WHITE PAPER

12 Requirements for a Modern Data Architecture in a Hybrid Cloud World

Why a Successful Hybrid Cloud Strategy Requires an Enterprise Data Strategy

HYBRID
CLOUD
WORLD

Table of Contents

A Modern Data Architecture

- Requires data-at-rest and data-in-motion
- Manages the entire lifecycle of data
- Spans across on premises, cloud and multi-cloud
- Processes and drives insight
- Requires consistent security, governance and operations

A **hybrid architecture** is a key requirement of a modern data architecture, and is composed of **on-prem + multi-cloud + edge**

Introduction

Businesses today are moving to the cloud at unprecedented speeds. Today, cloud seems to be the quintessential way to do business—and for good reason. It provides access to an unlimited pool of compute and storage resources, agility in provisioning them, ease of operations, cost efficiencies, and more. Nine out of ten companies will have some part of their applications or infrastructure in the cloud by 2019, and the rest expect to follow by 2021, according to IDG's 2018 cloud computing research[1] .

But moving to the cloud is not an all-or-nothing strategy. In fact, most organizations will settle on a hybrid cloud strategy, deploying applications, data, and infrastructure on a combination of on-premises and cloud resources. Hybrid cloud brings substantial advantages in terms of rapid deployment and reduced infrastructure costs, but it also comes with a new set of IT and data management challenges—namely, managing data-intensive applications as well as storing, processing, securing, and governing data when it's distributed across on-premises, multiple cloud, and edge environments. A modern data architecture designed for the hybrid cloud solves these challenges and helps deliver ongoing business value.

This paper is designed to help you make the right decisions for delivering a modern data architecture for hybrid cloud environments—an architecture that uses open source software to protect against cloud vendor lock-in and that enables you to manage, secure, and govern your data across multi-cloud and on- premise environments.

You are likely already charting the course for how to effectively lead your organization to a managed, secured, and governed hybrid cloud. But, with a massive and complex ecosystem of disparate legacy systems running on premise, this is no small feat. So, we'll begin by reviewing strategic and business considerations for success in a hybrid cloud world and follow this up with 12 technical requirements for building an enterprise data strategy for the hybrid cloud.

Strategic & Business Considerations for Success in Hybrid Cloud

Most enterprises have existing on-premises technology investments that need to work seamlessly with your new investments in the cloud. It takes time and significant financial investment to migrate legacy IT systems; therefore, you need to have a consistent architecture that can support both on-premises and cloud applications in the hybrid cloud.

Given the global hybrid cloud market is projected to more than double from $44.6 billion worldwide in 2018 to $97.64 billion by 2023[2], your enterprise data strategy must aim to establish a modern data architecture with hybrid architecture as a key requirement. To achieve this, you need to deploy an enterprise data platform that complements your cloud provider's infrastructure with a layer of consistent data services while giving you the freedom to move data, metadata, and workloads across hybrid cloud environments.

In summary, your enterprise data strategy must advocate for a modern data architecture that allows you to:

_ Combine the benefits of cloud infrastructure with your on-premises resources

_ Leverage a consistent architecture across on-premises and public cloud environments that eliminates cloud vendor lock-in and ensures application portability without any re-work

_ Deploy an enterprise data platform that ingests, stores, processes, and analyzes all data regardless of volume, velocity and variety

_ Implement a data fabric that extends across your entire enterprise and provides a single pane of glass to locate, view, manage, and access disparate enterprise data assets—no matter where the data resides

_ Apply and enforce a consistent set of security and governance policies across hybrid cloud environments—including fine-grained access controls, data lineage, and audit logs

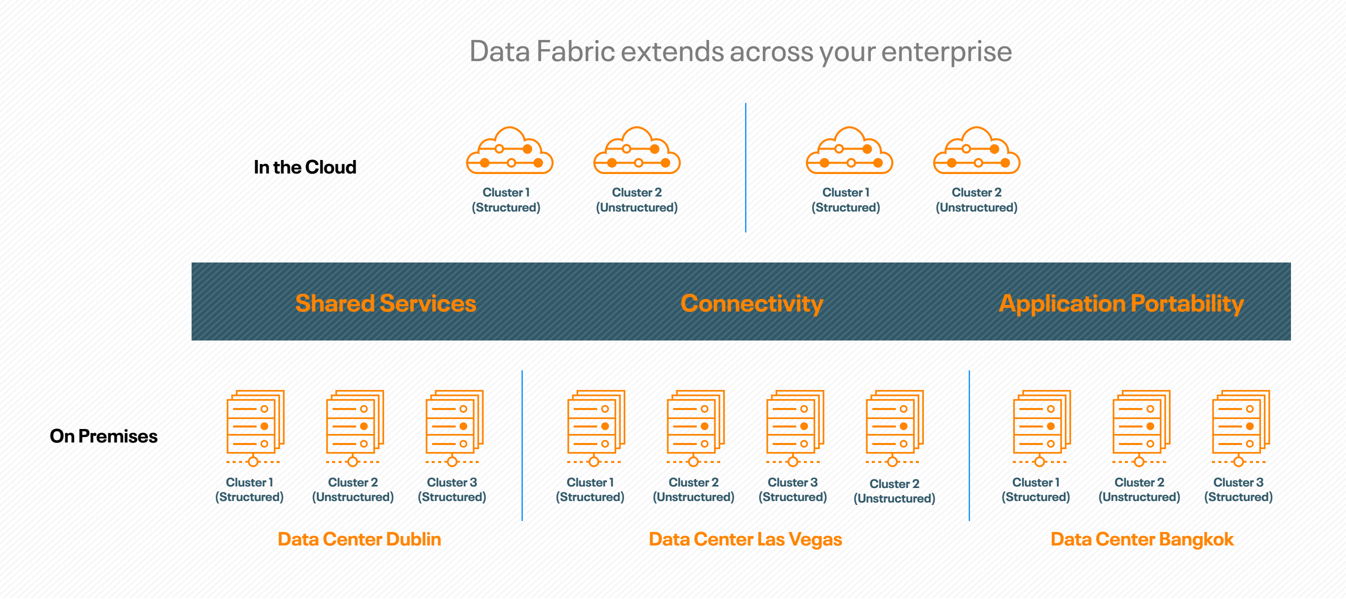
_ Continue your ongoing efforts to plan and rationalize which workloads, applications, and datasets are suitable for the cloud and which need to stay on premise

What to Look for in an Open Source Enterprise Data Platform

Hybrid cloud is the new reality, which requires a versatile enterprise data platform based on a modern data architecture. As you embark on a journey to select an enterprise data platform, think of your enterprise data assets at a global scale—spanning multiple on-premises, cloud, and multi-cloud environments, forming a data fabric. To truly grasp the potential of your data and turn it into a strategic asset, look for an open source enterprise data platform that provides a single pane of glass for all of your enterprise data assets regardless of where the data resides, with consistent security and governance as one of the key tenets.

Remaining true to open source will help you:

- alleviate vendor lock-in concerns,
- benefit from the rapid pace of open source software community innovation,
- take advantage of the open source ecosystem partnerships,
- and ensure that your business success is not tied to any proprietary technology.



Whereas a data management platform is organized around a single data repository, and is deployed for each repository, data fabric combines data management platforms for one global view of data.

Treat Cloud as IT Infrastructure

Many early cloud adopters started their journey with a single cloud vendor for good reasons—regional availability, pricing, service levels, and other factors—but now realize that their enterprise cloud strategy and their global business presence calls for multiple cloud providers. At the end of the day, cloud is providing you with access to compute and storage infrastructure much like a power utility is providing customers with access to power. Businesses should be able to take advantage of cloud compute and storage resources without fear of vendor lock-in or of a proprietary application layer hindering their progress in the hybrid cloud world.

Cloud vendors offer critical infrastructure to power your cloud needs, but they don't provide the necessary foundational capabilities required to establish an enterprise-level hybrid data architecture. Most organizations today are managing a very complex application and data ecosystem that requires a full suite of enterprise features such as security and governance, operational controls, application portability, and enterprise support.

Your cloud vendors are simply not focused on enabling a hybrid data architecture because they don't have any footprint on premise, and they're naturally motivated to get you locked into their ecosystem of compute, storage, and applications. Therefore, your data strategy for a hybrid cloud requires additional technology that addresses your global data management requirements while fully leveraging your investment in cloud infrastructure.

Why Have an Enterprise Data Strategy?

Leading your IT organization through the business transformation to a well-managed hybrid architecture requires new ways of thinking about your data strategy. In fact, you need a data strategy more than a cloud strategy. Your data is a strategic asset and needs to be treated as such. Cloud provides an immense opportunity for a scalable and robust delivery model, but it's the well-planned data strategy that lets you control costs and reduce risks while enforcing consistent security and governance across your enterprise data assets.

As the new currency of business, data provides value that will transform business, geopolitical, and human landscapes. Whether it's mapping cancer genes, reducing hunger by improving crop yields, understanding customers, or diagnosing medical conditions before they occur, it all starts with data. In fact, the more historical data you have at your disposal, the higher the likelihood that you can make predictions and act upon them in ways that can significantly improve business outcomes.

In the end, it's not so much about the cloud strategy. It's the data strategy that counts—and it should allow you to answer the following questions about your data regardless of its location, whether it's distributed across hybrid cloud and on-premises data lakes:

Remaining true to open source will help you:

- What data do you have?
- Where does it reside?
- How do you govern and secure it?
- How do you derive value from it?

What you risk without a data strategy

Your data strategy allows you to mitigate risks by focusing on data storage, management, and protection. It delivers security and governance, limits fraud, prevents theft, and ensures compliance—without which your business could face steep fines. It also ensures the integrity or accuracy of the data flowing through your enterprise, making data known, discoverable, available, trusted, and compliant. Your data strategy should also support your business objectives—like increasing revenue, improving customer satisfaction, and driving profitability. Discovering and delivering data allows your lines of business to gain insights quickly, accelerate improvements to products or customer experiences, and understand how to gain cost efficiencies.

Finally, you need a data strategy that allows you to gain flexibility and agility while remaining cloud-agnostic in order to avoid future vendor lock-in. Therefore, having a well-defined data strategy is critical to your success delivering a hybrid architecture.

No business operates without data—and no business's IT department should go without an enterprise data strategy.

Balance business and IT needs

As you start thinking about an enterprise data strategy, it's imperative to balance the needs of your line-of-business (LOB) practitioners with enterprise IT standards. Your data strategy must accelerate LOB practitioners' time to insights with on-demand self-service access to data and a broad set of analytics tools. On the other hand, your enterprise IT stakeholders still need to retain control of the enterprise data platform, so they can reduce risk and ensure compliance with enterprise IT standards for security, data governance, and operational reliability.

Developing an enterprise data strategy that gives your LOB practitioners self-service access to the tools they need without creating shadow IT or duplicating data and analytics silos is not a trivial task. In today's world where data leaks and breaches can cripple any business, enterprise IT must have a long-term vision for the enterprise data strategy, one that simplifies on-boarding and operations without hindering progress.

In the rush to cloud and the agility, simplicity, and efficiencies it delivers, it is paramount you understand and rationalize your data assets and workloads for cloud readiness based on security, sensitivity, ownership, and other concerns. This is especially important since you've probably invested years into implementing, optimizing, and operationalizing systems of record in on-premises-only environments.

Manage costs and optimize resources

Understanding cloud service costs can be daunting considering that every service uses a different pricing model consisting of multiple components. For instance, some services are priced based on consumption of compute resources for virtual machines and storage capacity. Depending on the type of storage medium used, storage can be priced based on the actual amount of space consumed or on provisioned storage regardless of actual consumption. Other services can be priced based on data being scanned on a per query basis, number of API requests, number of bytes transferred over a network, and many other factors that are often difficult to estimate up front. If your big data environment is sitting idle or if your jobs, queries, and pipelines aren't properly designed and tuned, you might, in fact, be paying significant overhead due to improper consumption of cloud resources while getting very little use or business value from that consumption.

Your big data pipeline may incorporate multiple steps, with each step requiring a different technology. For example, an end-to-end data processing pipeline may require tools such as Apache Kafka for streaming data, Apache NiFi for managing data flows, Apache Spark for data science and machine learning (ML), Apache Hadoop for storing the data using a Hadoop distributed file system (HDFS), and Apache Hive for running SQL queries to answer business intelligence questions from your data.

Implementing this entire pipeline with disparate cloud-native services is challenging from both operational and budgetary perspectives. While some cloud vendors may offer native services that provide similar functionality, there are three major challenges:

- Cloud providers are mostly packagers of Open Source Software (OSS) and typically lack the open source committers to projects needed to support your big data workloads in production
- It's up to you to orchestrate and integrate multiple cloud-native services into an end-to-end data processing pipeline, without any enterprise-level support from the vendor
- Cloud vendor billing models for each service may be completely different, making the overall cost extremely difficult to estimate up front

When moving data analytics workloads into the cloud, it's important to understand the key resource metrics for running your workload and the associated cloud billing model and service costs. One of key cost control strategies is to keep long-running workloads on-premises while moving ephemeral workloads that need resource elasticity to the cloud.
Having the right operational controls in place for optimizing use of cloud resources for your big data analytics workloads can save you significant expense.

Balancing Enterprise Requirements for Cloud Data Strategy

LINE OF BUSINESS PRACTITIONERS

DATA ANALYSTS, DATA ENGINEERS AND DATA SCIENTISTS

TIME TO INSIGHT

Access to analytics tools, on-demand, self-service access, single pane of glass

ENTERPRISE IT STAKEHOLDERS

BIG DATA PLATFORM OWNERS

REDUCE RISK

Consistent security and governance; manage cloud and shadow spend; retain data context and visibility; operational reliability and portability; cloud agnostic

Keep sensitive data secure and compliant

Determining which workloads should remain on premise and which should move to the cloud also requires a thorough understanding of your data assets. For instance, to reduce risk of data breaches, you may decide to keep sensitive data such as personally identifiable information (PII), payment card industry (PCI) data, HIPAA-protected health information (PHI), and other regulatory types of data on-premises while moving other less-sensitive data and workloads into the cloud.

As the trend is to move more and more sensitive data to the cloud, it becomes important to have a unified security model. To reduce your risk of data exposure and unauthorized access, you need to have consistent security and governance controls in the cloud that allow you to apply fine-grained security policies with full end-to-end data provenance and lineage as well as an audit trail to track who has accessed the data. After all, in the case of unauthorized access, you need to be able to immediately identify which data assets were exposed, what information was potentially leaked, and identify the perpetrators by analyzing a system-wide access audit log.

Reduce Risk with Enterprise-Level Support

Mission-critical applications must be backed by an enterprise support agreement from a vendor that you trust. Just as you wouldn't deploy mission critical applications on-premises without having proper support from your software and hardware vendors, you shouldn't risk running applications in the cloud without establishing enterprise-level support.

Make sure your enterprise support vendor is not only able to respond within specified SLAs but can also provide security and fix patches in a timely fashion and deliver them back into the open source code. Your vendor should provide not just OSS but intelligent support that proactively analyzes your big data environments both on premise and in the cloud to maximize performance and reduce risk.

Your enterprise support vendor must be a true partner that communicates early and often and keeps your business interests aligned with the OSS community vision and product roadmap. Innovating in the open source community is done by consensus, and voting determines whether to include code modifications. Look for a vendor that has gained favorable votes from the Apache community and has those with a "committer" status (meaning they've been given write access to the Apache codebase) or "contributor" status who have been voted in by the community and Apache PMC members and can influence OSS direction to ensure your enterprise success.

Bottom line, make sure your enterprise support vendor isn't just an OSS packager or service provider, but actually employs those with a "committer" status (meaning they've been given write access to the Apache codebase) or "contributor" statuscommitters and contributors in the community. who can influence OSS direction and ensure your enterprise success.

12 Technical Requirements for Building an Enterprise Data Strategy for the Hybrid Cloud

Based on our experience working with customers around the world, we've identified a key set of technical requirements that should be an essential part of your enterprise data strategy and are critical to establishing a modern data architecture in a hybrid cloud world.

1. Create a Data Fabric Layer

For years, enterprises have been trying to bring data residing in disparate silos together into a single, scalable platform. The objective of consolidating data is to deliver a more complete 360-degree view of all activity across the enterprise so you can act and respond with more agility and drive increased revenue, improved service, and reduce costs and/or risk.
With data growing at a rate of 40 percent per year, driven by new types like Internet of Things (IoT) and social data, a new wave of opportunities has been unleashed for businesses and people around the world. And achieving that "360-degree view" goal is more important than ever. As you expand your data landscape in the hybrid cloud and collect more data acros s on-premises and public cloud silos, your need for a better global data platform grows exponentially.

The use of cloud for data storage introduces additional challenges for companies to manage data scattered across multiple public clouds, private clouds and on-premises environments. In addition, the rapid expansion and adoption of cloud and edge computing, further necessitates a more modern, comprehensive, scalable, and sustainable architecture.

Key to this modern data architecture in a hybrid cloud environment is a data fabric that manages the entire lifecycle of data and provides a single pane of glass for storing, processing, securing, and analyzing your data no matter where it resides. Data fabric is an essential element of your hybrid data architecture. It provides consistent security, governance, and data management capabilities across on-premises and multiple cloud environments. Data fabric is a unified, secure platform that connects diverse data repositories to help manage, move, protect, govern, and analyze data residing anywhere in the hybrid cloud.

TECH TIPS

Your data fabric must:

_ Deliver a unified data management environment that integrates disparate data sources into a single logical view of all global data assets.

_ Enable a consistent data management experience across all data repositories in the hybrid cloud, regardless of how individual systems view and access data at the edge, on-premises, and/or in a public cloud.

_ Provide consistent security and governance for all data, no matter where it resides: on premise, at the edge, or in the cloud.

_ Support provisioning, management, and orchestration of ephemeral big data workloads by leveraging on-premises, private cloud, or public cloud resources:

_ Support segregation of compute and storage

_ Leverage container-based platforms for application portability across hybrid cloud environments

_ Deliver shared services for security, governance, and metadata, which are necessary to maintain state across ephemeral workloads

_ Support the entire data lifecycle across edge, on-premises, and multiple cloud platforms.

_ Support a variety of tools and applications that access "data at rest" and "data in motion" using an array of interfaces that abstract the underlying system complexities.

_ Support all types of workloads for application processing as the data is flowing at the edge, in the data center, and in the cloud.

2. Capture, store, and process your data anywhere, in any format

Your data management landscape is likely becoming more complex as new sources, types, and formats of data become available and add significant value to your traditional systems of record. Your ability to capture, store, and analyze these new types of data will enable you to uncover hidden business insights and create new revenue streams from the newly captured data assets. Traditional data management systems were not designed to deal with the volume, velocity, variety, and veracity of data today, and they often fail to scale to meet the demands of modern data applications.

Today's reality is that some data originates at the edge, in the cloud, or on-premises as your business, customers, and supply chain all interact and exchange information. You will need to capture, store, and process all of this data in its original format without losing any contextual information about the data and its source of truth. And you'll need real-time visibility into this data to respond, remedy, and/or make real-time adjustments. All this requires your data fabric to seamlessly span hybrid cloud environments and deliver a single view through which your users can easily store, find, and process data wherever it

TECH TIPS

- Make sure you have the flexibility to capture, store, find, and process data across edge, on-premises, and multiple cloud environments.
- Make sure your data can be easily ported across cloud boundaries by decoupling storage from compute resources and leveraging cloud object stores such as Apache Ozone O3, Amazon S3, Azure Data Lake Storage (ADLS), or Google Cloud Storage (GCS).
- Look for an enterprise data platform that offers a consistent API layer with pluggable storage connectors to your desired storage medium in the cloud or on-premises.
- Look for a solution that allows you to collect, curate, and store the data in its original format—whether it's unstructured, semi-structured, or structured—without losing any contextual information about the data and source of truth.
- Make the data available in full fidelity to your analysts and data scientists so they can run SQL on Hadoop, machine learning, AI, NoSQL, and real-time complex event processing applications that address your most pressing business needs.

3. Establish a single pane of glass

Organizations that are successful at managing their global data assets from a single view—or "pane of glass"—will be able to glean important insights that can dramatically change the course of their business. Establishing this single pane of glass can seem like a lofty goal when your data is spread and locked across various data silos spanning on-premises, cloud, and hybrid cloud environments.

Start by establishing strong data stewardship that addresses how to ingest, store, catalog, secure, discover, and track data lineage as it moves throughout its lifecycle. Understand which data assets reside where, and create an enterprise data catalog that contains business glossary tags and metadata describing schemas, location, security policies, and lineage details.

In most cases, you will find your data residing in various data management systems that were not designed to work together, creating information silos. To break down those silos, bring relevant data together into multiple distributed data lakes that can be purpose-built for a specific ephemeral data pipeline and/or as a general-purpose shared data lake. In either case, the goal is to establish a data fabric that provides a single pane of glass to search, discover, and understand your global data assets while enabling self-service access to it.

TECH TIPS

- Look for a platform that lets your data stewards understand and govern data across enterprise data lakes in the hybrid cloud. They should be able to organize and curate data globally based on business classifications, purpose, and sensitivity.
- Make sure you have the flexibility to move data and metadata across the hybrid cloud without losing sight of and context for your data assets. Your metadata should contain information about database schemas, security policies, business catalog, audit logs, data lineage, and provenance that can be viewed and managed from a single pane of glass.
- Your enterprise data platform should enable self-service data discovery by allowing:
- data stewards to curate data assets and group relevant datasets together into collections representing business entities that business users can search
- data engineers to understand how data is created and modified by visualizing upstream lineage and downstream impact, understand schema and data evolution over time
- business users to find data of interest by searching the catalog to locate relevant data assets and collections (e.g., sensitive data, commonly used data, high-risk data, etc.), understand data shape and distribution characteristics
- data auditors to understand how data access is secured, protected, and audited by being able to see who has accessed what data from a forensic audit/compliance perspective and visualizing data access patterns to identify anomalies

4. Apply consistent security and governance

As you work to define your hybrid data architecture, data governance and security requirements can pose a key challenge and must be at the center of your enterprise data strategy.

Regulations like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act of 2018[3] will continue to change the way businesses collect, store, and use personal customer data. By making sure that data protection and governance are built into your hybrid data architecture, you can leverage the full value of advanced analytics without exposing your business to new risks and comply with mandatory regulations.

Just as you wouldn't leave your data center exposed to cyber threats, you wouldn't want to have your data assets left unprotected. Whether you are dealing with sensitive data such as PII, PCI, PHI, or other types of company data, it's imperative to have consistent security and governance controls to guarantee data protection and compliance no matter where that data resides.

Consistent security and governance controls allow you to identify and classify sensitive data, secure it at the appropriate level of granularity, track lineage, authorize access, and audit who has actually accessed the data throughout its lifecycle in the hybrid cloud. Having these controls will help ensure your data is protected, trusted, and compliant at all times.

TECH TIPS

Look for a data management solution that enables you to bring security and governance together by providing:

- A single administrative console to manage all security policies across all data assets and all workloads
- A catalog of business terms and classifications to help organize and discover data stored inside your enterprise data platform
- Dynamic attribute-based security policies for fine-grained access control and dynamic data masking policies using classification, prohibition, time and location attributes
- Column-level security (filtering/masking/ obfuscation) and row-level filtering policies based on dynamic data tagging and classification
- Detailed audit log that tracks all user access requests in real time
- Improved data access security using Attribute-Based Access Control (ABAC) in addition to traditional Role-Based Access Control (RBAC).

5. Track data lineage and provenance

Your hybrid data architecture is likely to encompass various systems of record running on-premises, systems of engagement running in the cloud, and potentially a variety of edge IoT devices deployed in the field, all interconnected by a data fabric. It's imperative that you have a complete view of data movement as it flows through this complex pipeline, starting at the edge and traversing multiple analytics engines.

Understanding data lineage and provenance is paramount to making sense of your data lifecycle and complying with regulatory requirements. It's important to know where the data originated, when it changed, and who changed it. This information gives you critical decision-making power and allows data stewards and data engineers to make your data known, available, trusted, and compliant.

The challenge is that your enterprise data landscape is likely to span multiple vendor technologies that aren't integrated, which makes it difficult to obtain a complete data lineage and provenance view. Having that view offers important benefits to data stewards and auditors.

For example, data that starts as event data coming from your edge devices or systems of engagement and is combined with enterprise data coming from your systems of record can be tracked and governed at every stage of its life cycle. This allows data stewards, operations teams, and compliance engineers to visualize a dataset's lineage and then drill down into operational, security, and provenance-related details. As this tracking is done at the platform level, any application that uses these engines will be natively tracked. This allows for end-to-end visibility of data provenance and lineage across the entire data lifecycle spanning multiple applications and analytics engines.

TECH TIPS

- Look for a data platform solution that provides complete data lineage and impact functionality telling you where data came from, how it was transformed along the way, and what assets are affected downstream.

- Standardize on an open source metadata repository solution that provides open API and allows various data management technologies to integrate their data lineage and provenance metadata for a complete view of your enterprise data assets.

- Your open metadata repository should provide out-of-the-box capabilities to track and visualize cross-component lineage, delivering a complete view of data movement across a number of data streaming and analytics engines such as Apache NiFi, Apache Spark, Apache Hive, Apache Sqoop, and Apache Kafka.

6. Enable easy data discovery for self-service analytics

In most enterprises, data is still locked in silos managed by centralized IT teams. This hinders visibility and reduces productivity for LOB users who require self-service access to data for greater agility.

The first step toward delivering self-service analytics and faster business insights is centralizing your enterprise data assets in an enterprise data lake or multiple distributed data lakes connected by a data fabric. Once your data is consolidated and silos are broken, you need to ensure it can be easily discovered and available for self-service provisioning and consumption.

In addition, before making your data available to a broad group of LOB users for self-service analytics, make sure you curate and organize it. This involves understanding where relevant data is located, defining your business taxonomy, securing access based on data tags and attributes, and making it discoverable to business users.

Making curated data available for self-service analytics while maintaining governance of your data assets is a significant challenge that requires a balance between enterprise IT stakeholders who want to reduce risk and LOB practitioners who want to accelerate time to insight. Business users need to be able to run ad-hoc analytics on curated data assets using tools of their choice in the hybrid cloud by having access to the data they are entitled to without relying on IT.

The Journey to Self-Service Analytics: It All Starts with Data

One of the more common uses cases that hybrid data architecture powers is a self-service analytics platform developed by IT for the business.

A common complaint today is that data scientists and analysts have to spend too much time looking for, prepping, and cleaning data and not enough time analyzing it. While some data cleansing and preparation will always be necessary, we can reduce the time data analysts spend on this task by enabling them to leverage a self-service analytics solution with a broad set of data discovery and analytics tools. This will allow data scientists to spend more of their time collaborating, sharing data models, and analyzing data. Success on the journey of data democratization requires a solid data infrastructure and a focus on data strategy. After all, it all starts with data.

TECH TIPS

- Your enterprise data platform must be flexible and adaptable to support data ingestion from a variety of data sources and formats, with appropriate data governance controls that allow data tagging, classification, and data lineage as data moves from ingestion to consumption.

- Automate data tagging and classification based on dynamic rules and machine learning algorithms, while allowing data stewards to review and approve data tagging and classification policies before datasets become available downstream. This will ensure a high quality of data and help establish organizational trust in the data assets you have under control.

- Your enterprise data platform must empower data stewards to understand, secure, and govern data across enterprise data lakes.

- Look for a data platform that enables business users to discover, access, and provision data and workloads for self-service analytics using SQL, Spark, or other tools for performing advanced analytics on-demand.

7. Remain Cloud-Agnostic to Avoid Application Layer Lock-In

Open source software arose from business concerns about proprietary software and vendor lock-in, since it often inhibits innovation, hinders integration with systems outside of the vendor's ecosystem, and inflates the cost of maintenance. Results from the Open Source Program Survey[4] confirm that large technology companies are leading the way in establishing open source programs to create and nurture best practices. This momentum has contributed to a wide adoption of Hadoop and other related open source big data technologies.

When charting your cloud strategy, vendor lock-in becomes somewhat inevitable since you are moving data and using the cloud service provider's infrastructure to run your applications. But the key concern is application layer lock-in. The native cloud services that cloud providers offer are often tightly coupled to the provider's infrastructure and lack the full breadth of open source features and ecosystem interoperability. For example, a cloud vendor's data analytics service may limit your ability to:

- Migrate your application workloads to other cloud platforms or back on-premises
- Move your data between multiple cloud services and/or on-premises solutions, since it was converted into custom data formats specific to your cloud vendor's analytic services
- Leverage the latest OSS innovations
- Integrate with other software and services designed to augment existing OSS capabilities
- Keep track of data lineage, metadata, and security policies as the data moves between different hybrid cloud environments

As you move to hybrid cloud, you'll want to leverage and bring into cloud deployments the same open source ecosystem of big data technologies running in your data center. That way, your common data fabric can consistently unify management, analytics, operations, security, and governance of data across on-premises, cloud, and hybrid cloud environments.

TECH TIPS

- Consider an open source enterprise data platform that provides cloud-agnostic capabilities and allows easy migration of data assets, metadata, and workloads across all environments.
- Look for a data platform that enables a consistent data fabric running on top of your cloud infrastructure, so you can more easily deliver portable business applications.
- Leverage a centralized provisioning platform that allows you to:
- Simplify deployment of data management services across multiple cloud environments and enables you to quickly run big data workloads in any cloud
- Provision big data workloads on the fly that fully leverage use of native cloud resources such as compute and storage while optimizing use of those resources

8. Ensure portability of data-driven applications

Cloud application portability is a key concern when implementing a hybrid cloud architecture. Consider that, according to IDC's 2018 Cloud and AI Adoption Survey[5], "80% of IT decision-makers have migrated either applications or data from public cloud environments to on-premises or private cloud solutions in the last year." Deciding whether an application or workload belongs in the cloud or on premise is not an obvious one and can frequently change depending on your business strategy and priorities. Having the ability to easily move applications with minimal integration issues will save you time and money.

Data has gravity: the more data you have, the more it will attract other applications and services around it. This ecosystem of applications and services that expands around your data in the cloud can become complex enough that at some point you will likely need to consider moving your data and applications between public cloud vendors or from cloud to on-premises and vice versa. Replicating data between on-prem and multiple cloud providers introduces additional challenges for data governance.

Therefore, make sure that your organization's data fabric enables you to easily port applications, data, and metadata across on-premises and cloud boundaries while keeping track of its movement and without requiring any changes to the application code.

TECH TIPS

- Implement an enterprise data platform that provides a consistent set of storage and application APIs that can be used to migrate your workloads across on-premises and different cloud providers.
- Make sure that your data access layer, application API, and metadata repositories provide the necessary abstraction to decouple your application code from the underlying storage and compute infrastructure.
- Select an enterprise data platform that offers native data and metadata replication capabilities across on-premises and hybrid cloud environments while keeping track of data governance and lineage.
- Look for a platform that allows you to leverage open source standards, tools, and technologies across any data no matter where it resides in the hybrid cloud.

9. Enable Edge Connectivity

Today's business interactions are becoming more complex than ever before. Customers are expecting an unprecedented level of service that requires you to have real-time visibility into the health of your business and deep insight into customer behavior, product performance, and customer satisfaction levels.

To deliver this level of service, your hybrid data architecture needs to support collecting data from the network's edge and pushing analytics and decision making closer to the edge, where IoT devices and sensors reside. To be effective, your data architecture should provide real-time responses based on real-time events measured and scored against machine learning models built and trained on historical insights.

Evolution of new cyber security threats and opportunities to improve process efficiencies in supply chain, transportation, manufacturing, oil and gas, and other industries requires edge connectivity to capture time-series data at the point of origination from edge IoT devices. In order to mitigate these threats and capture new business opportunities, your hybrid data architecture needs to support data flow, stream processing, and messaging technologies to build a true streaming data architecture across the edge, cloud, and on-premises environments.

TECH TIPS

_ Implement an enterprise data platform that provides a consistent set of storage and application APIs that can be used to migrate your workloads across on-premises and different cloud providers.

_ Make sure that your data access layer, application API, and metadata repositories provide the necessary abstraction to decouple your application code from the underlying storage and compute infrastructure.

_ Select an enterprise data platform that offers native data and metadata replication capabilities across on-premises and hybrid cloud environments while keeping track of data governance and lineage.

_ Look for a platform that allows you to leverage open source standards, tools, and technologies across any data no matter where it resides in the hybrid cloud.

Leveraging Data from the Edge

The growth of internet-connected devices has led to a vast amount of easily accessible data. Increasingly, companies are interested in mining this data to derive useful insights and make data-informed decisions. Recent technology advancements have improved the efficiency of collecting, storing, and analyzing time-series data, spurring an increased appetite to consume it.

Around the world, investment in IoT is skyrocketing:

- IDC forecasts[6] worldwide IoT spending to reach $1.2 trillion in 2022.
- In recent research, Bain predicted[7] B2B IoT segments would generate more than $300 billion annually by 2020, including about $85 billion in the industrial sector.
- Discrete manufacturing, transportation and logistics, and utilities will lead all industries in IoT spending by 2020[8], averaging $40B each.

Companies looking at leveraging IoT often encounter serious hurdles along the way. These include building and deploying new applications quickly, accessing the data from edge devices, and scaling to support the massive volumes of data generated.

To address these concerns, companies need to look at open source technologies that provide high-performance analytics of data stores for event-driven and time-series data. Think of leveraging data from the edge as a four-step process:

1. Connect the remote sensors and devices that produce the data
2. Collect data from the edge, deliver it to your data lake, and then process and organize the data
3. Analyze the edge data, looking for nuances and correlations between data events
4. Act: Combine the analytics model with a stream processing tool to apply its predictive capability to events happening in real time. Then inject this back into the original workflow between the Connect and Collect stages. Now, as rivers of data flow in, they flow through these trained models in order to evaluate each event that is happening and allow you to act when an event is about to occur.

Your stream processing is now constantly on the lookout for events and will alert you at a moment's notice as to what is taking, or about to take place.

10. Deploy Ephemeral Workloads in the Cloud to Optimize Costs

Cloud provides many benefits for optimizing management and cost of your IT infrastructure. Depending on your cloud provider, resources can be made instantly available with up-to-the-second billing based on consumption. Cloud provides the fundamental building blocks for your big data processing needs: data storage at massive scale and access to infinite compute resources such as CPU, GPU, and memory.

One of the biggest advantages of using cloud resources for your big data workloads is separation of compute and storage as part of your cloud infrastructure. This fundamental design principle allows independent scaling of storage and compute resources depending on your workload requirements.

When deploying big data applications in the cloud, you need to understand your resource requirements based on workload type and usage pattern. No matter whether your application is batch, in-memory, interactive, or real-time, it can be deployed as a long-running or ephemeral workload depending on your business need. Your ability to deploy ephemeral workloads that can take advantage of cloud resource elasticity in response to a business need will allow you to accelerate time to insights while reducing the overall cost. After all, considering that most cloud billing models are on a consumption basis, you'll want to optimize your use of cloud resources based on actual workload utilization vs. resource consumption.

TECH TIPS

_ Your enterprise data platform should simplify provisioning of big data workloads in the cloud in a consistent and repeatable fashion.

_ Look for a solution that allows IT operators to define workload blueprints for automated provisioning of cloud resources and workload deployment, which can be made available to line of business users via self-service access.

_ Look for automated workload scaling based on business SLAs and actual resource utilization.

_ Your enterprise data platform should let you optimize cloud resource usage by seamlessly adjusting the cluster as workload and activity changes.

_ And it should enable you to respond faster to new business requirements and align costs with utilization and business value rather than paying for provisioned cloud infrastructure.

11. Leverage Shared Metadata Services for Ephemeral Workloads

Underpinning cloud-native architecture is a paradigm shift from a single long-running, centralized data lake containing all supporting metadata services to a distributed collection of data lakes and ephemeral workloads connected via your data fabric. This paradigm shift requires rethinking metadata services (database schemas, security policies, audit, governance, lineage) that must be shared across ephemeral workloads.

Your data-intensive applications require rich metadata providing technical and business characteristics as well as stateful information. This is what makes the data useable and shareable across multiple ephemeral workloads. Business users who want to explore data on demand with different analytics engines need to have a layer of shared services that provide all the necessary metadata, context, and state information for a consistent view of data assets.

By incorporating a layer of shared metadata services, the resulting modern data architecture allows you to have both long-running and ephemeral workloads accessing the same data as part of a distributed data lake in the cloud. As your workloads execute, the defined security context, audit controls, and schema should be applied automatically and consistently.

TECH TIPS

- Your modern data architecture should incorporate a layer of shared schema, security, and governance services that provide a consistent view of metadata across multiple ephemeral workloads accessing the data in your shared data lake.
- Once you define your schema, security, and governance policies, the metadata you defined should be available to any ephemeral workload that can be attached to the shared data lake.
- As your data and workloads move across hybrid cloud, the architecture should automatically carry and propagate all of the associated metadata to ensure a consistent application experience no matter where the data and applications reside.

12. Implement a Consistent Hybrid Architecture for Containerized Workloads

Just like shipping containers have revolutionized international trade, containerization is changing the way we package, deploy, and consume software applications today. Containers provide a lightweight, consistent, and repeatable way to package software that includes everything needed to run it.

Containerization allows you to manage different versions and multiple instances of various big data technologies, application code, and dependencies without interfering with one another. Containers provide the needed isolation and dependency packaging to enable easy deployment of big data applications across hybrid cloud environments. They allow you to move workloads between different on-premises and cloud container platforms without making any application code changes.

As hybrid cloud is quickly becoming the new reality, organizations need a unified hybrid architecture for on-premises, multiple cloud, and edge environments that provides a consistent fabric for deploying containerized workloads. Your enterprise data platform for hybrid cloud needs to enable a consistent experience and interaction model built on top of a container platform. This architecture allows you to seamlessly move data and containerized workloads across on-premises and multiple cloud environments.

TECH TIPS

- Look for an enterprise data platform that delivers a consistent hybrid architecture in the cloud and on premise.
- Your enterprise data platform should support industry standard container technologies—e.g., Docker, Kubernetes, Red Hat OpenShift—to ensure portability of containerized applications across hybrid cloud environments.
- Your data platform should provide a consistent user interaction model that allows you to seamlessly move data and workloads across hybrid cloud environments as well as deploy containerized workloads regardless of the environment.
- Look for a data platform that provides DevOps orchestration tools for managing services and workloads via the "infrastructure as code" paradigm to allow spin-up and down of ephemeral container workloads.
- Look for shared services for metadata, governance, and security to support containerized applications running as ephemeral workloads based on data stored in the shared data lake.

Next Steps

At Cloudera, we believe the time has come to reimagine the modern data architecture, with hybrid cloud as a key requirement. Enterprises need a consistent and unified hybrid architecture for on-premises, multi-cloud, and edge environments with these four key components:

1. Consistent security and governance across all data no matter where it resides
2. Separate storage and compute layers
3. Containerized workloads
4. Ability to orchestrate and manage workloads

For more information and a framework for ensuring that happens, download "Power your Business Transformation with an Enterprise Data Cloud."[9]

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises. Learn more at cloudera.com

Connect with Cloudera

About Cloudera: cloudera.com/more/about.html

Read our VISION blog: vision.cloudera.com/ and Engineering blog: blog.cloudera.com/

Follow us on Twitter: twitter.com/cloudera

Visit us on Facebook: facebook.com/cloudera

See us on YouTube: youtube.com/user/clouderahadoop

Join the Cloudera Community: community.cloudera.com

Read about our customers' successes: cloudera.com/more/customers.html

Sources

[1] https://resources.idg.com/download/executive-summary/cloud-computing-2018?utm_campaign=Cloud%20Computing%20Survey%202018&utm_source=Forbes%20-%20Louis

[2] https: //www.marketwatch.com/press-release/hybrid-cloud-market-worth-9764-billion-by-2023-2018-08-22

[3] https://www.caprivacy.org/

[4] https://thenewstack.io/survey-open-source-programs-are-a-best-practice-among-large-companies/

[5] https://www.crn.com/businesses-moving-from-public-cloud-due-to-security-says-idc-survey

[6] https://www.idc.com/getdoc.jsp?containerId=prUS43994118

[7] https://www.bain.com/insights/choosing-the-right-platform-for-the-industrial-iot

[8] https://www.statista.com/statistics/666864/iot-spending-by-vertical-worldwide/

[9] https://www.cloudera.com/content/dam/www/marketing/resources/whitepapers/power-your-business-transformation-with-edc.pdf.landing.html

CLOUDERA

Cloudera, Inc. 395 Page Mill Road Palo Alto, CA 94306 USA cloudera.com

© 2019 Cloudera, Inc. Cloudera and associated marks and trademarks are registered trademarks of Cloudera, Inc. All other company and product names may be trademarks of their respective owners.