AWS LAKE FORMATION

Introduction

Lake formation is a service provided by Amazon that automates the process of creating data lakes in AWS. In other words, AWS lake formation allows you to build, manage, and secure data lake automatically. It means you don’t have to use and set up different required services manually. To setup new data lake in AWS and get it ready to use for process. You can create data store, access it ,use analytics and machine learning services. To accomplish these task you can you different services such as Amazon S3 storage, AWS Glue Crawlers classifiers for crawling and data categorization, and AWS Glue jobs for data transformation.

Then, for analysis and data visualization, you can use tools such as Athena, QuickSight, and EMR. Using these services and tools individually and then configuring them manually to form your data lake is a difficult process and involves many time-consuming tasks.

Lake formation simplifies this process by automating many of the required steps, which otherwise need to be done manually. In essence, lake formation can help you to

1. Register a data lake and the paths where it will reside
2. Orchestrate data flows
3. Create and manage a centralized data catalog containing meta-data
4. Define fine-grained permissions to access data

## **What is a Data lake?**

A data lake is a **centralized repository**, where you can store all of your data without thinking about its structure and scaling restrictions. It means that you can store your data as it is, without structuring and running different types of analytics.

There is a common misunderstanding that data lake tools are like a powerful database. In fact, data lakes provide multiple functions such as data ingestion, processing, data visualization, and the use of machine learning algorithms. More importantly, unlike databases the storage and processing mechanisms of a data lake are decoupled. This allows the use of different AWS data processing and data visualization services.

Data lakes are also different from data warehouses, as you can store both structured and unstructured data in a data lake. Contrary to a data warehouse, there is no need to define the schema or structure of data in data lakes. This means you do not need to worry about the design of the data or the data-related questions that might arise in the future. You can also apply different types of analytics to your data such as real-time analytics, machine learning, SQL queries, and text searches to gain insights.

A data lake usually consists of the following elements:

Storage

Data ingestion

Analytics and machine learning

Real-time analysis

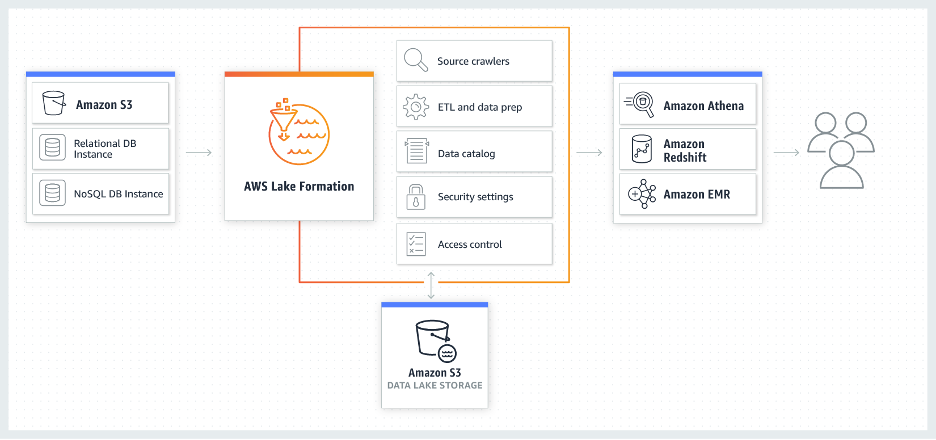
### **Data lakes characteristics**

The most important characteristic of data lakes is that they should be data agnostic. This means that they should allow storing of both structured and unstructured data and should not be limited to just one type of file or data structure.

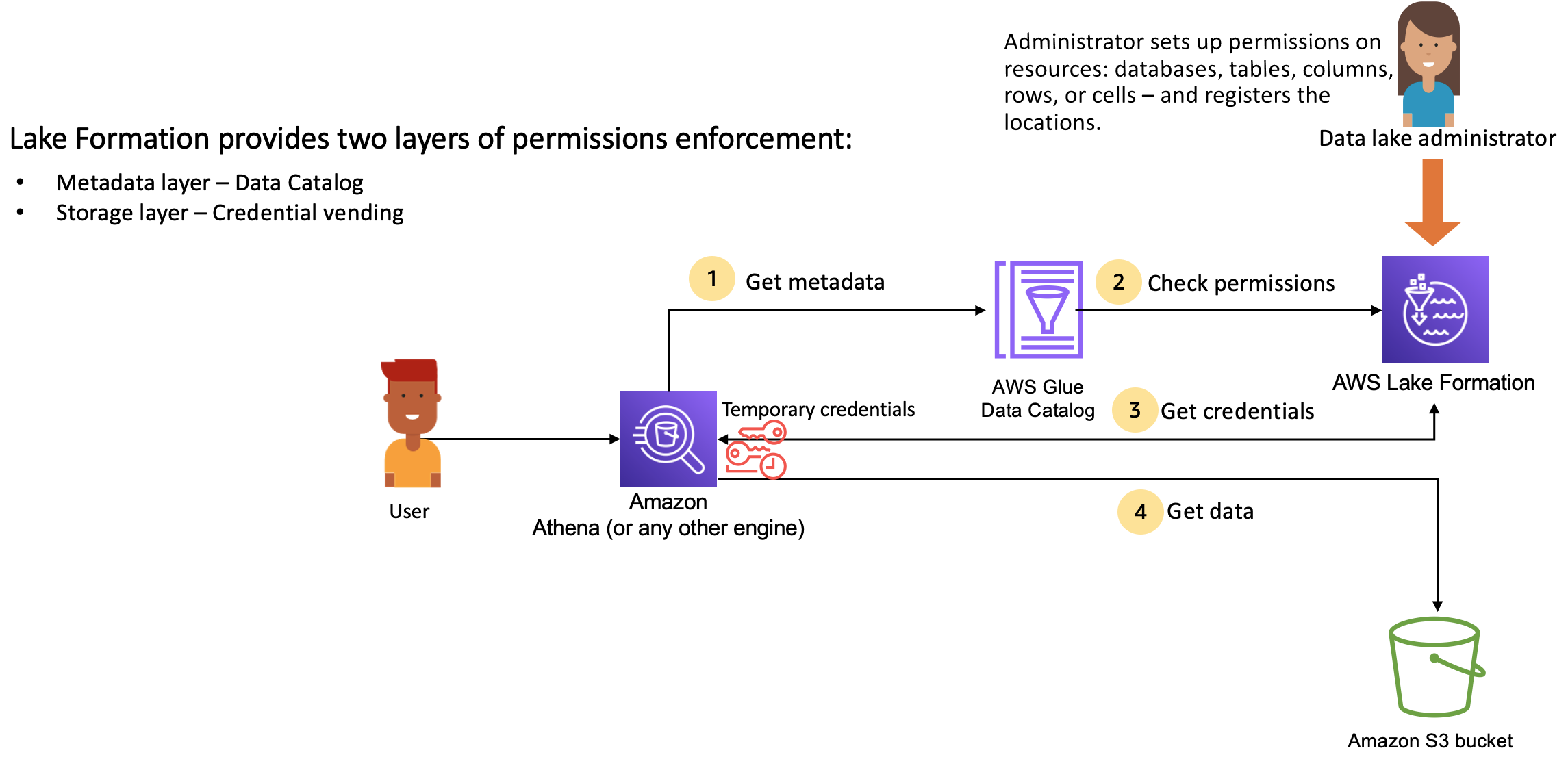
In addition, all the data should be stored in a centralized place to break down data silos (which means that data of one type should not be stored separately from data of another type).

How it Works?

You can build a data lake by using different services offered by Amazon. These services can be categorized into three types:



1. Data storage and cataloging servicesThe preferred AWS services for the data layer are **Amazon simple storage service (S3)** and **S3 Glacier**. These services are used to build the data lake itself and not to populate or analyze the data lake. S3 uses the concept of “Bucket” to store all types of content in a central repository.In S3, you can use different storage tiers to categorize your data. These tiers are based on the frequency of the data access and you will be charged accordingly. So, for frequently accessed data, you can select a tier that costs less for access requests and more for data storage.**AWS Glue** is a data processing and cataloging service you can use to catalog data in your data lake and prevent it to become a data swamp. Likewise, **AWS Glue Data Catalog** is the metadata repository that represents your data. You can use Glue “Crawler” to populate **AWS Glue Data Catalog** with tables.
2. Data movement servicesThese services are used to ingest data into your data lake. For real-time data ingestion, you can use **Kinesis Data Streams** or **Kinesis Firehose**. Whereas you can use **Amazon API Gateway** to ingest restful data using standard HTTP calls.On the other side, you can use **AWS data exchange** to extract public data from third-party and store it with your data. Similarly, **App flow** can be used for collecting your data from third-party software and securely transferring it to your data lake.
3. Analytics and processing servicesAmazon provides a great serverless service called Athena via which you can run SQL queries against data in S3. You can also use **Amazon EMR** to process and analyze data in batches, whereas to transform raw data collected in real-time, the **AWS Lambda** service is available.



###### **Lake Formation permissions management high-level steps**

In Lake Formation we can provide access controls for data in your data lake, a [data lake administrator](https://docs.aws.amazon.com/lake-formation/latest/dg/initial-LF-setup.html#create-data-lake-admin) or a user with administrative permissions sets up individual Data Catalog table user policies to allow or deny access to Data Catalog tables using Lake Formation permissions.

Then, either the data lake administrator grants Lake Formation permissions to users on the Data Catalog databases and tables and registers the Amazon S3 location of the table with Lake Formation.

1. **Get metadata** – A principal (user) submits a query or an ETL script to an [integrated analytical engine](https://docs.aws.amazon.com/lake-formation/latest/dg/working-with-services.html) such as Amazon Athena, AWS Glue, Amazon EMR, or Amazon Redshift Spectrum. The integrated analytical engine identifies the table that is being requested and sends a request for metadata to the Data Catalog.
2. **Check permissions** – The Data Catalog checks user's permissions with Lake Formation, and if the user is authorized to access the table, returns the metadata that the user is allowed to see to the engine.
3. **Get credentials** – The Data Catalog lets the engine know if the table is managed by Lake Formation or not. If the underlying data is registered with Lake Formation, the analytical engine requests Lake Formation to provide data access by granting temporary access.
4. **Get data** – If the user is authorized to access the table, Lake Formation provides temporary access to the integrated analytical engine. Using the temporary access, the analytical engine fetches the data from Amazon S3, and performs necessary filtering such as column, row, or cell filtering. When the engine finishes running the job, it returns the results back to the user. This process is called [credential vending](https://docs.aws.amazon.com/lake-formation/latest/dg/using-cred-vending.html).

If the table is not managed by Lake Formation, the second call from the analytic engine is made directly to Amazon S3. The concerned Amazon S3 bucket policy and IAM user policy are evaluated for data access

AWS Lake Formation Console:

1] Data Catalog - In AWS, a Data Catalog is a centralized repository that stores metadata about your data assets. It helps organize, discover, and manage metadata, making it easier to understand and use your data. In the context of AWS Lake Formation, the Data Catalog is a key component that integrates with AWS Glue.

* Database - In AWS Lake Formation, the term "database" is often associated with the AWS Glue Data Catalog. The AWS Glue Data Catalog is a managed metadata repository that integrates with AWS Lake Formation and provides a central location to store metadata about your data assets, such as tables, partitions, and schemas.
* Tables - AWS Lake Formation, tables refer to structured datasets or collections of data organized in tabular form. These tables are registered in the AWS Glue Data Catalog, which is integrated with Lake Formation
* View - A view is a virtual table in which the contents are defined by a query that references one or more tables. You can create a view that references up to 10 tables using SQL editors for Amazon Athena, Amazon Redshift, or Amazon EMR. Underlying reference tables for a view can belong to the same database or different databases within the same AWS account.
* Data Filters - You can implement column-level, row-level, and cell-level security by creating data filters. You select a data filter when you grant the SELECT Lake Formation permission on tables. If your table contains nested column structures, you can define a data filter by including or excluding the child columns and define row-level filter expressions on nested attributes.
* Data Sharing - You can use the AWS Lake Formation data sharing feature to grant and manage permissions on data stored in locations other than Amazon S3, and metadata stored in locations other than the AWS Glue Data Catalog. With the data sharing capability, you can set up and manage permissions on datasets in Amazon Redshift without migrating the data into Amazon S3. You can also use the Data Catalog federation feature to connect to external megastores
* Crawler - AWS Glue Crawlers are used to discover datasets, extract schema information, and populate the AWS Glue Data Catalog. Before this integration, you needed to set up Amazon IAM and Amazon S3 bucket policies for crawler access to S3 data lake targets. Customers who use Lake Formation to manage these targets preferred having all permissions centralized in Lake Formation instead of setting up direct S3 access for the crawler role. With this Glue Crawler and Lake Formation integration, you can now use Lake Formation permissions for the crawler's access to your Lake Formation managed tables.

Permissions -

In AWS Lake Formation, permissions refer to the access controls and privileges assigned to users or roles for interacting with data stored in a data lake. AWS Lake Formation provides fine-grained access control to manage who can access, modify, or perform specific actions on the data within the data lake.

**Data Permissions:**

* Permissions in AWS Lake Formation are applied at the data object level, such as databases, tables, and columns. This means you can control access to specific data assets within your data lake.

[LF-Tags and Permissions:](https://us-east-1.console.aws.amazon.com/lakeformation/home?region=us-east-1#policy-tags)

To use the Lake Formation tag-based access control (LF-TBAC) method to secure Data Catalog resources (databases, tables, and columns), you create LF-Tags, assign them to resources, and grant LF-Tag permissions to principals.

Before you can assign LF-Tags to Data Catalog resources or grant permissions to principals, you need to define LF-Tags. Only a data lake administrator or a principal with LF-Tag creator permissions can create LF-Tags.

Hybrid access mode:

AWS Lake Formation hybrid access mode supports two permission pathways to the same AWS Glue Data Catalog databases and tables.  In the first pathway, Lake Formation allows you to select specific principals, and grant them Lake Formation permissions to access databases and tables by opting in. The second pathway allows all other principals to access these resources through the default IAM principal policies for Amazon S3 and AWS Glue actions.

Data Locations:

Data location permissions in AWS Lake Formation enable principals to create and alter Data Catalog resources that point to designated registered Amazon S3 locations. Data location permissions work in addition to Lake Formation data permissions to secure information in your data lake.

Administration

Administrative roles and tasks: