AWS re:Invent

GPSTEC347

Accelerating data lake creation using AWS Lake Formation

Shashi Raina
Partner Solutions Architect
Amazon Web Solutions

AWS re:Invent

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Agenda

Why data lakes?

What is hard about building data lakes?

Why AWS Lake Formation?

What is Lake Formation?

How it works

Why data lakes?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Decision making used to …

… revolve around the

Enterprise data warehouse



Data no longer fits



There is **more data** than people think

Data is **more diverse**

Data	Data platforms need to	
grows **> 10x** every five years	live for **15** years	scale **1,000x**

*IDC, Data Age 20215: The Evolution of Data to Life-Critical Don't Focus on Big Data, Focus on the Data That's Big, April 2017.

Data lake: The new information hub

A **centralized secure repository** that enables you to **govern**, **discover**, **share**, and **analyze structured and unstructured data** at any scale

What is hard about building data lakes?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

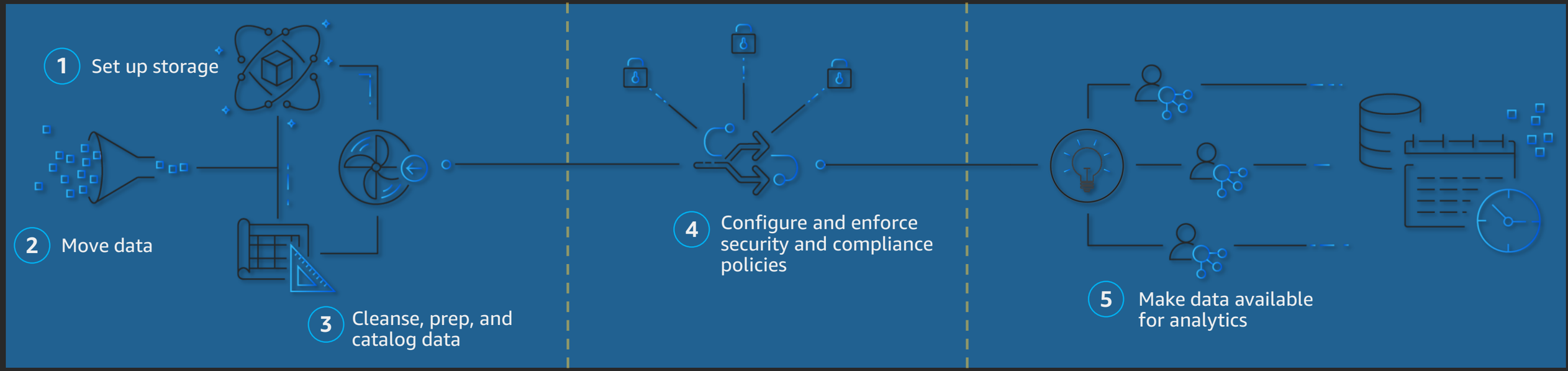
Manually building secure data lakes is **hard**

Typical steps of building a data lake

Ingestion and cleaning

Security

Analytics and ML





Data engineer



Data security officer



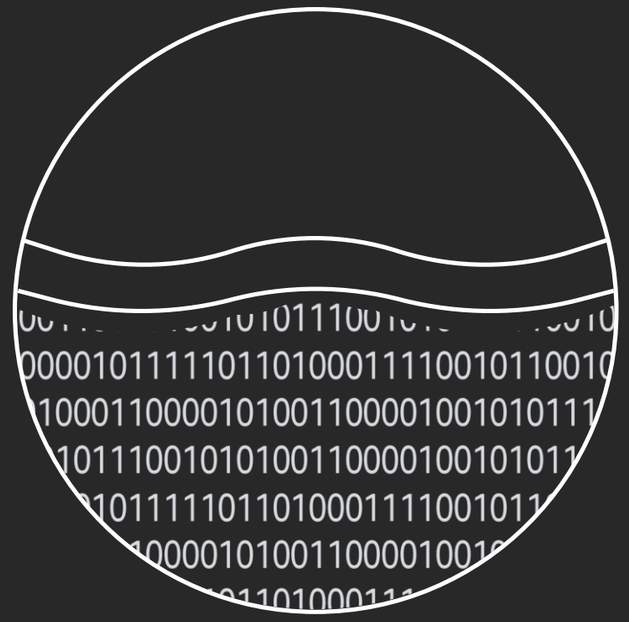
Data analyst

Why AWS Lake Formation for data lakes?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Lake Formation lets you build secure data lakes in *days*

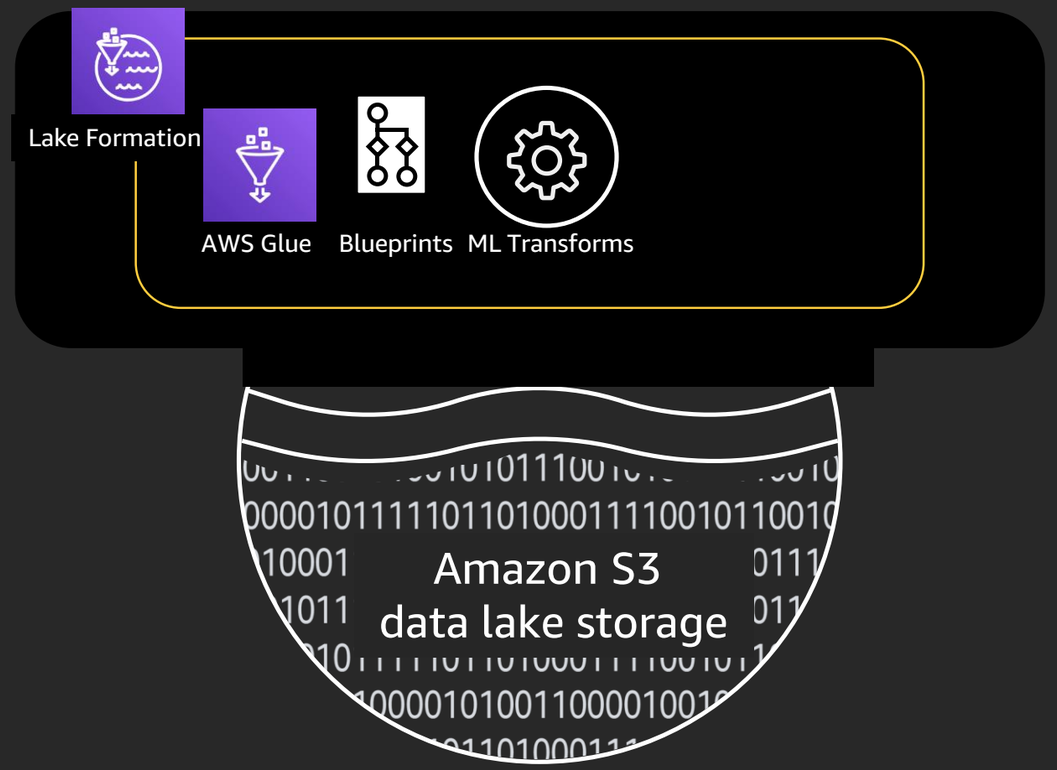
Built on Amazon S3 a robust data lake infrastructure



Amazon S3
data lake storage

Cost-effective, durable storage with global replication capabilities

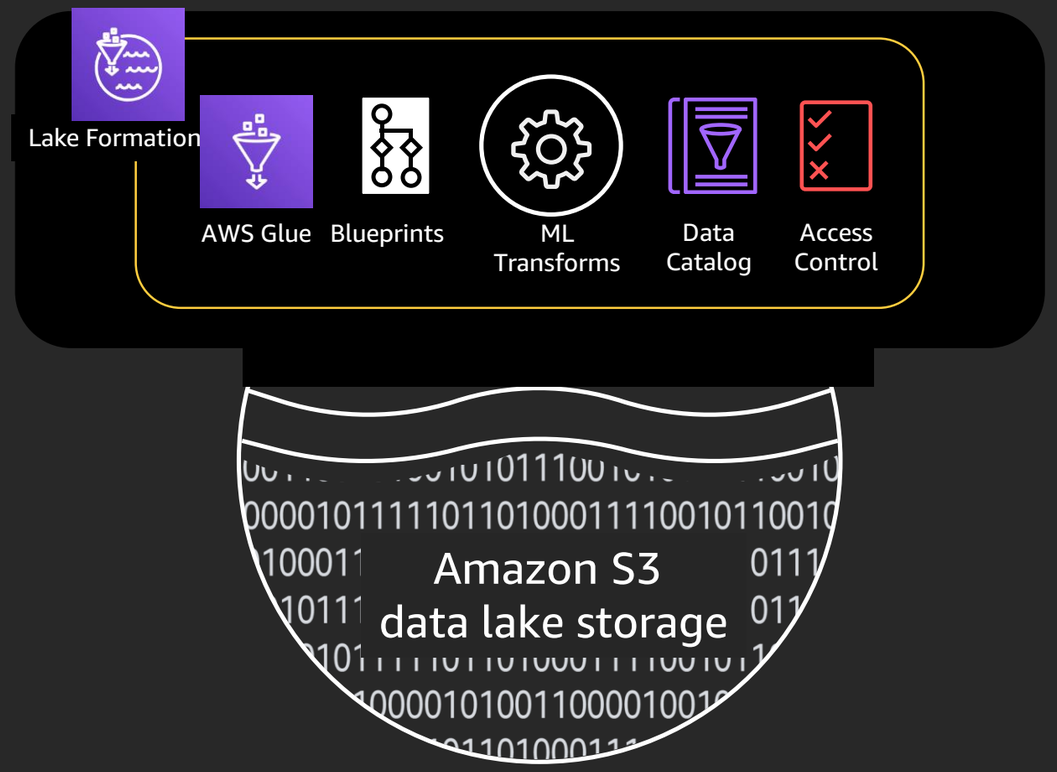
Automates manual, repetitive, low-value tasks



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

Cost-effective, durable storage with global replication capabilities

Provides a central locus of control

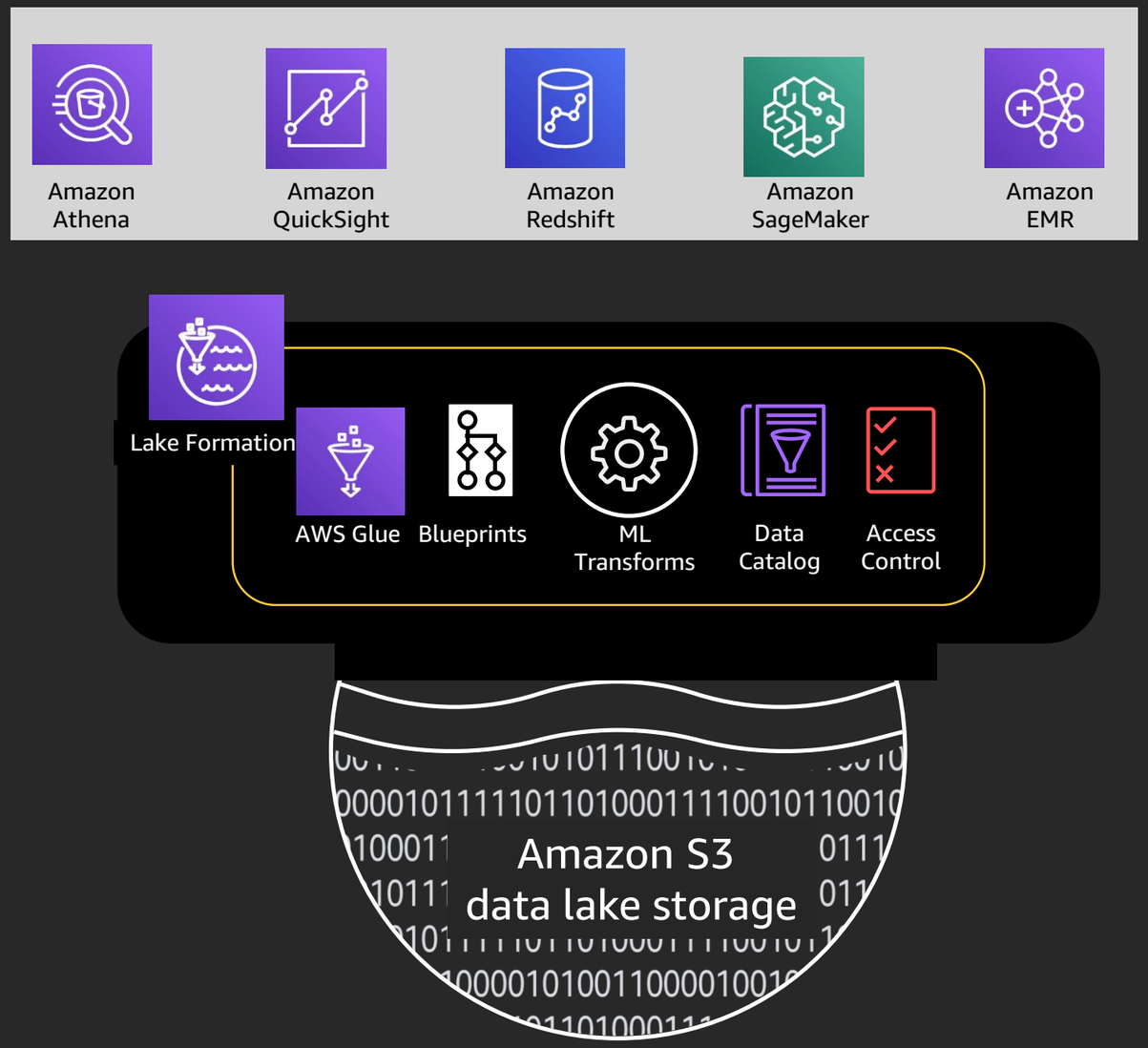


Centralized management of **fine-grained permissions** empower security officers

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

Cost-effective, durable storage with global replication capabilities

Enables all your data users



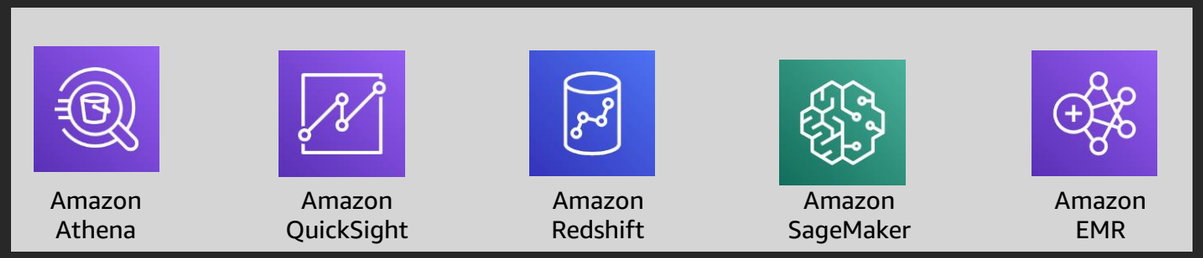
Comprehensive set of **integrated tools** enable every user equally

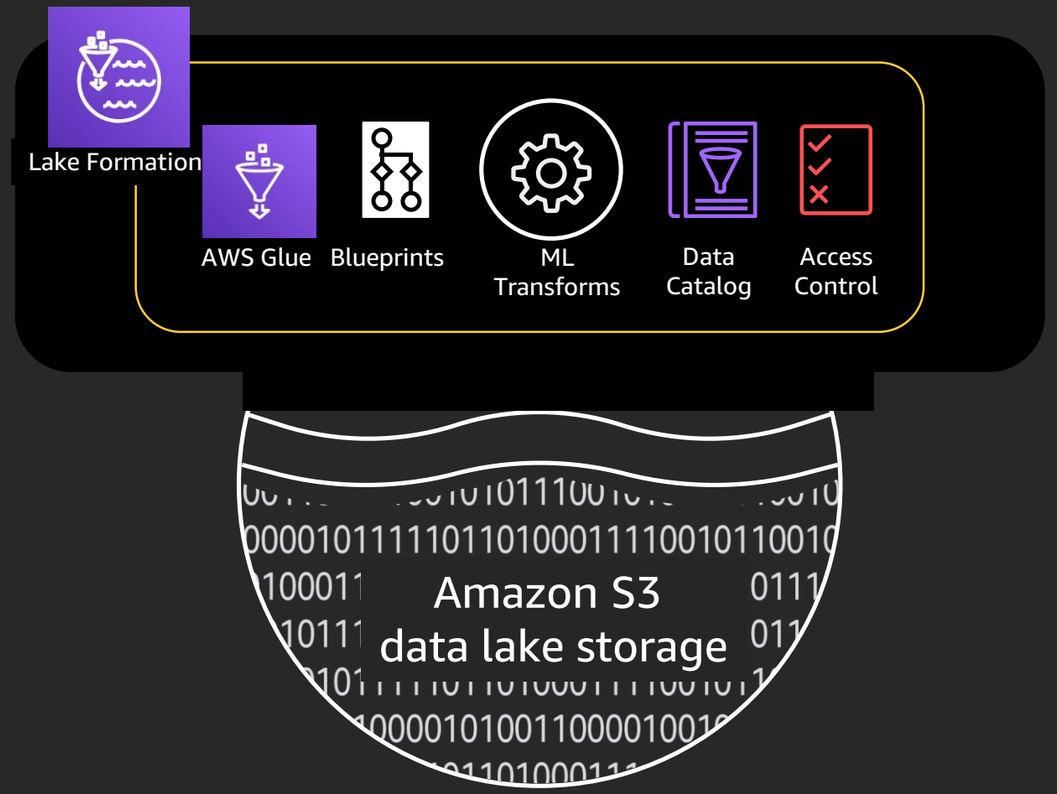
Centralized management of **fine-grained permissions** empower security officers

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

Cost effective, durable storage with global replication capabilities

Fastest way to build secure data lakes





Enables all your users to run any **analytics workload**, at any scale, in a secure and cost-effective manner

What is Lake Formation?

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Tools that enable data engineers, security officers and data analysts

Building data lakes with Lake Formation

Ingestion and cleaning



Data Engineer

AWS Glue

Serverless Spark

Blueprints

ML Transforms

Security



Data security officer

Data catalog

Centralized permissions

Real-time monitoring

Auditing

Analytics and ML

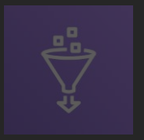


Data analyst

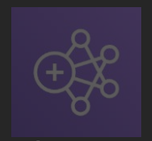
Comprehensive portfolio of integrated tools



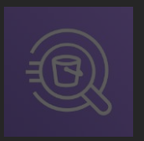
Amazon Redshift



AWS Glue

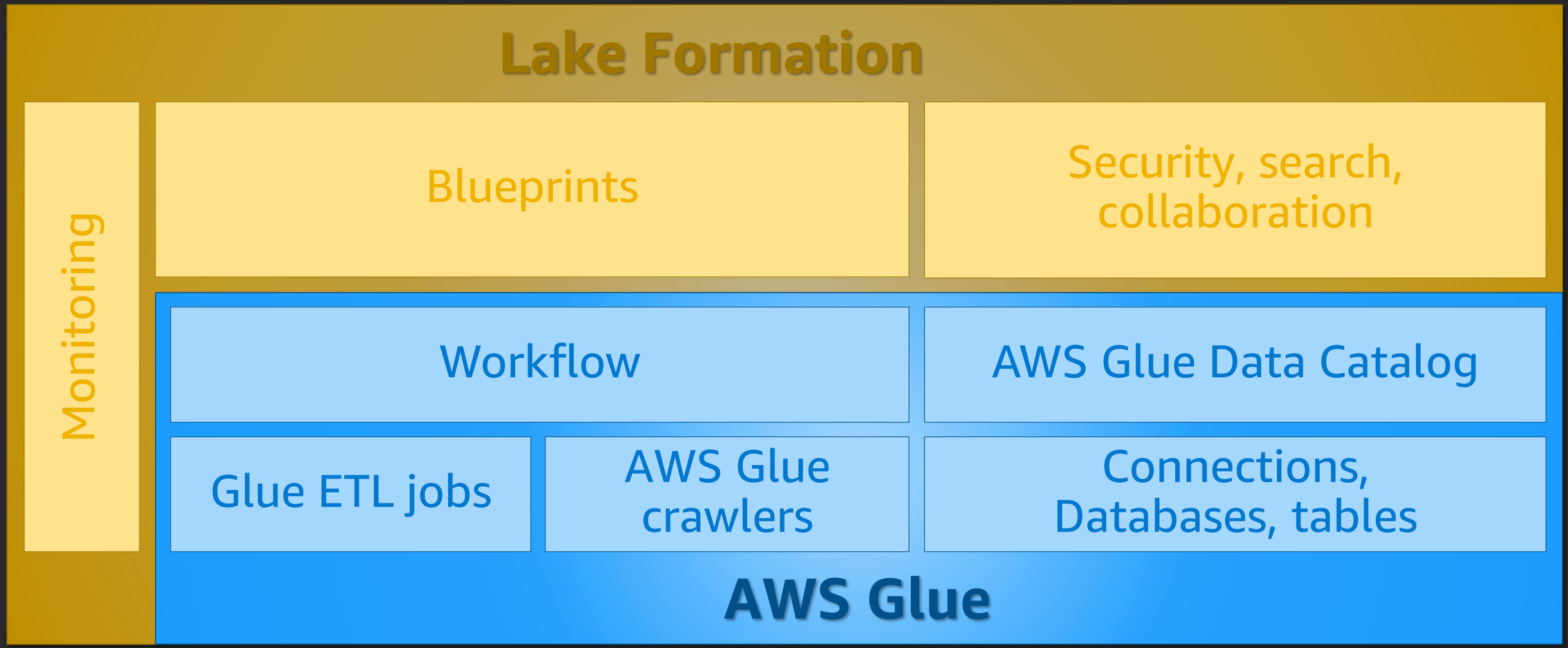


Amazon EMR



Athena

AWS Lake Formation is fully integrated w/ AWS Glue



AWS Glue components



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

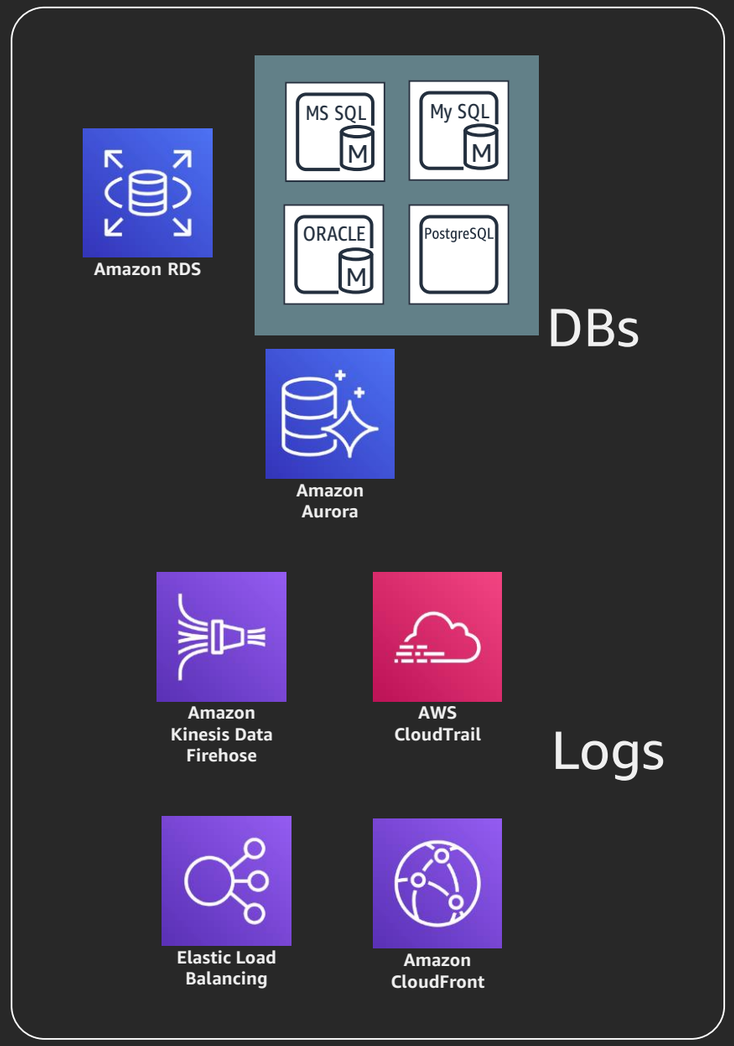
Deploy

Flexible workflows

Monitoring and alerting

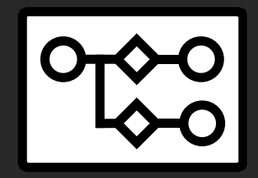
External integrations

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

Securing data lakes with Lake Formation

Ingestion and 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 and ML

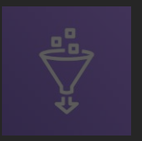


Data analyst

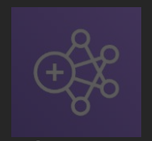
Comprehensive portfolio of integrated tools



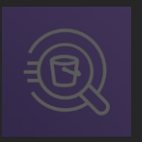
Amazon Redshift



AWS Glue



Amazon EMR



Athena

Data catalog and permissions

Permissions are set on data catalog objects

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



- Choice of using the **AWS Glue** or the **Lake Formation** permissions system
- For backward compatibility, the default settings enable the **AWS Glue** permissions system
- Existing **AWS Glue** crawlers, jobs, triggers and workflows will not change
- Existing access to **AWS Glue** resources will still be governed by **IAM and Amazon S3 policies**





Upgrading to the Lake Formation permissions model

Not using the AWS Glue Data Catalog?

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

Using the Data Catalog?

Explicitly upgrade each data location, database, and table when ready

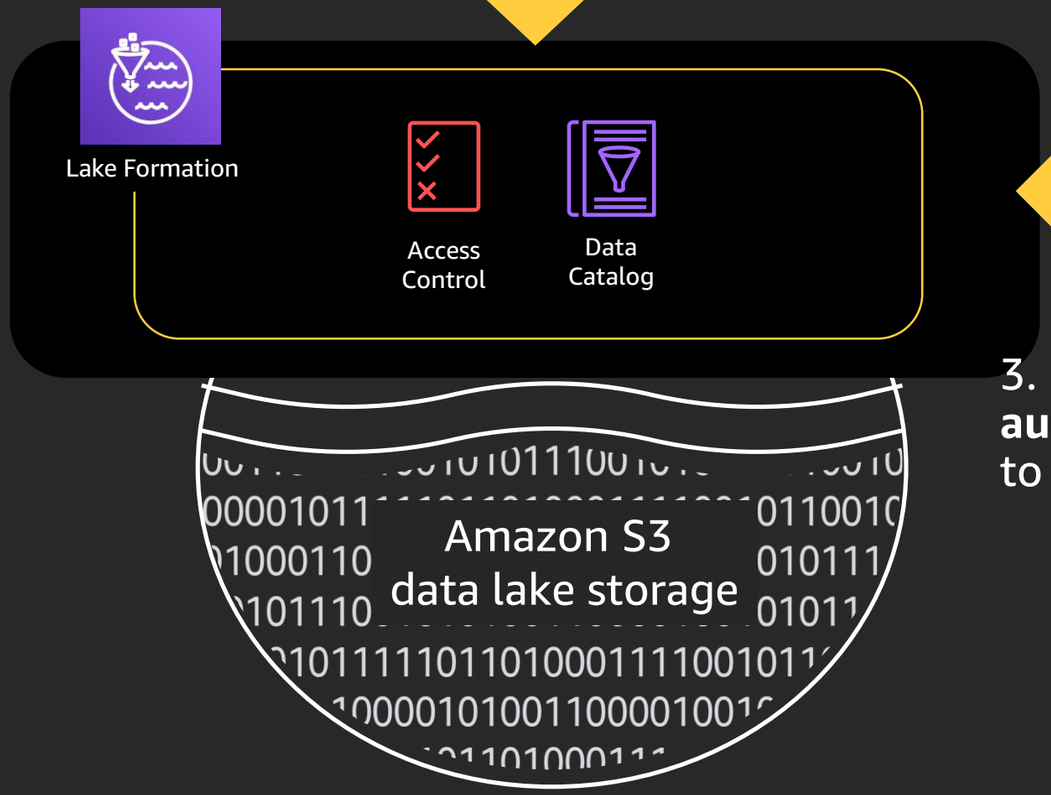
1. Understand existing policies/access/usage
2. Configure corresponding Lake Formation policies
3. Remove the Glue permissions system by changing the default settings
4. Turn on the Lake Formation permissions system by registering the location

Centralized permissions

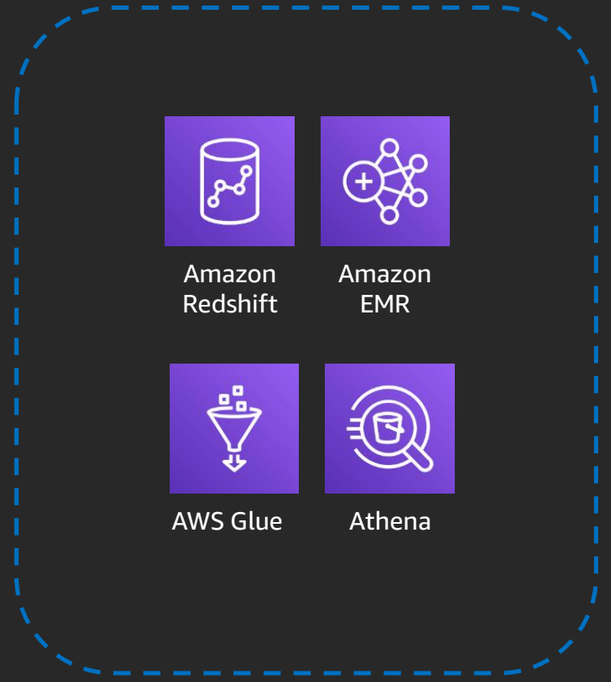


Data security officer

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



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





Data Analyst

2. User access data via integrated services

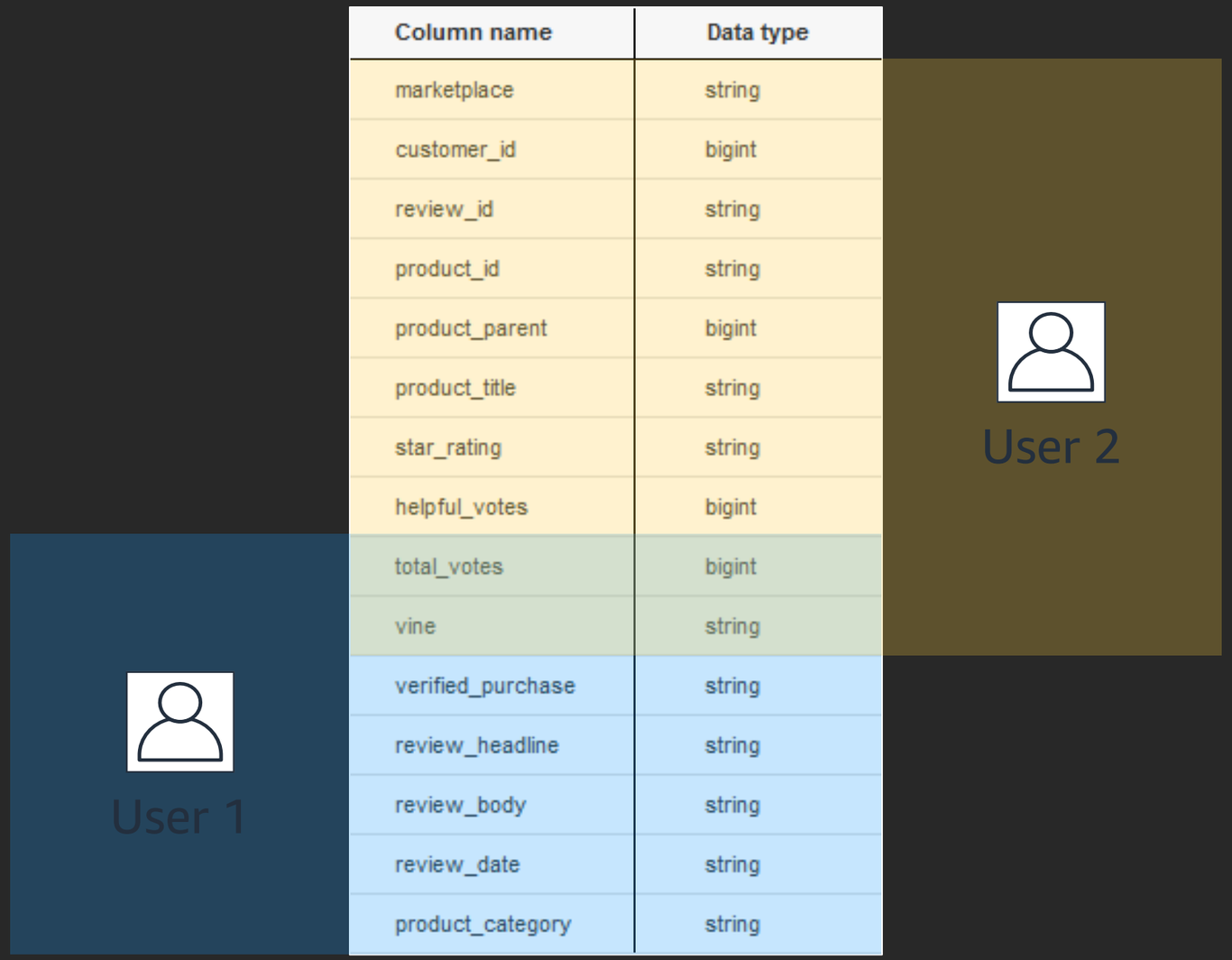
Security permissions in Lake Formation

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

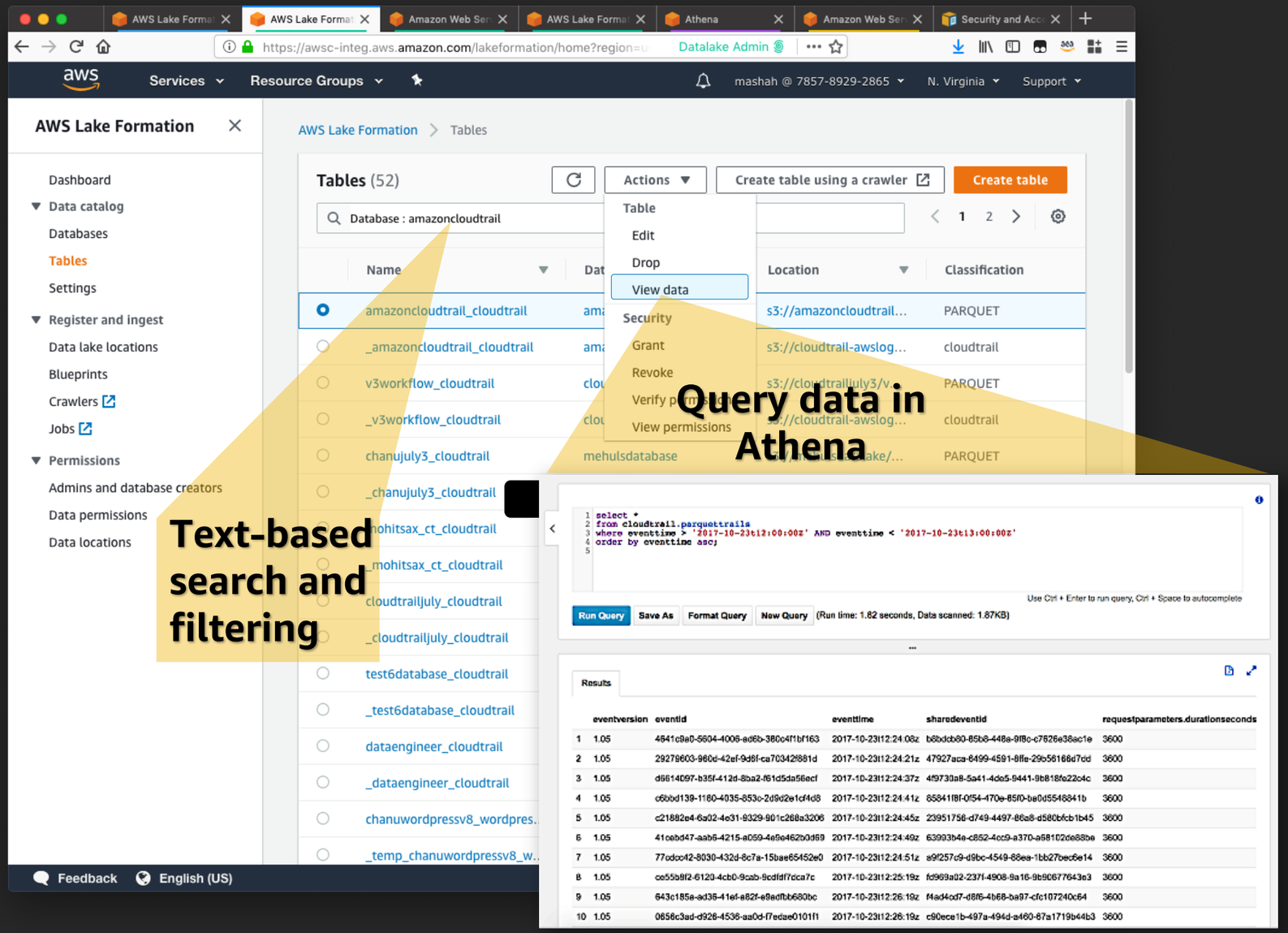


Data catalog and metadata management

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**

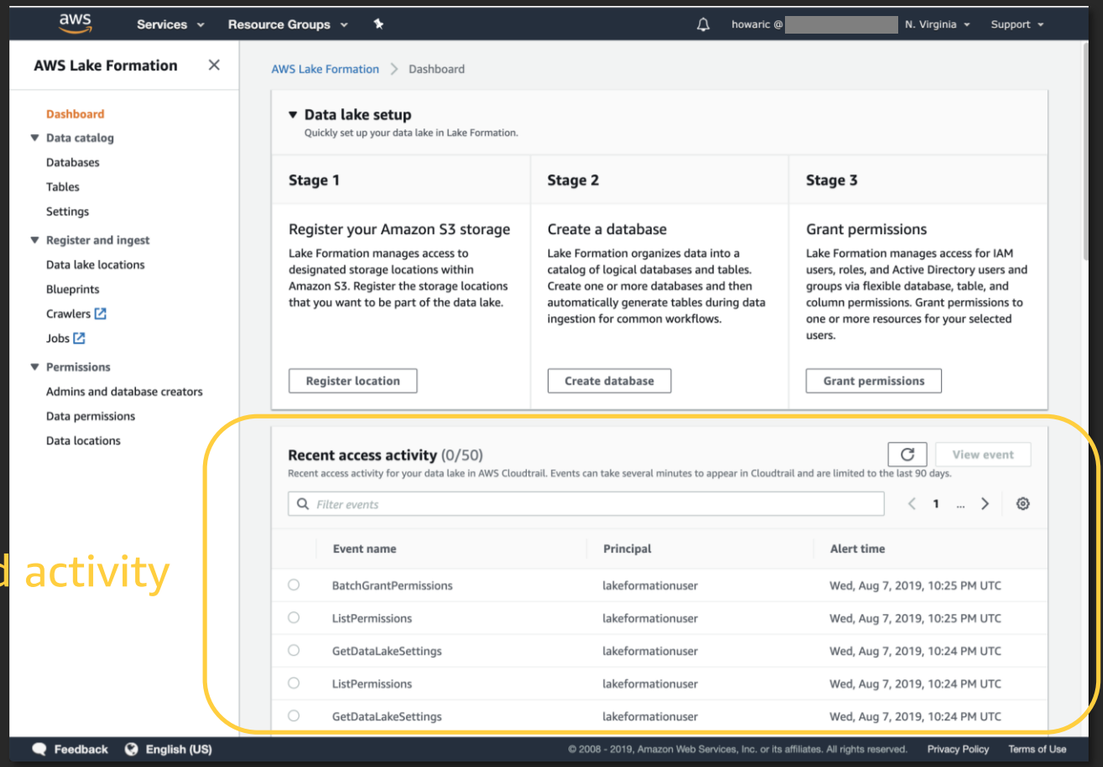


Audit and monitor in real time

See detailed activity in the console

Analyze audit logs in CloudTrail using Athena

Data ingest and catalog notifications also published to Amazon CloudWatch Events



Detailed activity

Accessing data lakes with Lake Formation

Ingestion and cleaning



Data engineer

Serverless Spark

AWS Glue

Glue ML transformations

Blueprints

Security



Data security officer

Data catalog

Centralized permissions

Real-time monitoring

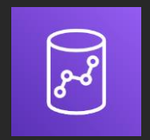
Auditing

Analytics and ML

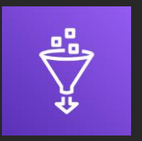


Data analyst

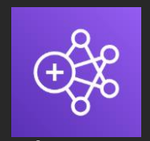
Comprehensive portfolio of integrated tools



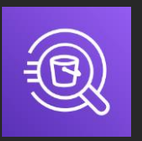
Amazon Redshift



AWS Glue



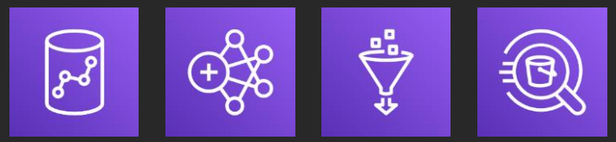
Amazon EMR



Athena

Comprehensive portfolio of integrated tools

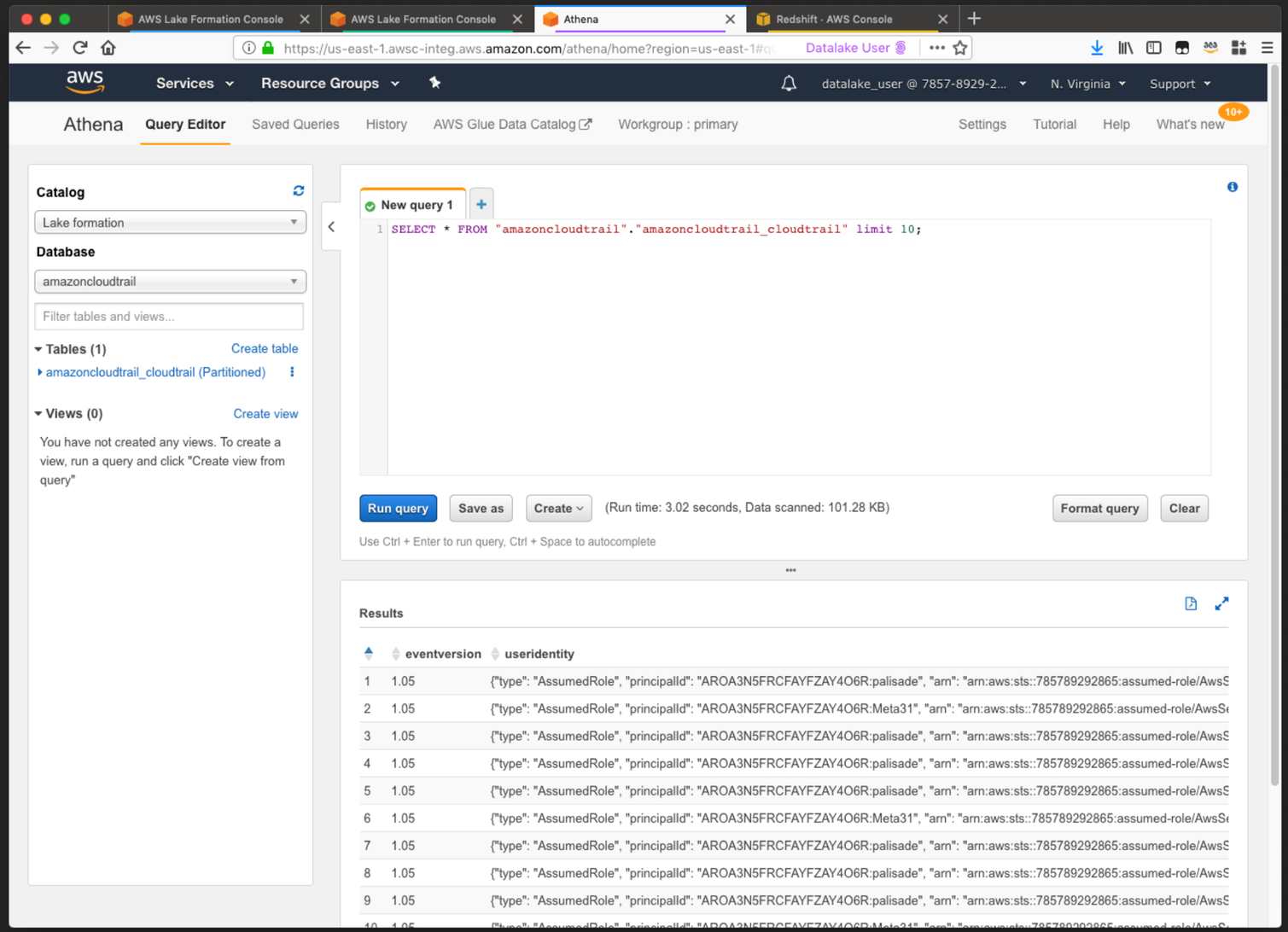
Compliant services honor Lake Formation permissions



Amazon Redshift | Amazon EMR | AWS Glue | Athena

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

All access is logged and auditable



How it works

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Demo

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Shashi Raina

shraina@amazon.com

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey in the mobile app.

AWS re:Invent

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws