Morris & Opazo

aws partner network

Advanced Consulting Partner

Data & Analytics Competency

Channel Partner

EBOOK:

Diving Deep Into Data Lakes

We Love the Cloud

www.morrisopazo.com contacto@morrisopazo.com

Contents

What is a Data Lake?

Organizations are tasked with managing greater volumes of data, from more sources, and containing more types of data than ever before. In the face of massive, heterogeneous volumes of data, many organizations are finding that to deliver timely business insights, they need a storage and analytics solution that offers more speed and flexibility than legacy systems. Enterprises do not only need increasing analytics functions but also require secure data access, making data governance easier and more efficient. Any environment that includes digital elements in it should also be considered a data-related environment, because at some point it will produce or consume data in whether structured or unstructured format.

A Data Lake is a new and increasingly popular way to store and analyze data that addresses many of these challenges by allowing an organization to store all of their data in one, centralized repository. It is not necessary to define structure nor rules until the data is required. Since data can be stored in its original form, there is no need to convert it to a predefined schema before ingestion, giving you a way to store all of your data, both structured and unstructured, with minimal lead time.

With a Data Lake on AWS, you no longer need to know what questions you want to ask of your data before you store it, giving you a flexible platform for data analysis. At the heart of an AWS-based Data Lake is Amazon Simple Storage Service (S3), which provides secure, cost- effective, durable, and scalable cloud storage. AWS also offers an extensive set of services to help you provide strong security for your Data Lake, including access controls, virtual isolation, 256-bit encryption, logging and monitoring, and more.

There's a variety of ways to transfer data to your Data Lake, including through services such as Amazon Migration Services (heterogeneous migrations between different database platforms), Amazon Kinesis, which enables you to ingest data in real-time, AWS Import/ Export, a service where you can send a portable storage device with your data to AWS, AWS Import/Export Snowball, a secure appliance AWS sends you for ingesting data in batches, AWS Storage Gateway, which enables you to connect on-premises software appliances with your AWS Cloud-base storage, or through AWS Direct Connect, which gives you dedicated network connectivity between your data center and AWS.

> *"If you think of a data mart as a store of bottled water – cleansed and packaged and structured for easy consumption – the data lake is a large body of water in a more natural state. The contents of the data lake stream in from a source to fill the lake, and various users of the lake can come to examine, dive in, or take samples."*
>
> ***James Dixon, CTO, Pentaho (2010)***

> "With a Data Lake on AWS, you no longer need to know what questions you want to ask of your data before you store it, giving you a flexible platform for data analysis."

Value of a Data Lake

A company aware of properly implement a data lake increases its probabilities of getting the most out of their related analytics and transforming them into valuable ROI. Therefore it is extremely important to build a Data Lake using the appropriate technologies and techniques available, supported by a strong Architecture.

There are a variety of benefits to hosting your Data Lake on AWS, including:

Cost-Effective Data Storage

Amazon S3 provides cost-effective and durable storage, allowing you to store nearly unlimited amounts of data of any type, from any source. Because storing data in Amazon S3 doesn't require upfront transformations, you have the flexibility to apply schemas for data analysis on demand. This enables you to more easily answer new questions as they come up and improve the time-to-value. Having an object-based storage expands the potential use cases for a data lake.

Easy Data Collection and Ingestion

There's a variety of ways to ingest data into your Data Lake, including services such as Amazon Kinesis, which enables you to ingest data in real-time; AWS Snowball, a secure appliance AWS sends you for ingesting data in batches; AWS Storage Gateway, which enables you to connect on-premises software appliances with your AWS Cloud-based storage; or through AWS Direct Connect, which gives you dedicated network connectivity between your data center and AWS.

Security and Compliance

When hosting your Data Lake on AWS, you gain access to a highly secure cloud infrastructure and a deep suite of security offerings designed to keep your data secure. As an AWS customer, you will benefit from a data center and network architecture built to meet the requirements of the most security-sensitive organizations. AWS also actively manages dozens of compliance programs in its infrastructure, helping organizations to easily meet compliance standards such as PCI DSS, HIPAA, and FedRAMP.

Most Complete Platform for Big Data

AWS gives you fast access to flexible and low cost IT resources, so you can rapidly scale virtually any big data application including data warehousing, clickstream analytics, fraud detection, recommendation engines, event-driven ETL, serverless computing, and Internet of Things processing. With AWS, you don't need to make large, upfront investments in time and money to build and maintain infrastructure. Instead, you can provision exactly the right type and size of resources you need to power big data analytics applications.

Considerations for Building a Data Lake

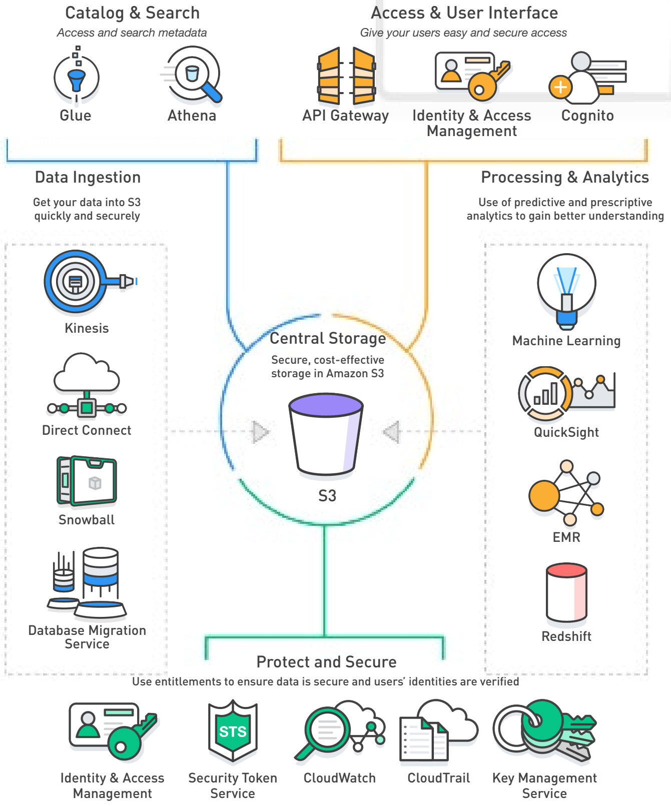
Data Lakes enable to properly separate storage and compute. A data lake solution on AWS, at its core, leverages Amazon Simple Storage Service (Amazon S3) for secure, cost-effective, durable, and scalable storage. You can quickly and easily collect data into Amazon S3 from a wide variety of sources by using services like AWS Snowball or Amazon Kinesis Firehose delivery streams. Amazon S3 also offers an extensive set of features to help you provide strong security for your data lake, including access controls & policies, data transfer over SSL, encryption at rest, logging and monitoring, and more.

For the management of the data, you can leverage services such as Amazon DynamoDB and Amazon ElasticSearch to catalog and index the data in Amazon S3. Using AWS Lambda functions that are directly triggered to Amazon S3 in response to events such as new data being uploaded, you easily can keep your catalog up to date. With Amazon API Gateway, you can create an API that acts as a "front door" for applications to access data quickly and securely by authorizing access via AWS Identity and Access Management (IAM) and Amazon Cognito.

For analyzing and accessing the data stored in Amazon S3, AWS provides fast access to flexible and low cost services like Amazon EMR, Amazon Redshift, and Amazon Machine Learning, so you can rapidly scale any analytical solution. Example solutions include data warehousing, clickstream analytics, fraud detection, recommendation engines, event-driven ETL, and internet-of-things processing. By leveraging AWS, you can easily provision exactly the resources and scale you need to power any big data applications, meet demand, and improve innovation.

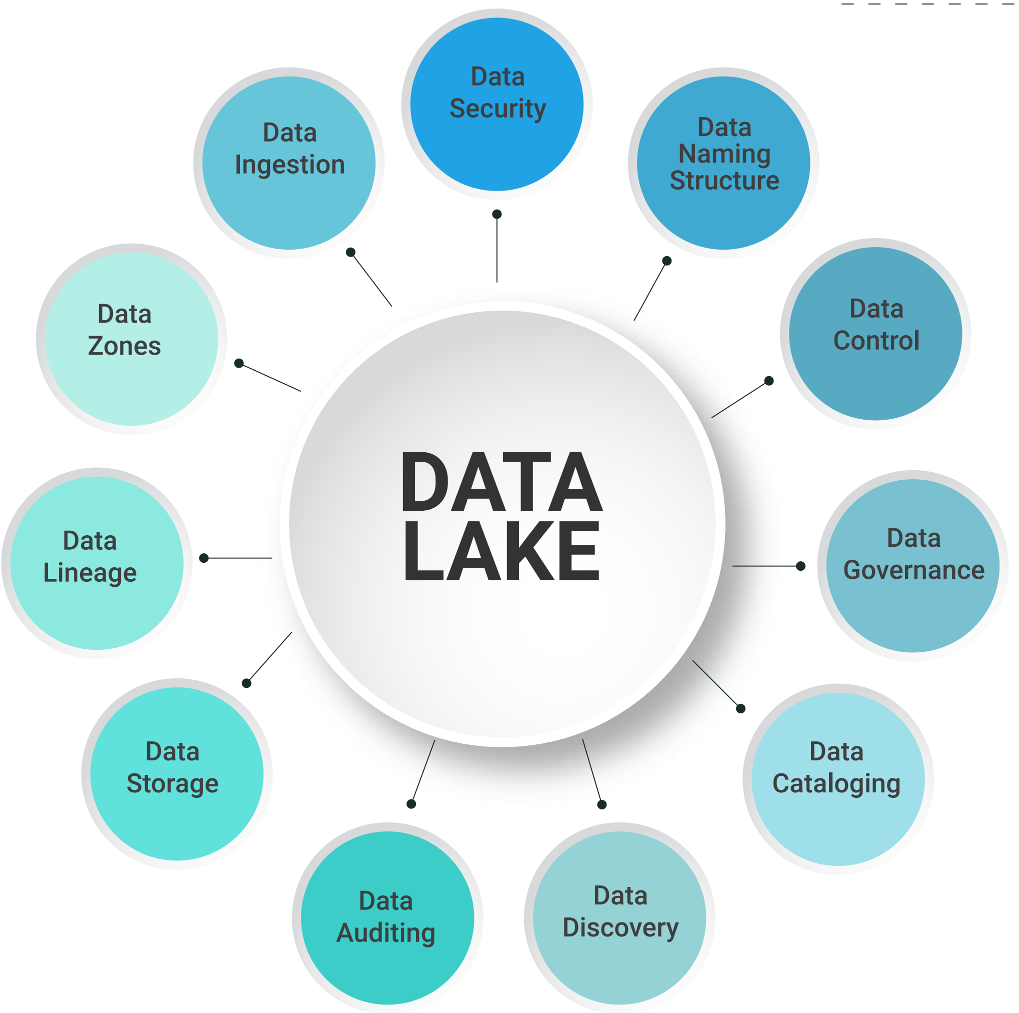
Regarding security of data, it is suggested to consider:

- Network isolation.
- Control access based on roles, at the most suitable level for each case.
- Masking and de-identification of data.
- End-to-end encryption

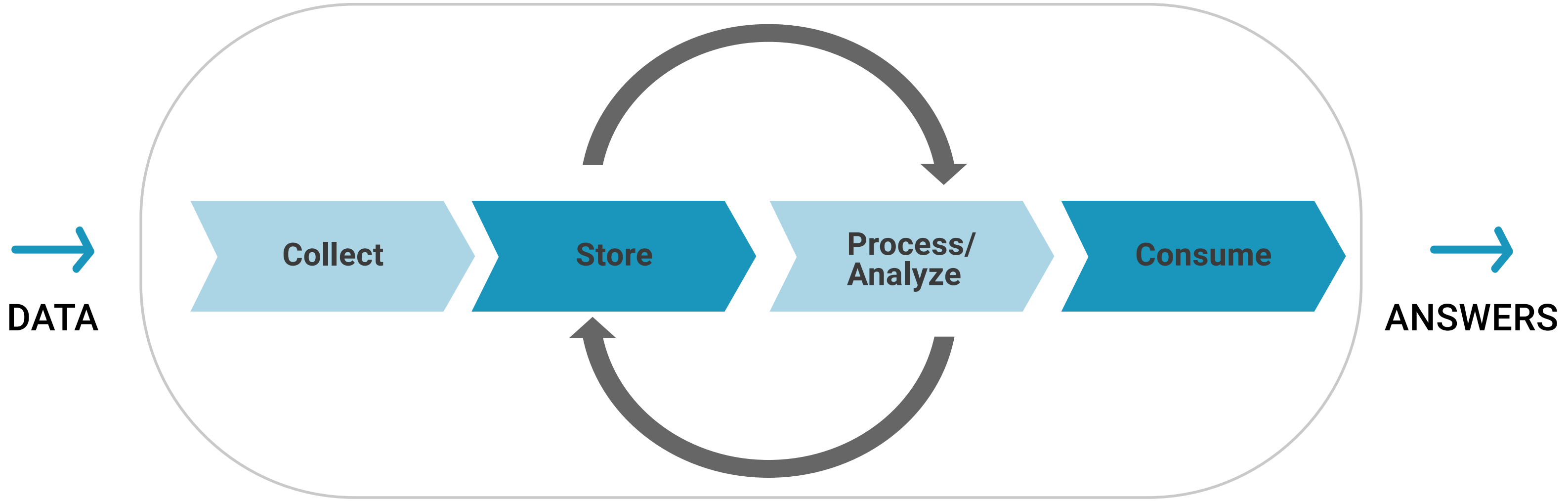

Catalog & Search
Access and search metadata
Glue
Athena
Access & User Interface
Give your users easy and secure access
API Gateway
Identity & Access Management
Cognito
Data Ingestion
Get your data into S3 quickly and securely
Kinesis
Direct Connect
Snowball
Database Migration Service
Central Storage
Secure, cost-effective storage in Amazon S3
S3
Processing & Analytics
Use of predictive and prescriptive analytics to gain better understanding
Machine Learning
QuickSight
EMR
Redshift
Protect and Secure
Use entitlements to ensure data is secure and users' identities are verified
Identity & Access Management
STS
Security Token Service
CloudWatch
CloudTrail
Key Management Service

Components of a Data Lake

- > Data Workflows
- > Data Ingestion-Batch
- > Data Storage
- > Data Lake Zones
- > DataTransformation
- > Data Control: Data Status / Tracking
- > Data Governance
- > Data Cataloging
- > HCatalog with AWS Glue
- > Exhaustive Data Catalog
- > Organization and Management of Data Lake

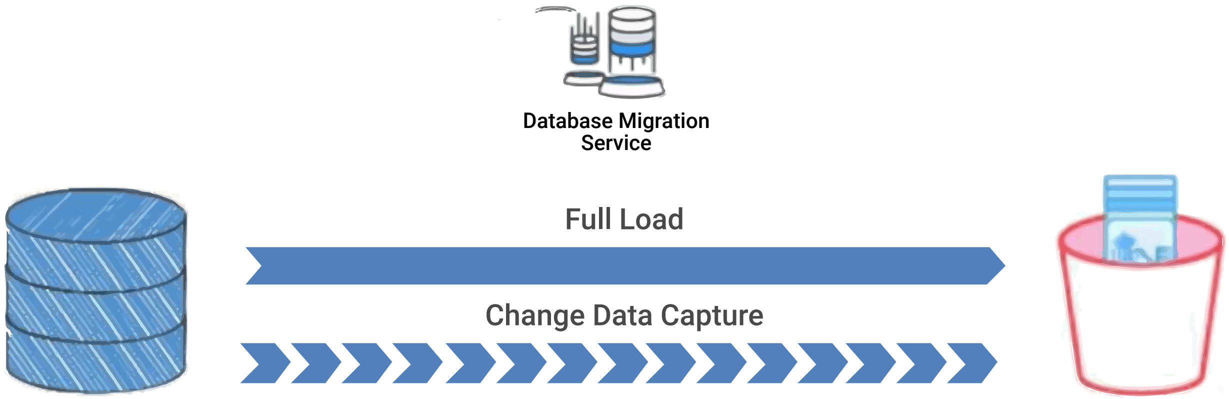


Data Workflows



Collect	Store	Process and Analyze	Consume and visualize
Collect raw data, such as transactions, records, mobile devices and more. Allow developers to receive a wide variety of data (structured and unstructured) at any speed, in real time or in batches.	A secure, scalable and durable repository is required to store the data before or even after processing. They could be temporary warehouses for data in transit.	In this step, the data is transformed from raw data to consumable data, usually by sorting, accumulating, joining and even performing more advanced functions and algorithms.	Access to data through data visualization tools and self-service business intelligence that allow you to explore the datasets quickly and easily. They can also be statistical "predictions" (predictive analysis) or recommended actions (prescriptive analysis).

Data Ingestion: Batch



<schema_name>/ <table_name>/ LOAD001. csv
<schema_name>/ <table_name>/ LOAD001. csv
<schema_name>/ <table_name>/ <time.stamp>.

It is very important to keep the Data Lake consistent with the data changes at the sources. Therefore the relevance of a good 'teaming' between the Data Collector (element in charge of pulling the data from a source), and the Data Integrator (ingests the data into the Data Lake).

Data Storage

S3 as the Factory of the Data Lake

> **Unlimited** objects number and volume

> 99.99% **availability**

> 99.999999999% **durability**

> Versioning

> **Storage by layers** through lifecycle policies.

> **SSL, Data Encryption** at the client and server at rest.

> **Low cost** (a bit lower than US $23 per month per 1TB).

> **Natively supported** by Big Data frameworks (Spark, Hive, Presto, etc.)

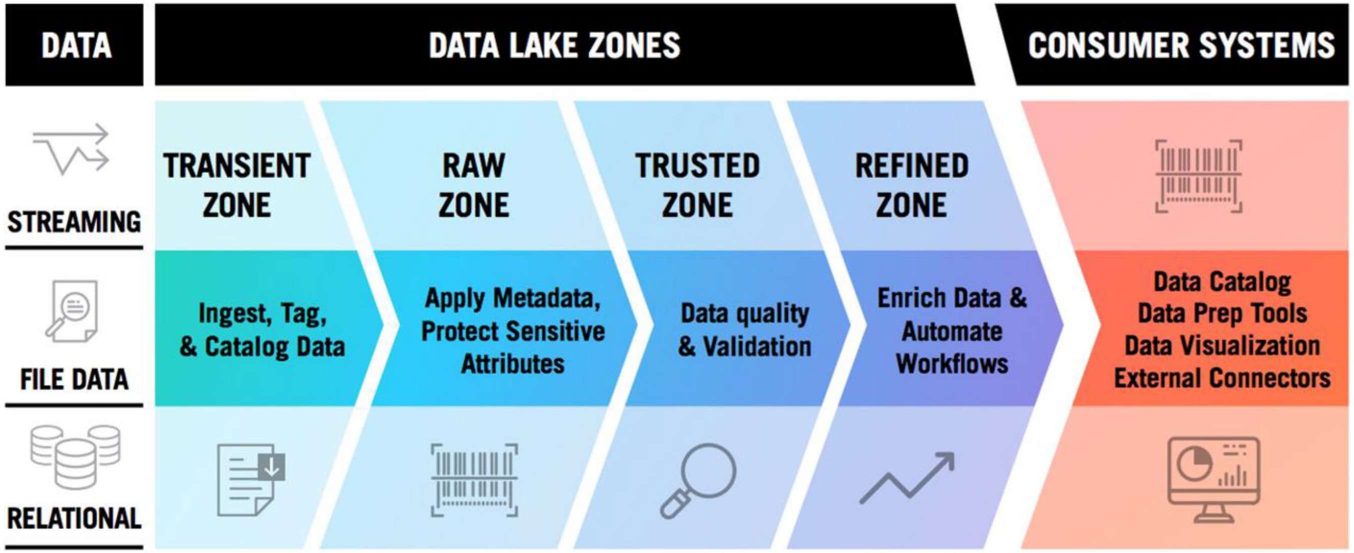
> **Decouples** storage and compute, optimizing costs, speeding up innovation, and enabling experimentation with various processing technologies.

> **Events** and **lifecycle** management.

> Data **protection.**

> Data management with objects **labeling.**

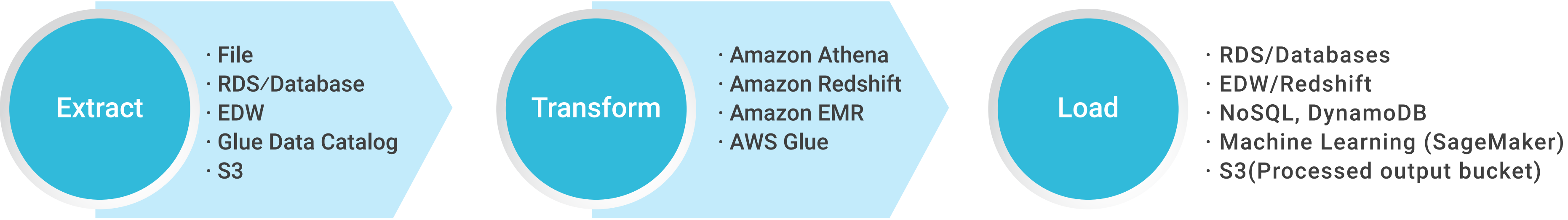
Data Lake Zones



It is recommended to use a separated quarantine zone to remove PII and sensitive data before moving it to the raw data zone.

Data Transformation

Extract Transform Load in AWS



Extract

- File
- RDS/Database
- EDW
- Glue Data Catalog
- S3

Transform

- Amazon Athena
- Amazon Redshift
- Amazon EMR
- AWS Glue

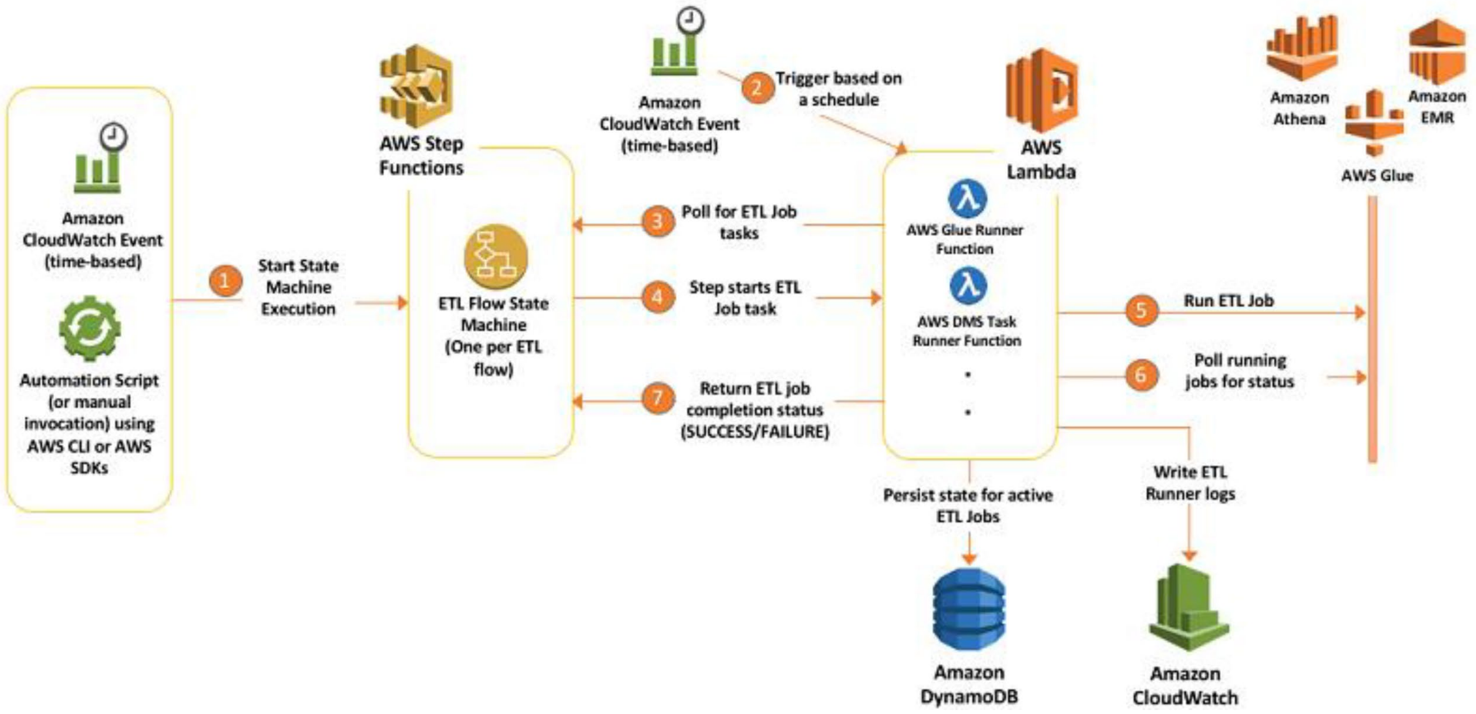
Load

- RDS/Databases
- EDW/Redshift
- NoSQL, DynamoDB
- Machine Learning (SageMaker)
- S3(Processed output bucket)

It is important to know where data comes from, what transformations are applied to it, and what different areas it moves from/to. This information could help engineers to solve possible issues during workloads.

Data Control: Data Status / Tracking

The ETL orchestration architecture and events



Amazon CloudWatch Event (time-based)
Automation Script (or manual invocation) using AWS CLI or AWS SDKs
1
Start State Machine Execution
AWS Step Functions
ETL Flow State Machine (One per ETL flow)
Amazon CloudWatch Event (time-based)
2
Trigger based on a schedule
3
Poll for ETL Job tasks
4
Step starts ETL Job task
7
Return ETL job completion status (SUCCESS/FAILURE)
AWS Lambda
AWS Glue Runner Function
AWS DMS Task Runner Function
Amazon Athena
Amazon EMR
AWS Glue
5
Run ETL Job
6
Poll running jobs for status
Persist state for active ETL Jobs
Write ETL Runner logs
Amazon DynamoDB
Amazon CloudWatch

Data Governance

Pillars of Data Governance

- **Data Catalog =** What data is available and where it is stored?
- **Data Lineage =** Where has data come from and what has happened to it?
- **Data Quality =** Is data accurate and fit for purpose?
- **Data Security =** Is access to data securely protected?

A solution with a de-identified data lake (DIDL) helps companies get to the main cause of risk in relation to data architectures and PII protection. A DIDL may help to discover, identify, catalogue, monitor and protect data.

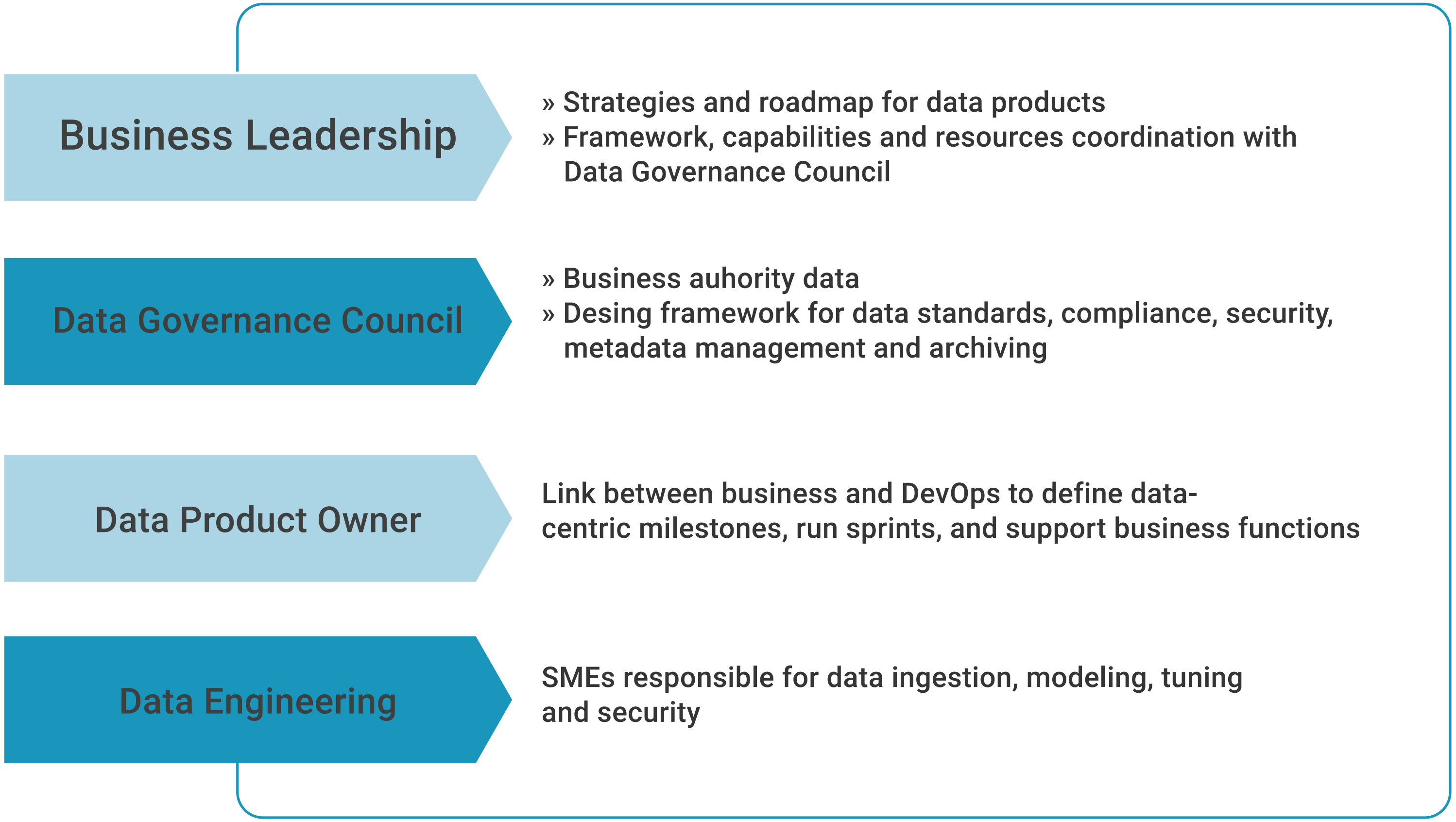
Advantages of AWS Data Governance for Data and Analytics

- **Data Catalog:** A data catalog management system that monitors each asset in the data lake and provides data stewards with the ability to manage access to data assets.
- **ETL:** services to Extract, Transform and Load that integrate with policies-based masking services.
- **Masking:** A policies-based solution that extracts and mask sensitive PII data even before they get to a data lake.
- **Matching and Unidentified Data Transfer:** Transfer with security third-party data using a centralized trust model

There are **several AWS services** with a particular meaning for clients focused on meeting the GDPR, including:

- **Amazon GuardDuty:** a security service that incorporates smart threats detection and constant monitoring.
- **Amazon Macie:** machine learning tool that helps to discover and secure personal data stored in S3.
- **Amazon Inspector:** automated security advisory service to help keep applications in compliance with best security practices.
- **AWS Config Rules:** monitoring service that dynamically verifies the cloud resources for compliance with security rules.

Data Governance Structural Organization



Business Leadership

» Strategies and roadmap for data products
» Framework, capabilities and resources coordination with Data Governance Council

Data Governance Council

» Business auhority data
» Desing framework for data standards, compliance, security, metadata management and archiving

Data Product Owner

Link between business and DevOps to define data-centric milestones, run sprints, and support business functions

Data Engineering

SMEs responsible for data ingestion, modeling, tuning and security

Data Governance Council

Responsible for reviewing data patterns and practices, to make Data Lake follows secure, steady and sustainable approaches. It should have representatives from all the business areas and IT, in order to approve policies, practices and standards for the program team.

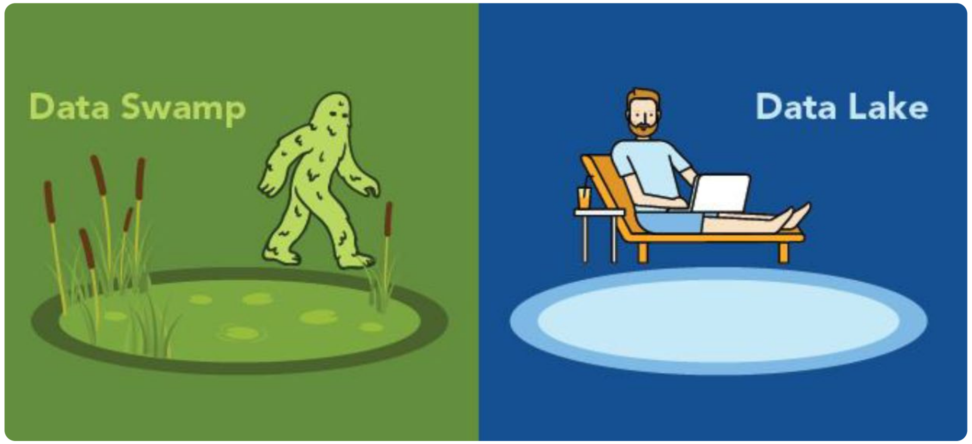
There should be a Data Governance Council for each business line and key functional area, considering that they must verify data sources, types and conversion (including other tasks) for successful Data Lake operations.

Data Councils should confirm authoritative sources and execute contracts between data owners and consumers.

Data Swamps

Among the multiple factors that benefit or affect the efficiency of a Data Lake, metadata management plays a key role.

If there is not a proper descriptive metadata and its corresponding management, the Data Lake turns into a set of disconnected data pools all in one single place.



Data Catalog

The data catalog provides a query interface for all the assets stored in the S3 buckets of a data lake. It allows to automatically crawl and compile both metadata and index data sets within a data lake, allowing them to be searchable. The data catalog is designed to provide a single source of truth about the data lake content. It also can be used for audit purposes or to dynamically drive data transformations.

.

There are two general forms of a data catalog:

Comprehensive Data Catalog: It contains information about all the assets that have been stored in the S3 datalake.

HCatalog with AWS Glue: It contains information on all assets that have been transformed into table formats and definitions that are usable by analytical tools such as Amazon Athena, Amazon Redshift, Amazon Redshift Spectrum, and Amazon EMR.

The two catalogs are not mutually exclusive and both can exist. The exhaustive catalog of data can be used to search all the assets in the data lake, and the HCatalog can be used to discover and consult data assets in the data lake.

Hcatalog with AWS Glue

Compatible with Hive metastore, metadata repository highly available:

> **Search** of metadata for data discovery

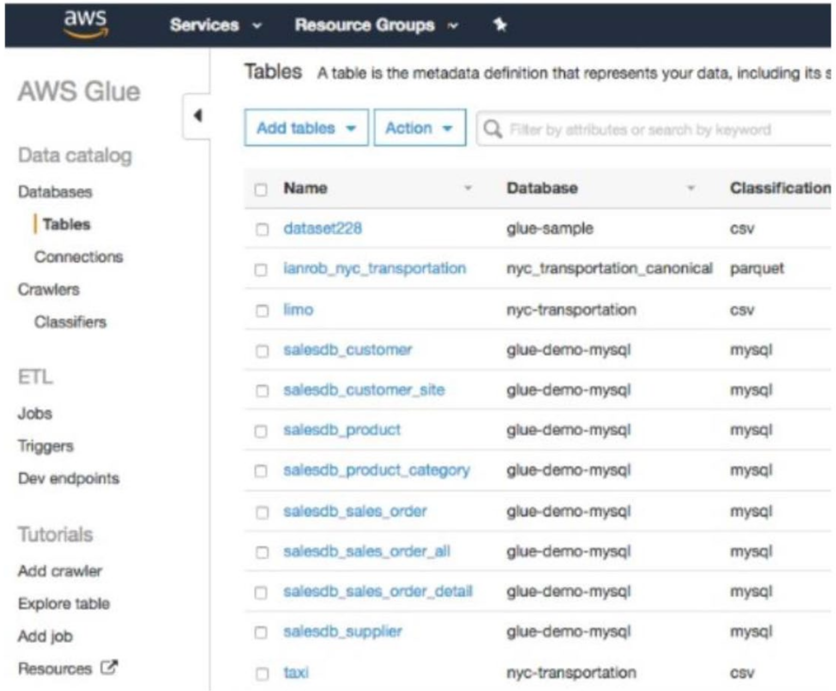
> **Connection information** - JDBCURLs,credentials

> **Classification** to identify and parsefiles.

> **Versioning** of tables metadata as schemas evolve and other metadata are updated

> **Tables definition** - usablebyRedshift,Athena,Glue,EMR.

Complete using Hive DDL, bulk import or automatically through **crawlers.**

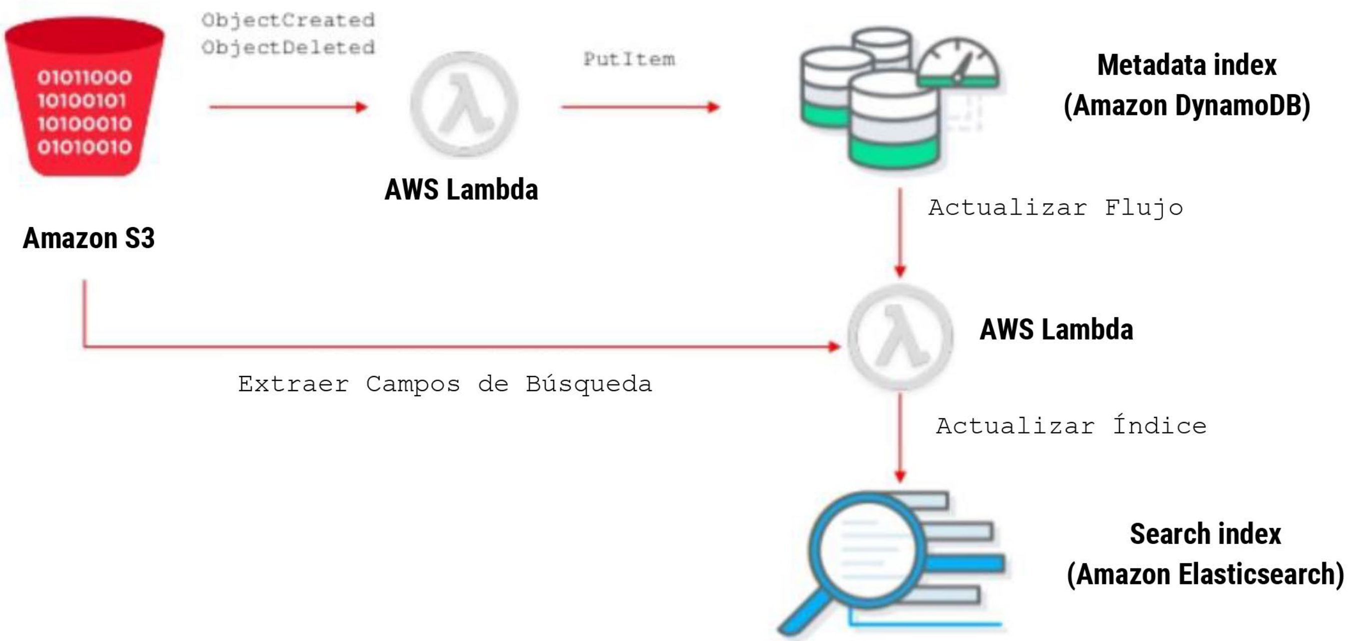


Tables A table is the metadata definition that represents your data, including its s

□	Name	Database	Classification
□	dataset228	glue-sample	csv
□	ianrob_nyc_transportation	nyc_transportation_canonical	parquet
□	limo	nyc-transportation	csv
□	salesdb_customer	glue-demo-mysql	mysql
□	salesdb_customer_site	glue-demo-mysql	mysql
□	salesdb_product	glue-demo-mysql	mysql
□	salesdb_product_category	glue-demo-mysql	mysql
□	salesdb_sales_order	glue-demo-mysql	mysql
□	salesdb_sales_order_all	glue-demo-mysql	mysql
□	salesdb_sales_order_detail	glue-demo-mysql	mysql
□	salesdb_supplier	glue-demo-mysql	mysql
□	taxi	nyc-transportation	csv

Comprehensive Data Catalog

Indexed and Searching Using Metadata



Organization and Administration of the Data Lake

Identity Management and Access Control

> Manage users, groups and roles

> Federation of identities with Open ID

>Temporary credentials with Amazon Security Token Service (Amazon STS)

> Stored policy templates

> Powerful policy language

> Amazon S3 bucket policies

Auditing and Monitoring

> AWS CloudTrail, S3 Access Logging, Cloud Watch

Naming Structure Model

> First thing to avoid is to consider Data Lake as one great centralized bucket to put everything.

> S3 prefixes - To separate and partition data

> Ex: s3://bucket-raw/origen/env/tabla

Auditing and Monitoring

> AWS CloudTrail, S3 Access Logging, Cloud Watch

source = Financial/Scheduling/Commercial/Activity

env = Staging/Production

table = tb_sales/tb_employees

Archiving

> S3Lifecyclepolicies

Classification

> Tags

Well-Architected Framework

Our solutions architecture considers the best practices to design and operate reliable, safe, efficient and profitable systems in the cloud.

This methodology was developed around the Amazon Web Services (AWS) Well-Architected Framework, which helps clients understand the pros and cons of decisions made when building AWS systems.

Pillars of the Framework

Operational Excellence

The operational excellence pillar focuses on running and monitoring systems to deliver business value, and continually improving processes and procedures. Key topics include managing and automating changes, responding to events, and defining standards to successfully manage daily operations.

The Operational Excellence pillar includes the ability to run and monitor systems to deliver business value and to continually improve supporting processes and procedures.

Security

The security pillar focuses on protecting information & systems. Key topics include confidentiality and integrity of data, identifying and managing who can do what with privilege management, protecting systems, and establishing controls to detect security events.

The security pillar includes the ability to protect information, systems, and assets while delivering business value through risk assessments and mitigation strategies.

Reliability

The reliability pillar focuses on the ability to prevent and quickly recover from failures to meet business and customer demand. Key topics include foundational elements regarding setup, cross-project requirements, recovery planning, and change management.

The reliability pillar includes the ability of a system to recover from infrastructure or service disruptions, dynamically acquire computing resources to meet demand, and mitigate disruptions such as misconfigurations or transient network issues.

Performance Efficiency

The performance efficiency pillar focuses on using IT and computing resources efficiently. Key topics include selecting the right resource types and sizes based on workload requirements, monitoring performance, and making informed decisions to maintain efficiency as business needs evolve.

The performance efficiency pillar includes the ability to use computing resources efficiently to meet system requirements, and to maintain that efficiency as demand changes and technologies evolve.

Cost Optimization

Cost Optimization focuses on avoiding un-needed costs. Key topics include understanding and controlling where money is being spent, selecting the most appropriate and right number of resource types, analyzing spend over time, and scaling to meet business needs without overspending.

The Cost Optimization pillar includes the ability to run systems to deliver business value at the lowest price point.

Morris & Opazo: Your Amazon Web Services Partner

We are a company specialized in providing Business Solutions in the field of Information Technologies. Morris & Opazo is part of the AWS Partner Network, a world-wide network of companies oriented to organizations that specialize in the design and management of platforms on the cloud-based infrastructure of Amazon Web Services.

aws partner network

Advanced
Consulting Partner

Channel Partner

AWS Competency

Benefits of Working with Us - Morris & Opazo

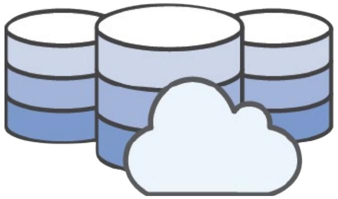
Let the experience of Morris & Opazo, an Advanced Consulting Partner of Amazon Web Services, help you in your journey to the cloud.

- Consulting, Proof-of-Concept and deployment of on-demand test environments in the cloud.
- Technical team with the highest level in all official certifications.
- Access to the global infrastructure of the AWS Cloud with the management and support of an expert partner.
- Constant optimization of your solution to maximize efficiency.
- Training to learn more about the possibilities the AWS Cloud can offer for you.
- Design and validate your solutions with our Certified Architects

aws CERTIFIED
Cloud Practitioner
aws CERTIFIED
Solutions Architect
Associate
aws CERTIFIED
Developer
Associate
aws CERTIFIED
SysOps Administrator
Associate
aws CERTIFIED
Solutions Architect
Professional
aws CERTIFIED
DevOps Engineer
Professional
aws CERTIFIED
Big Data
Specialty

Case Study: "Social Networks Insights"

Morris & Opazo | aws



SIM Partners

Case Study

"Social Networks Insights"

> Morris & Opazo has been an important collaborator in supporting our initiatives to leverage cloud-based resources and getting the most out of cloud technologies for Big Data and Analytics.
>
> **Francois Toubol**
> **VP of Technology**
> **SIM Partners**

About SIM Partners

SIM Partners is a company based on Chicago (US) which offers a complete set of scalable digital marketing technologies, supported by SaaS solutions of automated local marketing.

THE CHALLENGE

The final solution had to be scalable enough to collect, store, process and visualize metrics from different social networks. This information should also be available for diverse platform consumers.

THE SOLUTION

The cost-effective Big Data solution represented an improvement in security, shorter development times, wide availability, more frequent updates, more elasticity, a greater geographic coverage.

THE BENEFITS

Speed

The data lake allowed us to import significative amounts of real-time data. The data was collected from different sources, and was moved to a data lake in its original format. This process let us scale to data of any size with a time saving.

Economy of Scale

AWS cloud-based services follow a model of prices per layer and volume discounts. This has allowed us to grow without reaching unexpected billing situations, keeping the clam of paying the lowest possible cost for services consumed.

Services integration

By implementing our Data Lake we were able to access multiple AWS cloud services, including Machine Learning, which allowed us to improve our products and deliver extra value in our services offer.

Getting Started

For more information about Data Lakes on AWS, visit:

> Amazon Web Service Website

> Big Data on AWS

> Building a Data Lake on AWS (Video)

> About AWS

About AWS

For 10 years, Amazon Web Services has been the world's most comprehensive and broadly adopted Cloud platform. AWS offers over 70 fully featured services for compute, storage, databases, analytics, mobile, Internet of Things (IoT) and enterprise applications from 33 Availability Zones (AZs) across 13 geographic regions in the U.S., Australia, Brazil, China, Germany, Ireland, Japan, Korea, and Singapore. AWS services are trusted by more than a million active customers around the world – including the fastest growing startups, largest enterprises, and leading government agencies – to power their infrastructure, make them more agile, and lower costs.

To learn more about AWS, visit aws.amazon.com

aws Partner Network

Morris & Opazo