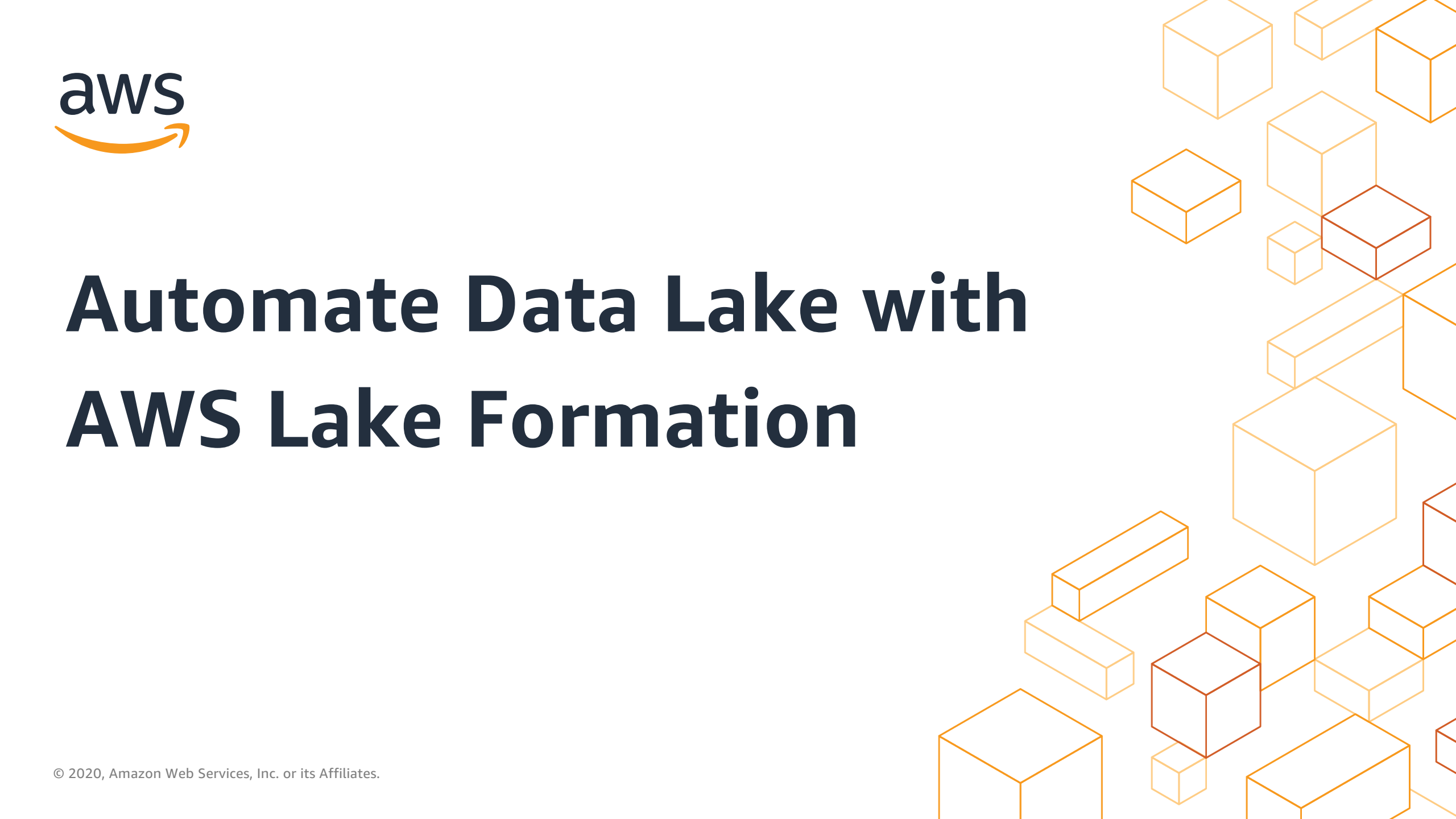
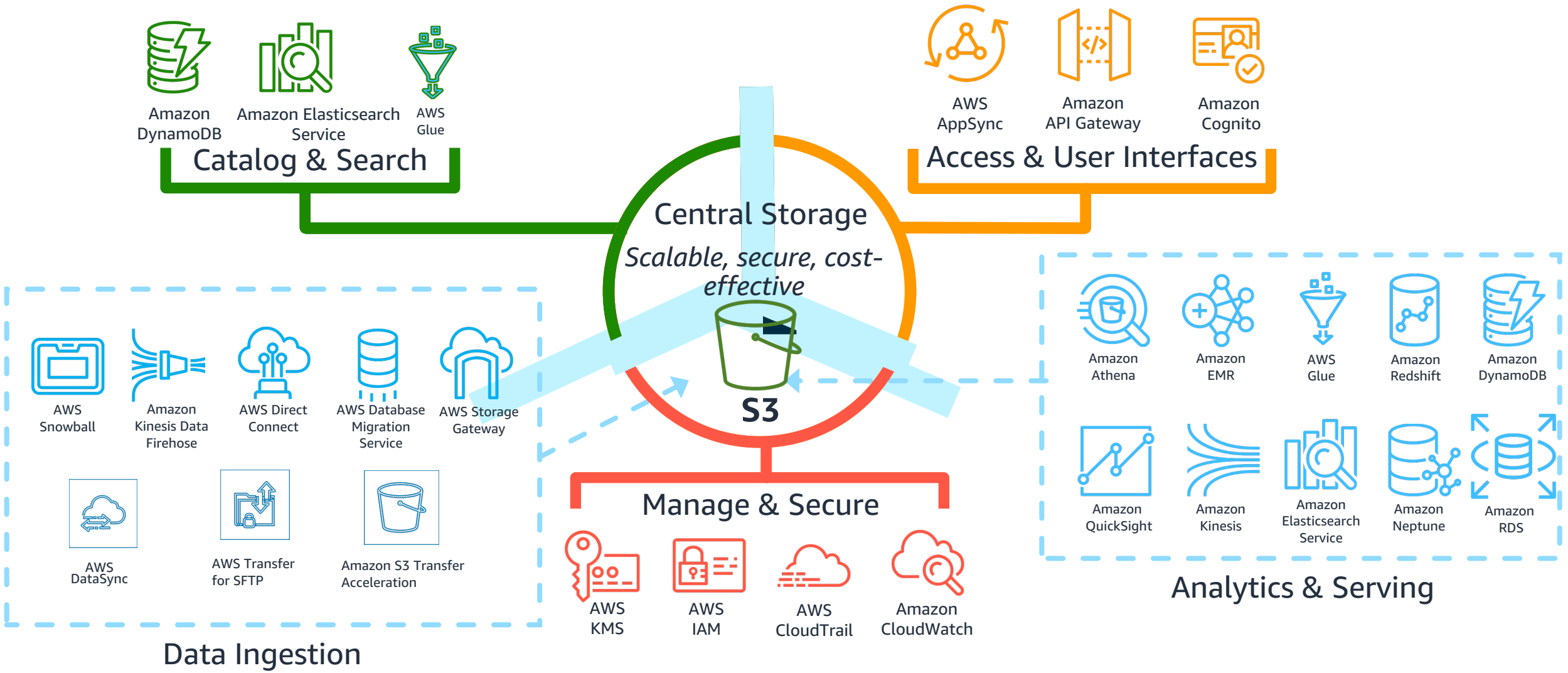


Automate Data Lake with AWS Lake Formation

Covering It All With Lake Formation



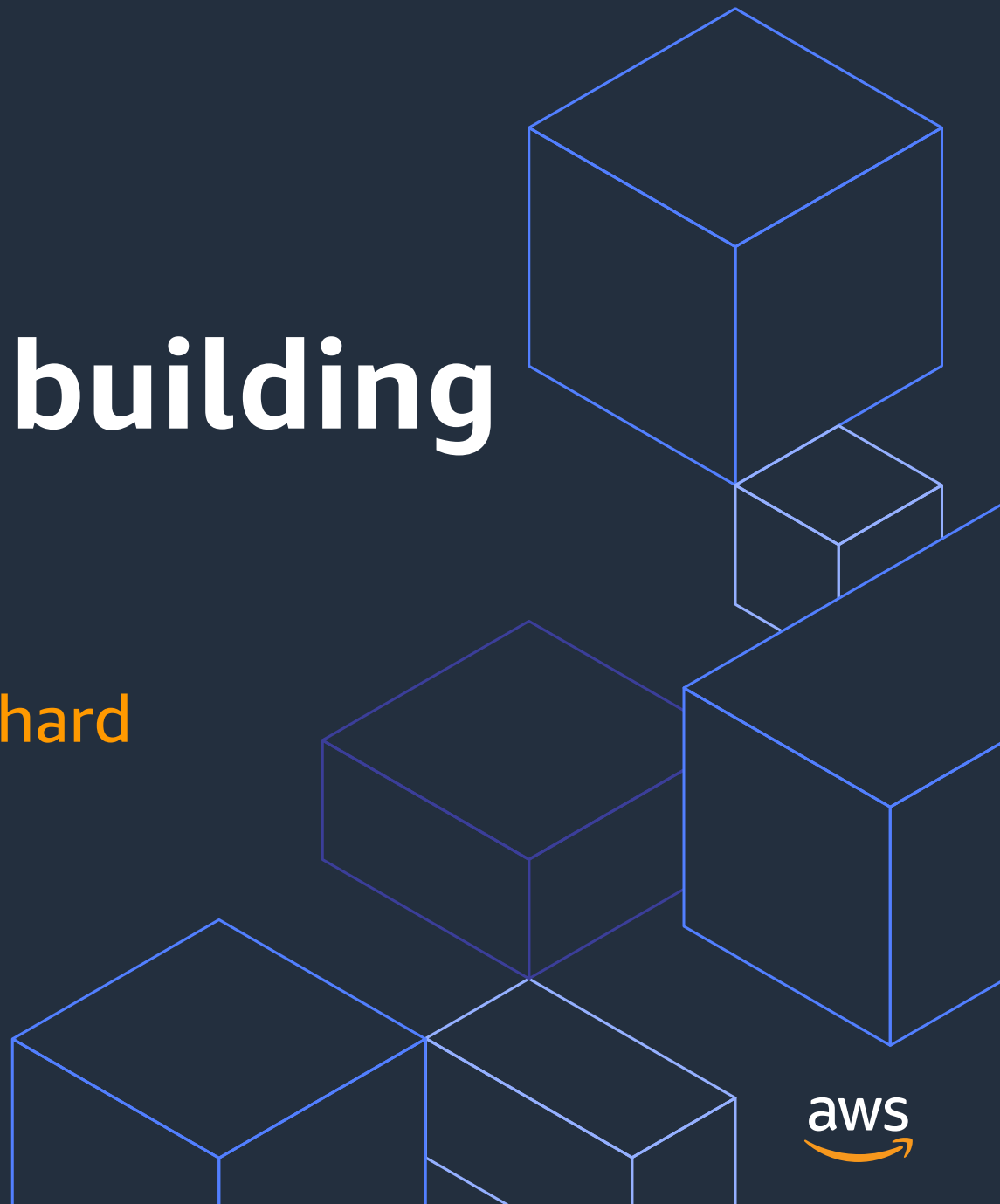
Agenda

- Lake Formation Value Proposition
- AWS Lake Formation
    - Steps to build Data Lake
    - Reference Architecture
- Lake Formation Console Walkthrough

aws

What is hard about building data lakes?

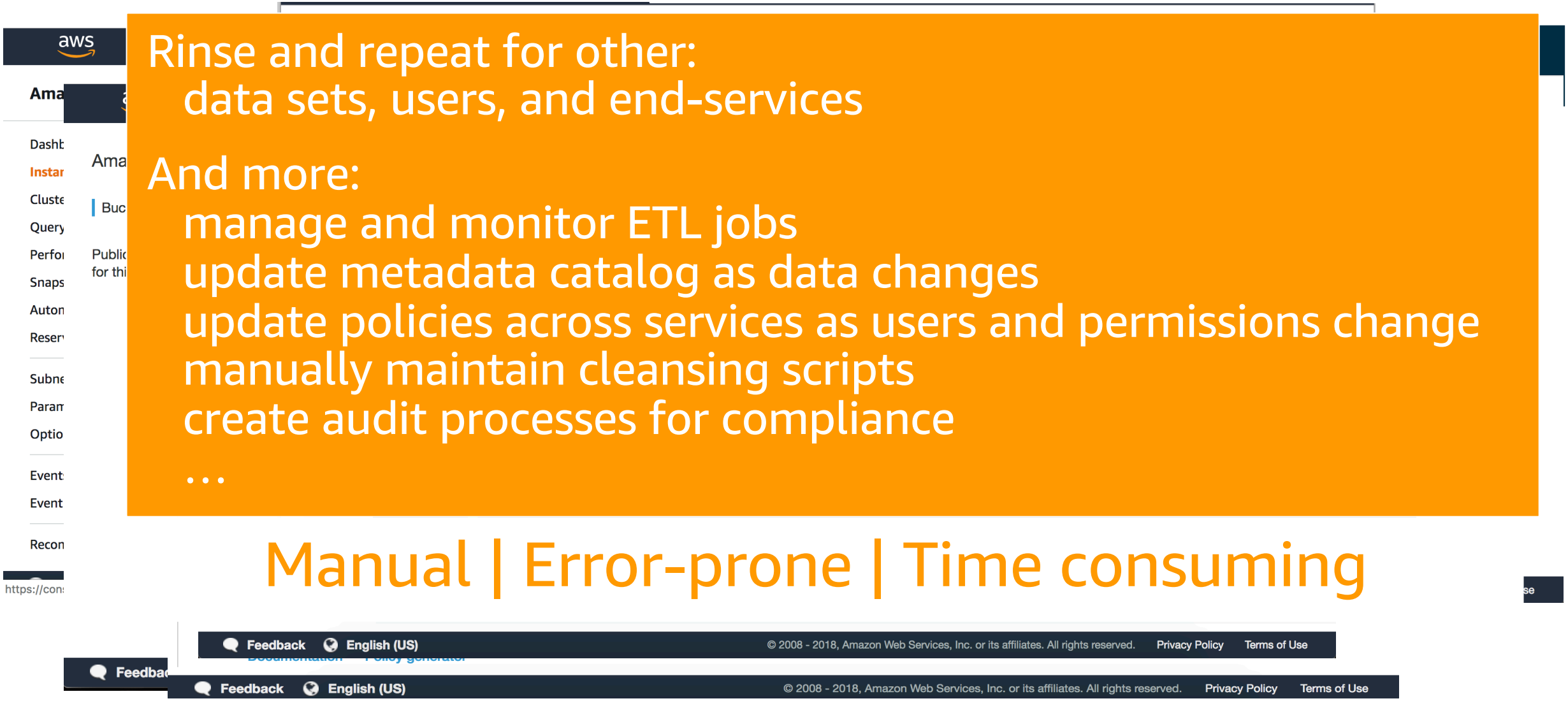
Manually building secure data lakes is hard



aws

Sample of steps required

Configure access from analytics services



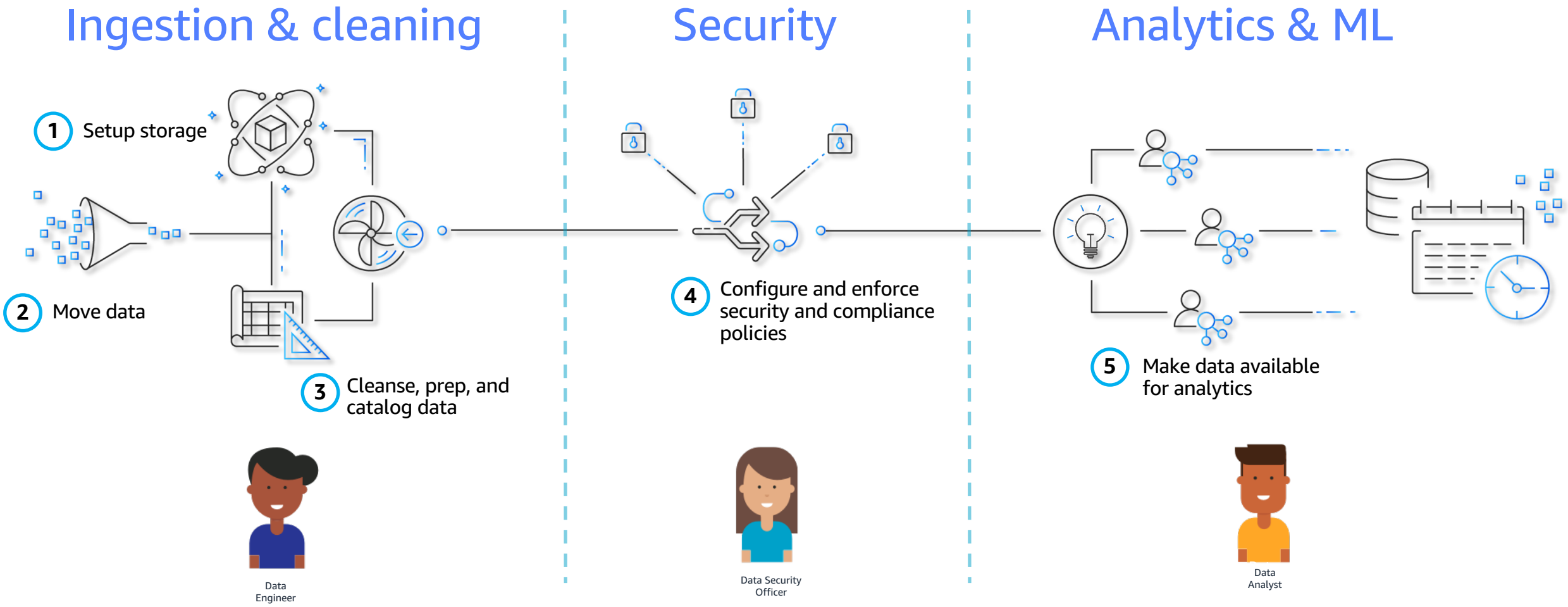
Rinse and repeat for other:
- data sets, users, and end-services

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

Feedback English (US) © 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

Typical steps of building a data lake

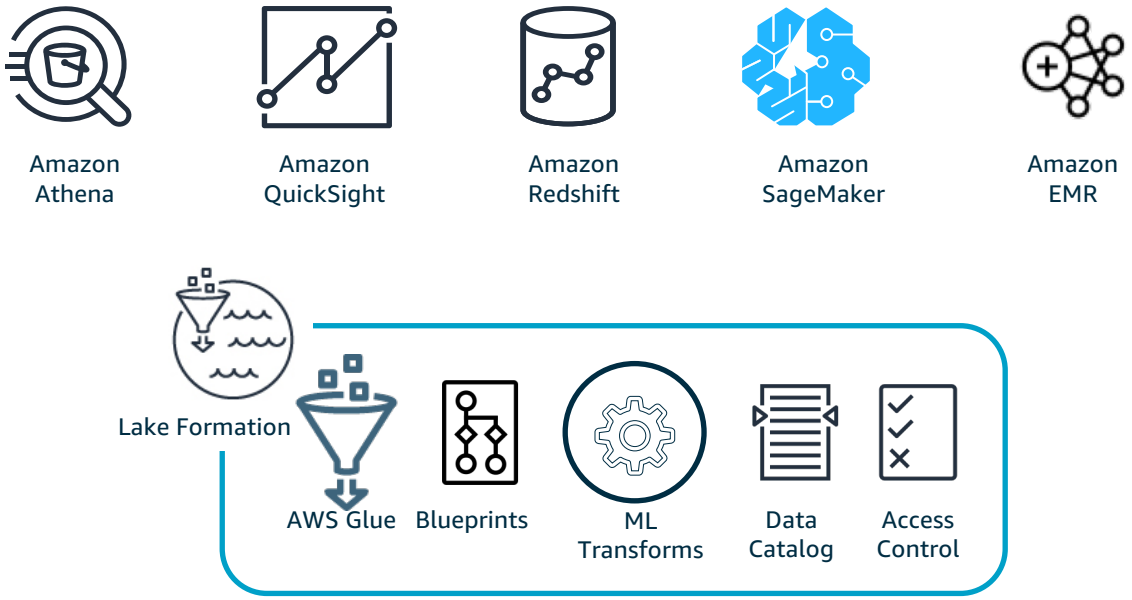


What is Lake Formation?

Automating Data Lake

aws

Features and Benefits of Lake Formation



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

Building data lakes with Lake Formation

Ingestion & cleaning

AWS Glue

Serverless Spark

Blueprints

ML Transforms

Security

Data catalog

Centralized permissions

Real time monitoring

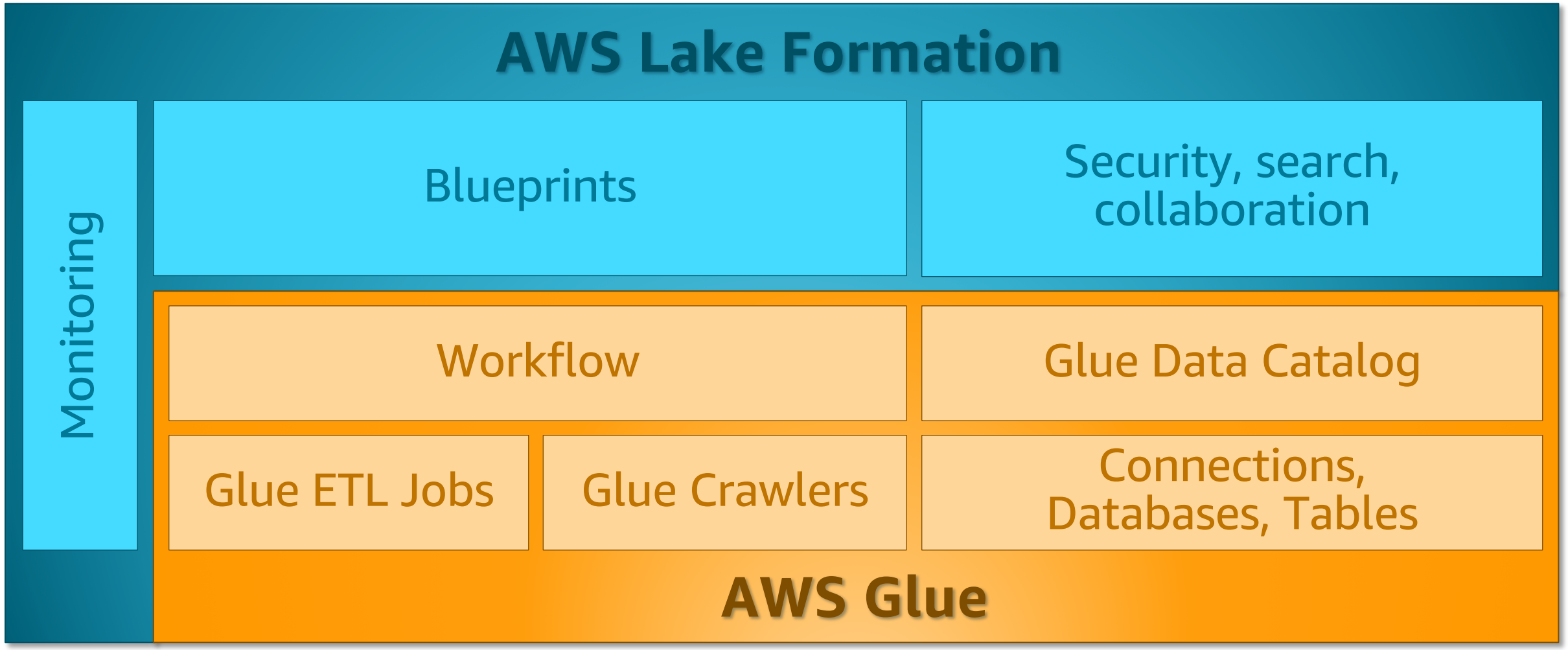
Auditing

Analytics & ML

Comprehensive portfolio of integrated tools

aws

AWS Lake Formation is fully integrated w/ AWS Glue



aws

AWS Glue Components (Recap)

Data Catalog

Discover

- Automatic crawling
- Apache Hive Metastore compatible
- Integrated with AWS analytic services

Serverless Engine

Develop

- Apache Spark
- Python shell
- Interactive and batch jobs

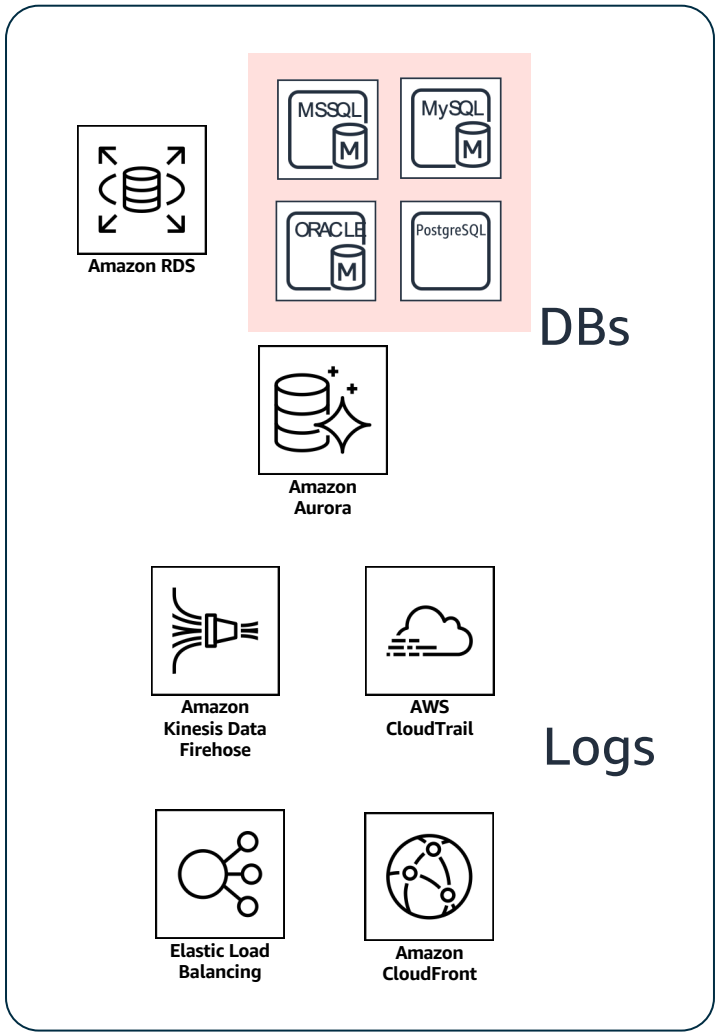
Orchestration

Deploy

- Flexible workflows
- Monitoring and alerting
- External integrations

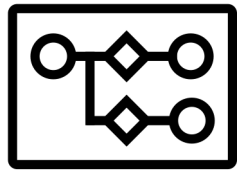
aws

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

aws

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

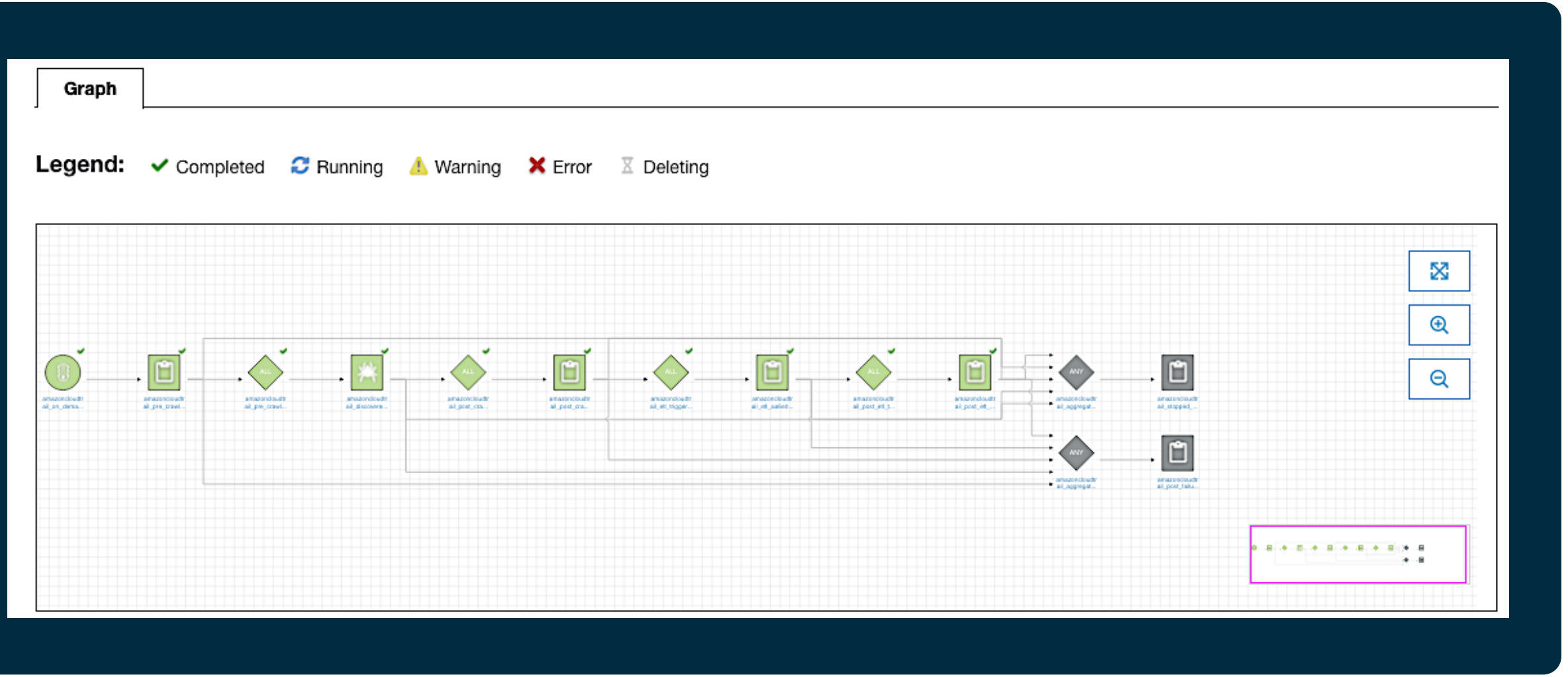
Partition data automatically

Track data that was already processed

Customize to your needs

aws

Blueprints create AWS Glue workflows



Securing data lakes with Lake Formation

Ingestion & cleaning

Serverless Spark

AWS Glue

Glue ML transformations

Blueprints

Security

Data catalog

Centralized permissions

Real time monitoring

Integrated auditing

Analytics & ML

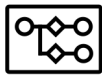
Comprehensive portfolio of integrated tools

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**

aws

Upgrading to the Lake Formation permissions model

Not using the Glue Catalog?

Change the default settings to start using the Lake Formation permissions system

Using the Glue Catalog?

Explicitly upgrade each data location, database and table when ready

1) Understand existing policies / access / usage

2) Configure corresponding Lake Formation policies

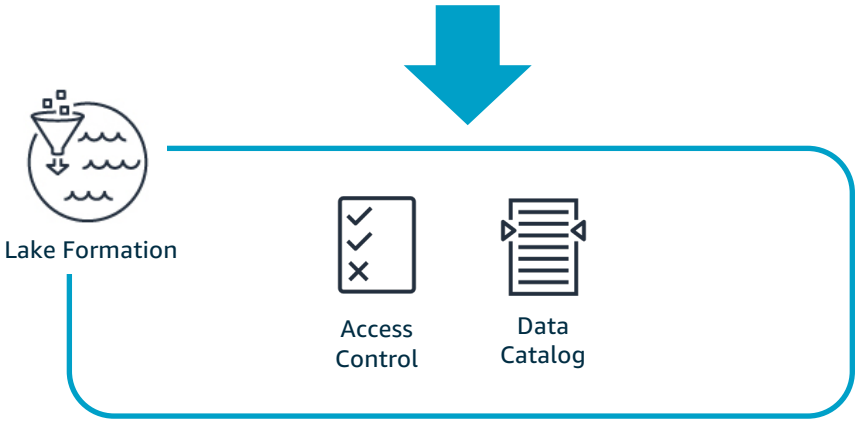
3) Remove the Glue permissions system by changing the default settings

4) Turn on the Lake Formation permissions system by registering the location

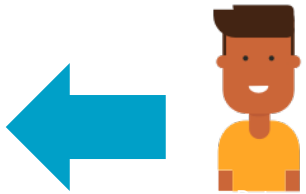
aws

Centralized permissions

1. Administrator sets up user permissions on lake resources: databases, tables, and columns



3. Lake Formation **authorizes** access to resources

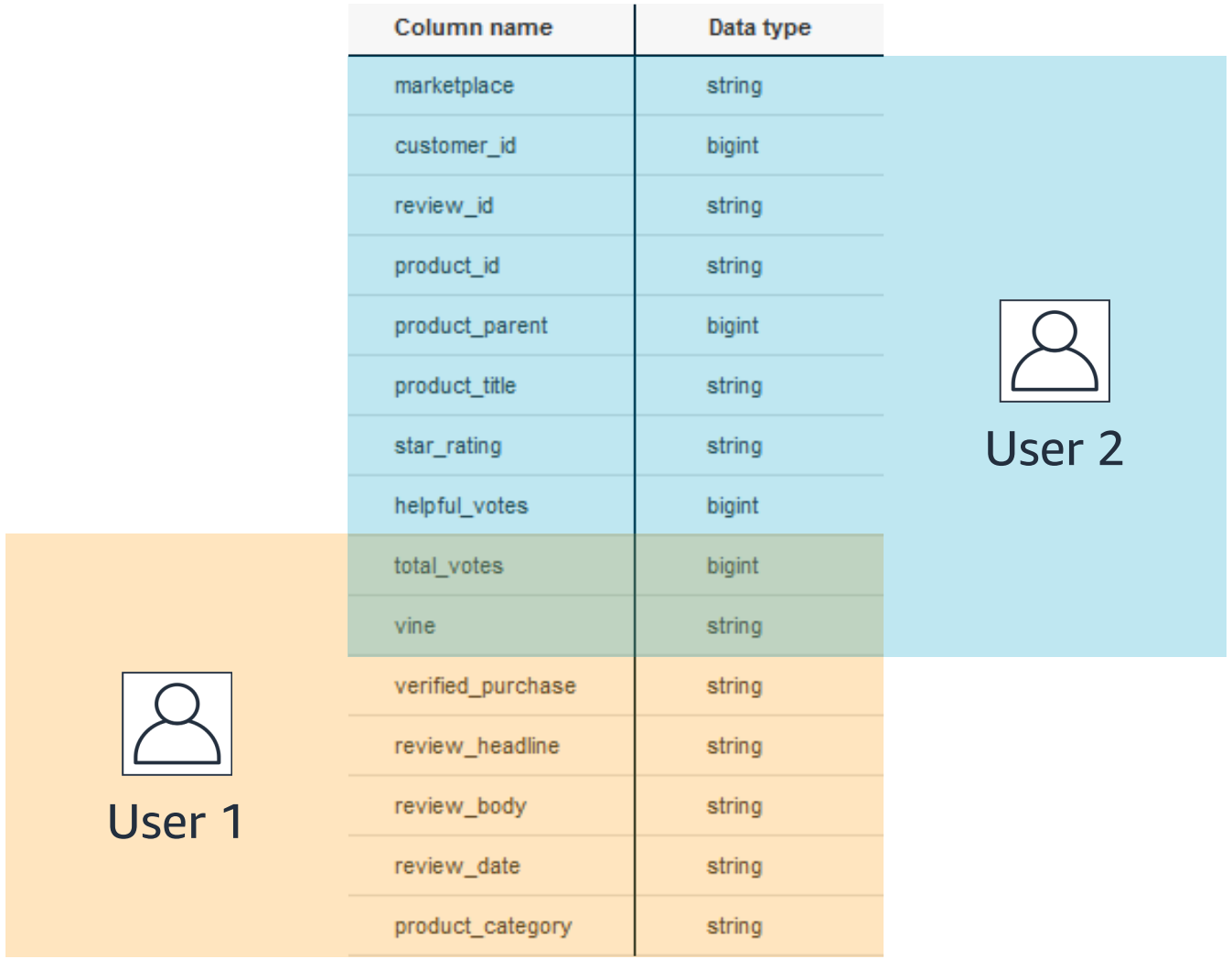


2. User access data via integrated services

aws

Security permissions in Lake Formation

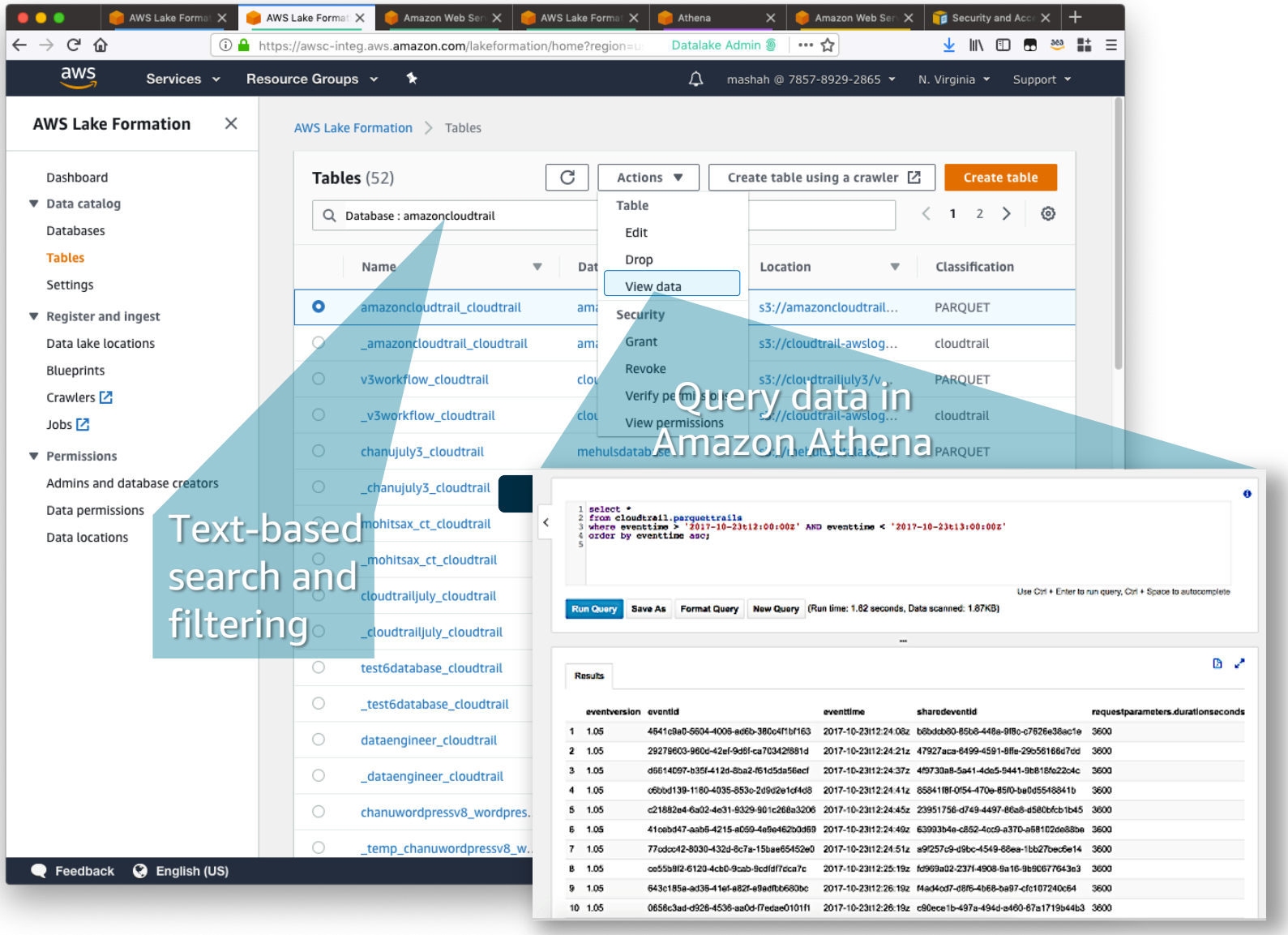
- Control data access with simple grant and revoke permissions
- Specify permissions on **tables** and **columns** rather than on buckets and objects
- Easily view permissions granted to a particular user
- Audit all data access in one place



aws

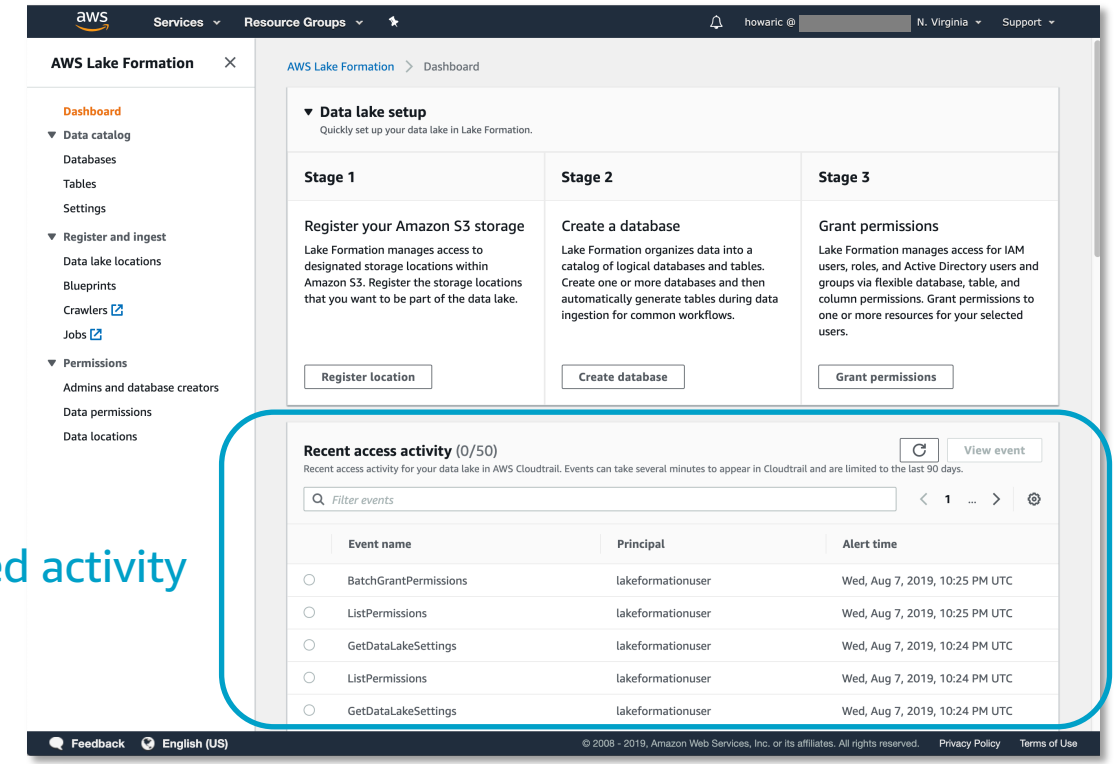
Data catalog and metadata management

- Text-based **search** 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

Serverless Spark

AWS Glue

Glue ML transformations

Blueprints

Security

Data catalog

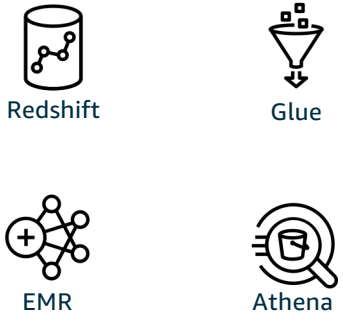
Centralized permissions

Real time monitoring

Auditing

Analytics & ML

Comprehensive portfolio of integrated tools



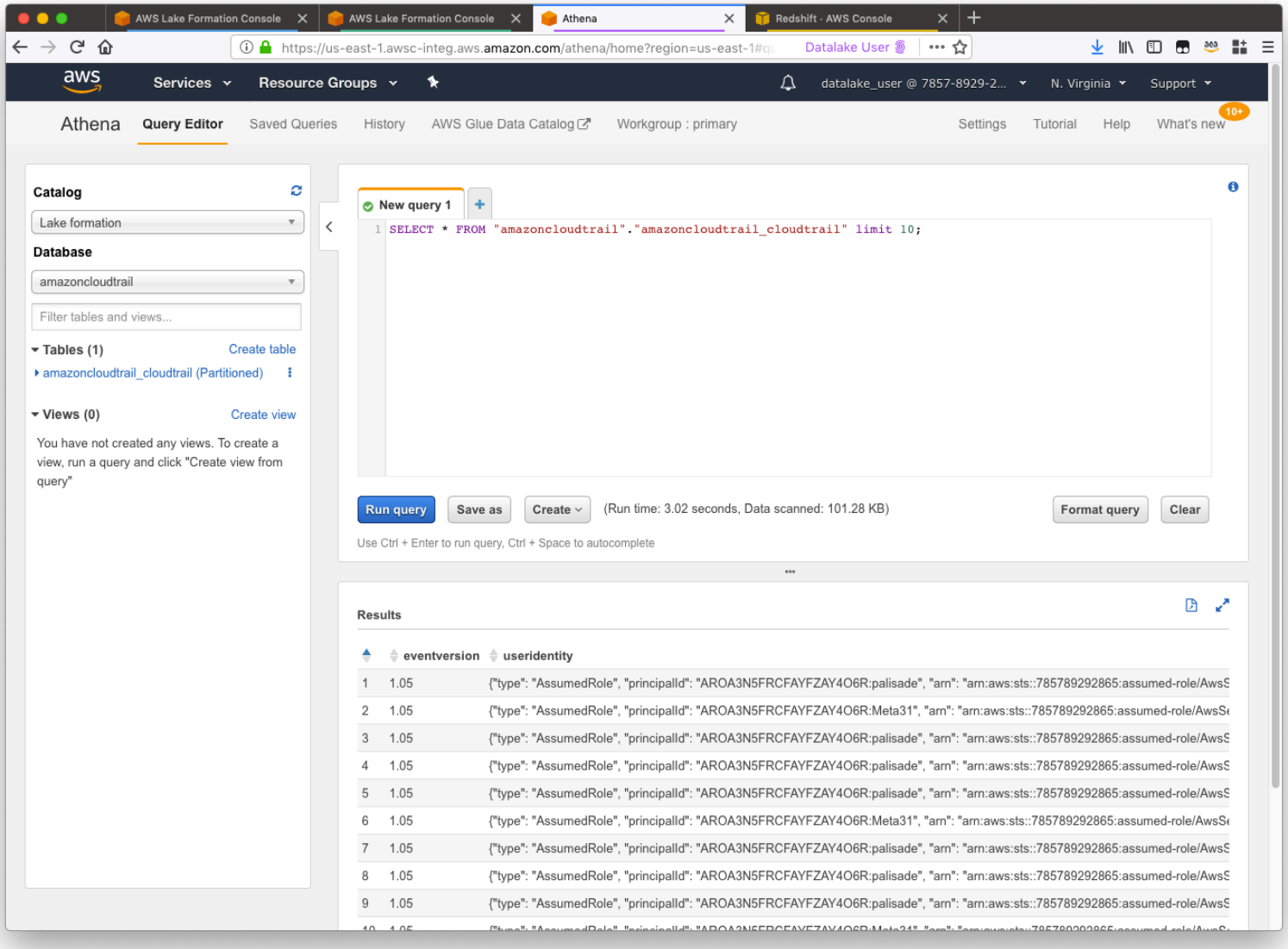
aws

Comprehensive portfolio of integrated tools

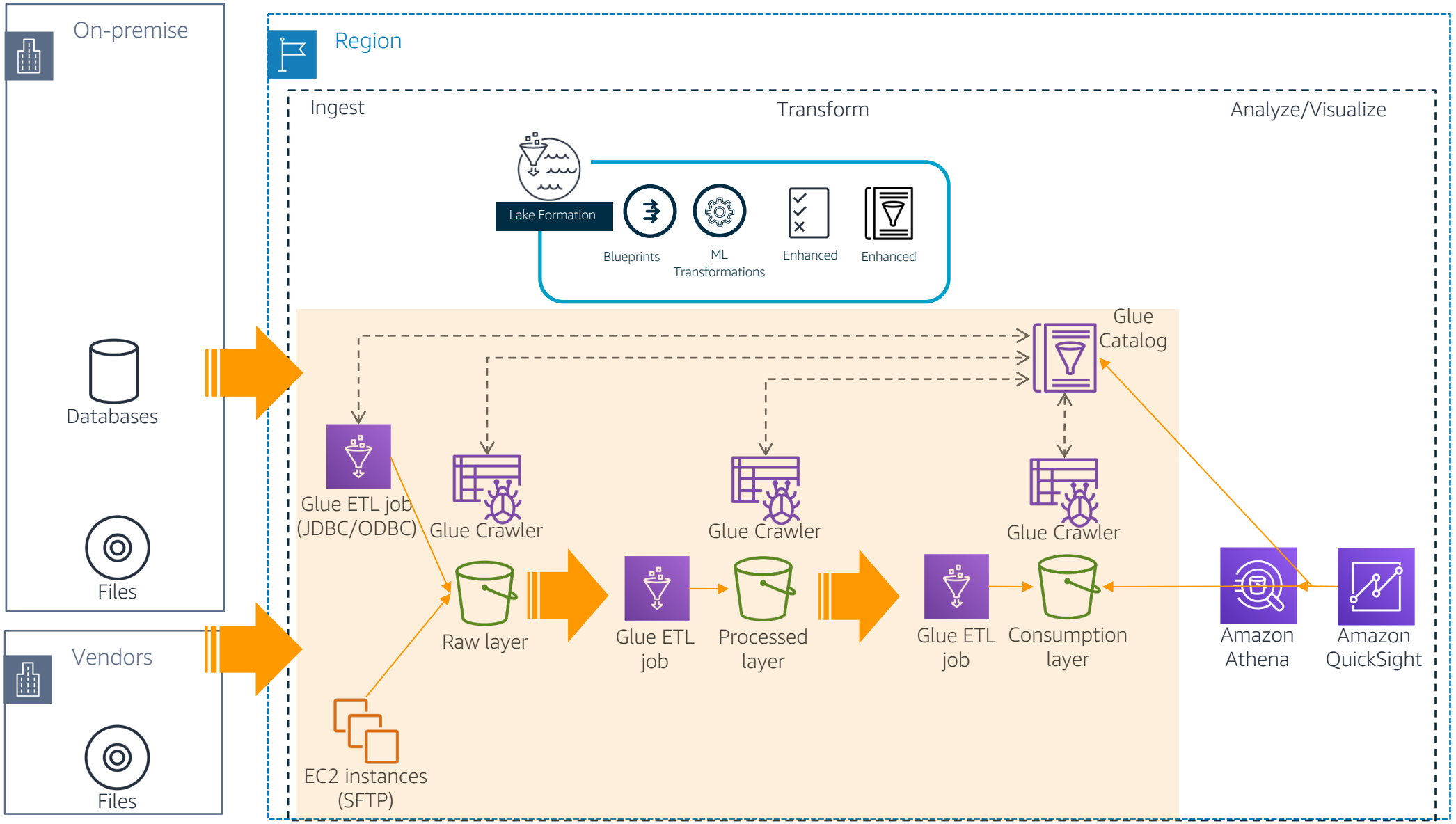
- Compliant services honor Lake Formation permissions

- They guarantee that users only see tables & columns

- All access is logged and auditable

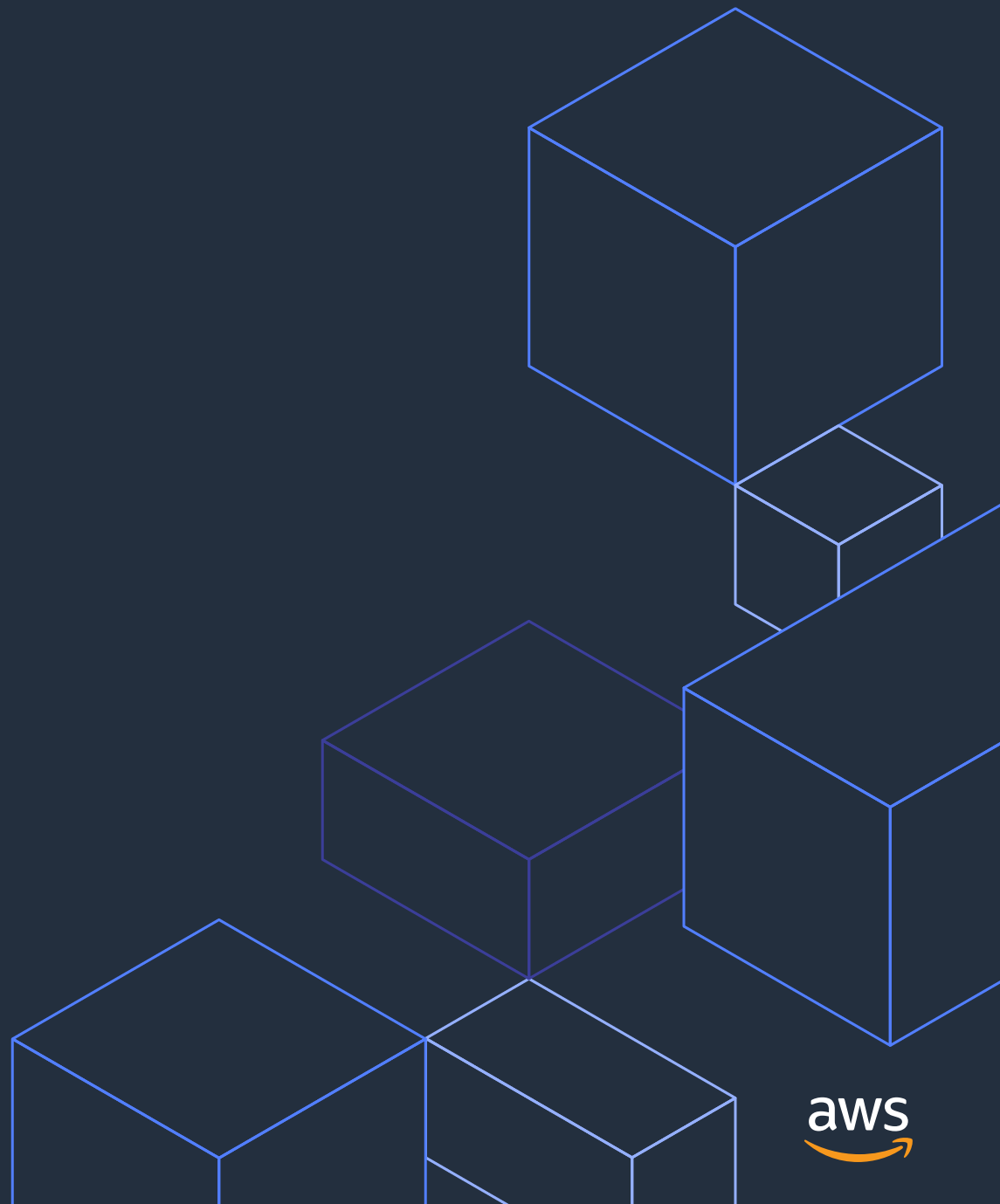


Serverless Lake Formation for Financial Services



How It Works

Console Walkthrough



Step 1: Create a data lake administrator

A user who can create databases and tables in the data catalog and grant fine-grained permissions for other users

AWS Lake Formation > Admins and database creators

Data lake administrators (0/2)

Administrators can view all metadata in the Data Catalog. They can also grant and revoke permissions on data resources to principals, including themselves.

Grant

Filter administrators

Name	Type
Admin	IAM role
pratik	IAM user

Database creators

Choose IAM principals permitted to create databases in your Data Catalog.

Revoke Grant

Filter database creators

1

Principal	Principal type	Permissions	Grantable
No permissions			

aws

Step 2: Register S3 path as data lake location

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.

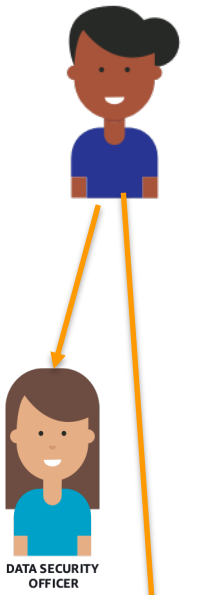
AWSServiceRoleForLakeFormationDataAccess

Cancel

Add storage

aws

Step 3: Grant permissions to users at tables and columnar level



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

Step 4: Load data with blueprints

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

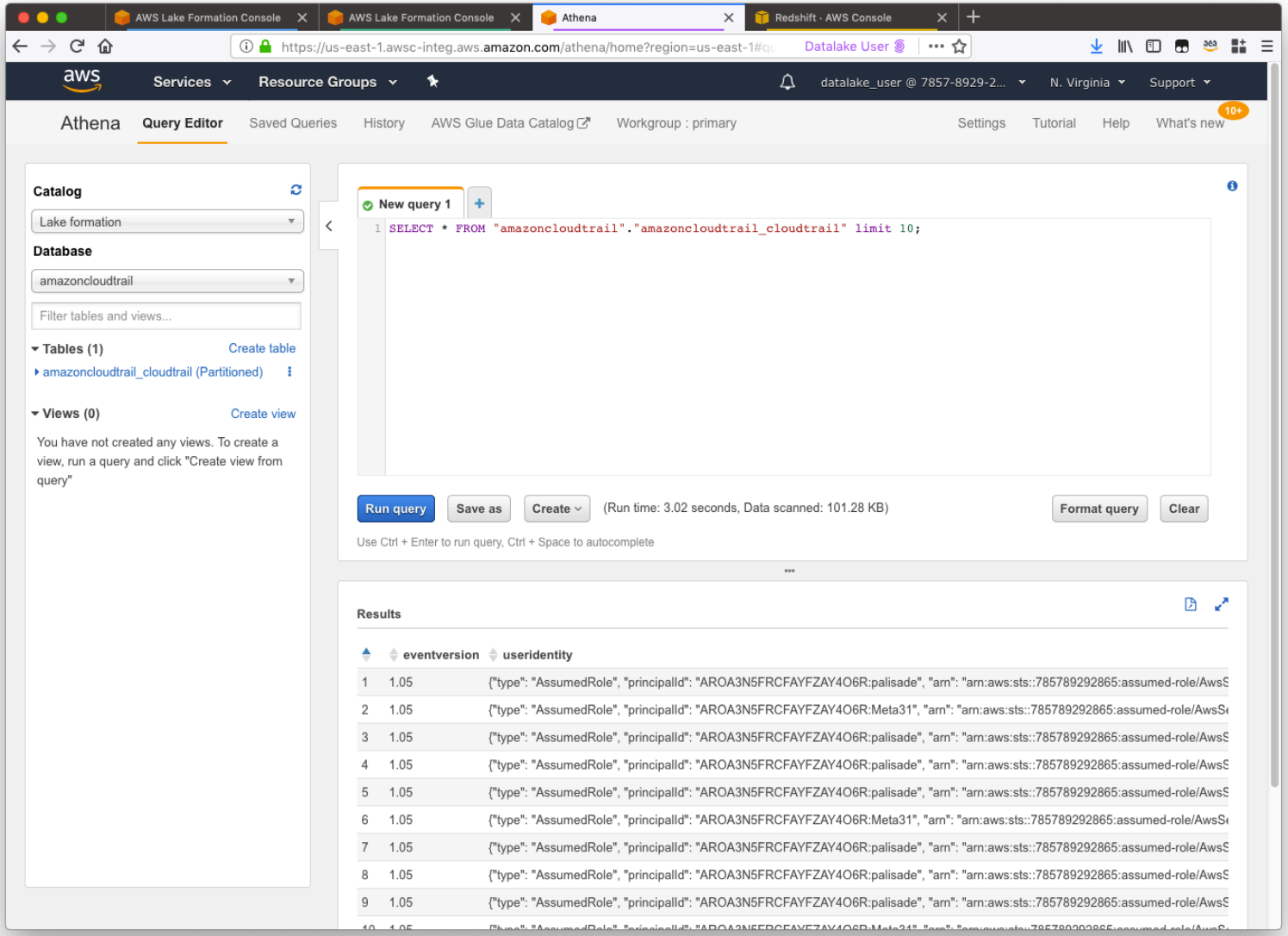
Table prefix
The table prefix that is used for catalog tables that are created.
v3chanucloudtrailtest

Maximum capacity - *optional*
Enter a maximum capacity

Cancel Create

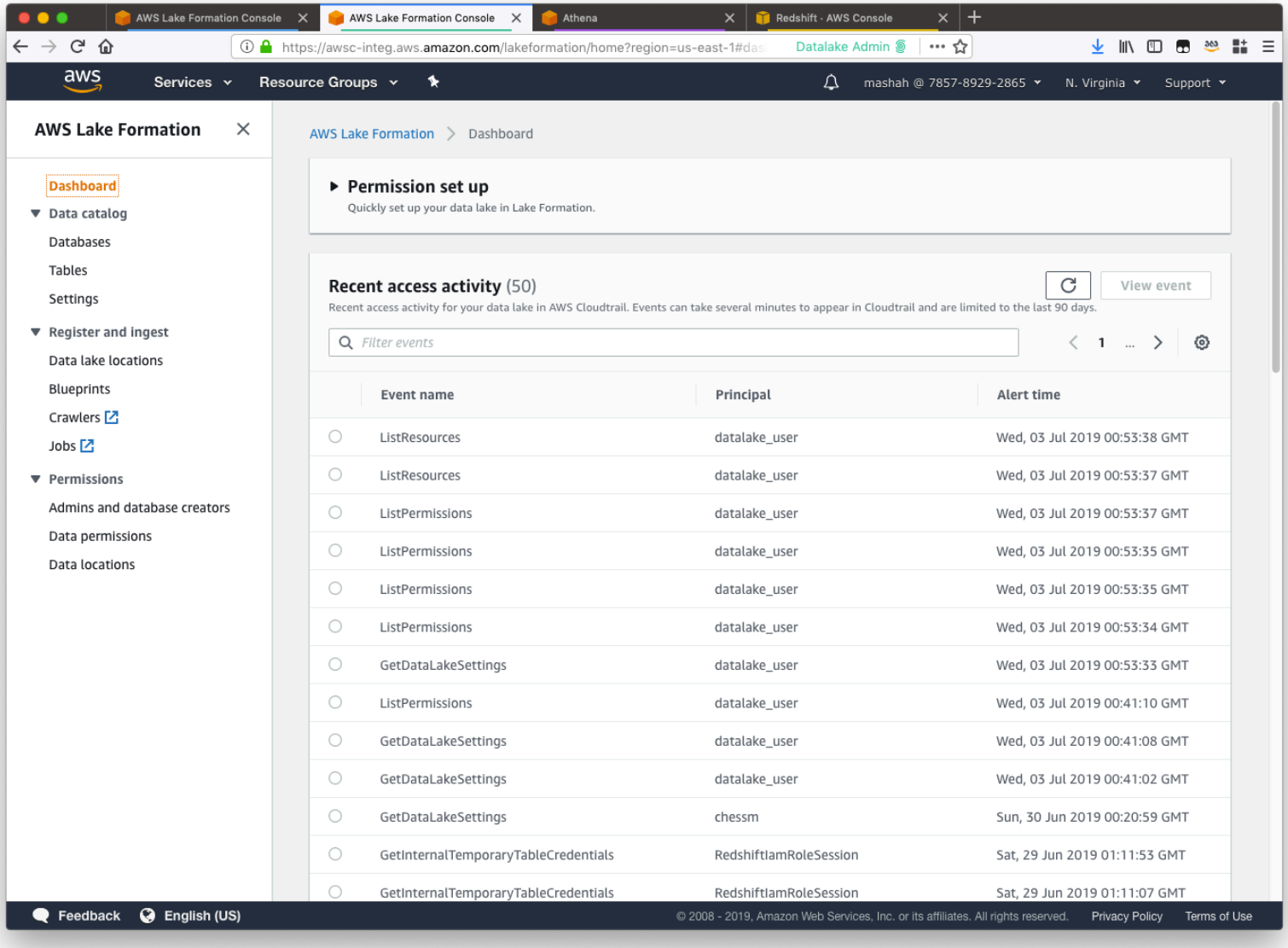
Step 5: Query data with compatible services

Data Analyst



aws

Step 6: Audit and monitor in real time



Thank you!

aws