CLOUDERA

WHITEPAPER

Five Steps for Aligning Data and Hybrid Cloud Architectures

How to ensure data leads your cloud transformation strategy

Table of Contents

Introduction

The meaning of "moving to the cloud" for businesses is very different today from its past connotation, because "the cloud" itself has changed. Cloud services can be deployed as public, private, or any mashup of the two. When businesses are moving to the cloud, they are really moving applications, data, and the interaction of those to a mix of infrastructures — some stay on-premises, others move to private clouds, still others to public clouds, or to a combination of all of the above. More than two out of three enterprises — or 69% — were forecast to have multi-cloud or hybrid cloud architectures this year[1].

A hybrid cloud architecture is more than simply a mix of public and private clouds. Gartner defines it as "policy-based and coordinated service provisioning, use and management across a mixture of internal and external cloud services." More simply, a hybrid cloud architecture — with "architecture" being the key term — is an assortment of public and private cloud deployments across the enterprise, often from more than one cloud provider, all operating separately, but still managed and controlled as one.

The biggest benefit of a hybrid cloud architecture is flexibility — as an IT leader, you can move workloads and data between private and public clouds as demands, needs, and costs change, giving you more options for deploying and using data to gain value. The biggest challenge is getting value and real-time insights from data and applications that live across different infrastructures, which requires managing, accessing, securing, and governing multiple clouds or a patchwork of cloud data silos. Each of those silos — whether from legacy applications or more recent point solutions — carries its own data and has a distinct framework for handling metadata, security, and governance. Creating and maintaining consistency on this front is challenging, and this is precisely why your business and data strategy need to drive your hybrid cloud strategy rather than the other way around.

This paper will outline five steps for aligning data and hybrid cloud strategies for IT leaders and their teams. You'll learn the importance of data context that is shared across multifunctional systems to deliver self-service business analytics and help turn insights into actions.

Step One: Understand your data

Start by understanding the disparate nature of your data — which data assets reside where. In most cases, you will find your data is located in various systems that were not designed to work together, creating information silos. Understand how the data is used by creating an (enterprise) data catalog that contains business glossary classifications and metadata describing schemas, location, security policies, and lineage details. Then place this comprehension within the context of the current regulatory environment — particularly regarding where information is held and who accesses it how.

As your systems are siloed, you will likely find subtle differences in how identical data is being used and perceived in different applications. Security and governance policies will show similar differences. To satisfy the needs of both IT and business users, this data 'context' should be identical for the same data. However, keeping it consistent between different systems and applications is a significant challenge and one that only gets more complex as different infrastructures and clouds are used. A shared, consistent data context, part of your data strategy, is critical for enabling the underlying hybrid cloud architecture's control of enterprise data and is the key to being able to move fluidly between different applications and their data, gaining insights as you go.

Data context consists of:

- Schema–how is data structured?
- Catalog–what is the business meaning of data?
- Security–who can access what data?
- Governance–how is data related and used?
- Lifecycle–where does data reside and how do data and policies change over time (from ingestion to deletion)?

Regulations like the General Data Protection Regulation (GDPR), the California Consumer Privacy Act of 2018, and the New York Privacy Act of 2019 will continue to influence the way you collect, store, and use personal customer data. Especially in a hybrid cloud environment, it is key to have consistent security and governance controls — or data context — to allow you to identify and classify sensitive data, secure it at the appropriate level of granularity, track lineage, authorize access, and audit who has accessed the data throughout its lifecycle. Having these controls will help ensure your data is always protected, trusted, and compliant no matter where it resides.

Step Two: Run a workload assessment

Workloads vary considerably in the digital age. Whether the workload is running ad infinitum or only for the time needed to work through a set of data, your business must understand the strains these demands place on infrastructure. You must also consider how workloads vary by or span across regions and how they cross internal, cloud, and hybrid resources. For instance, in the graphic below, the data warehouse is a permanent workload running on-premises and using a traditional RDBMS; the data science application is transient, only used when needed, and operating on data in a public cloud.

Remaining true to open source will help you:

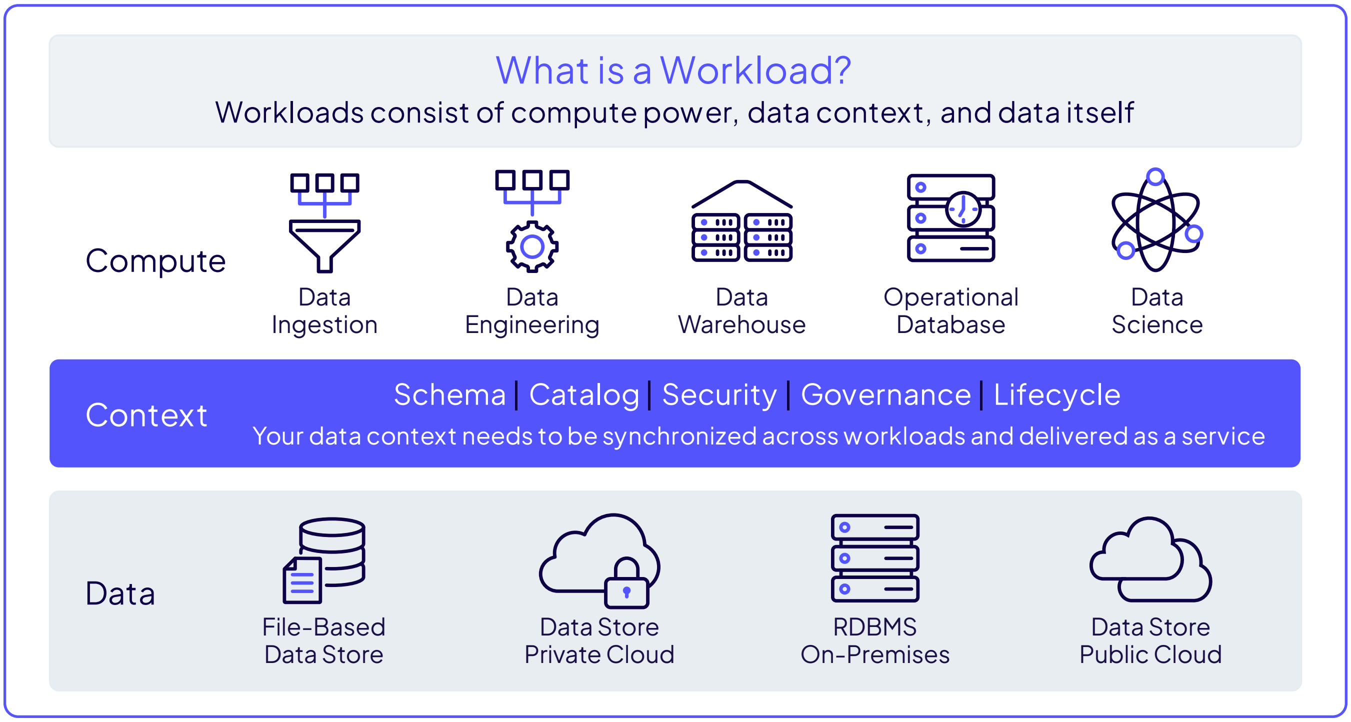
- alleviate vendor lock-in concerns,
- benefit from the rapid pace of open source software community innovation,
- take advantage of the open source ecosystem partnerships,
- and ensure that your business success is not tied to any proprietary technology.

It's also not just the location of data. It's equally about the suitability of the workload and its data for a particular deployment infrastructure and use case. Not evaluating the workload and use case properly against that background can have significant, costly impacts. For instance, not all data and workloads are suited for public cloud and may need subtle changes to make them suitable. Some may be restricted on the grounds of regulatory issues or intellectual property concerns; making the data anonymous may, however, make the move possible.

Organizations that do not address their data strategies ahead of their cloud strategy will find they'll have been perhaps too "public cloud happy." In a recent IDC survey[2], 80% of decision makers have repatriated 50% or better of the applications and/or cloud data over the last year to private cloud solutions. And when they do, they also need to migrate associated security and governance policies. Data gravity is something to be aware of and may work either for or against a deployment preference.

When deploying data analytics applications in the cloud, you need to understand your resource requirements based on the workload type and usage pattern. No matter whether your application is batch, in-memory, interactive, or real-time, you have a choice to deploy it as either a long-running or an ephemeral (transient) workload depending on your business needs. Your ability to deploy ephemeral workloads that can take advantage of cloud resource elasticity in response to a business need will allow you to take advantage of cloud bursting as well as adaptive scaling in order to always provide the appropriate performance and end user experience, accelerating time to insights at an optimal 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 rather than resource consumption.


What is a Workload?

Workloads consist of compute power, data context, and data itself

Compute

Context

Schema | Catalog | Security | Governance | Lifecycle

Your data context needs to be synchronized across workloads and delivered as a service

Data

Step Three: Classify your data and analytics systems, infrastructure, and controls

Most organizations have grown their IT landscape over time, sometimes on a departmental basis, or sometimes because they've acquired a company and inherited their systems and applications in turn. This next step is to identify your organization's data and application silos, the use cases for which they were created and for which they could continue to apply.

For instance, if you were to set up yet another system or application today, similar to one you already have, you would likely need to manually recreate the data context rather than synchronize it from an existing one. When systems run in an ephemeral or transient fashion, on any infrastructure, how do you preserve the data context when they terminate?

Classifying these issues will give you an idea of:

- Which systems are siloed and often proprietary?
- How much effort is involved in keeping the data context synchronized between them?

Consider this example: You want to develop a model that predicts maintenance required on trucks. For that, you'd need to:

- Capture real-time maintenance data from the vehicles
- Uncover structures through data engineering
- Store the data for initial, rough-cut analysis
- Create data features (data engineering) and develop a model (data science)
- And push this model back out to the vehicles

Now, as an organization with a siloed set of systems and applications, you needed to involve quite a few of those to deliver this use case. How did you ensure consistent data context across all of them?

Often this step requires a lot of manpower and is done manually, and introduces too many risks. If a mistake is made in setting the security policy for a particular set of data, it could result in a business user not having access to certain information. As a result, they would be basing decisions on incomplete or erroneous data with significant consequences. More significant is when, as a result, confidential or private data is somehow exposed or used inappropriately, resulting in fines under regulations and loss of reputation.

The lack of a single platform that brings together all these different kinds of analytics and that can be deployed flexibly with consistent security, governance, and control policies is the crux in addressing challenges like this. Fundamentally, organizations should consider using an enterprise data cloud to alleviate these problems.

Step Four: Identify how your data platform achieves business goals

Though they are the ones managing the data systems and applications, IT does not work in isolation. Your business can use the cloud for technology services on demand, so your combined data and cloud strategy must identify how your business can meet its current data demands in a more flexible and cost-effective manner while also moving toward becoming a truly data-driven organization. That means figuring out how to leverage — rather than being hindered by — a hybrid cloud environment in a sensible and smart way.

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: on-demand data storage at massive scale and access to infinite compute resources.

But to empower the business with self-service capabilities, you need a multi-disciplinary platform that:

- Delivers the variety of analytics you need to gain insight from data
- Can take advantage of any infrastructure, including cloud
- And that provides consistent data security, governance, and control

IT can finally deliver on the business demand of self-service. With that, business can use the resources (data and analytics) they need when they need them. Systems can be made to automatically scale as well as run where they are most efficient.

The status quo is that 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. Also, business risk increases as you manually replicate data context — or security and governance policies — between the siloed workloads. Manpower and operational costs are higher because of the variety of systems needing to be supported.

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 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.

Step Five: Define and implement a hybrid cloud data strategy

If Step Four reveals what you'll need in a data platform to achieve business goals, this step defines the plan to go from where the organization is now to where it needs to be. Your business objectives should drive your data and cloud strategies, and your hybrid cloud architecture should deliver:

- The flexibility to run modern analytics workloads anywhere, regardless of where data resides;
- The ability to move your workloads to the cloud environment of your choice — public or private — and prevent vendor lock-in addition to facilitating portability;
- Agility, elasticity, and ease-of-use of public clouds;
- A multi-disciplinary platform that unifies metadata, security, and governance across all environments — to eliminate silos and deliver the self-service analytics the business requires.

With all the other considerations in place, think about how you can use the hybrid cloud — and a mix of private and public resources — to run your workloads in an optimized manner. With a hybrid setup, you can make sure data is stored correctly in relation to both governance requirements and ongoing cost management.

Conclusion

In the end, your hybrid cloud data strategy should:

1. Catalog your enterprise data in a data catalog with business glossary classifications and metadata in the context of regulations
2. Ensure workloads and their data are appropriate for their deployment infrastructures and use cases and can accelerate time to insights while reducing overall costs
3. Leverage an enterprise data platform that ensures consistent data context across all silos and workloads, and
4. Deliver self-service analytics for faster business insights Ultimately, it comes down to finding the right collection of cloud technologies and orchestration tools to make the organization run faster and more efficiently given its past, current and future business needs.

Next Steps

Clearly, your enterprise data platform is a key component in speeding time to insights regardless of where your data lives and on what workload it runs. To understand more about what's needed in an enterprise data platform operating in a hybrid cloud environment, download "12 technical requirements for your enterprise data cloud platform[3]."

Sources

[1] https://451research.com/images/Marketing/press_releases/Pre_Re-Invent_2018_press_release_final_11_22.pdf

[2] Businesses Moving From Public Cloud Due To Security, Says IDC Survey, https://www.crn.com/businesses-moving-from-public-cloud-due-to-security-says-idc-survey

[3] https://www.cloudera.com/content/dam/www/marketing/resources/whitepapers/12-requirements-for-a-modern-data-architecture-in-a-hybrid-cloud-world.pdf.landing.html

About Cloudera

Cloudera is the only true hybrid platform for data, analytics, and AI. With 100x more data under management than other cloud-only vendors, Cloudera empowers global enterprises to transform data of all types, on any public or private cloud, into valuable, trusted insights. Our open data lakehouse delivers scalable and secure data management with portable cloud-native analytics, enabling customers to bring GenAI models to their data while maintaining privacy and ensuring responsible, reliable AI deployments. The world's largest brands in financial services, insurance, media, manufacturing, and government rely on Cloudera to be able to use their data to solve the impossible — today and in the future.

To learn more, visit Cloudera.com and follow us on LinkedIn and X. Cloudera and associated marks are trademarks or registered trademarks of Cloudera, Inc. All other company and product names may be trademarks of their respective owners.