Big Data on AWS

Gustavo Veloso
Senior Solutions Architect, AWS
velosog@amazon.com

December, 2019

Data is a strategic asset for every organization

> The world's most valuable resource is **no longer oil, but data**.*

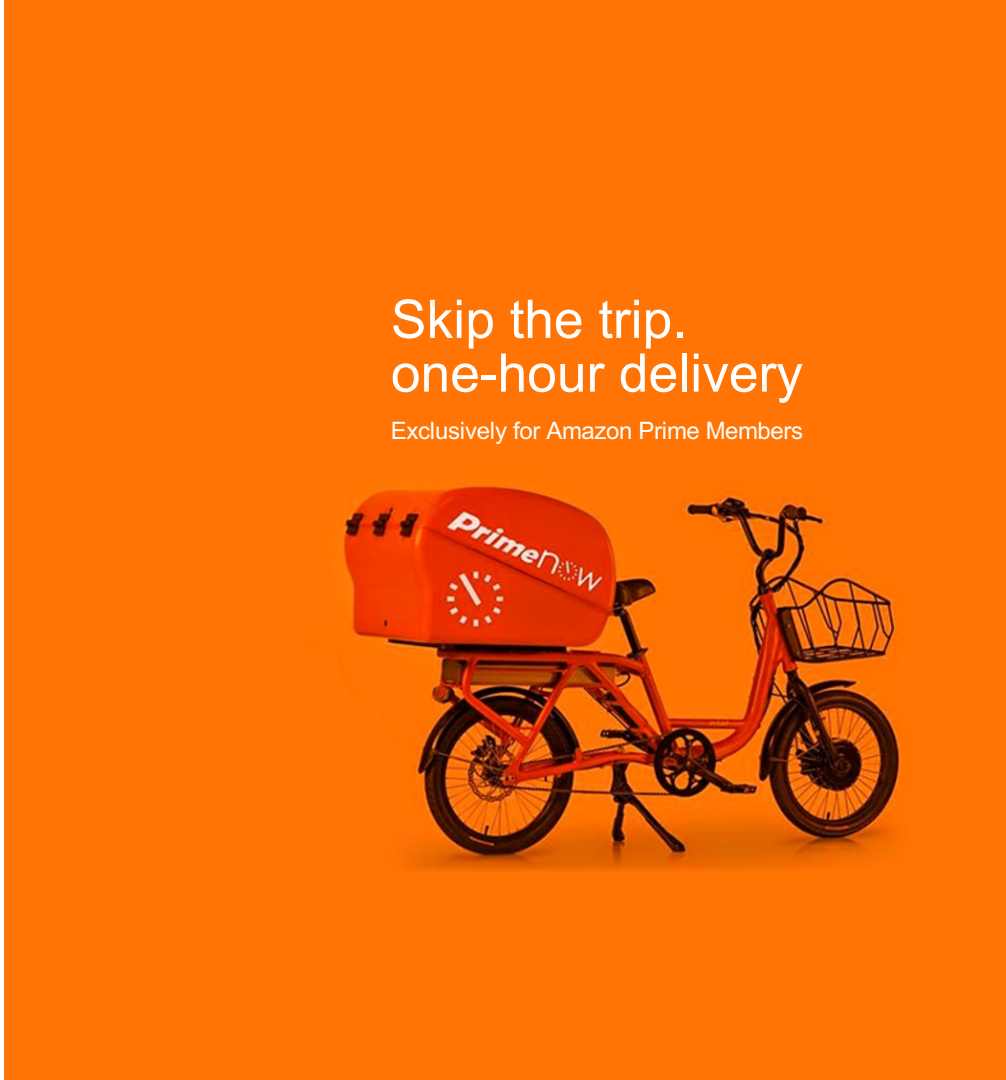
*Copyright: The Economist, 2017, David Parkins

We need to rethink what we mean by data and analytics

This is data

This is data

This is data



Data no longer fits

There is **more data** than people think

Data is **more diverse**

Data	Data platforms need to	
grows **>10x** every 5 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 Apr 2017

aws

Broader workloads

Data Scientists

Business Users

Analysts

Applications

machine learning

SQL analytics

scientific

real-time, streaming

There are **more people** accessing data

That want to **analyze it in different ways**

And there are **more rules** around data use

aws

Decision making used to…

…revolve around the

Enterprise Data Warehouse

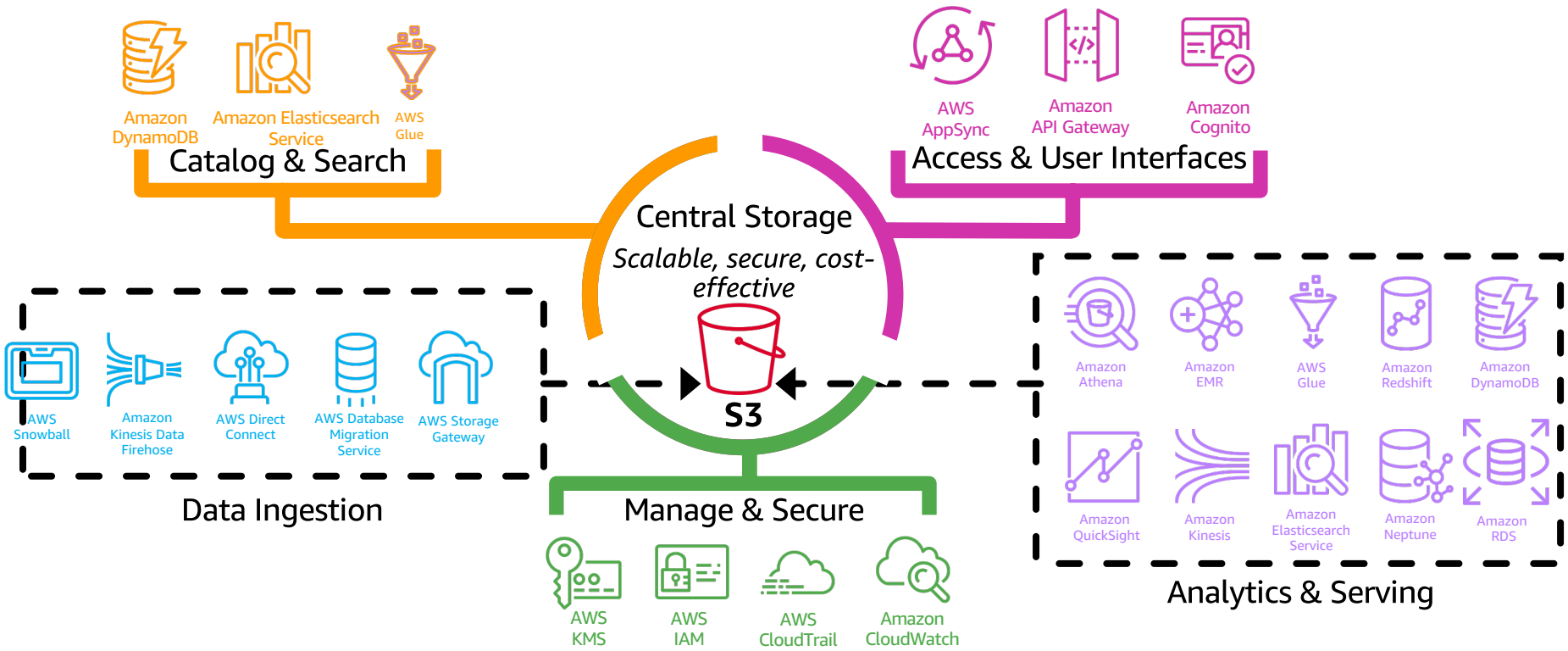
aws

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

aws

Data lake on AWS



Tier 1 Data Lake: Ingestion

Amazon S3

Single source of truth for raw data

Use least transformations

Use lifecycle policies to Amazon Simple Storage Service (Amazon S3) IA or Amazon Glacier

aws

Tier 2 Data Lake: Analytics

Amazon S3

Use columnar formats – Parquet/ORC

Organized into partitions

Coalescing to larger partitions over time

Optimized for analytics

aws

Tier 3 Data Lake: Analytics

Amazon S3

Domain level DataMart

Organized by use cases

Optimized for specialized analysis

aws

Amazon Redshift

Data Warehouse:

Fast speeds over structured schemas

Serves dashboards and reports

Fine-grained access controls

Supports joining native and external tables

Lifecycle back to S3 Data Lake

aws

EQUINOX

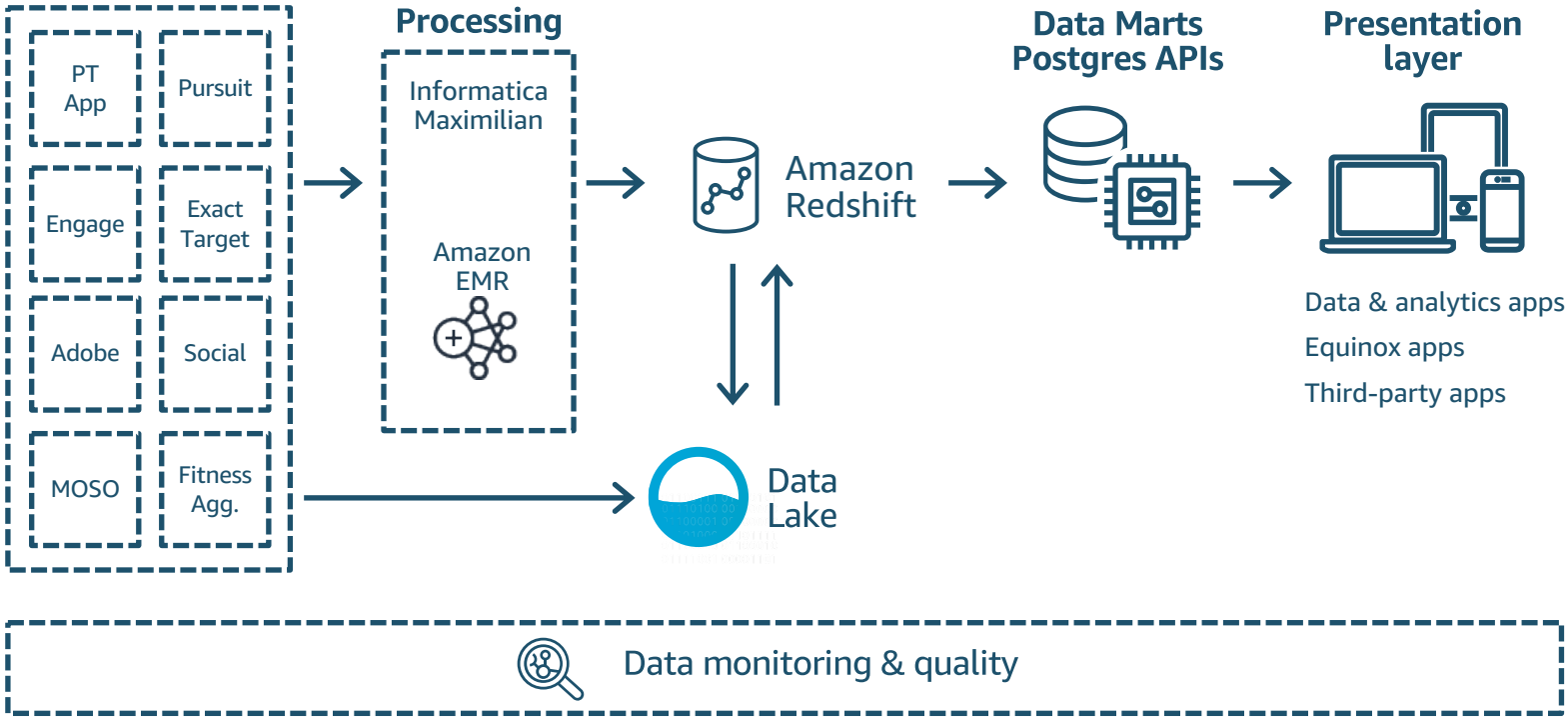
Equinox Fitness Cubs is a company with integrated luxury and lifestyle offerings centered on movement, nutrition and regeneration. Equinox built connected experiences using applications that connect to Apple Health and built data collection in their exercise equipment.

blink FITNESS | PURE YOGA | SOULCYCLE | FURTHERMORE FROM EQUINOX | INTRODUCING EQUINOX HOTELS

More than 200 locations within every major city across the U.S., London, and Canada

aws

Data lake architecture



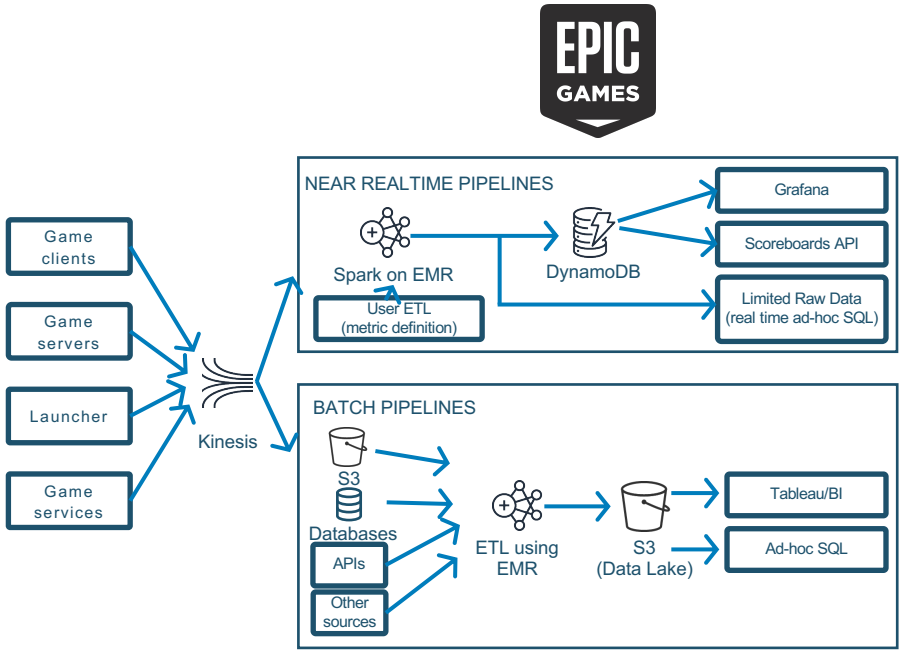
Fortnite | 125+ million players

CHALLENGE

Need to create constant feedback loop for designers

Gain up-to-the-minute understanding of gamer satisfaction to guarantee gamers are engaged, thus resulting in the most popular game played in the world

Epic Games uses Data Lakes and analytics

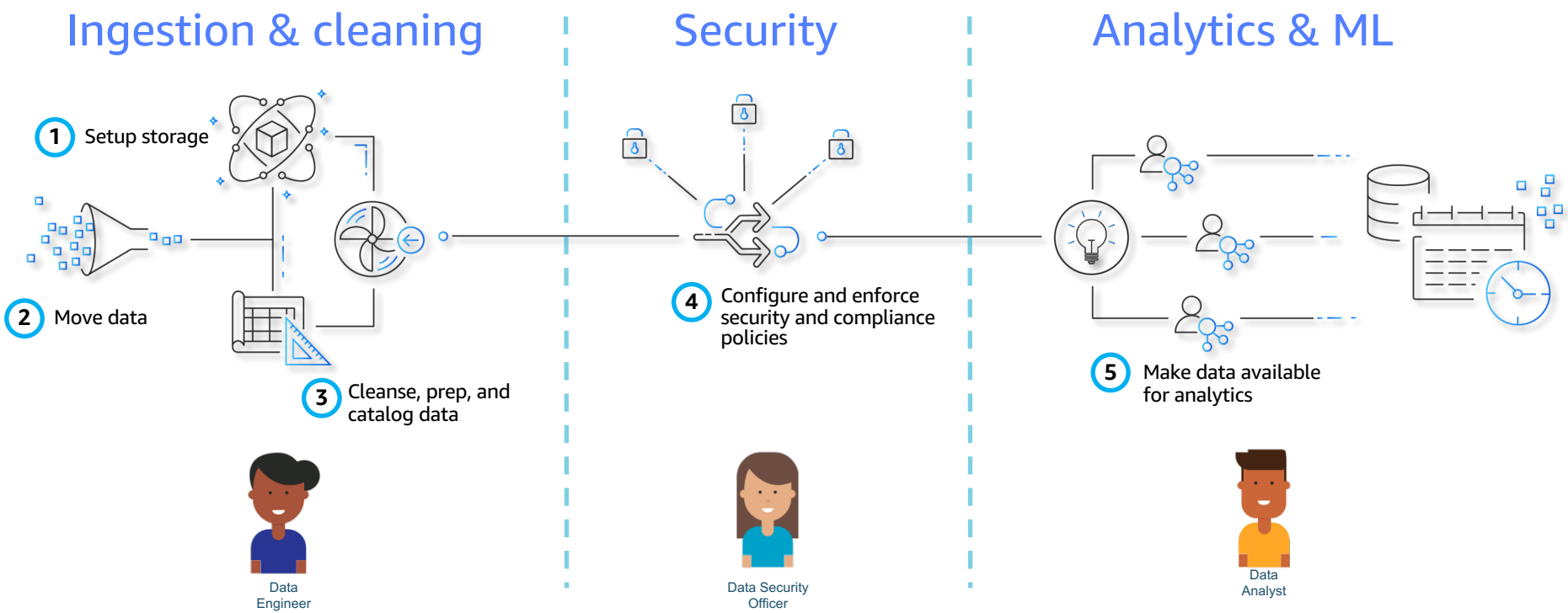


- Entire analytics platform running on AWS
- S3 leveraged as a Data Lake
- All telemetry data is collected with Kinesis
- Real-time analytics done through Spark on EMR, DynamoDB to create scoreboards and real-time queries
- Use Amazon EMR for large batch data processing
- Game designers use data to inform their decisions

aws

Manually building secure data lakes is **hard**

Typical steps of building a data lake



aws

Sample of steps requ

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

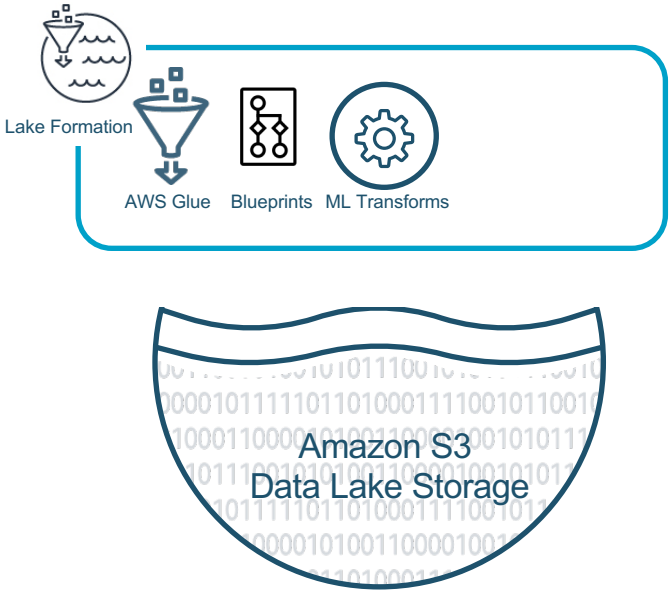
Lake Formation lets you
build secure data lakes in **days**

Built on Amazon S3 a robust data lake infrastructure

Cost effective, durable storage with global replication capabilities

aws

Automates manual, repetitive, low value tasks

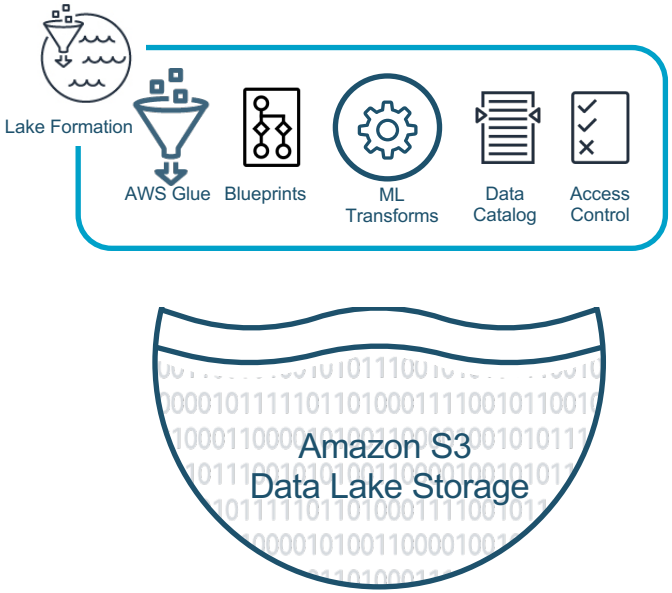


Simplified **ingest & cleaning** enables data engineers to build faster

Cost effective, durable storage with global replication capabilities

aws

Provides a central locus of control



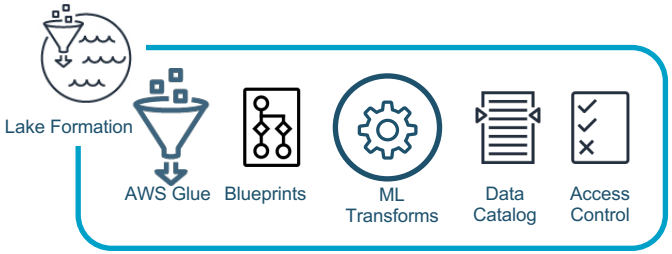
Centralized management of **fine grained permissions** empower security officers

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

Cost effective, durable storage with global replication capabilities

aws

Enables all your data users



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

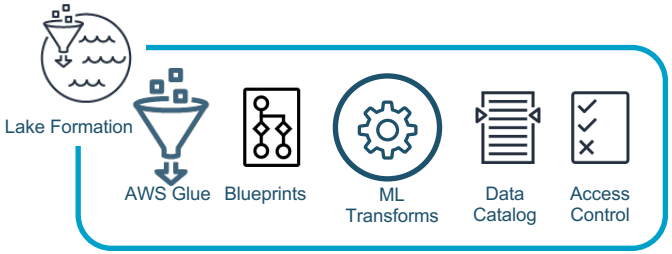
Centralized management of **fine grained permissions** empower security officers

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

Cost effective, durable storage with global replication capabilities

aws

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

aws

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

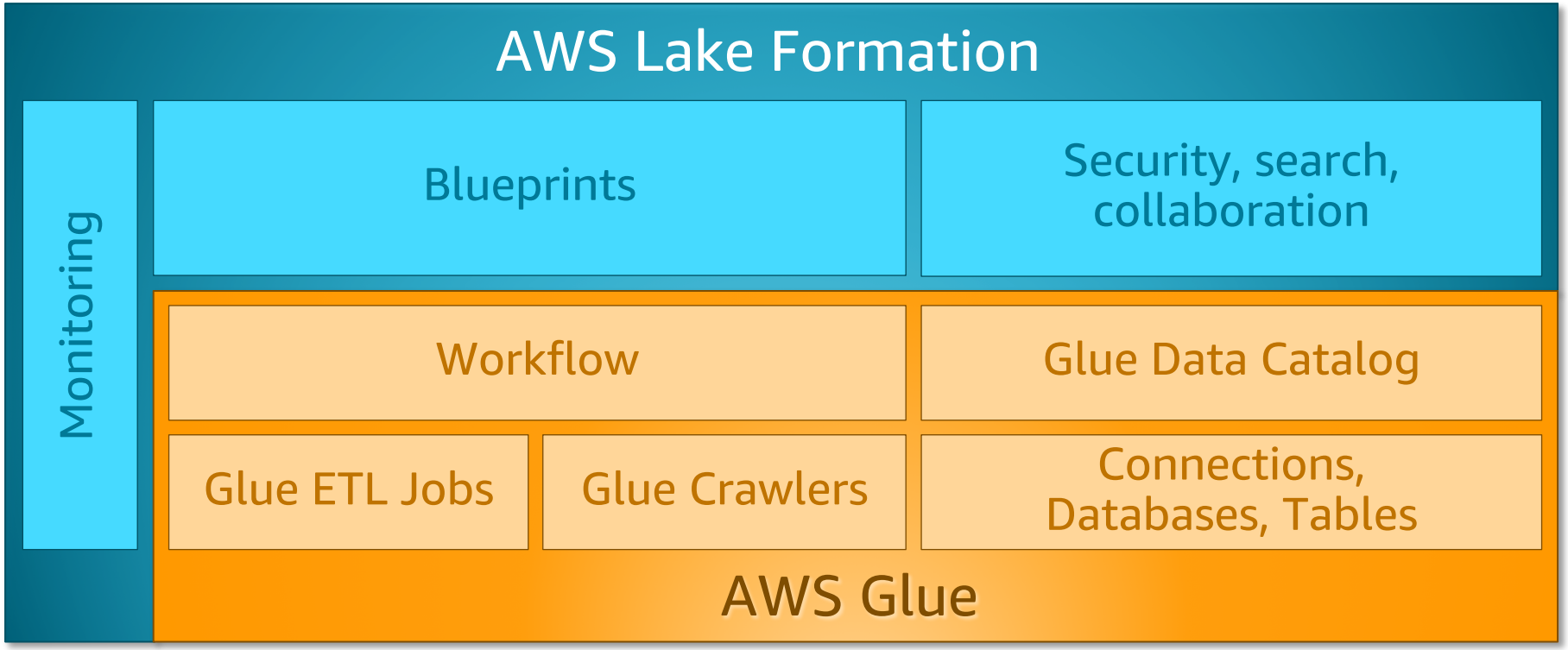
Glue

EMR

Athena

aws

AWS Lake Formation is fully integrated w/ AWS Glue



aws

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

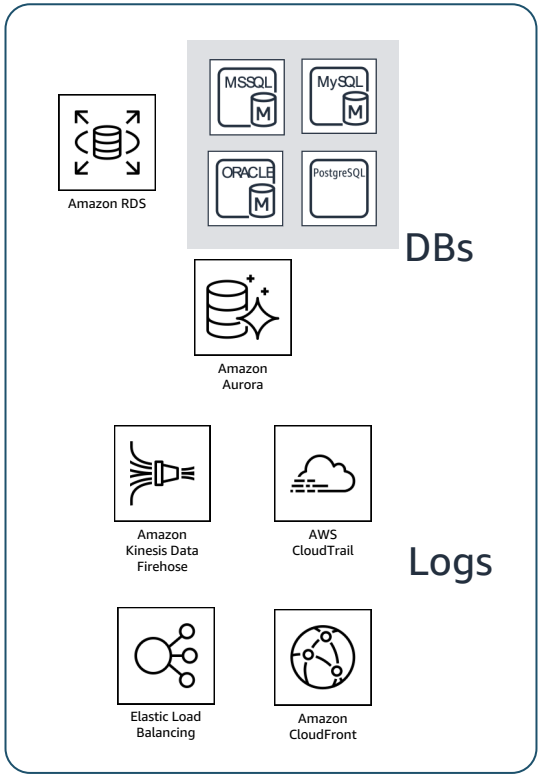
Flexible workflows

Monitoring and alerting

External integrations

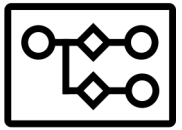
aws

Easily load data into your data lake w/ blueprints



Prebuilt templates to serve common ingestion use cases

Automatically build AWS Glue workflows



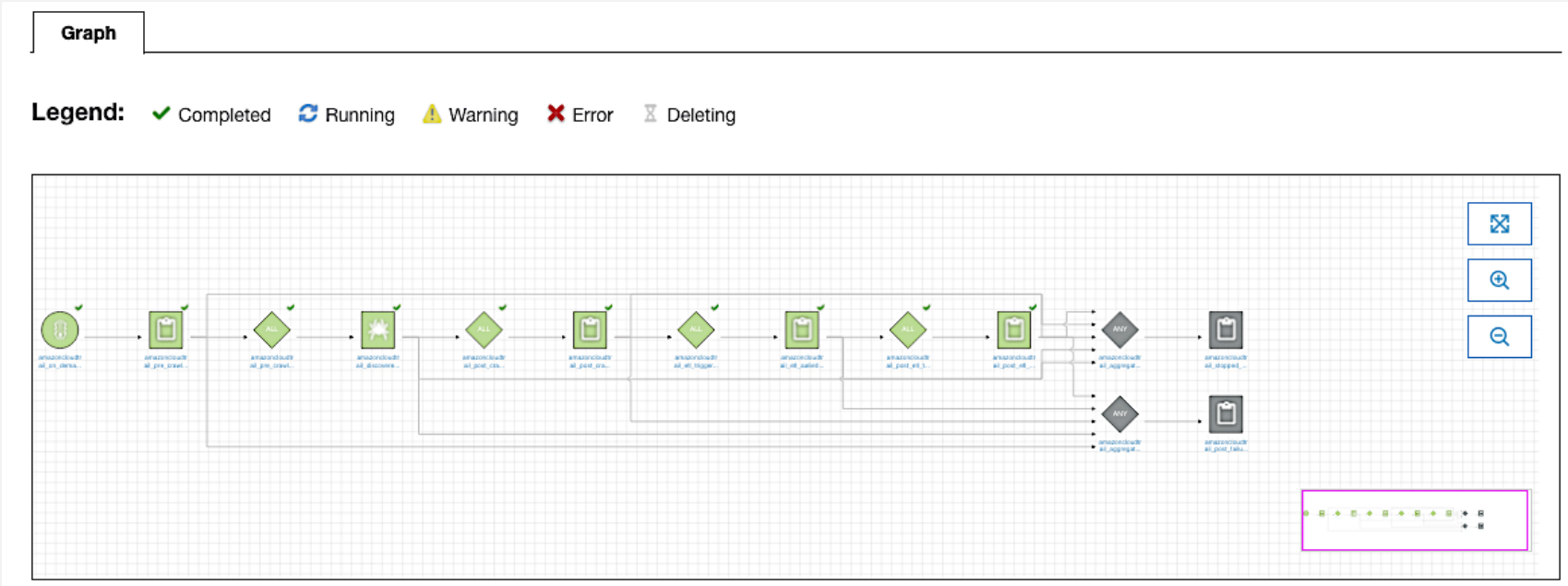
AWS Glue jobs and crawlers discover, transform and structure data

Automatically populate the Data Catalog

Load data incrementally or in full

aws

Blueprints create AWS Glue workflows



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

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

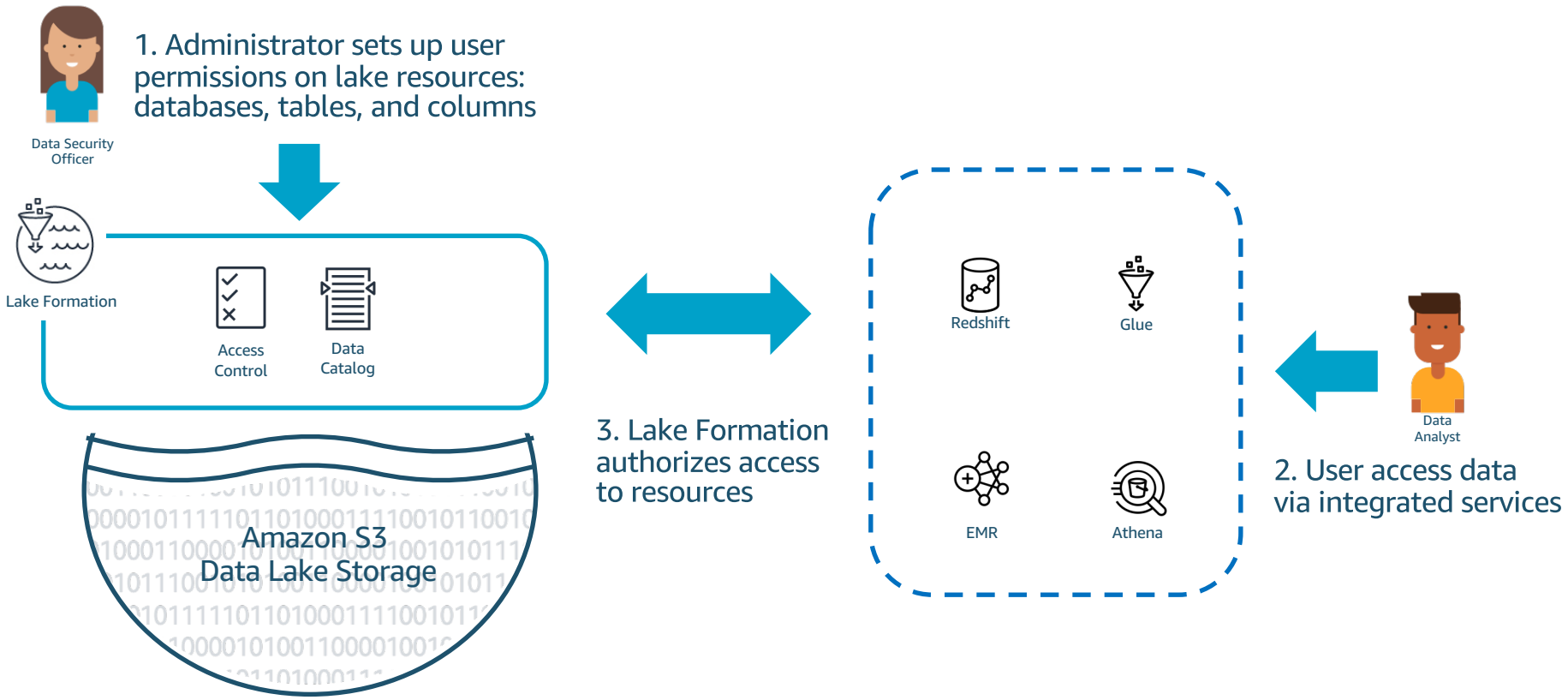
Glue

EMR

Athena

aws

Centralized permissions



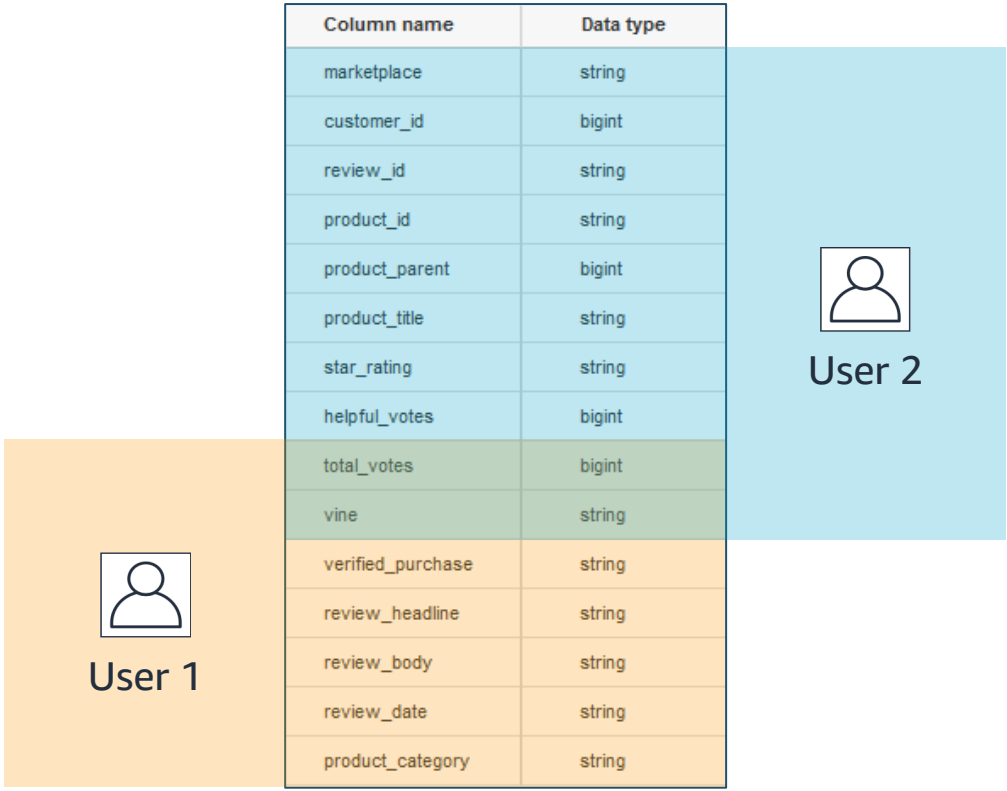
Security permissions in Lake Formation

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



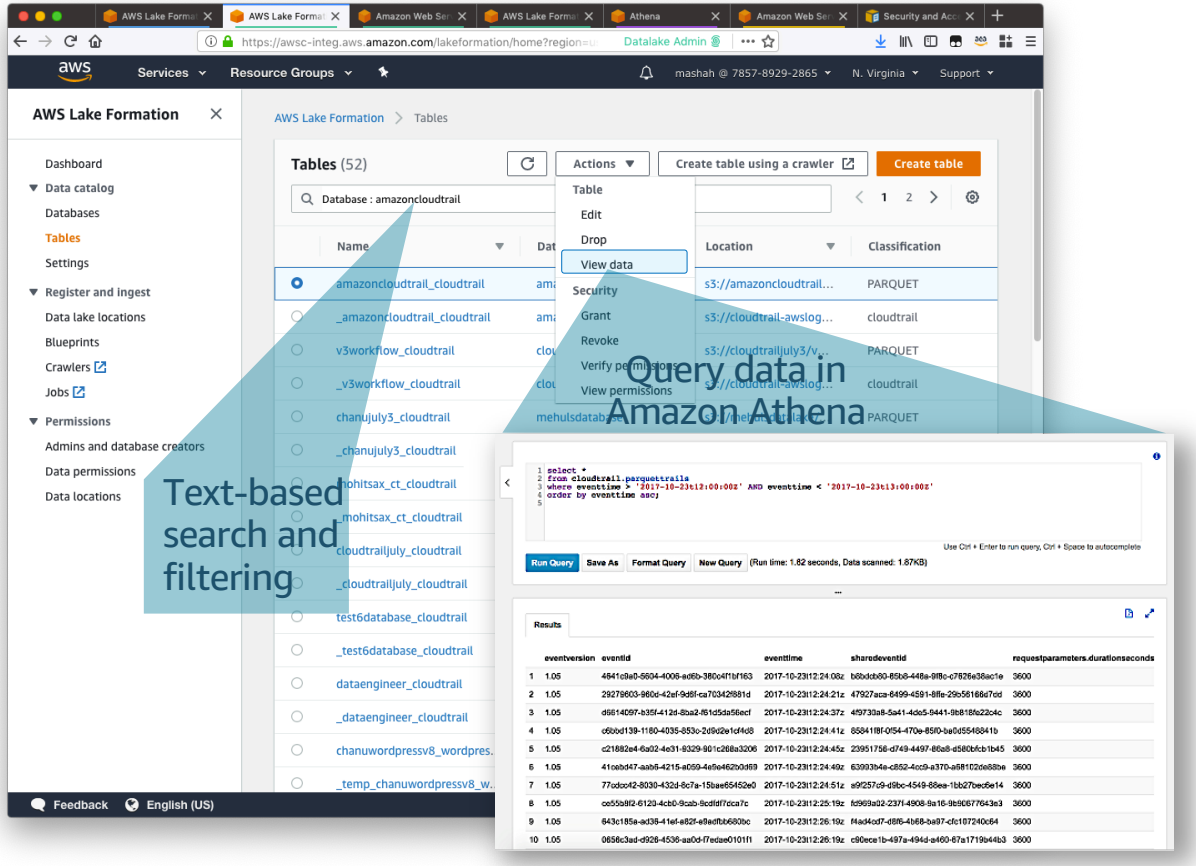
aws

Data catalog 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



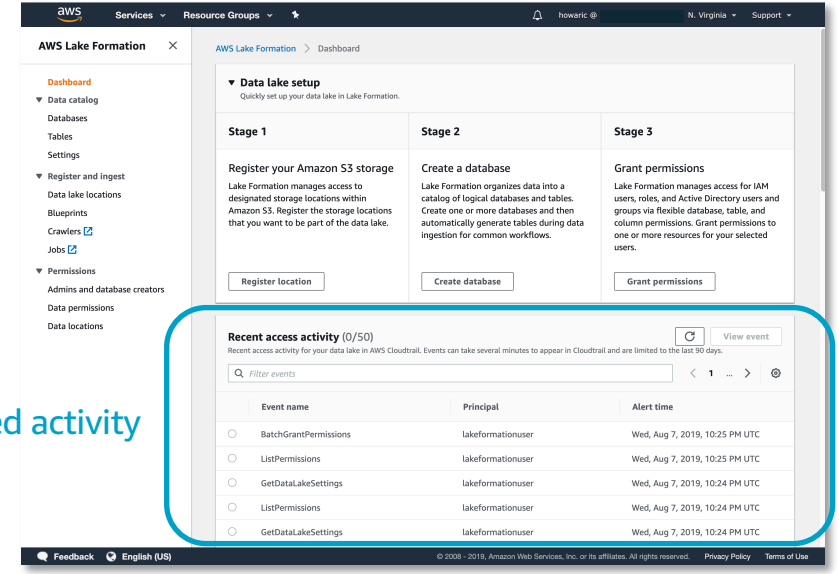
Audit and monitor in real time

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

Redshift

Glue

EMR

Athena

aws

Comprehensive portfolio of integrated tools

Compliant services honor Lake Formation permissions

Amazon Redshift

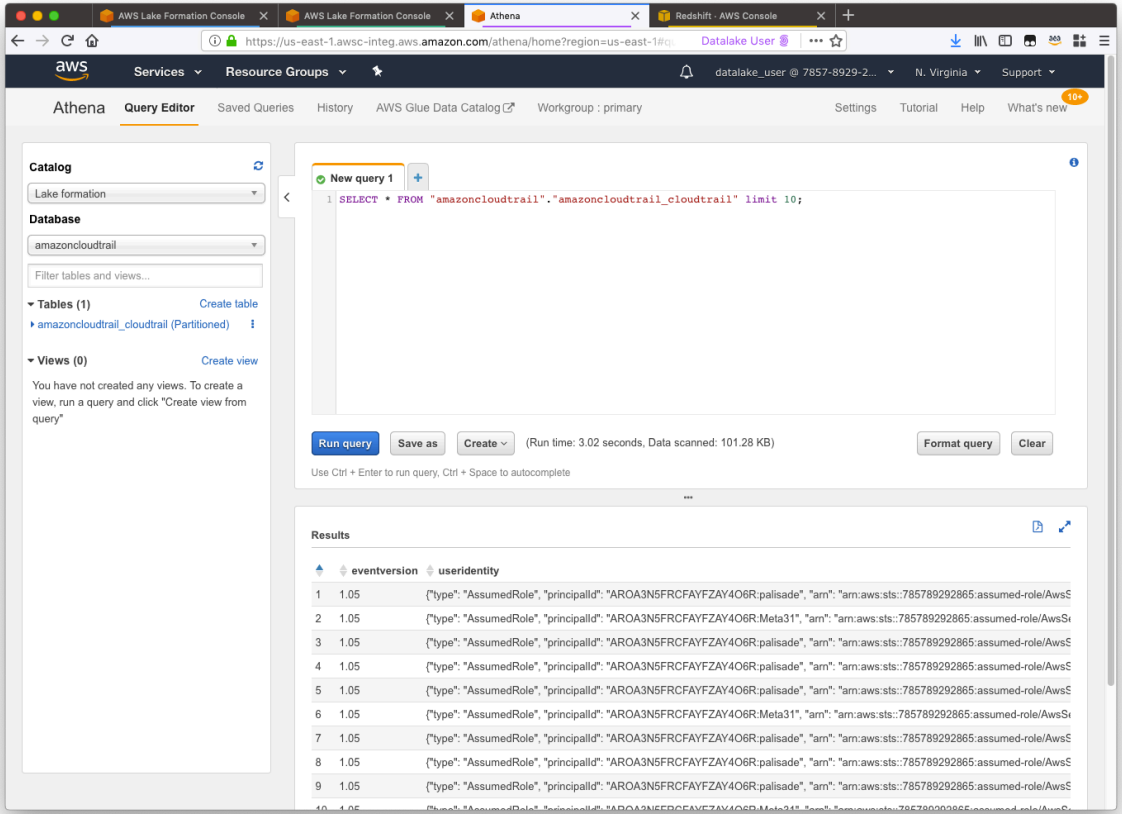
Amazon EMR

AWS Glue

Amazon Athena

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

All access is logged and auditable



Thanks

velosog@amazon.com