AWS Lake Formation

Introduction

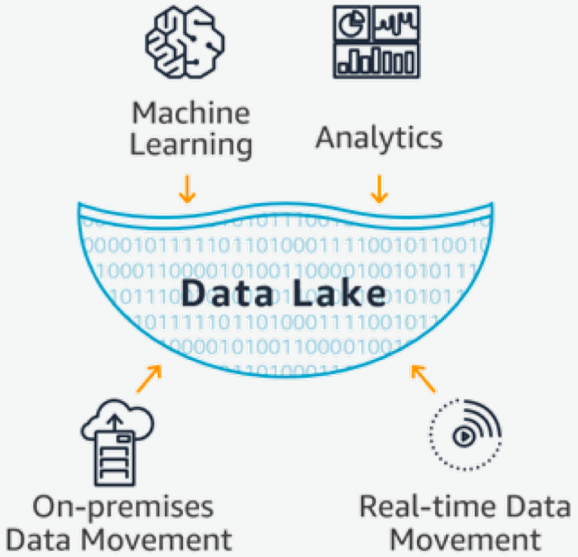
David Bayard, Solutions Architect

October 2019

Why Data Lakes?

What is a Data Lake?

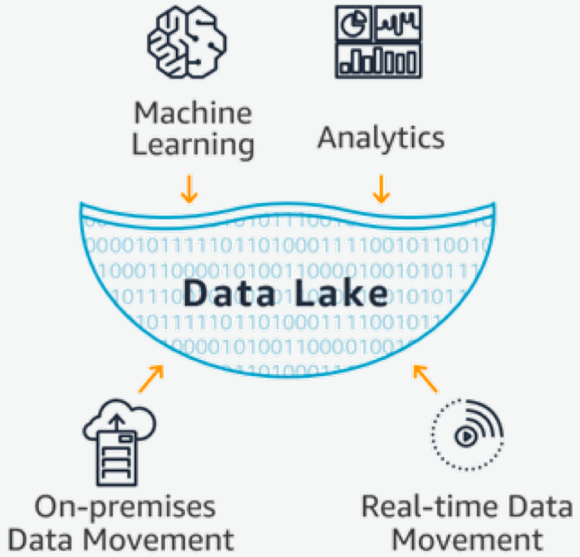
- A centralized repository for both structured and unstructured data
- Store data as-is in open-source file formats to enable direct analytics



aws

Why a Data Lake?

- Decouple storage from compute, allowing you to scale
- Enable advanced analytics across all of your data sources
- Reduce complexity in ETL and operational overhead
- Future extensibility as new database and analytics technologies are invented



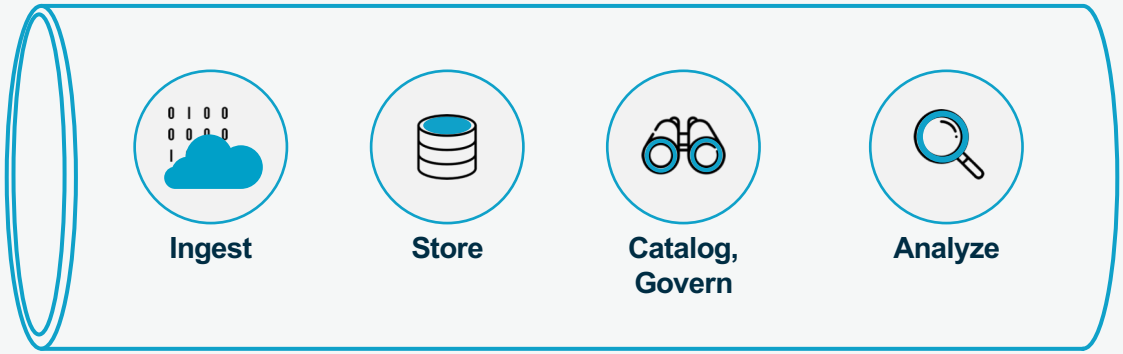
aws

Building a Data Lake on AWS

More data lakes & analytics on AWS than anywhere else

Realizing customer and operational insight from your data requires a robust Data Pipeline

Customer and Operational **Data**



Customer and Operational **Insights**

Data Lake Reference Architecture

Ingest

Kinesis
streaming data

Direct Connect
link to data center

Snowball
mail-in bulk data

Database Migration Service
import from Oracle, Netezza, and others

and many more…

Store

Amazon S3
secure, cost-effective storage

Catalog, Govern

Glue
data catalog & ETL

Lake Formation
security & blueprints

Analyze

Redshift
data warehouse

Athena
Ad-hoc analysis

EMR
unstructured data

QuickSight
visualization

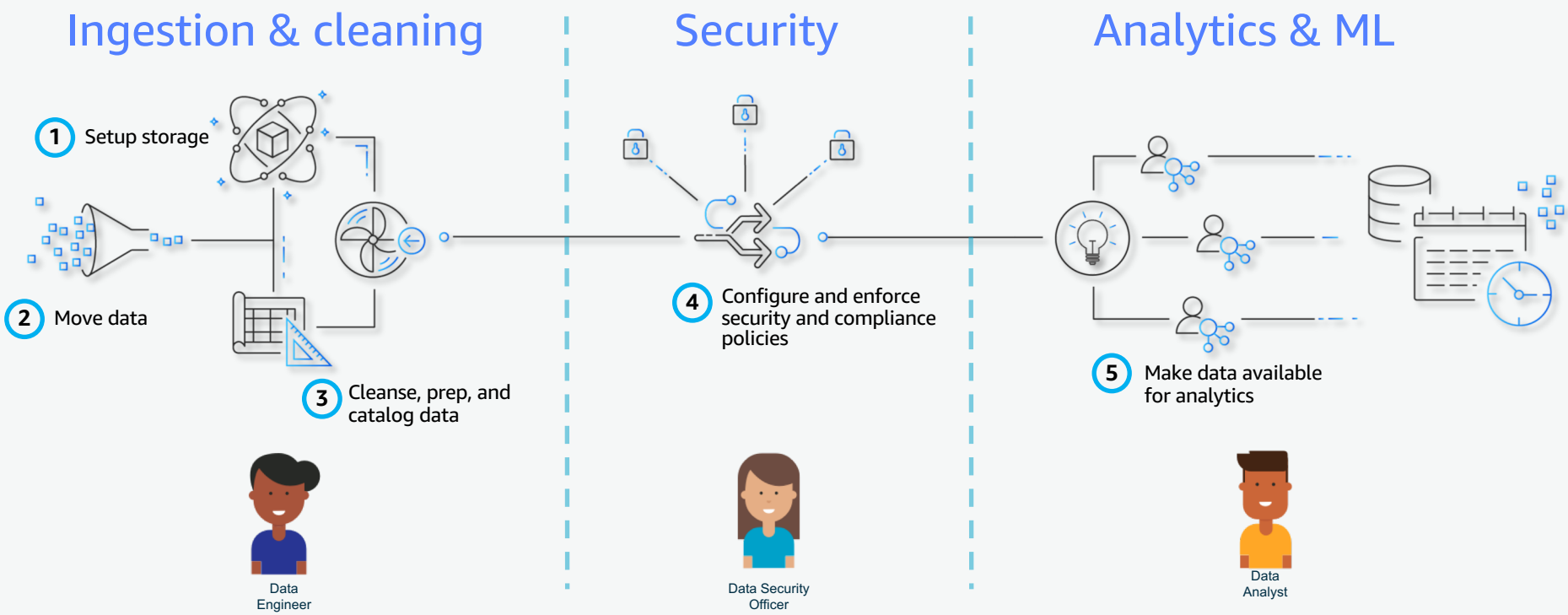
Sagemaker
AI/Machine Learning

and many more…

AWS Lake Formation

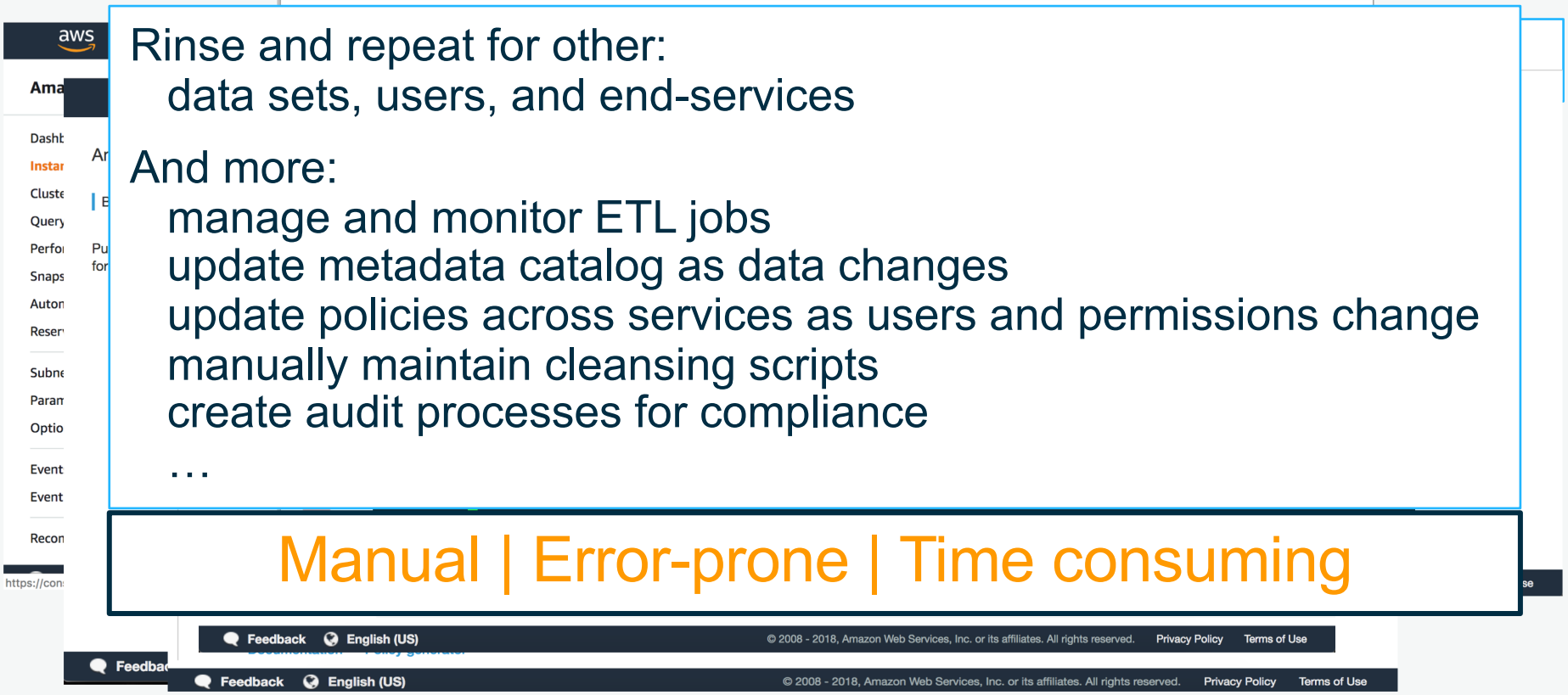
Manually building secure data lakes is **hard**

Typical steps of building a data lake



Sample of steps required

Configure access from analytics services



Rinse and repeat for other:

- data sets, users, and end-services

And more:

- manage and monitor ETL jobs
- update metadata catalog as data changes
- update policies across services as users and permissions change
- manually maintain cleansing scripts
- create audit processes for compliance
- …

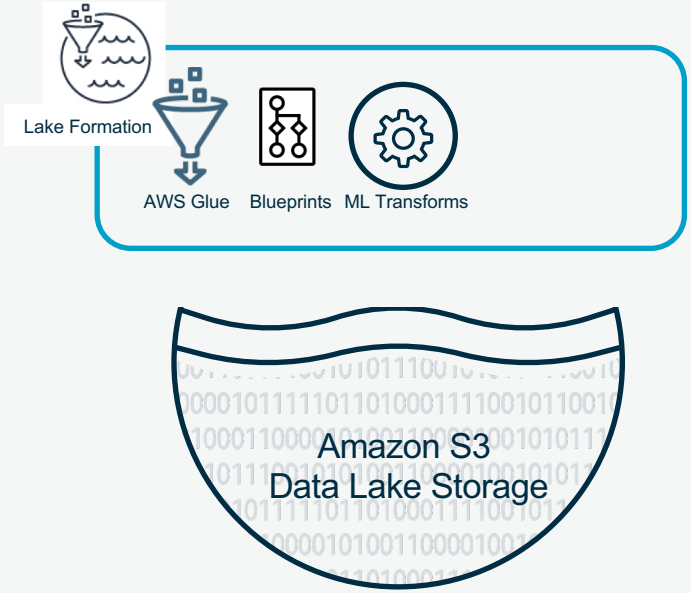
Manual | Error-prone | Time consuming

Lake Formation simplifies building secure data lakes

Built on Amazon S3 a robust data lake infrastructure

Cost effective, durable storage with global replication capabilities

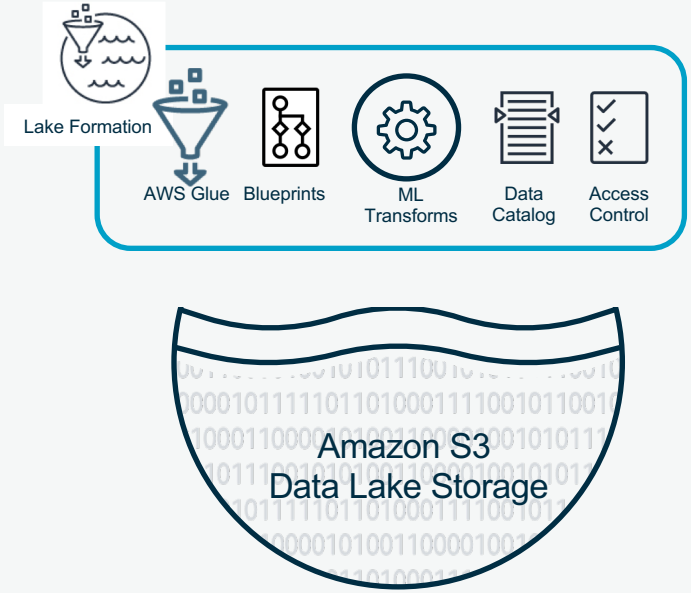
Automates manual, repetitive, low value tasks



Simplified **ingest & cleaning** enables data engineers to build faster

Cost effective, durable storage with global replication capabilities

Provides a central locus of control

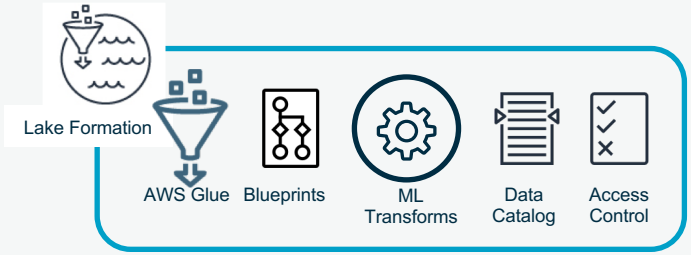


Centralized management of **fine grained permissions** empower security officers

Simplified **ingest & cleaning** enables data engineers to build faster

Cost effective, durable storage with global replication capabilities

Enables all your data users



Comprehensive set of **integrated tools** enable every user equally

Centralized management of **fine grained permissions** empower security officers

Simplified **ingest & cleaning** enables data engineers to build faster

Cost effective, durable storage with global replication capabilities

aws

AWS Lake Formation Pricing

No additional charges – Only pay for the underlying services used.

Tools that enable data engineers, security officers & data analysts

Building data lakes with Lake Formation

Ingestion & cleaning

Data Engineer

AWS Glue

Serverless Spark

Blueprints

ML Transforms

Security

Data Security Officer

Data catalog

Centralized permissions

Real time monitoring

Auditing

Analytics & ML

Data Analyst

Comprehensive portfolio of integrated tools

Redshift

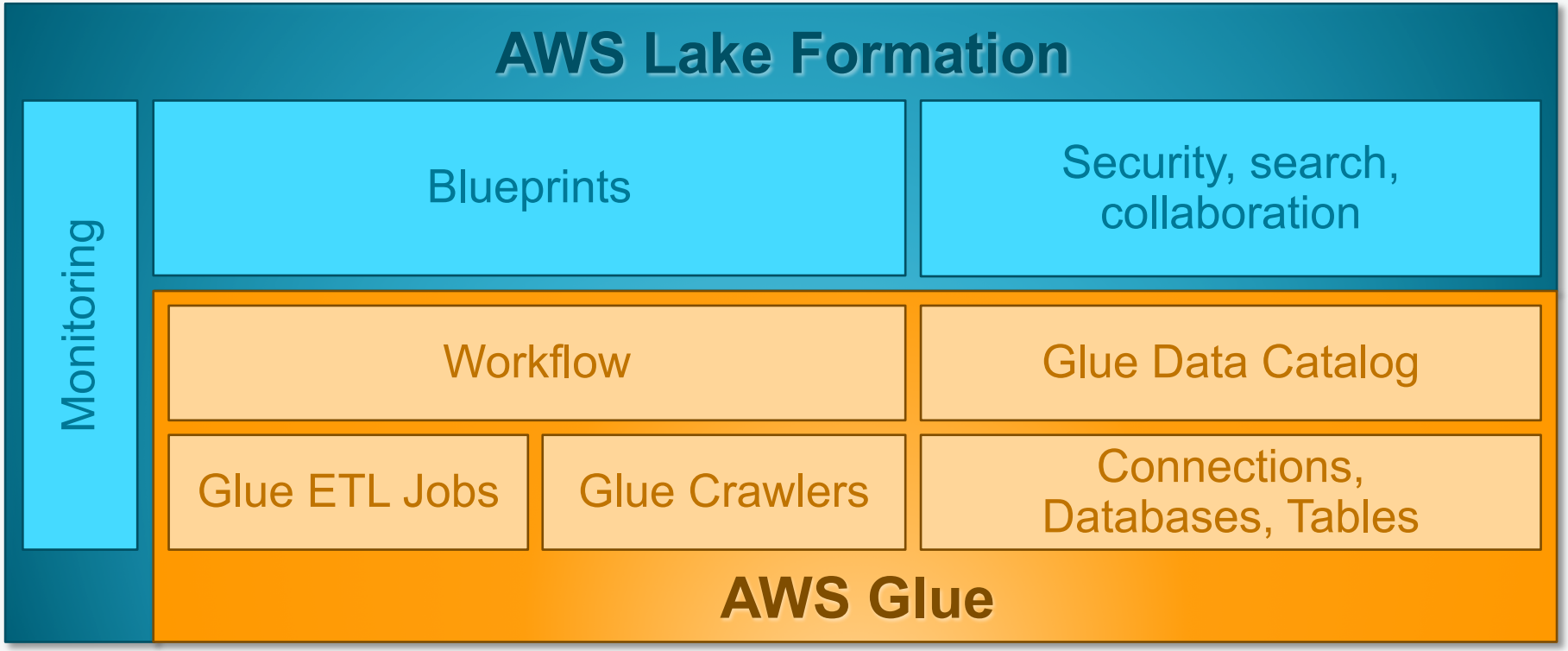
Glue

EMR

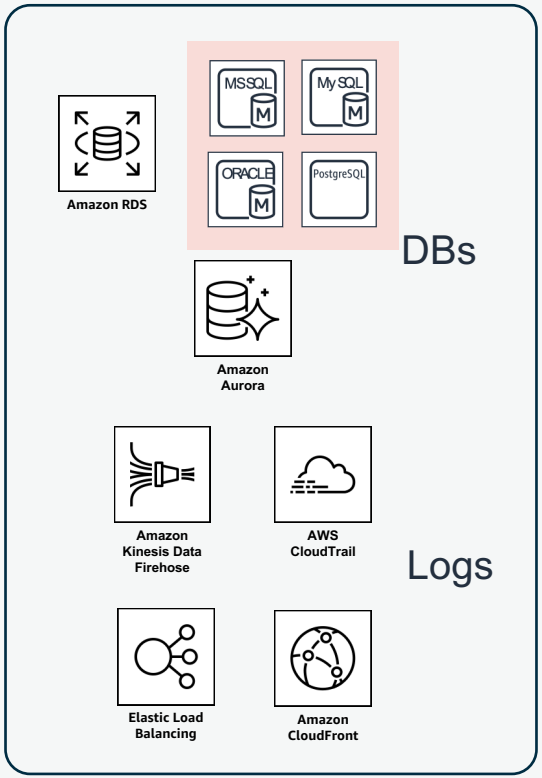
Athena

aws

AWS Lake Formation is fully integrated w/ AWS Glue

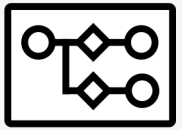


Easily load data into your data lake w/ blueprints



Prebuilt templates to serve common ingestion use cases

Automatically build AWS Glue **workflows**



AWS Glue jobs and crawlers discover, transform and structure data

Automatically populate the Data Catalog

Load data incrementally or in full

With blueprints

You

Point to data **source**

Specify data lake **location**

Specify data load **frequency**

Blueprints

Discover source table(s) schema

Convert to target data format

Partition data automatically

Track data that was already processed

Customize to your needs

aws

Glue ML Transformations

Deduplication

Transforming a dataset that has multiple rows referring to the *same actual thing* into a dataset where no two rows refer to the *same actual thing*

Record matching

Finding the relationships between multiple datasets, even when those datasets do not share an identifier (or when their identifier is unreliable)

Securing data lakes with Lake Formation

Ingestion & cleaning

Data Engineer

Serverless Spark

AWS Glue

Glue ML transformations

Blueprints

Security

Data Security Officer

Data catalog

Centralized permissions

Real time monitoring

Integrated auditing

Analytics & ML

Data Analyst

Comprehensive portfolio of integrated tools

Redshift

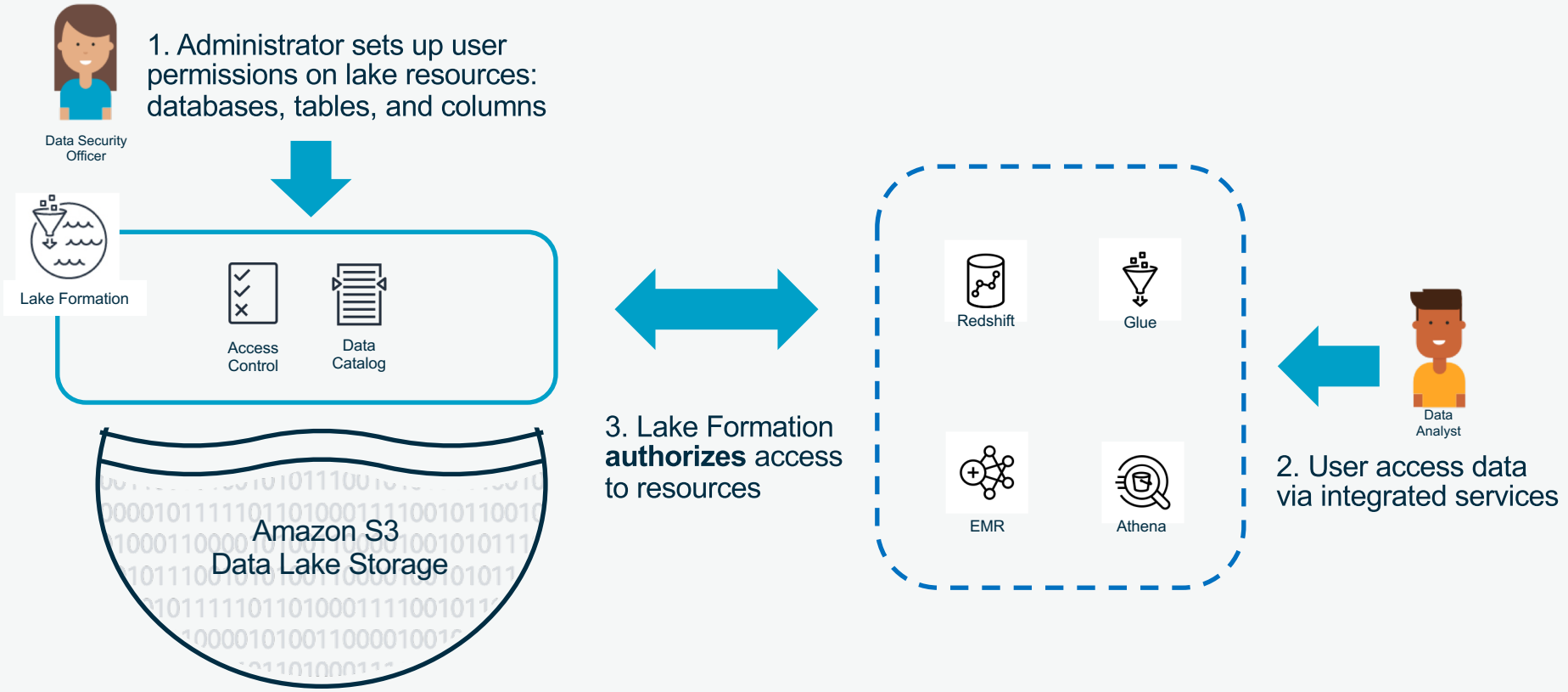
Glue

EMR

Athena

aws

Centralized permissions



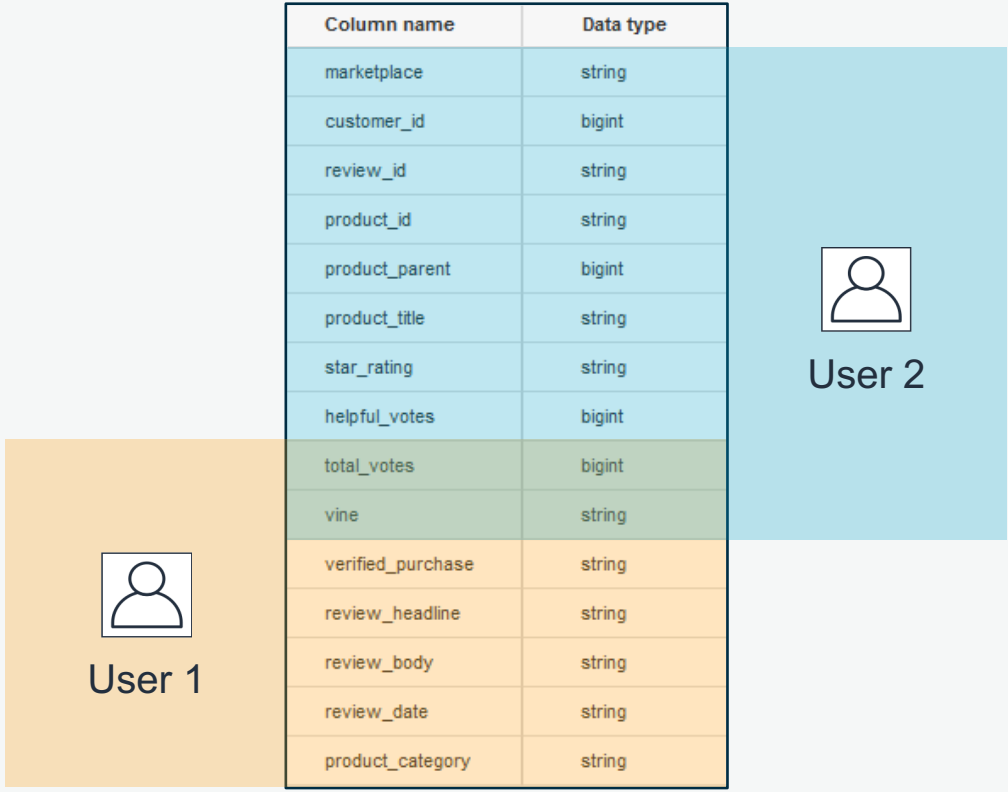
Security permissions in Lake Formation

Control data access with simple grant and revoke permissions

Specify permissions on **tables** and **columns** rather than on buckets and objects

Easily view permissions granted to a particular user

Audit all data access in one place



aws

Data Catalog & Permissions

Permissions are set on data catalog objects

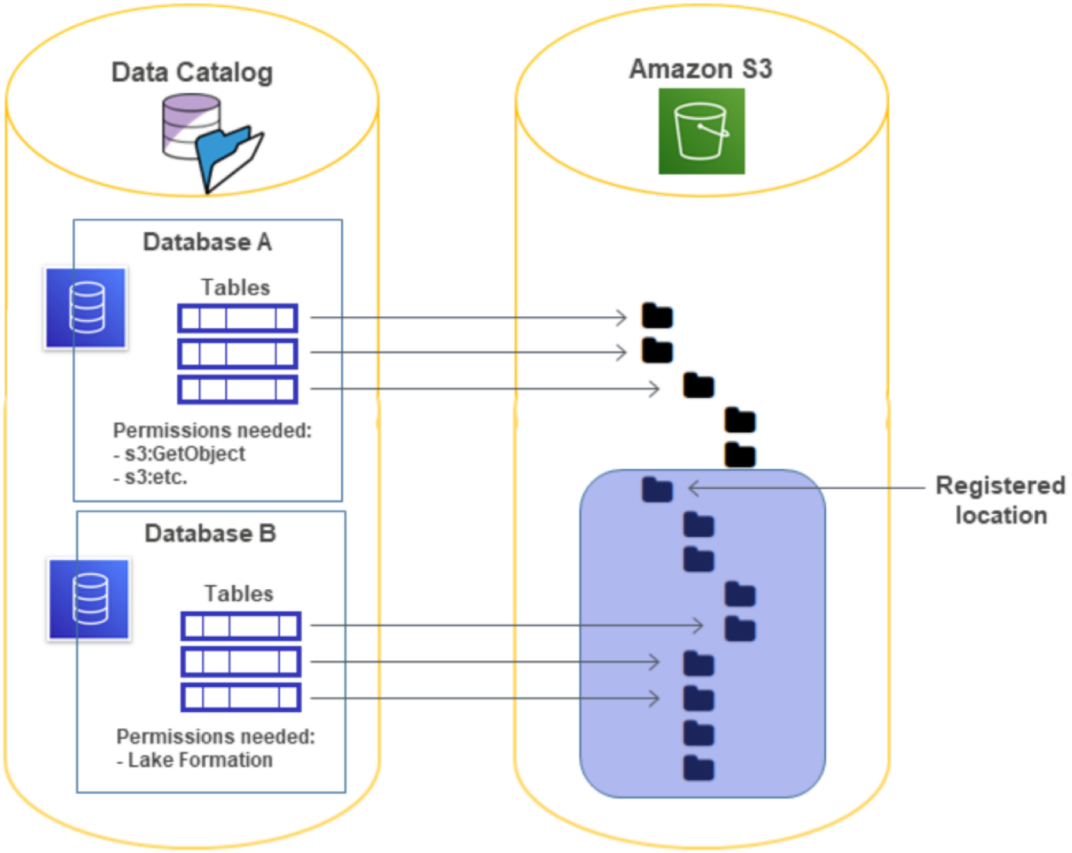
Lake Formation & **AWS Glue** use the same **Data Catalog**

Choice of using the **Glue** or the **Lake Formation** permissions system

For backwards compatibility, the default settings enable the **Glue** permissions system

Existing **Glue** crawlers, jobs, triggers and workflows will not change

Existing access to **Glue** resources will still be governed by **IAM & S3 policies**



Database A is an example of the old permissions model.

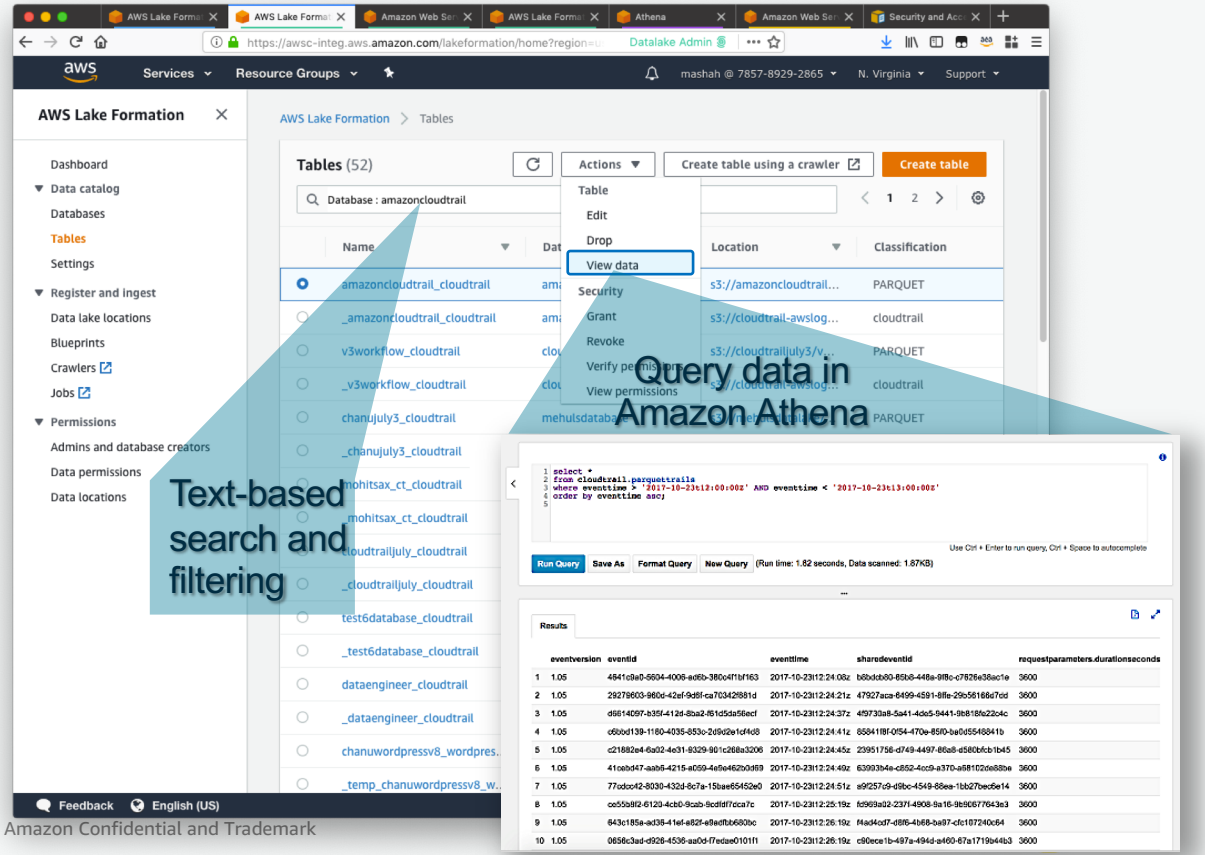
Database B is an example of the new Lake Formation permissions model.

Data catalog and metadata management

Text-based **search** across all metadata

Add attributes like data owners, stewards, and others as **table properties**

Add **data sensitivity** level, column definitions, and others as **column properties**

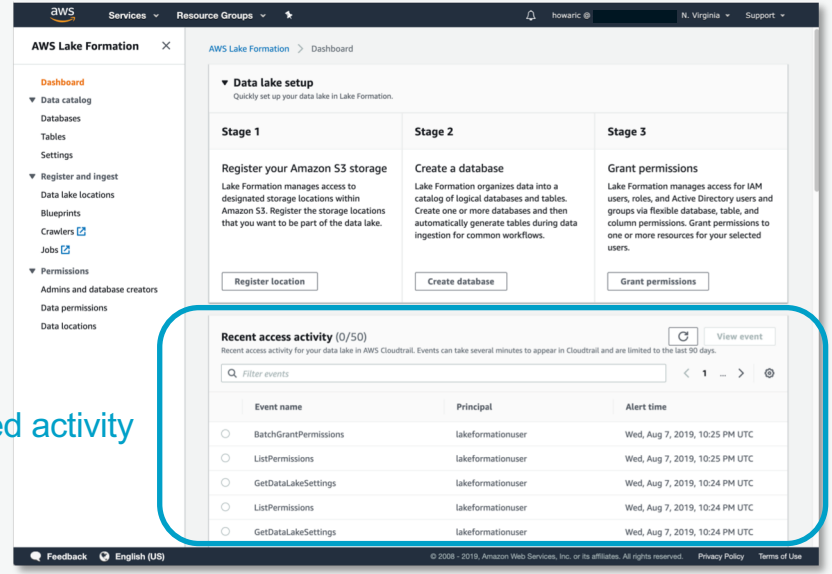


Audit and monitor in real time

See detailed activity in the console

Analyze audit logs in CloudTrail using Amazon Athena

Data ingest and catalog notifications also published to Amazon CloudWatch events



Detailed activity

Accessing data lakes with Lake Formation

Ingestion & cleaning

Data Engineer

Serverless Spark

AWS Glue

Glue ML transformations

Blueprints

Security

Data Security Officer

Data catalog

Centralized permissions

Real time monitoring

Auditing

Analytics & ML

Data Analyst

Comprehensive portfolio of integrated tools

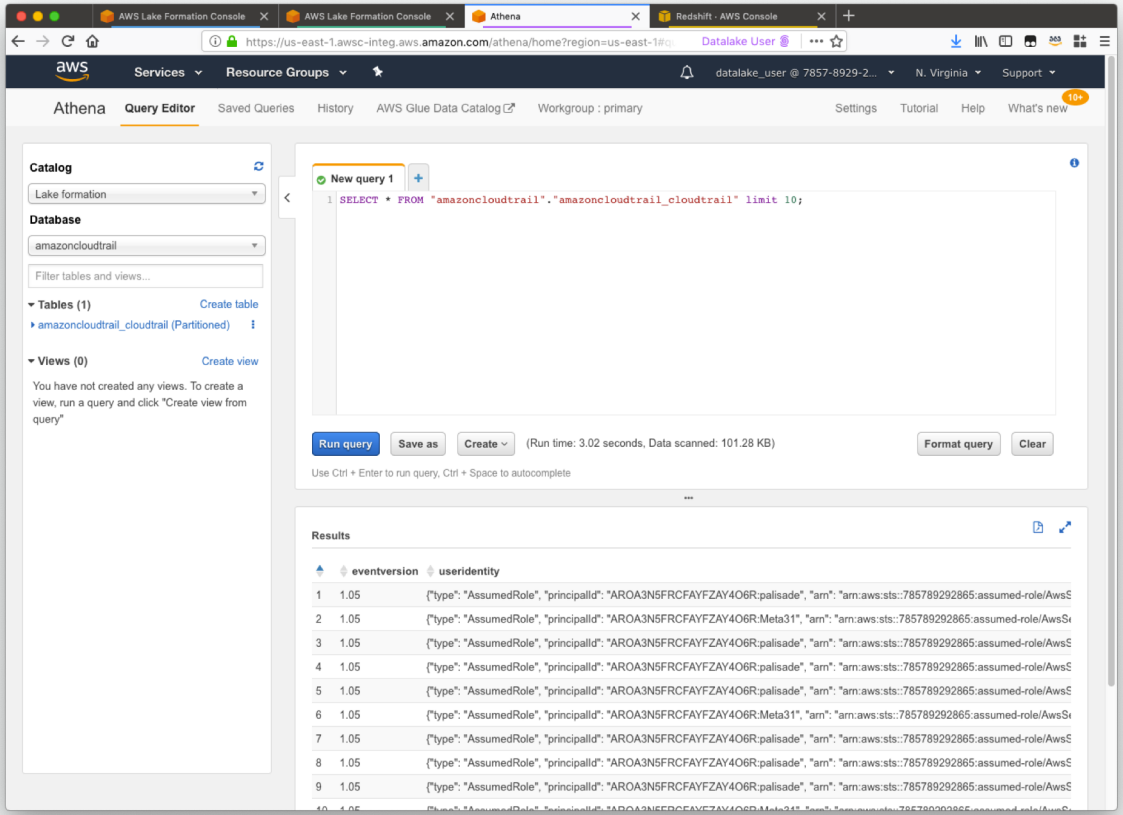
aws

Comprehensive portfolio of integrated tools

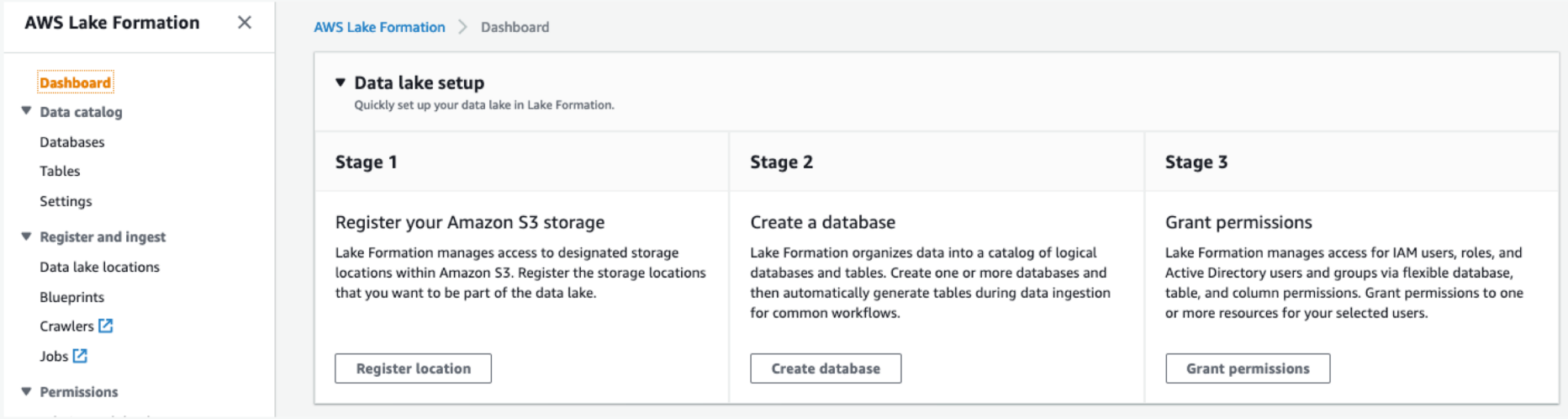
Compliant services honor Lake Formation permissions

They guarantee that users only see tables & columns they have access to

All access is logged and auditable



How does Lake Formation work?



Step 1: Register S3 path as data lake location

Data Engineer

Add storage

Amazon S3 storage

Register Amazon S3 storage as your data lake.

Amazon S3 path

Choose an Amazon S3 location for your data lake.

s3://v3-datalake-cloudtrail

Browse

Existing location permissions - *optional*

Registering the selected location may result in your users gaining access to data already at that location. Before registering the location, we recommend that you audit permissions on resources in the location.

Audit location permissions

IAM role

To add or update data, Lake Formation needs read/write access to the chosen Amazon S3 path. Choose a role that you know has permission to do this, or choose the **AWSServiceRoleForLakeFormationDataAccess** service-linked role. When you register the first Amazon S3 path, the service-linked role and a new inline policy are created on your behalf. Lake Formation adds the first path to the inline policy and attaches it to the service-linked role. When you register subsequent paths, Lake Formation adds the path to the existing policy.

AWSServiceRoleForLakeFormationDataAccess

Cancel

Add storage

aws

Step 2: Load data with blueprints

Data Engineer

Blueprint type

Configure a blueprint to create a workflow.

- ○ Database snapshot — Bulk load data to your data lake from MySQL, PostgreSQL, Oracle, and Microsoft SQL Server databases.
- ○ Incremental database — Load new data to your data lake from MySQL, PostgreSQL, Oracle, and SQL Server databases.
- ◉ AWS CloudTrail — Bulk load data from AWS CloudTrail sources.
- ○ AWS ELB logs
- ○ AWS ALB logs

Import source

Configure the workflow source.

CloudTrail name
Choose a CloudTrail source.
IsengardTrail-DO-NOT-DELETE

Start date
Choose a CloudTrail source start date.
2019/06/17

Import target

Configure the target of the workflow.

Target database
Choose a database in the Data Catalog. Create database
v3_chanu_cloudtrail

Data lake location
Choose where to import from your data lake storage locations.
s3://v3-datalake-cloudtrail

Data format
Choose the output data format.
Parquet

Import frequency

Schedule the workflow.

Frequency
Choose how often to run the workflow.
Run on demand

Import options

Configure the workflow.

Workflow name
v3chanucloudtrailtest
Names may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and must be less than 256 characters long.

IAM role
v3_datalake_service_role

Table prefix
The table prefix that is used for catalog tables that are created.
v3chanucloudtrailtest

Maximum capacity - *optional*
Enter a maximum capacity

Cancel Create

aws

Step 3: Grant permissions to users

Grant permissions

Grant access permissions to specific users and roles.

IAM users and roles
Add one or more IAM users or roles.
Choose IAM principals to add
datalake_user_redshift ✕
User

Database
Add one or more databases.
Choose databases
amazoncloudtrail ✕

Table - *optional*
Add one or more tables.
Choose tables
amazoncloudtrail_cloudtrail ✕

Column - *optional*
Grant permissions to:
The include columns

Include columns
Add one or more columns to include.
Choose columns
useridentity ✕ string
eventsource ✕ string
eventname ✕ string
sourceipaddress ✕ string

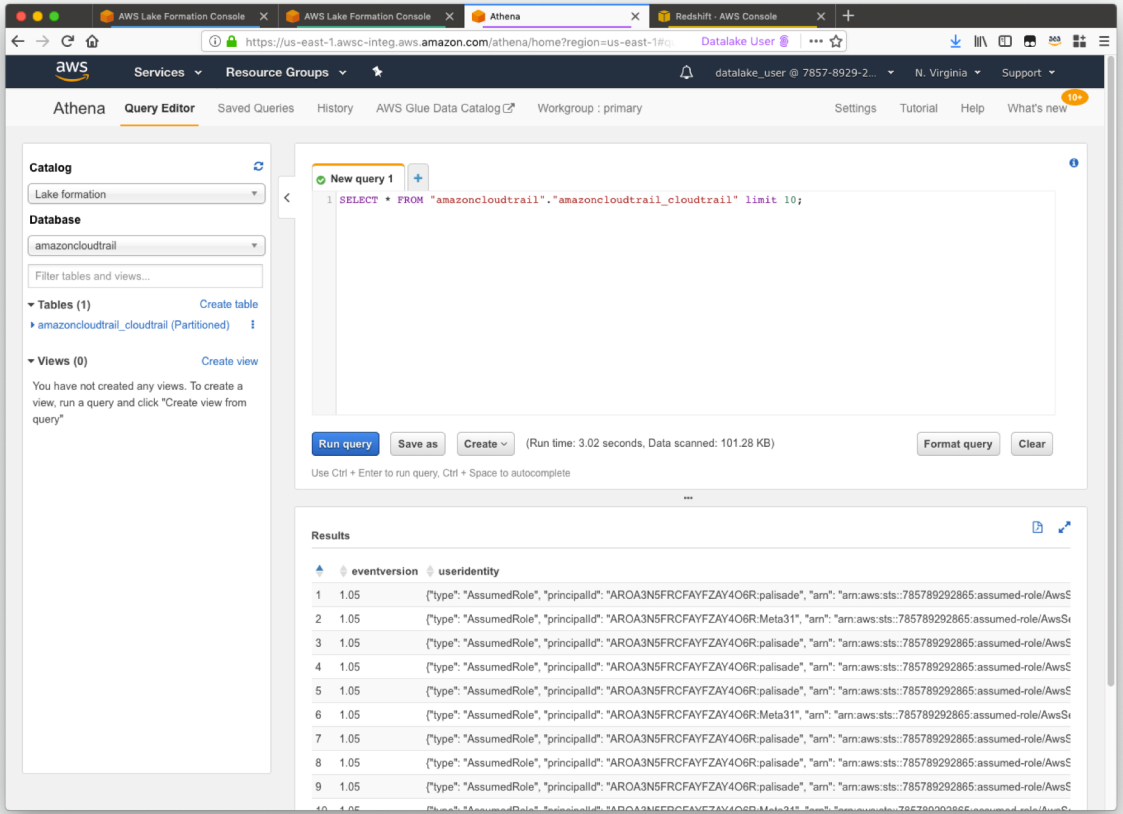
Table permissions
Choose the specific access permissions to grant.
☑ Select all ☑ Alter ☑ Insert ☑ Drop
☑ Delete ☑ Select
☐ Grant all
Enabling this permission grants full access to the specified resources while still logging access requests based on the individual permission settings above. It is typically used for debugging. Specific individual permissions set above will take effect when Grant all is later removed.

Grantable permissions
Choose the specific permissions that may be granted to others.
☐ Select all ☐ Create table ☐ Alter ☐ Drop
☐ Grant all
Enabling this permission grants full access to the specified resources while still logging access requests based on the individual permission settings above. It is typically used for debugging. Specific individual permissions set above will take effect when Grant all is later removed.

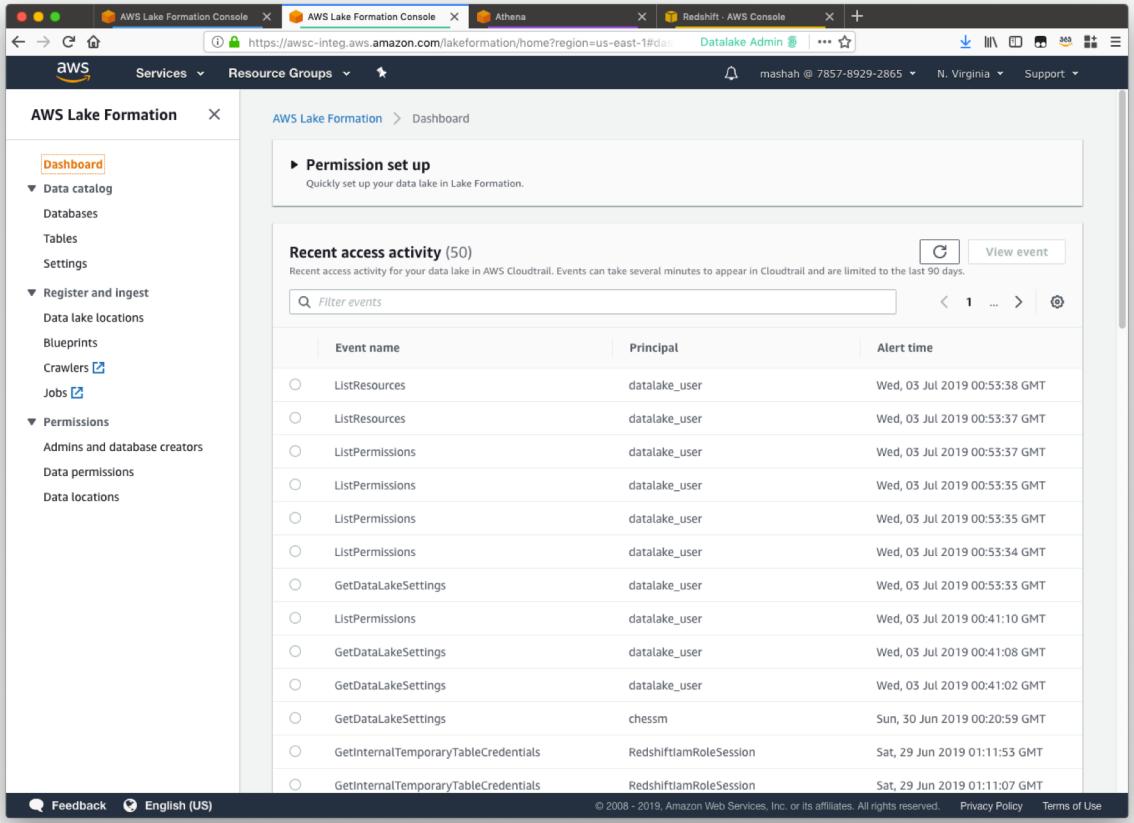
Cancel Grant

aws

Step 4: Query data with compatible services



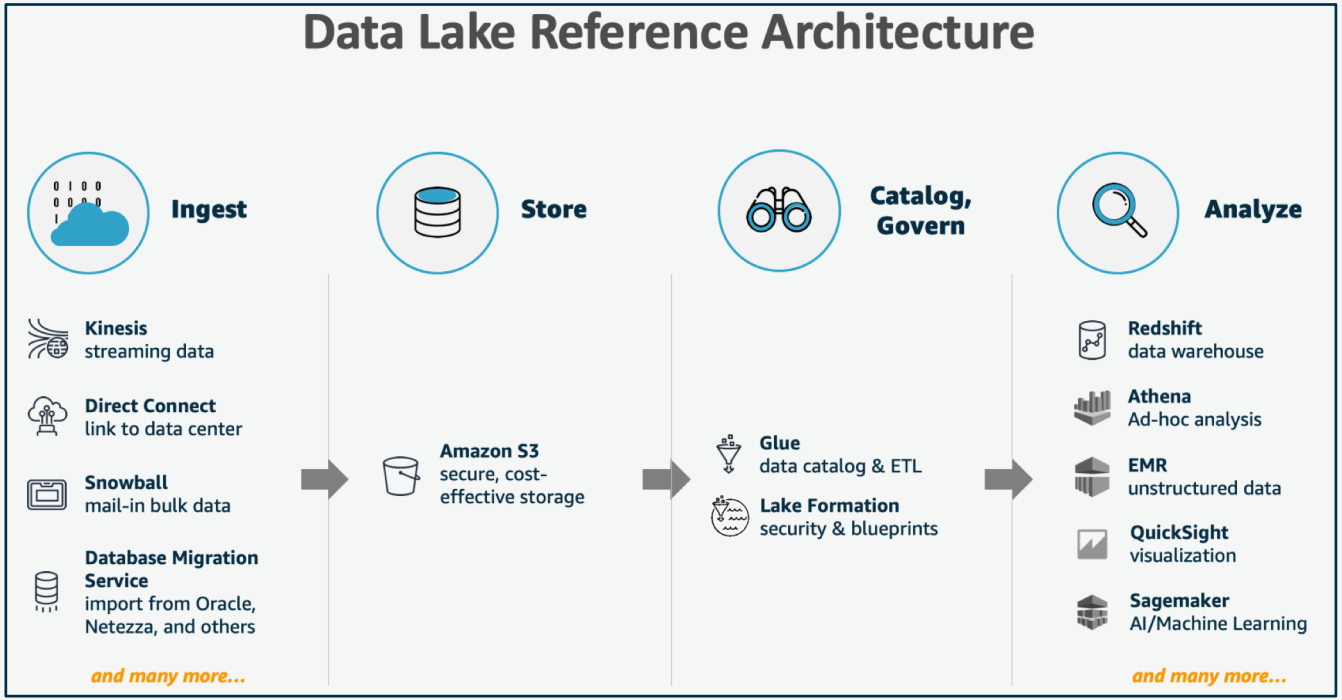
Step 5: Audit and monitor in real time



AWS Lake Formation Pricing

No additional charges – Only pay for the underlying services used.

Thank you!



Thank you!

Learn more:
https://aws.amazon.com/lake-formation/