acc
applied cloud computing

Data Lake

Agenda

01 What is Data Lake?

02 Why do you need Data Lake?

03 AWS Solution

04 Data Lake on AWS

05 Data Lake Architecture

What is Data Lake?

A data lake is a centralized repository that allows you to store all your structured and unstructured data at any scale. You can store your data as-is, without having to first structure the data, and run different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning to guide better decisions.

Why do you need Data Lake?

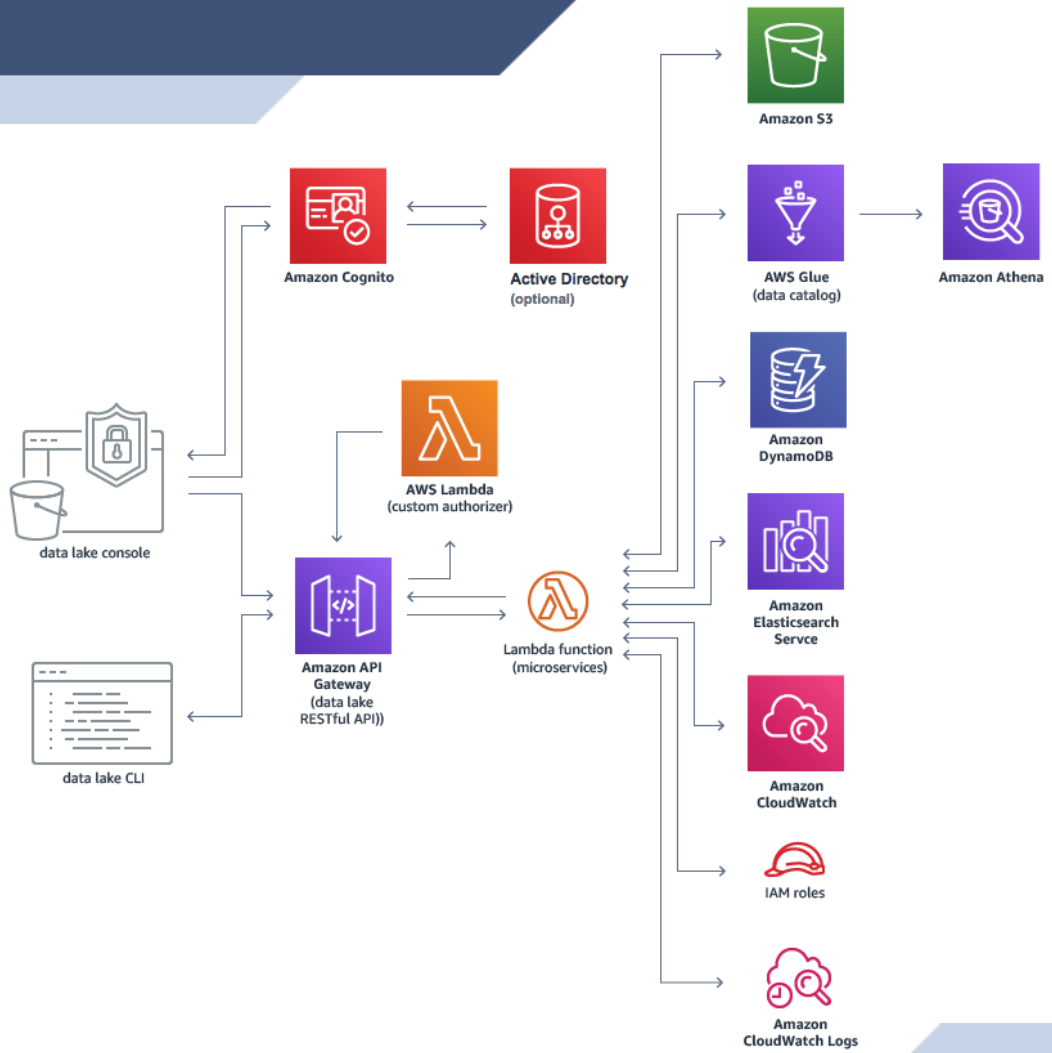
The main objective of building a data lake is to offer an unrefined view of data to data scientists.

Reasons for using Data Lake are:

- With the onset of storage engines like Hadoop storing disparate information has become easy. There is no need to model data into an enterprise-wide schema with a Data Lake.
- With the increase in data volume, data quality, and metadata, the quality of analyses also increases.
- Data Lake offers business Agility
- Machine Learning and Artificial Intelligence can be used to make profitable predictions.
- It offers a competitive advantage to the implementing organization.
- There is no data silo structure. Data Lake gives 360 degrees view of customers and makes analysis more robust.

AWS Solution

AWS offers a data lake solution that automatically configures the core AWS services necessary to easily tag, search, share, transform, analyze, and govern specific subsets of data across a company or with other external users. The solution deploys a console that users can access to search and browse available datasets for their business needs. The solution also includes a federated template that allows you to launch a version of the solution that is ready to integrate with Microsoft Active Directory.



Data Lake on AWS

What does this AWS Solution do?

Many **Amazon Web Services** (AWS) customers require a data storage and analytics solution that offers more agility and flexibility than traditional data management systems. A data lake is a new and increasingly popular way to store and analyze data because it allows companies to manage multiple data types from a wide variety of sources, and store this data, structured and unstructured, in a centralized repository. The AWS Cloud provides many of the building blocks required to help customers implement a secure, flexible, and cost-effective data lake. These include AWS managed services that help ingest, store, find, process, and analyze both structured and unstructured data. To support our customers as they build data lakes, AWS offers the data lake solution, which is an automated reference implementation that deploys a highly available, cost-effective data lake architecture on the AWS Cloud along with a user-friendly console for searching and requesting datasets.

AWS Lake Formation

AWS Lake Formation is a service that makes it easy to set up a secure data lake in days. A data lake is a centralized, curated, and secured repository that stores all your data, both in its original form and prepared for analysis. A data lake enables you to break down data silos and combine different types of analytics to gain insights and guide better business decisions.

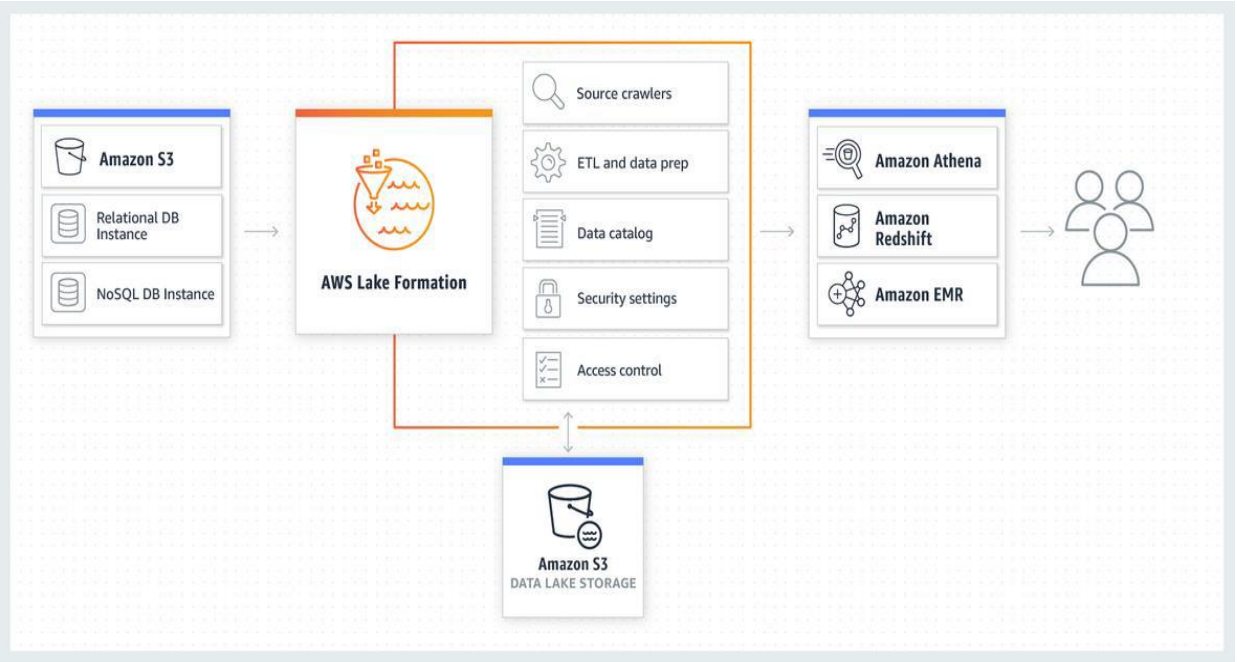
However, setting up and managing data lakes today involves a lot of manual, complicated, and time-consuming tasks. This work includes loading data from diverse sources, monitoring those data flows, setting up partitions, turning on encryption and managing keys, defining transformation jobs and monitoring their operation, re-organizing data into a columnar format, configuring access control settings, de-duplicating redundant data, matching linked records, granting access to data sets, and auditing access over time.

Creating a data lake with Lake Formation is as simple as defining data sources and what data access and security policies you want to apply.

Lake Formation then helps you collect and catalog data from databases and object storage, move the data into your new Amazon S3 data lake, clean and classify your data using machine learning algorithms, and secure access to your sensitive data. Your users can access a centralized data catalog which describes available data sets and their appropriate usage. Your users then leverage these data sets with their choice of analytics and machine learning services, like Amazon Redshift, Amazon Athena, and (in beta) Amazon EMR for Apache Spark. Lake Formation builds on the capabilities available in AWS Glue.

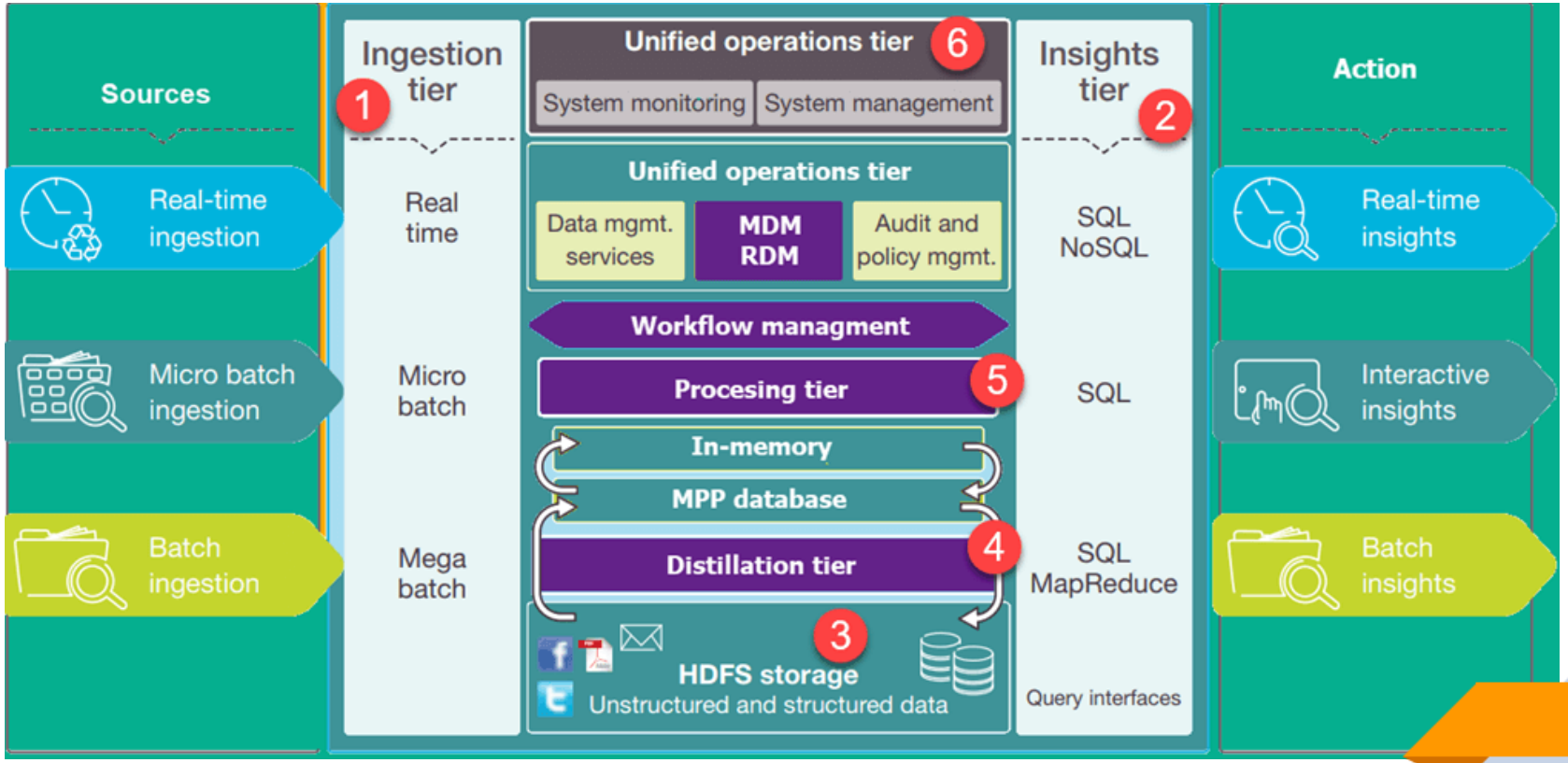
How it Work

Lake Formation helps to build, secure, and manage your data lake. First, identify existing data stores in S3 or relational and NoSQL databases, and move the data into your data lake. Then crawl, catalog, and prepare the data for analytics. Then provide your users secure self service access to the data through their choice of analytics services. Other AWS services and third-party applications can also access data through the services shown. Lake Formation manages all of the tasks in the orange box and is integrated with the data stores and services shown in the blue boxes.



Data Lake Architecture

The figure shows the architecture of a Business Data Lake. The lower levels represent data that is mostly at rest while the upper levels show real-time transactional data. This data flow through the system with no or little latency

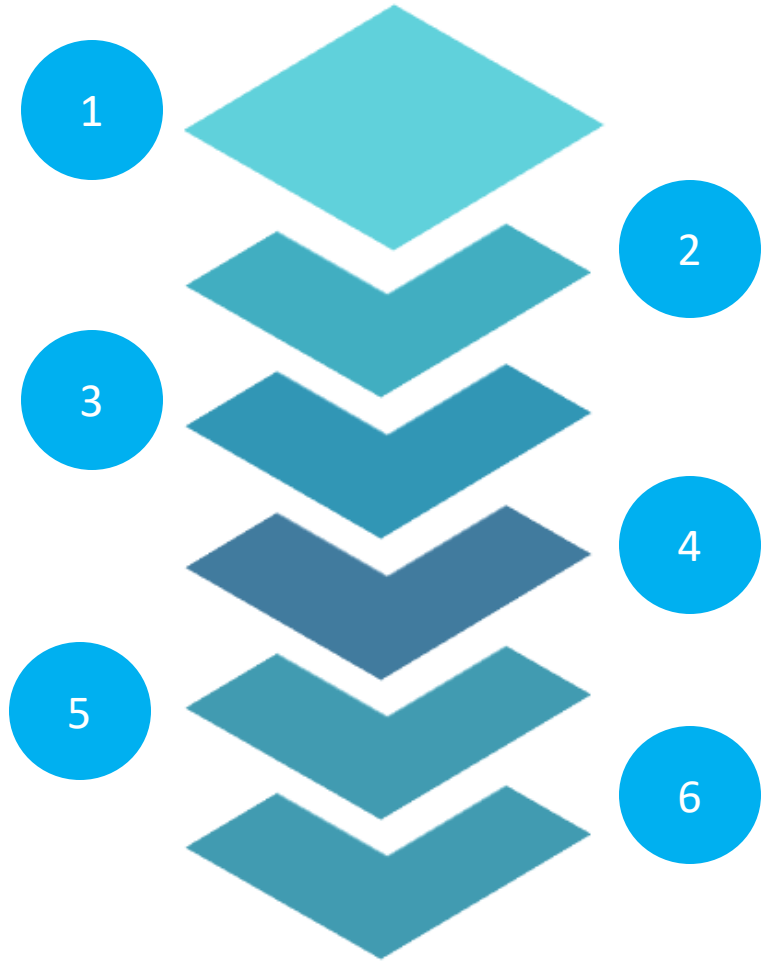


Data Lake Architecture

Ingestion Tier: The tiers on the left side depict the data sources. The data could be loaded into the data lake in batches or in real-time

HDFS is a cost-effective solution for both structured and unstructured data. It is a landing zone for all data that is at rest in the system.

Processing tier run analytical algorithms and users queries with varying real time, interactive, batch to generate structured data for easier analysis.



Insights Tier: The tiers on the right represent the research side where insights from the system are used. SQL, NoSQL queries, or even excel could be used for data analysis.

Distillation tier takes data from the storage tire and converts it to structured data for easier analysis.

Unified operations tier governs system management and monitoring. It includes auditing and proficiency management, data management, workflow management.

Key Data Lake Concept

- Security
- Data Governance
- Data Quality
- Data discovery
- Data Auditing
- Data Storage
- Data Lineage
- Data Exploration
- Data Ingestion
- Teamwork

Thanks!

Mumbai
508/509, New Era Business Park
Road No. 33, Wagle Industrial Estate,Thane, 400604

Nilesh Satpute
(+91) 8655423607
nilesh@acc.ltd

Shubho Pramanik
(+91) 9029720294
shubho@acc.ltd

Gautam Ahuja
(+91) 8600074444
gautam.ahuja@acc.ltd

www.appliedcloudcomputing.com