analytics efforts will make a real difference to an organization's success.

Risk in Choice of Cloud

By jumping into a cloud deployment with a "swipe and go" mentality, organizations face implementation and operational risks that far outweigh the benefits. These risks include:

- Corporate data assets stagnating in isolation and hard to find for analysis or audits
- Lack of visibility and inconsistent controls across data assets spread through multiple cloud vendors
- Issues with data movement between disparate services, even within single providers when coordination is required

IN THE KNOW

WHO: Executives, architects, and business stakeholders from data-driven organizations.

WHEN: Making decisions on cloud-based implementations associated with big data analytics and machine learning.

WHAT: Guidance on strategies and choices that prevent vendor lock-in and rework for CIO and data architects, as well as a lack of visibility for the CDO and data stewards.

Coordinated Strategy is Key

With a sensibly considered coordinated data environment strategy, organizations enjoy the benefits of hybrid and multi-cloud deployments without the risks. An organization's analyst community, including business analysts, data science teams, and data engineers, speeds the development and deployment of analytical initiatives including traditional, predictive, and ML analytical applications. With key data assets under management and accessed within a coordinated environment, this strategy provides:

- A single shared storage layer
- The availability of multiple processing engines to reduce replication, or siloing
- Access to a choice of tools and toolsets that enable analytical iteration and creativity
- Enterprise-wide visibility and control for data management, curation, security, privacy, and governance

DRIVING TOWARD ANALYTIC INITIATIVES IN THE CLOUD

As the rate of change in analytics initiatives expands, analyst communities within organizations need flexible and efficient ways to test, validate, deploy, and operationalize their deployments. Cloud implementations provide a better way to enable these analytics deployments, including powerful dashboards, predictive and prescriptive analytics, and ML algorithms over traditional practices. The cloud offers an increased speed of deployment. It also offers improved self-service, lower infrastructure administration, and scalability opportunities to push the advancement of analytics in ways difficult to match with traditional methods and in on-premises data centers.

Speed of Implementation

It is no secret that cloud-based deployments have a quicker provisioning speed than traditional, on-premises implementations based on bare metal installation and configuration. Cloud solutions help organizations spin-up sandbox environments quickly to discover or validate data-driven scenarios in hours, with a very low level of administration effort. This is opposed to the timeframe of weeks or months for a traditional deployment.

Bring Your Own Tools

Data-driven analyst communities like to have a wide range of tools at their disposal. Business analysts prefer their "own" visualization or exploration tools based on look and feel. Data scientists develop favored algorithms and techniques to find the best solutions. Analytical application architects draw upon chosen development tools to facilitate performance.

Cloud providers allow organizations to bring their preferred choice of software and tools to their infrastructure. Many of these providers will include a wide variety of these preferred tools in their prepackaged services. Along these lines, some of the providers of these preferred software tools have designed and optimized their solutions for cloud environments.

BIG DATA CLOUD MYTH #1

MISCONCEPTION: Cloud isn't built for "big data."

TRUTH: Cloud-based deployments of big data can be challenging due to technical constraints, such as data gravity or corporate culture issues relating to security and privacy concerns. However, 67.5 percent of big data projects in EMA end-user research were cloud-based deployments. Nearly 76 percent of respondents were implementing cloud-based strategies within their organization.

Self-Service Opportunities

Along with the self-selection of tools for analytical initiatives, an organization's analyst community desires self-service. Self-service further removes friction from the analysis and processing of data assets. Waiting for IT teams to approve, provision, and deploy those tools and environments delays and slows the analytical process.

IT-provided solutions stifle this analytical process with the traditional flow requests and approvals, procurement of software and hardware, and deployment delays. Self-service allows for the decoupling of timing and location from traditional IT processes. Business analysts and data scientists can take up analytic initiatives whenever the opportunity presents itself. Frontline employees can access analysis and key data assets wherever it suits the business. However, the key to enabling safe self-service is having a unified catalog defining appropriate business context of data, backed by a common security framework—both of which most cloud services lack.

Lower Administration and Scalability

Reducing the level of effort to support analytical initiatives is vital for data-driven organizations. Leaving behind aspects of the system and platform deployment and maintenance duties allows data engineering and administration teams to scale better. Freed from low-level activities, IT teams become strategic providers of new data sets and assets, and can offer guidance to improve productivity and drive results.

Cloud-based deployments give IT teams the ability migrate quickly from test and validation sandboxes to production environments. Scaling without having to rebuild or add additional physical infrastructure pays benefits in speed, quality, and business outcomes.

BLINDLY JUMPING INTO CLOUD HAS ITS DRAWBACKS

While the cloud presents great opportunities, it does have risks. Jumping into a cloud deployment without properly understanding how a cloud vendor fits into a greater corporate strategy and data management program means organizations face significant operational risks.

Business teams desperate for resolutions will often select a cloud solution, type in their credit card number, and start a cloud project, effectively performing uncontrolled shadow IT. This impulse to "swipe and go" creates dead-end deployments, lacks visibility to coordinate data assets, and increases the risks of data replication.

Limitations of a Bad Choice

Many organizations believe getting data out of a cloud solution is as easy as putting their data in. However, for many cloud providers, once data is uploaded and manipulated, the data is often difficult to extract and move to another location. These exit barriers are called "vendor lock-in."

The first vendor lock-in issue relates to the storage and/or schema of the information after it is imported into a cloud solution. Cloud providers manipulate data and place it in disparate storage formats, which can be variations of open-source standards or proprietary object structures, or combinations of both. These storage and file formats can be incompatible between cloud providers, or sometimes even between a single cloud provider's different services. For example, separating a cloud provider's offerings for data warehousing and data engineering may require conversion and/or transformation to move data between those offerings, often at additional cost. The schemas are designed and configured to support the cloud vendor's operational requirements rather than transportability. Such an approach leads to undocumented (to the customers) data schemas and processing workarounds. While it encapsulates the underlying complexity of data operations, this lock-in barrier prevents data removal later.

The other vendor lock-in issue is the accumulated weight of data once it exists within a cloud provider. Even if the data assets are formatted and available for export, the time and effort required to remove the information is prohibitive. This is called "data gravity." Compounding the issue of time to export data is that cloud providers often will charge fees for the migration of data in or out of their solution, which makes the removal of data time-consuming due to gravity and expensive due to fees.

BIG DATA CLOUD MYTH #2

MISCONCEPTION: I can just "swipe and go" for a cloud implementation.

TRUTH: "Swipe and go" cloud selections provide instant gratification for business teams frustrated with IT departments. However, by taking this approach, data governance and management issues arise that cause more pain in the long run than the short-term benefits they provide.

Lack of Data Coordination

With "swipe and go" deployments, data resides outside the normal chain of command of the Chief Data Officer (CDO), data steward, or data architect. This lack of coordination, security, privacy, and governance becomes an issue. Members of an organization's analyst community require a full understanding of corporate data assets to effectively choose current and relevant sources. Data scientists need a complete inventory to build accurate predictive and prescriptive models. Business analysts need to combine information from disparate departments to find business opportunities. App developers desire a single point of interface to include contextual data in their analytical applications. This makes it especially hard to build a multi-disciplinary analytics application.

At best, without the proper visibility, data siloes are created and the information within is either duplicated or undocumented with the wider data management strategy. At worst, unaccounted data creates risks with internal corporate compliance standards or government regulatory requirements. An example of the risks associated with replicated customer information are the penalties connected with violations of European Union data protection (GDPR) regulations.

Opportunity Cost of Data Movement

Even in the event that the data is properly inventoried, silos create problems with replication and data access. Anyone in the analyst community who wants access will experience increased costs, which are based on the time required to request data access and effort to replicate data between silos. The delays add time to the analytic processes of data scientists, business analysts, and frontline employees.

THE KEY TO CLOUD SUCCESS IS A PLANNED APPROACH

A judiciously considered coordinated data environment strategy to implementation provides a range of advantages over other deployment options. A coordinated data environment strategy loses the provisioning and deployment headaches of managing traditional infrastructure, but actually magnifies their data management and governance problems. Such an approach also forsakes the instant gratification of "swipe and go" deployments, but still has the speed and flexibility of cloud implementations and can offer a more curated self-service experience. With all of an organization's data within a coordinated data environment, community members speed the development and implementation of an organization's multi-functional analytical initiatives. This data environment strategy features a shared storage layer, consistent management capabilities, and comprehensive security, and maintains choices of implementation options and cloud providers.

Coordinated Data Environment

With all of an organization's data within a coordinated data environment, a shared data storage layer is created. This shared storage layer allows multiple processing engines to have access to data without replication. With a single point of access based on a holistic catalog and standardized policies for these processing engines, exploration, analytical, and machine learning workloads can be accomplished in a single, consolidated location. This allows for reporting and dashboard presentation, predictive modeling and workloads, and ML algorithms to all utilize a single set of corporate data assets instead of a collection of silos or replica of each workload.

In addition to a consolidated data layer for workloads, data access and security is configured, monitored, and managed from the same data layer. Data access and security teams can focus on a single location, and if there is a breach of security or data, access to the assessment of the incident(s) is confined to that location. Ideally, security controls are implemented consistently and comprehensively everywhere using the same tools.

BIG DATA ANALYTICS MYTH #3

MISCONCEPTION: Once I have my data in the cloud, I am "good."

TRUTH: "Swipe and go" cloud implementations can provide instant value for organizations. However, long-term "swipe and go" has limited value for deployments that need to have enterprise attributes such as data governance, low numbers of replicas, and company-wide access.

Choices on Implementation Strategy

With a coordinated data environment, organizations can make decisions for the technical implementation of data based on their strategic business requirements. Organizations with a cloud-first implementation strategy can use public cloud resources and migrate on their own timeline, knowing they have the same platform capabilities everywhere. Companies who decide to maintain their own data centers for security or privacy can utilize the consolidated data lake approach in a bare metal installation or a private cloud implementation, with many of the provisioning and administration advantages of public cloud. Enterprises that want to bridge the two worlds can utilize a hybrid architecture with part of their data assets within their data center or private cloud, and other components within the public cloud.

Choice of Cloud Providers

In addition to being able to spread this coordinated data layer across multiple technical implementation choices, organizations can also select solutions and services from a number of cloud providers to deploy their consolidated data layer. Some cloud vendors deliver a quick and easy option to deploy. They represent an excellent avenue for analytical initiatives without enterprise requirements for uptime, availability, and security. Each of these providers will have their own strengths in terms of pricing. This creates opportunities for mixed, or hybrid, cloud environments, especially for bursting and transient workloads. The choice of cloud provider infrastructure can be enabled by having the same analytics platform run in multiple clouds.

SPEED AND FLEXIBILITY FOR BUSINESS STAKEHOLDERS

The stakeholders associated with an organization's analyst community are diverse. Data scientists, business analysts, and frontline data consumers all have specific requirements for their analytical initiatives. Self-service aspects of cloud deployments, the application of multiple processing engines on a single consolidated data layer, and the multitude of analytical toolsets available across a modern technology base all stand to benefit from a coordinated data environment.

Advantage of Self-Service

One of the key advantages of an analytics initiative deployed through a coordinated data environment is the availability of self-service facilities. The ability to access, configure, and utilize analytics with nothing more than a web browser empowers internal resources across the organizational chart and provides access to external consumers such as customers, partners, and suppliers. The availability of this standardized security and governance layer across environments, services, and data consumers enables a consistent and reliable self-service experience.

Without the barrier of a particular physical location (laptop or server) or analytical environment (data warehouse, data mart, or discovery repository) to run analytical applications, analyst community members connect with data no matter where they are located around the globe. Customers interact with their accounts, shipments, and payments without depending on members of the customer service team. Partners and suppliers monitor and manage their relationships with engagement or delay from operations and supply chain. Data scientists can train on much bigger data sets than a download to a laptop would allow, increasing both accuracy and security.

BIG DATA CLOUD MYTH #4

MISCONCEPTION: Business stakeholders don't care about the "hows" of an analytics implementation.

TRUTH: Business stakeholders are concerned with the availability and stability of business outcomes. If those platforms work, then business stakeholders don't care about the "hows" of deployment. However, when the applications are unavailable or unstable, they become very interested in the "hows" of deployments that are preventing them from success.

Supports Multiple Workloads

One of the main considerations for a coordinated data environment strategy is supporting multiple processing engines and workloads. Many business analysts will utilize different workloads and thus different processing engines as part of their validation of an analytical application. Starting with exploration and discovery, the ability to search, profile, and perform light statistical validation on new data sets gives business analysts the ability to understand and evaluate new data assets. Business analysts also analyze how new and existing data sets are related and correlated, without the rigid limitations and cost of a traditional data warehouse. Finally, these business analysts validate their assumptions with a particular data asset by simulating or trial-running operational workloads.

A single coordinated data environment capable of running all those disparate workloads allows analyst community members to move seamlessly from one type of analysis to another. All of this takes place with that single consolidated layer and the appropriate guardrails.

Supports a Wide Range of Tools

Just as a coordinated data environment brings multiple workloads to the business analyst, it also empowers the data scientist with multiple analytical toolsets they can use when developing and validating their predictive, descriptive, and ML models.

In the past, the technical limitations would restrict data scientists to a single toolset. Now, they work with multiple tools to develop the most effective model or use multiple models together to find the best combined model.

COORDINATION AND FRICTIONLESS DEPLOYMENT FOR TECHNOLOGISTS

Data-driven organizations are seeking new insights and information that will give then be difficult for traditional data management options to store and access, let alone re Supporting the requirements for the various levels of an organization's analyst community is a daunting task. CIOs, IT departments, and data architects all feel the impacts of analytical initiative growth. However, coordinated data environment deployments, such analytical repositories, data lakes, and data warehouses, give these corporate technologists the tools they need. Corporate technologists have the ability to support more projects in less time because they are focused on the strategic implementation of analytical initiatives instead of the tactical provisioning and maintenance activities.

Reduced Friction

The appeal of the "swipe and go" approach is that organizations can quickly provision an analytical application. Short-term, this approach seems attractive. However, in the long term, organizations need to be more mindful of how operational friction in the form of management, replication, and visibility impacts the IT organization.

Coordinated data environment deployments provide the ability to reduce, if not eliminate, iterative friction tasks that stand between the analyst community and the corporate data assets that make their jobs possible. This friction comes from the centralized management and governance of data within an uncoordinated, unintegrated, complex environment. Inventory, data movement, and overall governance tasks are simplified with a consolidated data layer and shared data context. Corporate technologists work with a more strategic allocation of resources and headcount to advise and support a company's goals, not just wrangle infrastructure and administration.

Multitenancy

Traditional implementation approaches, and to a certain extent "swipe and go" deployments, are one-off implementations. Over time, organizations develop a hodgepodge of different implementations, which leads to extreme inefficiencies for deployments that tend to be singular in nature and long-term management aspects of supporting and maintaining these environments.

BIG DATA ANALYTICS MYTH #5

MISCONCEPTION: There is no administration in the cloud because they handle it for us.

TRUTH: Many cloud vendors will handle the administration of your data and data assets in the cloud. However, many of these run the risk of vendor lock-in with proprietary data schemas and formats. Other cloud providers with a more DIY approach to data management will handle the systems and platform administration for your deployment, but leave the DBA and data modeling work for your team to handle.

A coordinated data environment strategy utilizes the deployment methodology of multitenancy. Multitenancy provides additional economies of operational scale by using cloud resources to speed deployment. While this approach takes more coordination and infrastructure at the start, multitenancy enables the configuration, support, and update of those analytics services from a single point of management.

Enterprise Security, Privacy, and Governance

A single point of data storage, management, and governance allows the CDO and the CSO to strategically monitor and manage an organization's data assets. Such a data storage layer provides economies of scale for data governance and overall data security and privacy. For data governance, the ability to have visibility and inventory of a company's complete data assets lets the CDO curate and govern them in accordance with corporate standards for quality, consistency, informational tagging, and linking across datasets. The CSO can view this same information and understand which data assets require additional protection in terms of personally identifiable information and payment compliance, or how members of the organization access the information.

With these two concepts together, the CDO and CSO can also achieve wider corporate strategic goals for customer privacy to support regulations like GDPR.

ABOUT OUR SPONSORS

cloudera®

Cloudera, Inc. offers a data management, machine learning, and analytics software platform worldwide. The company's Cloudera Enterprise platform delivers an integrated suite of capabilities for data management, machine learning, and analytics to customers for transforming their businesses, optimized for cloud. The company serves organizations like banks, technology companies, telecommunications, and healthcare and life sciences through its direct sales force. Cloudera, Inc. has a strategic partnership with Intel Corporation. The company was founded in 2008 and is headquartered in Palo Alto, California.

Cloudera Altus

Cloudera Altus is a cloud service platform that enables organizations to use Cloudera Enterprise software to analyze and process data at scale within a public cloud infrastructure. It is designed to provision clusters quickly and make it easy for users to build and run data workloads in the cloud. Altus works within the cloud service provider architecture and creates clusters using flexible compute resources. Altus jobs read input from and write output to Microsoft ADLS. Altus provides a data engineering service that creates clusters and runs jobs specifically for data science and engineering workloads, including batch processing jobs. Altus is also expanding to offer data warehouse (Cloudera's Analytic DB) and a data science platform, supported by Cloudera SDX.

Cloudera SDX

Cloudera SDX (shared data experience) is a powerful software framework that makes multifunction data applications easier to develop, quicker to deploy, more cost-effective, and more secure. By applying stateful, centralized, consistent data context services that reside with the persistent object storage, not the transient compute nodes, SDX enables hundreds of different workloads to run against shared or overlapping sets of data. Catalog, security, and governance services are central to solving the problems of cloud-based analytics.

Microsoft Corporation develops, licenses, and supports software products, services, and devices worldwide. Its intelligent cloud segment licenses server products and cloud services, such as Microsoft SQL Server, Windows Server, Visual Studio, System Center, and related CALs, as well as Azure, a cloud platform with computing, networking, storage, database, and management services. The company markets and distributes its products through original equipment manufacturers, distributors, and resellers, as well as through online and Microsoft retail stores. The company was founded in 1975 and is headquartered in Redmond, Washington.

Azure Data Lake Service

Azure Data Lake Store (ADLS) is an enterprise, large-scale repository for big data analytic workloads. Azure ADLS lets users capture data of any size, type, and ingestion speed in one single place for operational and exploratory analytics. It is specifically designed to enable analytics on the stored data and is tuned for performance for data analytics scenarios. ADLS includes all the enterprise-grade capabilities, such as security, manageability, scalability, reliability, and availability, required for enterprise use cases. Analysts can easily evaluate data stored in the Data Lake Store using Hadoop analytic frameworks, such as MapReduce or Hive. Hadoop clusters can be provisioned and configured to directly access data stored in the Data Lake Store. The Azure Data Lake Store provides industry-standard availability and reliability. Data assets are stored durably by making redundant copies to guard against any unexpected failures. Enterprises can use the Azure Data Lake in their solutions as an important part of their existing data platform.

About Enterprise Management Associates, Inc.

Founded in 1996, Enterprise Management Associates (EMA) is a leading industry analyst firm that provides deep insight across the full spectrum of IT and data management technologies. EMA analysts leverage a unique combination of practical experience, insight into industry best practices, and in-depth knowledge of current and planned vendor solutions to help EMA's clients achieve their goals. Learn more about EMA research, analysis, and consulting services for enterprise line of business users, IT professionals, and IT vendors at www.enterprisemanagement.com or blogs.enterprisemanagement.com. You can also follow EMA on Twitter, Facebook, or LinkedIn.

Corporate Headquarters:
1995 North 57th Court, Suite 120
Boulder, CO 80301
Phone: +1 303.543.9500
Fax: +1 303.543.7687
www.enterprisemanagement.com

3690.032118