zaloni

THE CLOUD DATAOPS eBOOK

Navigating cloud data management through end-to-end data operations

Contents

CHAPTER - 1

What is a Cloud Data Lake?

Many organizations that we talk to are interested in leveraging cloud infrastructure as their data lake. They're smart to consider it. It's a highly flexible deployment where you only pay for the compute and storage. For companies that have highly varying levels of processing needs, this paradigm can offer a significantly lower price point and shifts management of hardware to a third party.

What is a cloud data lake?

Contrary to what some organizations have been led to believe, a cloud-based data lake is not an S3 bucket where data is dumped. A data lake is a maintainable, functioning infrastructure that maintains governance across all of the data. It provides access to the correct people at the appropriate stages of the data lifecycle and can adhere to a zone-based architecture specific to an organization's needs. A data lake should also provide self-service access to end users reducing overhead on IT.

Benefits of running in the cloud

Cloud providers have developed a plethora of services and tools that can be used by organizations in multiple ways. This means cloud subscribers have lots of pieces they can build their infrastructure upon. The cost to try (and potentially fail with) a number of options that could work is minimal.

An organization can develop upon the tools the cloud providers have to develop a fully functioning data lake. They can start small and scale out if necessary. In short, the cloud provides a scalable architecture with low upfront cost. As the needs of the organization increase, they can scale their compute, storage, and application requests.

In a cloud-based infrastructure, an organization only pays for the amount they use. For example, if an organization has high compute needs, but for short bursts, they are ideal candidates for savings.

Downside of the Cloud

Although cloud vendors offload a lot of the risk associated with data storage and security, those risks are still very real. The cloud vendor of choice needs to take this sort of risk into account. Data access pipelines need to be accounted for – can the customer send/receive the data at the speed necessary?

An oft-forgotten issue is the risk associated with choosing only one cloud vendor. If a cloud vendor suddenly decides to increase its prices by 20%, this can wreak havoc on an organization's IT budget. Many organizations are now realizing the reality of vendor risk and are seeking solutions that provide multi-cloud support to eliminate that risk.

An Ideal Option

Arena orchestrates the ingestion, transformations, tokenization and masking of sensitive/PII data, and provisioning to databases. Zaloni's data lake management system provides an abstraction layer leveraging the native compute and storage of underlying infrastructure. Thanks to its flexible architecture, it can natively work with multiple cloud (or on-premises) infrastructures.

As of this writing, it is the only data lake management solution that can provide a layer on top of a multi-cloud environment. This is a key win for companies who have appropriate risk minimization goals.

CHAPTER - 2

Elastic Compute and DataOps to Optimize Your Cloud Environment

Elastic Compute and DataOps

Many companies move to the cloud for cost-effectiveness and scalability. Still, the cloud journey can be difficult and costly if companies don't leverage elastic compute capabilities or don't have proper data management processes in place. In this blog, I'll cover how elastic compute can help organizations optimize their cloud data environment and manage and govern data in the cloud along with a couple use case examples.

The Benefits of Platform as a Service vs. Infrastructure as a Service in the Cloud

Initially, organizations would lift and shift workloads from on-premises deployments to the cloud. These projects were mostly driven by IT departments and focused on infrastructure. By default, the best choice was to start with infrastructure as a service (IAAS).

Although these offloads may be cost-effective compared to on-premises environments, the IAAS approach does not leverage cloud computing's full benefits. With the increasing demand for data processing, it soon becomes prohibitive to maintain (IAAS) in the cloud. In most cases, it becomes even more expensive to manage vs. in house deployment.

Most cloud providers encourage organizations to start using their platform as a service (PAAS) services and native cloud tools. These approaches provide huge benefits compared to IAAS services. The PAAS services save costs for workloads, which do not require full-time usage of a service.

For example, if an application requires MySQL database, instead of creating and

managing a cloud VM in an AWS EC2 Machine, you can opt to use the RDS service, or in the case of Microsoft Azure, you can use the SQL Database service.

The PAAS service changes the billing model to usage-based instead of a dedicated machine. Additionally, the PAAS service saves the effort required to patch, manage, and scale the database service since it is taken care of by the cloud provider.

Challenges Processing Data in the Cloud

Today we are faced with an entirely new set of challenges when trying to process data in the cloud. Some of these challenges include:

- Is my data secure in the cloud?
- How much does it cost to manage data in the cloud?
- I would like to access the data when I need it and without burdening my IT team.
- I would like to lease my data to other departments or third parties and revoke the lease when done.
- We would like to boost data processing capacity during high demand times and tone it down during low season.
- How do I ensure I can govern the cost spent on data processing for my projects?
- I would like to be able to approve the requests to access, lease or use my data.
- How can I measure or control the ROI from my data?
- Can I control the infrastructure required for my data processing on demand without IT or my Cloud management team?

How to Overcome Cloud Data Challenges with Elastic Compute and DataOps

So let's look at how to solve these problems by changing the data management mindset. I like to think about the analogy of changing from creating pet servers to herding cattle. You can read more about Pets Vs Cattle analogy here. This means that we develop our data pipelines in a transient way. We are able to fire up the data processing infrastructure on demand and destroy it once it is no longer required. Companies like DataBricks address this by autoscaling existing running clusters when needed and reducing them to

minimal when not needed. Some solutions may tie you up with services from a specific cloud provider.

Another way to solve this would be to have a single control plane, which can be cloud-agnostic, to catalog, control, and consume data from sources to any destination in a governed and managed way—giving organizations the ability to create and destroy the required infrastructures on demand. Zaloni's Arena is a distributive solution which addresses the challenges of cloud sprawl head-on.

Decoupling Storage from Compute

Most of the big data solutions based on Hadoop collocate compute with the storage. Although the distributed file system allows scalability, it's impossible to scale compute separately from storage, becoming even more apparent when implementing big data solutions in the cloud. As more and more organizations migrate their analytical workloads to the cloud, it is now possible to avail object stores like Amazon S3 and Azure ADLS, allowing independent compute and storage scaling.

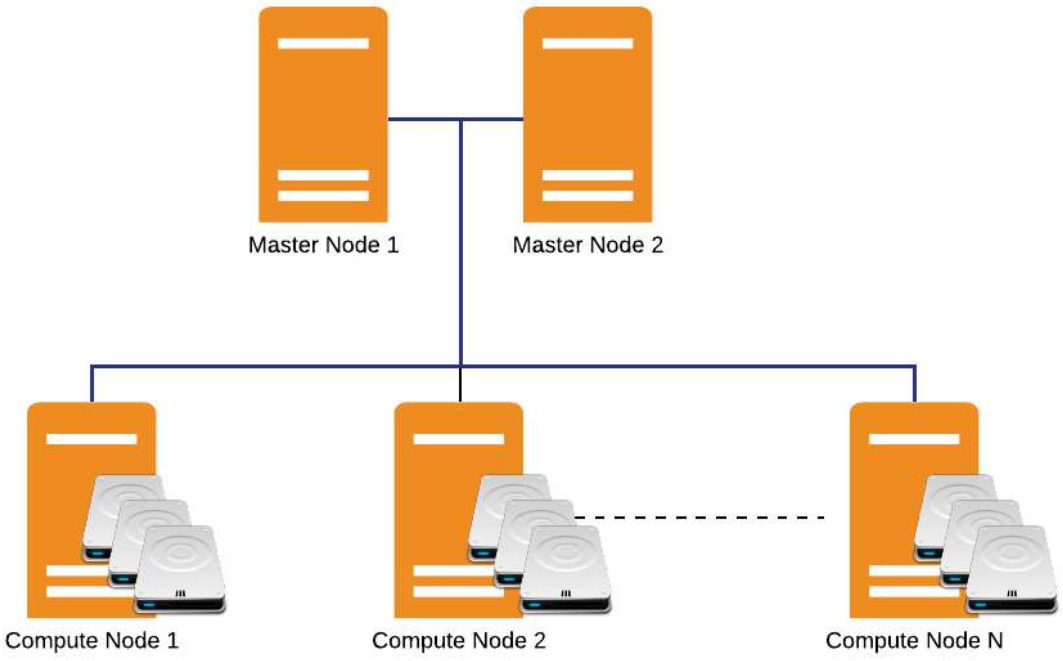


Fig. 1 Storage is Collocated within Compute Nodes

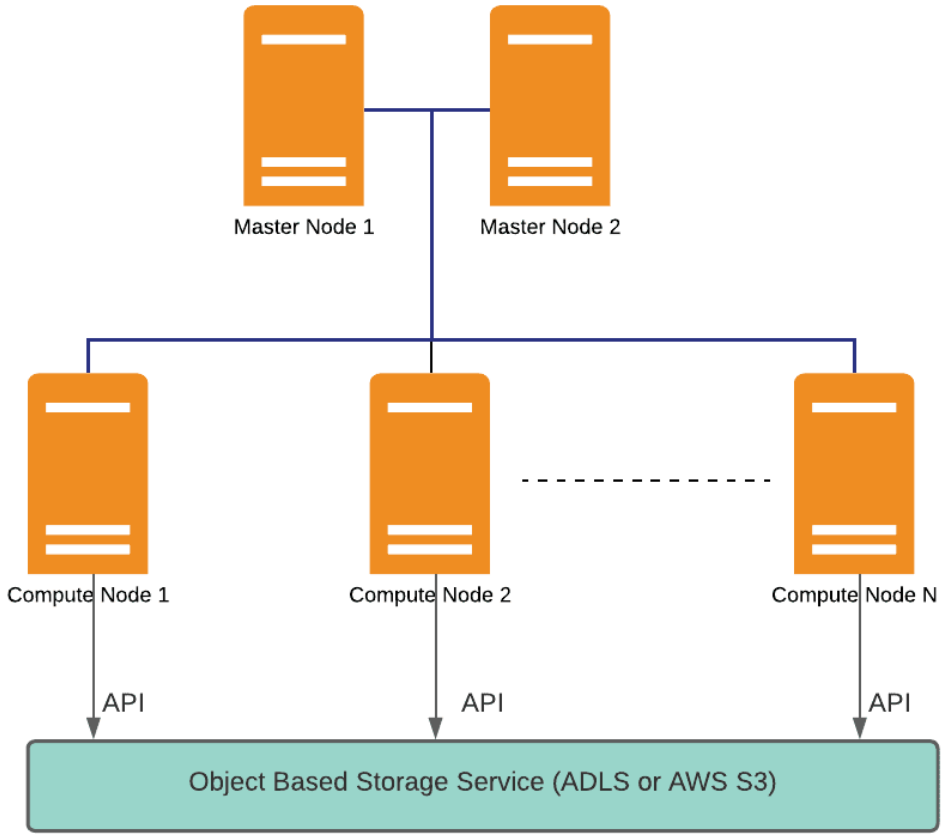


Fig. 2 Compute can be scaled separately from storage because the data is not stored within the compute nodes but on an object store service like ADLS or S3.

Elastic Compute Solutions using Azure

Now that we have been able to disassociate storage from compute nodes, let us look at a few options by which we can achieve elastic compute for big data workloads. The following are examples of implementing elastic compute in Azure:

HDInsight with Autoscaling

Microsoft Azure provides the capability to provision HDInsight clusters attached to Azure Data Lake Storage (ADLS). The HDInsight cluster can be scaled up and down via API calls to Azure services. Since the data is stored out of the cluster and in ADLS, it is possible to terminate the cluster without losing the data and recreating the compute cluster on demand. We do however, need to make sure the metadata is also stored in an external SQL database.

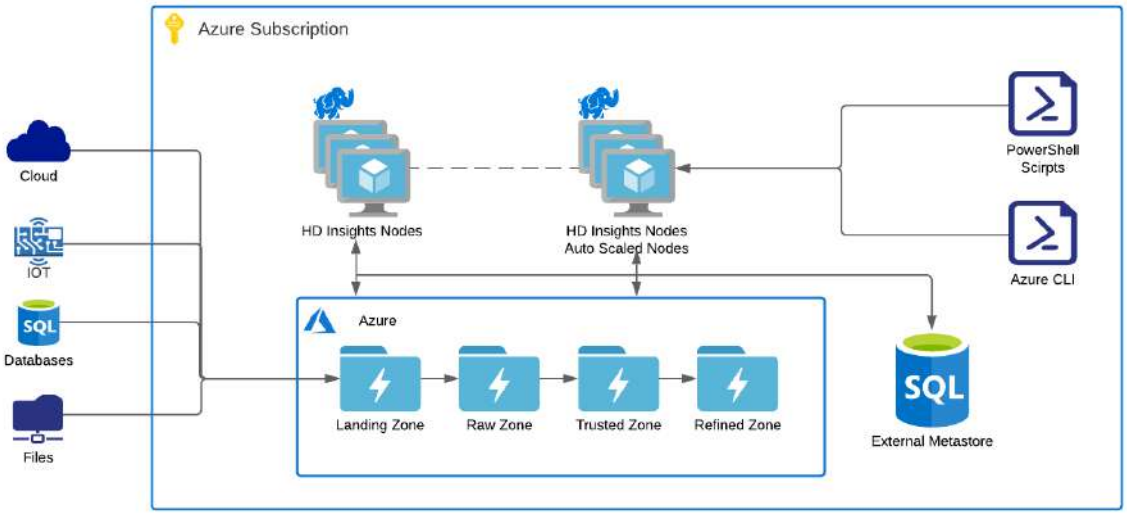


Fig. 3 HDInsight with Autoscaling

There are a few drawbacks in this scenario, as the HDInsight cluster is designed to be permanently running. It also takes time to spin up and spin down a cluster. On the other hand, if there are long running processes requiring the compute cluster to be available around the clock, this scenario will be a better fit.

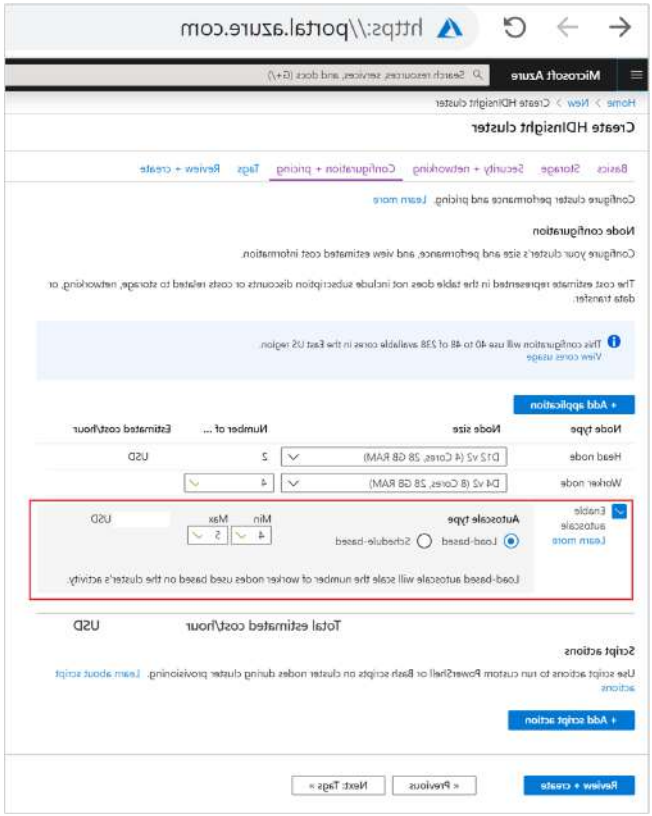


Fig. 4 Creating an HDInsight Cluster

Databricks with ADLS

Another approach is to use Databricks instead of running Spark on HDInsight. Azure Data Bricks does not require a permanently running cluster and the compute capacity can be instantiated on demand. Azure Databricks can be useful Data Science types of use cases where the experiments may require on-demand compute service when required and not necessarily a fully running cluster all the time.

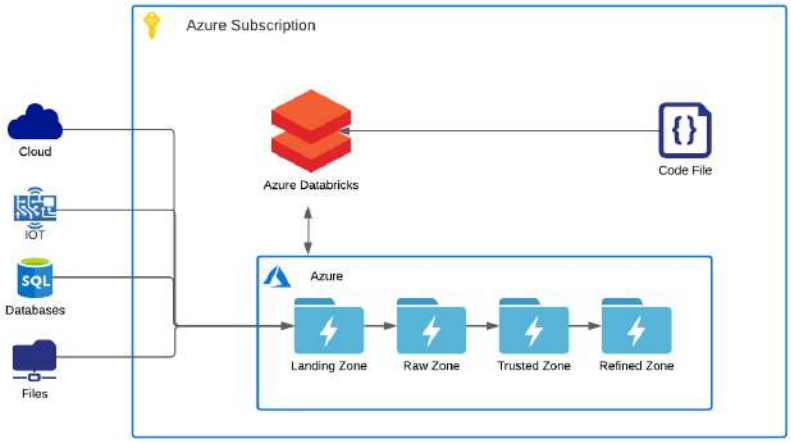


Fig.5 DataBricks with ADLS

Azure Kubernetes Services

Docker and Kubernetes bring yet another dimension to compute services. With Kubernetes, it's possible to package the processing logic in Docker-based applications, making it cloud independent with the elastic capabilities of the Kubernetes architecture.

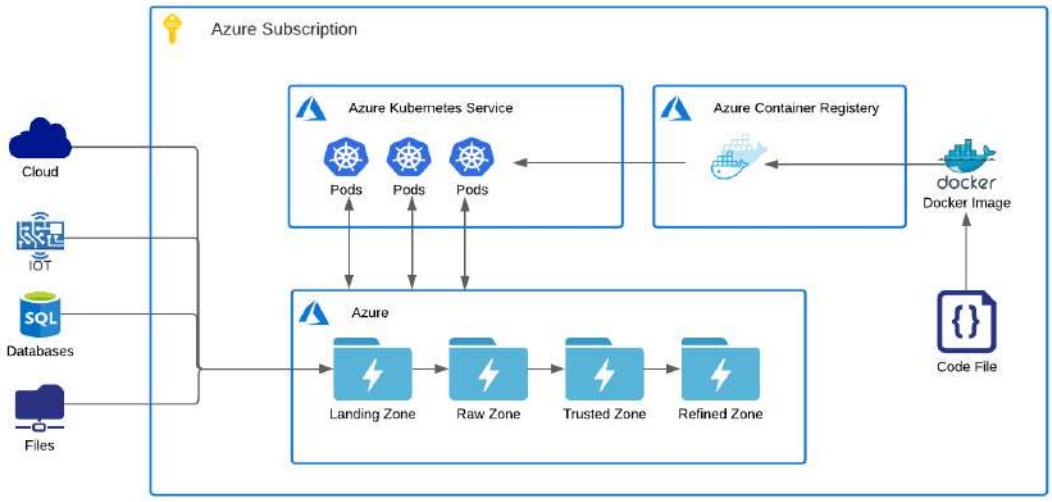


Fig. 6 Azure Kubernetes Services

Unlike Hadoop (HDInsight) or Spark (Databricks), Kubernetes allows the application developer to choose the language, libraries, and execution environment for each application and does not have to follow a particular stack necessarily. The application environment and the processing code is defined in Docker files and a Docker image is created. This image is then deployed to a Kubernetes cluster creating multiple execution pods. This method allows at scale processing of data while the data still resides within Azure Data Lake Storage.

Elastic Compute Solutions in AWS

The following are examples of implementing elastic compute in AWS:

AWS EMR AND S3 with Auto Scaling

The first option is to use the AWS flavor of Hadoop, AWS EMR. In order to achieve the separation between compute and storage we can use AWS S3 buckets for data storage and EMR to process the data. S3 can be mounted to EMR as additional storage. In this case the metadata can be stored in RDS service so that if the EMR cluster is destroyed and recreated the metadata is not lost. The cluster can be set to auto scaling the EMR cluster.

Using AWS Glue and Athena

Another option is to use Athena, Athena is a serverless service which queries S3 data without having to spin up dedicated compute nodes. It scales as per the requirements of the query language. It supports standard SQL. Athena is integrated with AWS Glue Catalog out of the box. Glue stores metadata information about the datasets in S3.

DataOps for Cloud Optimization

So how do you manage your data to support elastic computing and only use the cloud when needed? The data management practice of DataOps, brings together concepts from agile software development and DevOps to provide end-to-end visibility and control across your data environments and the supply chain.

Wikipedia defines DataOps as an automated process-oriented methodology, used by analytic and data teams, to improve the quality and reduce the cycle time of data analytics. By applying the concepts of agile development, DevOps, and data

management together we can start solving some of the most challenging cloud data problems.

With a DataOps platform, you can connect to your data, catalog the data, run data quality, create pipelines, then version and process precisely how you would manage a mobile application's development. DataOps also apply the concepts of continuous integration and delivery into data management.

Arena by Zaloni

Arena by Zaloni is a DataOps platform which includes an active data catalog, standardized governance and enables self service data consumption and enrichment. Using Arena's provisioning capabilities it is not only possible to provision data but also a compute service. With Arena you can bring the power of infrastructure as code, DevOps, and data management within the same platform.

Arena integrates with multiple cloud providers and on-premises data systems from a single control plane. The workflows in Arena allow dynamic provisioning of compute nodes in AWS and Azure. It is possible to use ARM Template or CloudFormation template to orchestrate the on demand deployment of any of the above scenarios. This brings you a multi-cluster and multi-cloud experience. A end-user such as a data analyst or data scientist can provision or lease datasets for his analysis using a self-service marketplace experience.

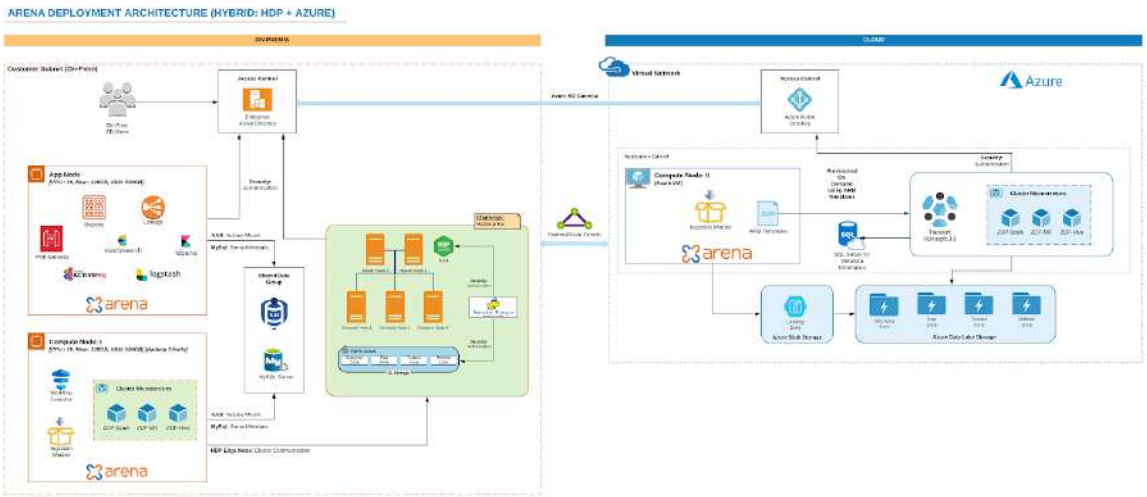


Fig. 7 Example of a Multi-Cluster Hybrid Deployment of the Arena Platform.

DataOps and Elastic Compute Use Cases:

Fighting Coronavirus with Data Management and Elastic Compute

Today we live in a world where we are facing new challenges every day. Let's use the example of one of the biggest challenges being faced by humanity today, the Coronavirus pandemic. Let's assume a government organization is tasked to contain and manage the spread of Coronavirus.

A telecommunications operator already has the big data stores in place. Now they need to create a special task force to extensively query the geo location data for people who came from overseas on a particular flight. This requires processing the data for each of the mobile devices which were active on the network for the last two weeks. In order to identify the potential spread, this needs to happen very quickly and requires extensive processing.

Procuring new hardware will take months to set up in an on-premise environment with traditional data processing. Even if this was all in the cloud, setting or extending an existing big data cluster has its own challenges of cost and maintainability. Plus you would still need to get budget approvals to expand in the cloud.

By leveraging a DataOps platform, like Arena, analysts are able to search and find data easily from the data catalog or marketplace experience then provision the data to a just-in-time, on demand elastic compute cluster in a self-service manner. This helps to reduce the time to insight while reducing costs through elastic compute.

In addition to the self-service data catalog and provisioning, Arena is able to add an approval process during data provisioning where the analyst submits an approval request providing clear business justification to access the data and spin up the cluster for a specific amount of time.

Providing Secure, External Access to Data

Let's assume you are an insurance provider and would like to share secured data with third-party research organizations across the globe but you do not want to give them access to your internal systems. You would like to lease the data along with the analytical tools to the third-party organization or subject matter experts for a specified period of time.

With Arena, you can provision the required data along with any compute or processing infrastructure to query the data in an isolated secured environment. Ability to get approvals from the dataset owners to share the data and then control how it is used and set the time limit on when the access should be revoked.

This capability within the Arena platform simplifies and secures external data sharing, enabling use cases such as external data marketplaces or sharing data with third-party organizations.

Conclusion

In the end, using on demand elastic compute along with a DataOps platform which can integrate with multiple cloud providers to enable elastic compute can make a significant impact on data security, time to insight and cost reduction.

CHAPTER - 3

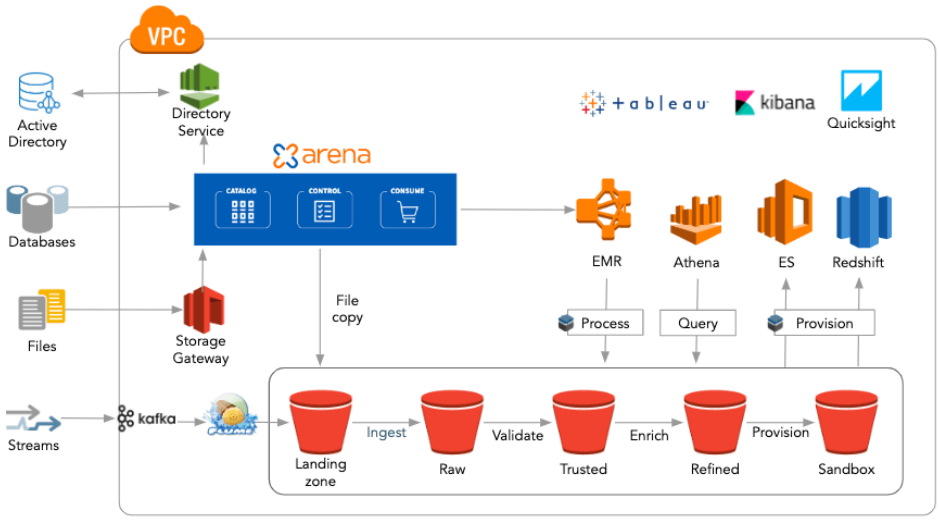
AWS Cloud Data Management: Conquer your Data Sprawl

Data and analytics success relies on providing data analysts and data scientists with quick, easy access to accurate, quality data. There's no better solution currently on the market to achieve this than Arena paired with AWS for better AWS cloud data management.

In a recent project, together with AWS, we helped the TMX Group (a Canadian financial services company that operates equities, fixed income, derivatives, and energy markets exchanges) manage their complex data sprawl into a consolidated and enriched self-service data catalog.

This allowed TMX Group to use their data for such cases as monetizing data for revenue growth and providing 360-degree customer views to improve customer experience and uncover cross-sell and up-sell opportunities.

Architecting for AWS Cloud Data



Zaloni Arena Solution Architecture for S3 Cloud Data Lake

When building a data lake on AWS, we recommend a zone-based architectural approach. This helps control how data is moved and processed while also providing governance and security controls through role-based access. This also provides data lineage that shows where data is coming from, where it's going and what's happened to it over time.

Understanding the data architecture is one thing, but what about actually deploying a data lake? How can you ensure success?

Data Lake Deployment Best Practices we Learned from TMX Group

1. Connect more data from more sources

Connecting to a variety of distributed and siloed data sources including cloud and on-prem data, and easily adding these sources to the catalog as they become available is essential to future-proofing your AWS data lake.

2. Catalog data for accurate, trusted, and repeatable use

To gain insights from your data, you need to know what data you have. A data catalog that focuses on automation with machine learning and artificial intelligence along with detailed and active metadata for easy consumption can help to get you answers fast so you can act accordingly.

3. Govern data for security and traceability

Data governance through role-based access control is critical for compliance with industry regulations around privacy and security along with masking and tokenization capabilities. With so much attention on protecting customer data, data governance is a must-have for any organization.

4. Provide business users with self-service AWS cloud data access

What good is a data catalog and data governance without allowing your business users access to the data they need? Granting self-service data access will allow them to see the

data they want, when they need it, without needing to request it from IT. That's a win-win!

Wish this chapter was more detailed? This was only a short overview of a much more in-depth version you can read on the AWS blog by visiting, https://aws.amazon.com/blogs/apn/turning-data-into-a-key-enterprise-asset-with-a-governed-data-lake-on-aws/.

CHAPTER - 4

Multi-Cloud Data Management: Greater Visibility, No Lock-In

Multi-Cloud Data Management of Today

Multi-cloud has evolved: what was once a relatively ad hoc practice of using public clouds to augment data storage and processing on an as-needed basis has now become a strategic focus for many enterprises. Companies are finding value in using multiple public clouds to manage costs and risk, and increase flexibility and agility. In fact, 84% of enterprises say they have a multi-cloud strategy in place, and 85% of senior IT decision-makers say they store data in two to five public clouds.

However, when you have "multi" anything, it brings complexity. A multi-cloud environment can result in siloed data and significant challenges for cohesive data management and governance – particularly when a multi-cloud environment has developed without a plan over time.

One global survey found that while more than 90% of respondents said they originally believed the public cloud would help simplify operations and provide greater data insights, the reality has been a different story. Of those who felt let down by the promise of the public cloud, 91% believed the problem was fragmented data in and across public clouds that would become extremely difficult to manage over time.

Why multi-cloud?

The appeal of multi-cloud ultimately comes down to minimizing costs and reducing risk. Companies use multiple public clouds to optimize costs by deploying workloads on different cloud providers as needed to be more efficient and/or leverage preferred tools. The use of multiple clouds also protects against the risk of vendor lock-in, which could have major business repercussions.

As the top cloud vendors – Amazon Web Services/AWS, Google Cloud Platform, and Microsoft Azure – continue to diversify (e.g., will Amazon make a foray into banking?), relationships with their customers may become more complicated – perhaps even competitive or a conflict of interest. Also, vendor lock-in can affect what cloud tools and services a company is able to use.

Even if a company isn't planning on implementing a multi-cloud strategy, multi-cloud can become a six-month-plus reality if a decision is made to transition from one cloud provider to another, or as a result of an acquisition or merger.

Preparing for a multi-cloud world

The many paths to multi-cloud all lead to the same need: a way to consistently and efficiently manage and govern all data, no matter where it resides – without disrupting the business or requiring huge investment to rip and replace systems and tools. To help evaluate your current needs, consider:

- Can you create a virtual "data lake" without having to move data from where it currently resides?
- Is your IT team spending too much time managing data and could automation help lighten the load?
- Are you at risk of vendor lock-in or hindered in any way from leveraging multiple cloud services or preferred tools?
- Can your business users easily discover and use data from across the enterprise (i.e., not only across public clouds but on-premises as well)?
- Do you feel confident that you will be able to scale the management and governance of your data into the future?

Multi-cloud data management platform essentials

What is the best solution in order to reduce complexity, not increase it? When working to determine what data management platform makes sense for a multi-cloud environment, here are some key features to consider:

1. Centralized visibility and governance

Multi-cloud makes "vendor agnostic" a must-have feature of a data management

platform. A platform that can talk to any system or tool provides the flexible connections that unify and integrate siloed data for centralized visibility and data governance – without the need for your IT team to move data around in order to find, manage and govern it. Centralized control using one, comprehensive platform allows administrators to set up consistent governance rules for data quality, compliance, privacy, and security.

From a business user perspective, a centralized platform that can crawl and recognize data across all cloud providers and create a self-service catalog that's a single source of truth can significantly accelerate time to insight.

2. Automation and machine learning

Unlike traditional data catalogs that simply provide an inventory of data, modern data catalogs that leverage machine learning and AI help reduce the complexity of multi-cloud by automating workflows along the entire data pipeline, from ingestion to provisioning.

Automation of repeatable tasks, such as metadata collection and data profiling, not only speeds up workflows and frees up team member time, it reduces errors and ensures consistent governance across multi-cloud data sources for data quality (change capture, duplicate identification), data privacy (tokenization and masking), as well as data security (enterprise-wide, role-based access rules).

3. Hybrid and cloud flexibility

The point of using the public cloud is to leverage its storage and processing power to reduce costs and increase speed; you need a data management platform that can flexibly manage this across cloud providers.

There are many data management platforms capable of data orchestration within a single cloud provider. What's rarer is a data management platform for multi-cloud – enabling governance in cloud-native storage, as well as flexible data transformations and delivery across multiple clouds (and hybrid environments, let's be real), without companies having to give up existing systems and tools.

For example, a platform for multi-cloud should automate repeatable, on-demand native processing, depending on where the data resides – if the data is in AWS, the platform should spin up processing power on AWS. In the near future, some platforms will

automate moving data between cloud providers to achieve even better efficiencies – so seamlessly that the business end-user won't even know.

A multi-cloud strategy can make a lot of sense for businesses – but it requires having the right data management and governance platform in place to simplify multi-cloud environments for both administrators and business users.

CHAPTER - 5

Migrating On-Premises Data Lakes to Cloud

In a previous article, we discussed some of the key drivers for a cloud data lake, such as the cost advantages of the elastic utility model of cloud and lower administrative and operational costs. There is also access to a range of compute and storage options beyond Hadoop, as well as advanced cloud services, geographical coverage and data availability guarantees

But how do enterprises that already have an on-premises data lake migrate to the cloud to realize those benefits?

Cloud Migration Objectives

Every cloud migration project has to begin with a clear statement of business as well as technical objectives. Cost reduction without loss of service levels and same or superior user experience tends to be the top business objective. While the current cost of on-premises data platform may be known, quantifying future costs of a data lake in the cloud can be done only in the context of architectural decisions made after sorting through and picking from a bewildering array of options across cloud providers.

Business objectives further clarify the scope and time frame for the migration, compliance requirements with respect to data security, physical location and longevity of data, and business continuity needs during and after migration. Scope has to do with which on-premises data sets, from which enterprise functions or departments, from which on-premises Hadoop clusters and from which data centers, will migrate to the cloud data lake. Compliance requirements are particularly important to highly regulated industries like healthcare and banking. It goes hand in hand with security needs founded on a well-researched threat model. Business continuity needs determine which are the critical applications that cannot tolerate any downtime during migration.

Technical objectives lay down the use cases that the data will be subjected to by a

variety of users. These users may belong to different business functions and come with varying skill sets and data access and processing needs. The enterprise on-premises data governance practice may have been spread across the infrastructure, application and security teams. But that enterprise governance model will change in the cloud based on the choice of Identity and access management services and tools. Similarly, the data security objectives will determine how data will be secured in the cloud while in motion and at rest. Technology choices in the cloud, in addition, may be guided by the requirements around feature parity and data processing performance by comparison with the on-premises data lake.

A data lake management application layer greatly facilitates the realization of the business and technical objectives. It does this by abstracting from the user the underlying data platform technologies, whether on-premises or in the cloud, and by providing a common metadata view.

Cloud Technology Choices and Migration Design

While there are many technology choices and cloud providers, we see three broad models of Data Lake migration to the cloud:

- Forklift migration of on-premises Hadoop cluster to cloud
- Migration to use Hadoop based cloud services and cloud-native storage
- Migration to a hybrid on-premises/cloud model, using a variety of cloud-native services, and establishing a seamless data fabric view with metadata

These are also reflective of the increasing levels of maturity of cloud data lake adoption. There are of course variations of these models making more or less use of cloud elasticity with the help of a management framework.

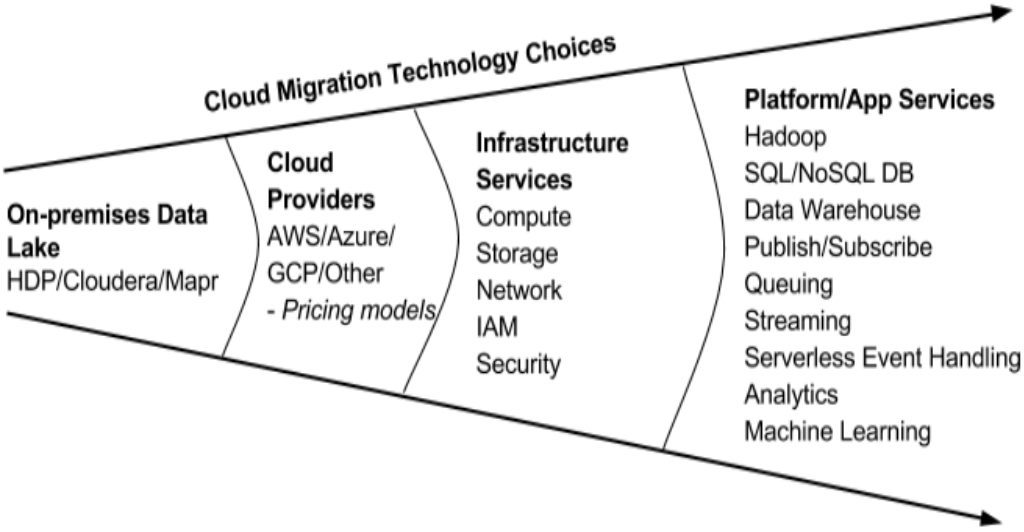
Forklift migration refers to moving an on-premises Hadoop cluster to one built ground up from basic compute instances in the cloud. This is the simplest migration model leveraging existing staff skill sets. It uses only the IaaS aspect of cloud with persistent compute instances, typically with instance local storage. Except for infrastructure access, security is entirely the cloud customer's responsibility, as is the creation, configuration, monitoring and maintenance of the cluster.

Moving from Hadoop on-premises to using Hadoop as a service from the cloud provider is the second model of migration. Much of the heavy lifting around Hadoop cluster setup

and configuration, and ensuring compatibility of Hadoop ecosystem components is left to the cloud provider. A data lake management application may aid in the creation and use of transient Hadoop clusters on demand and interface directly to cloud-native persistent storage.

The third model of data lake migration involves a gradual transition from Hadoop on-premises to hybrid architectures – on-premises/cloud, using a variety of cloud-native storage options and services in addition to the Hadoop ecosystem tools, adopting cloud service patterns for processing event streams, real-time analytics, and machine learning. This model presupposes a metadata management layer to remove any mismatch between the underlying technologies and provide a seamless data fabric view across all the data regardless of storage location.

Between the three aforementioned migration models, the major Hadoop distributions (Cloudera, Hortonworks, MapR), the ever-expanding Hadoop ecosystem tool variations they support, and the big three cloud service providers (AWS, Azure, GCP) each with unique service offerings and pricing, the options for migration are too numerous to list here. Meaningful comparisons will need to be done in the context of specific business and technical requirements.



Cloud Migration Technology choices

A good migration design requires deep expertise in Data Lake and cloud technologies, and data pipeline design patterns, either developed internally or bought from a service provider.

Migration Planning and Execution

Data Lake migration planning typically starts with a proof of concept pilot to validate the technical choices, feature parity, and performance in the cloud. This is followed by a phased approach consistent with the chosen migration model that takes into account:

- Infrastructure migration decisions – storage and compute, sizing, scaling, networking
- Security of data and governance of data access, and resource usage in the cloud
- Retooling data ingestion for sending to the cloud data lake data that is currently received by the on-premises platform from different sources
- Detailed inventory of on-premises data lake, and mapping to cloud platform
- Data transformation pipelines and corresponding translation to cloud mechanisms
- Application migration – forklift vs rewrite, processes for development, test, and production
- Data extraction tools and processes in the cloud for visualization, insights, or predictions
- Migration options for historical data
- Versions of cloud tools and application compatibility
- Data Lake management applications

An execution plan which defines the transition process from on-premises to the Cloud data lake, testing, performance monitoring, and business continuity during and after the cutover, are critical to a successful migration.

The benefits of migrating a data lake from on-premises to the cloud are achieved only through a careful specification of business and technical objectives, a validated set of migration design choices, planning and phased execution. A metadata management application layer is invaluable during the transition as well as for future proofing the data lake solution in the cloud.

Find Your Data Success With Zaloni

Learn more
About Zaloni

www.zaloni.com

About Zaloni

At Zaloni, we believe in the unrealized power of data. Our DataOps software platform, Arena, streamlines data pipelines through an augmented catalog, automated governance, and self-service consumption to reduce IT costs, accelerate analytics, and standardize security. We work with the world's leading companies, delivering exceptional data governance built on an extensible, machine-learning platform that both improves and safeguards enterprises' data assets.