Naya College

Data Lake on Amazon S3

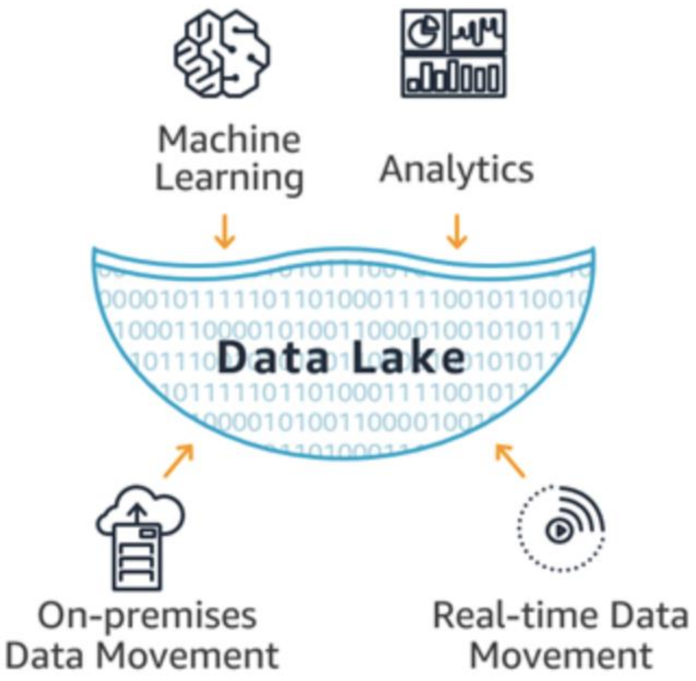
NAYA College

Data Lake on Amazon S3

Whats is a BigData Data Lake?

- A centralized repository that allows you to store all your structured and unstructured data at any scale

- Data can be stored 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

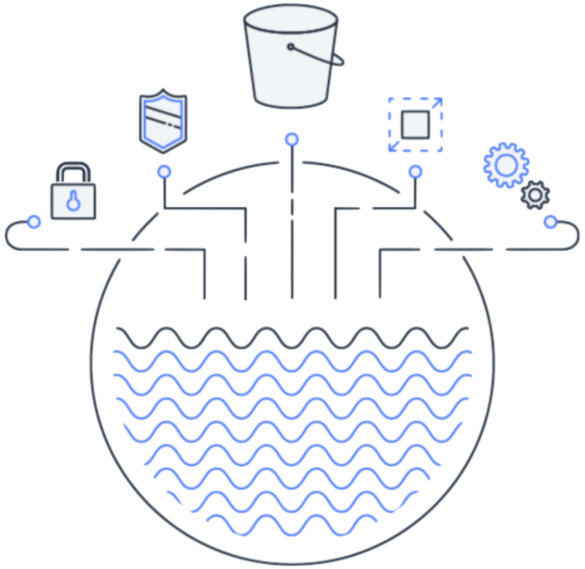


Data Lake on Amazon S3

- Amazon S3 (Simple Storage Service) is a largest object storage service for structured and unstructured data which can serve to build a complete data lake

- S3 offers cost-effectively building scalable data lake of any size secure environment where data is protected by 99.999999999% SLA

- Data lake built on Amazon S3 offers full integration with AWS computing and analytics services, AI, ML, high-performance computing and media data processing applications to gain insights from unstructured dataset



Data Lake on Amazon S3

Data Lake on Amazon S3 - Advantages

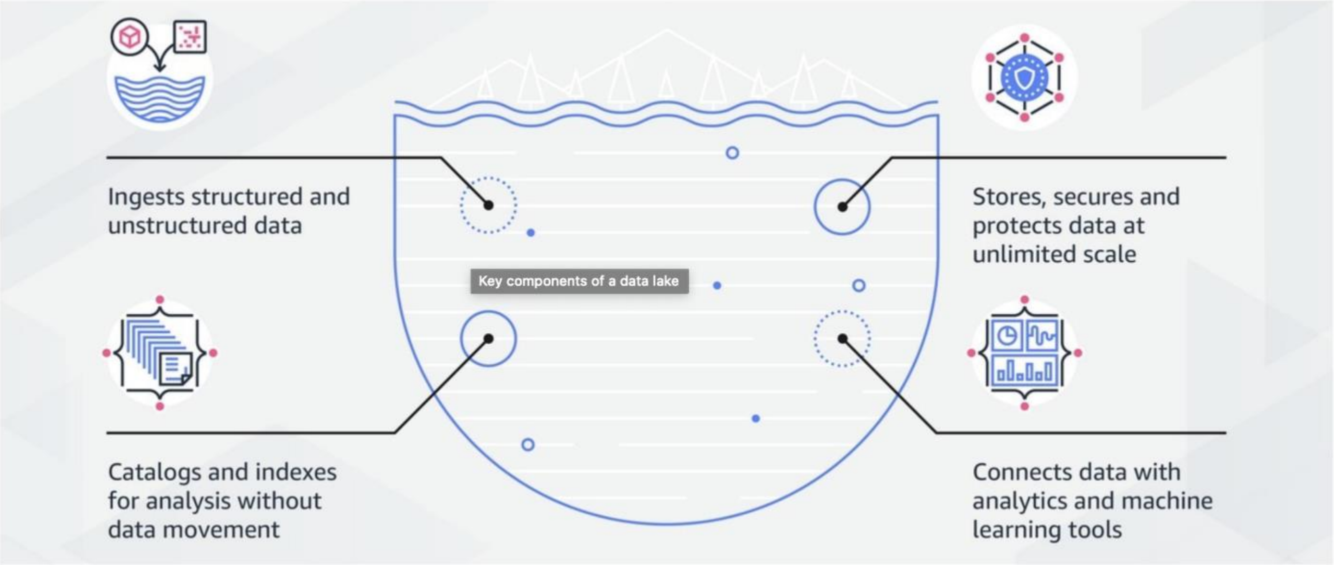
- **Data Durability**
  - Data is available when needed and protected against failures, errors and threats
- **Data Scalability**
  - Scale up storage capacity, without lengthy resource procurement cycles
- **AWS Services Integration**
  - Use AWS native services to run applications on your data lake (AI, ML, and media data processing)
  - Integrations with third-party service providers
- **Data Management Support**
  - Comprehensive flexibility to operate at an object level while configuring access and audit

Data Lake on Amazon S3

BigData Data Lake to Insights

- Amazon S3 allows to migrate, store, manage and secure all structured and unstructured data at unlimited scale, breaking down data silos

Key components of a data lake



Ingests structured and unstructured data

Stores, secures and protects data at unlimited scale

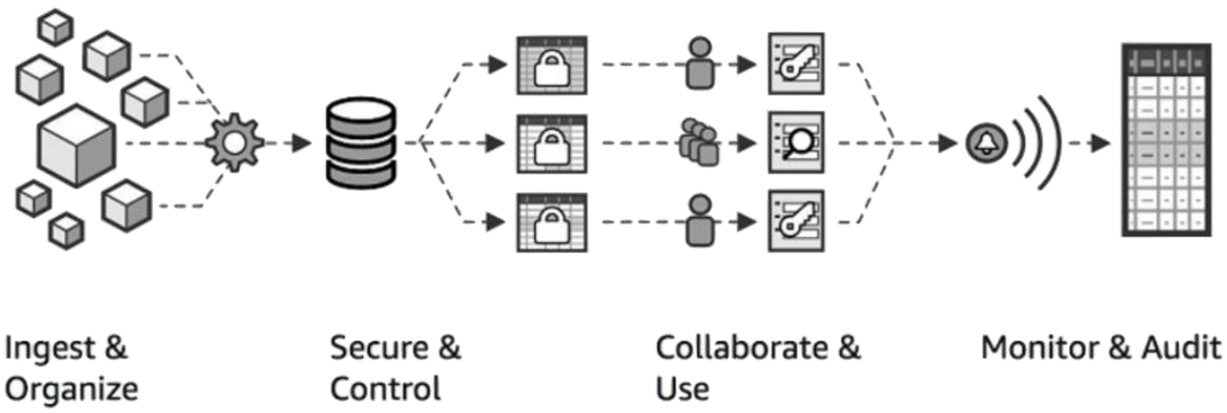
Catalogs and indexes for analysis without data movement

Connects data with analytics and machine learning tools

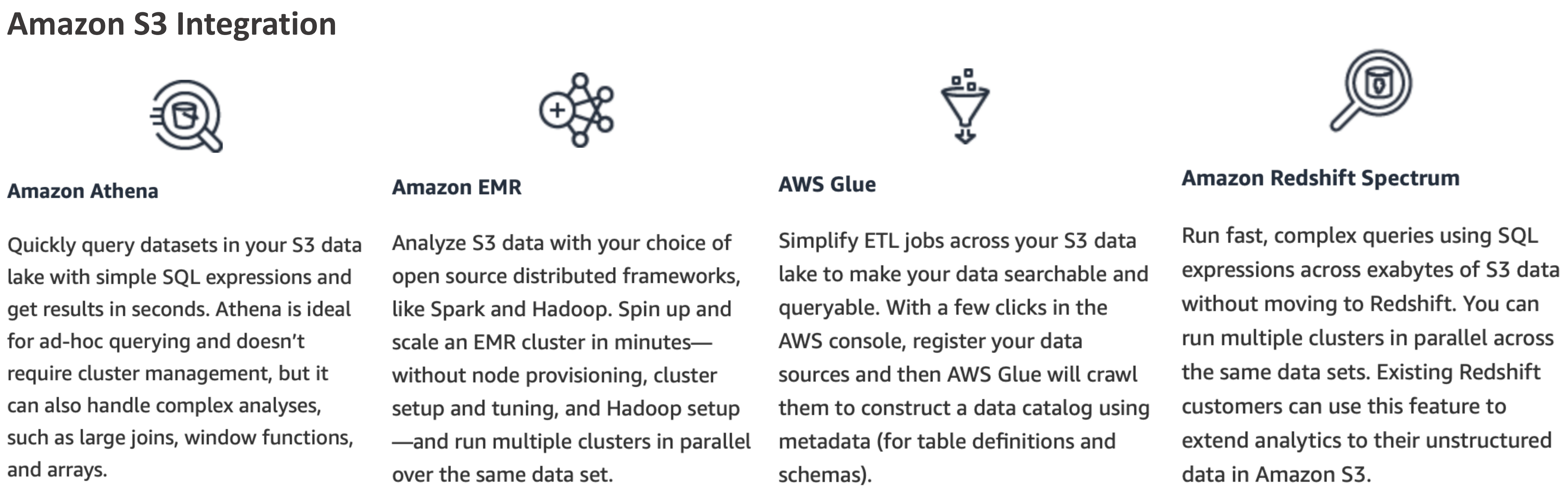
Data Lake on Amazon S3

Amazon S3 Integration

- **AWS Lake Formation**
  - AWS Lake Formation serves as a centralized data lake management platform
  - Lake Formation collects data from different sources and moves it into a new data lake in Amazon S3
  - The service cleans, catalogs and classifies data using ML algorithms and define access control
  - Users can then access a centralized catalog of data which lists available data sets and their usage terms



Data Lake on Amazon S3



Amazon S3 Integration

Amazon Athena

Quickly query datasets in your S3 data lake with simple SQL expressions and get results in seconds. Athena is ideal for ad-hoc querying and doesn't require cluster management, but it can also handle complex analyses, such as large joins, window functions, and arrays.

Amazon EMR

Analyze S3 data with your choice of open source distributed frameworks, like Spark and Hadoop. Spin up and scale an EMR cluster in minutes—without node provisioning, cluster setup and tuning, and Hadoop setup—and run multiple clusters in parallel over the same data set.

AWS Glue

Simplify ETL jobs across your S3 data lake to make your data searchable and queryable. With a few clicks in the AWS console, register your data sources and then AWS Glue will crawl them to construct a data catalog using metadata (for table definitions and schemas).

Amazon Redshift Spectrum

Run fast, complex queries using SQL expressions across exabytes of S3 data without moving to Redshift. You can run multiple clusters in parallel across the same data sets. Existing Redshift customers can use this feature to extend analytics to their unstructured data in Amazon S3.

Data Lake on Amazon S3

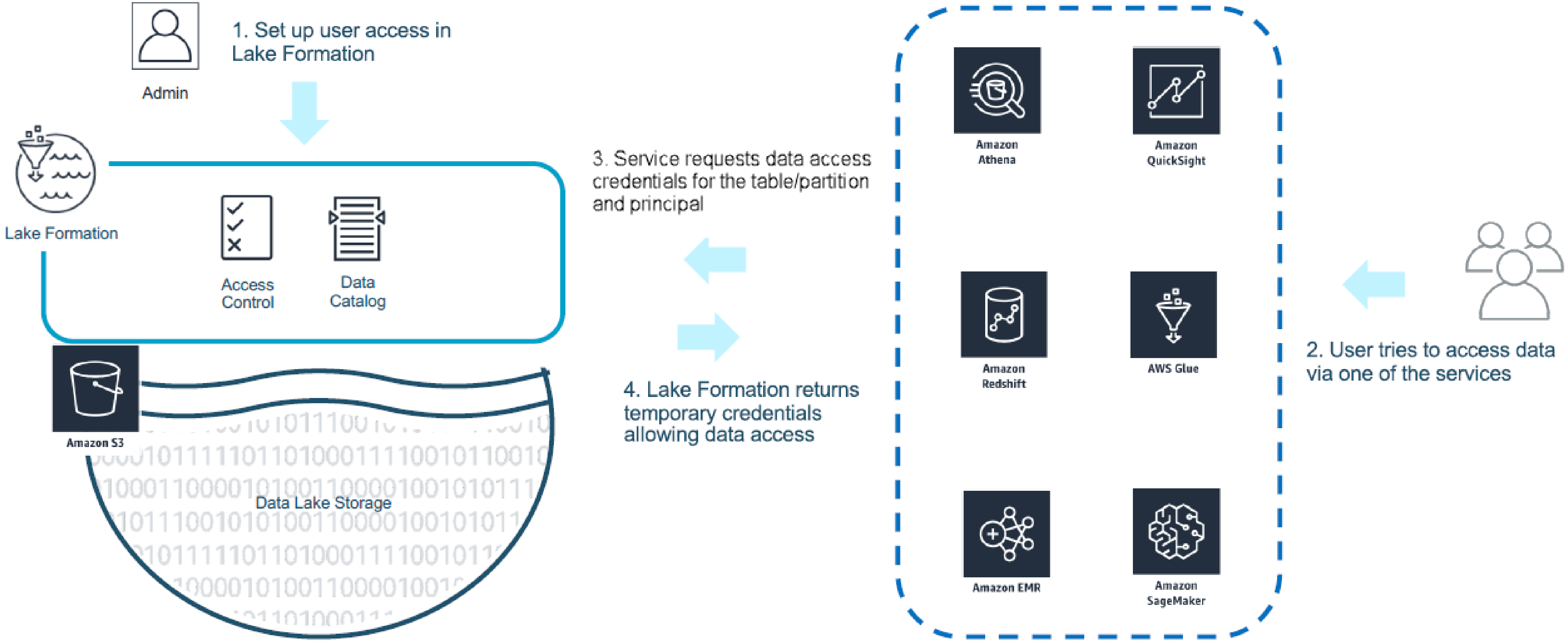
Amazon S3 Integration

- **SQL & Amazon S3**
  - S3 Select enables applications to offload the heavy lifting of filtering and accessing data inside objects to S3
  - With S3 SELECT, object's metadata can be queried without moving the object to another data store
  - S3 SELECT can improve the performance of most applications that frequently access data from S3 by up to 400% and reduce querying costs as much as 80%

Amazon S3 Select and Amazon Glacier Select

Select subset of data from an object based on a SQL expression

Data Lake on Amazon S3



Admin

1. Set up user access in Lake Formation

Lake Formation

Access Control

Data Catalog

Amazon S3

Data Lake Storage

3. Service requests data access credentials for the table/partition and principal

4. Lake Formation returns temporary credentials allowing data access

Amazon Athena

Amazon QuickSight

Amazon Redshift

AWS Glue

Amazon EMR

Amazon SageMaker

2. User tries to access data via one of the services

Data Lake on Amazon S3

Data Lake on S3 Use-Cases

- Amazon S3 hosts tens of thousands of data lakes for household brands such as Netflix, Airbnb, Sysco, Expedia, GE and FINRA, who are using them to securely scale with their needs and to discover business insights every minute