PUBLIC SECTOR SUMMIT ONLINE

Turn data to insights with data lakes, analytics and machine learning services from AWS

Mona Mona

AI/ML Specialist Solution Architect, AWS

aws PUBLIC SECTOR SUMMIT ONLINE

Agenda

Trends driving the revolution

What is a data lake?

Why AWS for data lake?

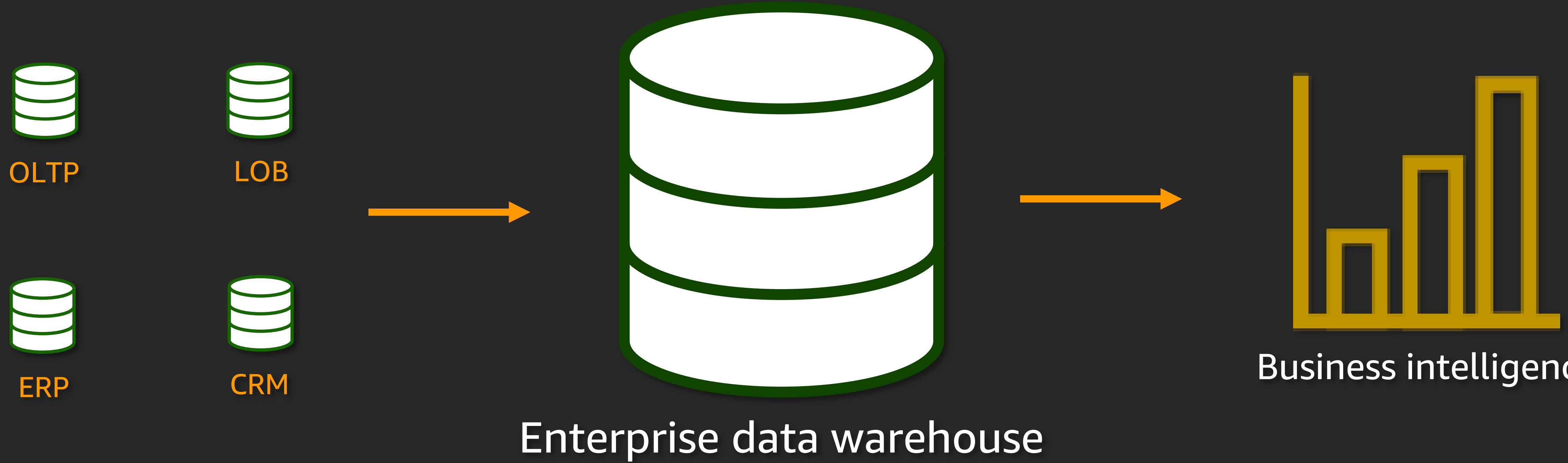
What is hard about building data lakes?

AWS Lake Formation and its blueprints makes data lakes easy.

Demo!

In the past, decision-making …

…revolved around the **enterprise data warehouse**



Data no longer fits

There is **more data** than people think

Data is **more diverse**

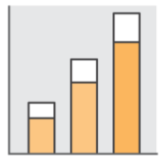
Data	**Data platforms need to**	
grows **>10x** every 5 years	live for **15** years	scale **1,000x**

* IDC, Data Age 2025: The Evolution of Data to Life-Critical: Don't Focus on Big Data, Focus on the Data That's Big, April 2017.

Broader workloads

Data Scientists

Business Users



Analysts

Applications

Machine learning

SQL analytics

Scientific

Real-time, streaming

There are **more people** accessing data

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

Agenda

Trends driving the revolution

What is a data lake?

Why AWS for data lake?

What is hard about building data lakes?

AWS Lake Formation and its blueprints makes data lakes easy.

Demo!

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

Agenda

Trends driving the revolution

What is a data lake?

Why AWS for data lake?

What is hard about building data lakes?

AWS Lake Formation and its blueprints makes data lakes easy.

Demo!

Why choose AWS for data lakes and analytics?

1. Easiest to build data lakes and analytics
2. Most secure infrastructure for analytics
3. Most scalable and cost-effective
4. Most comprehensive and open

The AWS analytics portfolio

Data, visualization, engagement & machine learning

NEW AWS Data Exchange | Amazon QuickSight | Amazon Pinpoint | Amazon SageMaker | Amazon Comprehend | Amazon Lex | Amazon Polly | Amazon Rekognition | Amazon Translate

+ Many more

Analytics

Amazon Redshift | Amazon EMR (Spark & Hadoop) | AWS Glue (Spark & Python) | Amazon Athena | Amazon Elasticsearch Service | Amazon Kinesis Data Analytics

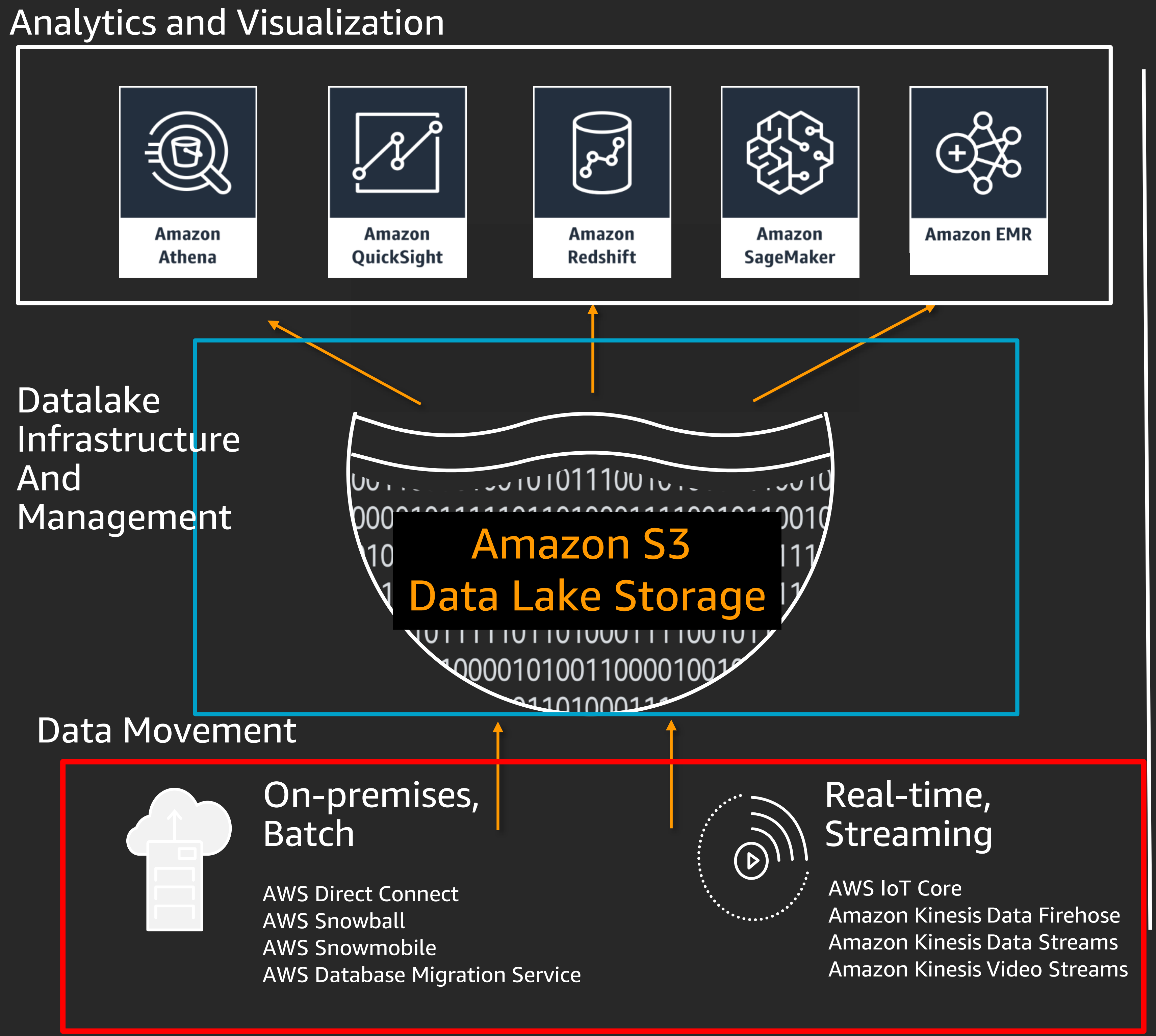
Data lake infrastructure & management

Amazon S3/ Amazon S3 Glacier | AWS Lake Formation | AWS Glue

Data movement

AWS Database Migration Service | AWS Snowball | AWS Snowmobile | Amazon Kinesis Data Firehose | Amazon Kinesis Data Streams | Amazon Managed Streaming for Apache Kafka

Game changer : data lakes on the cloud



Many scalable analytics engines available on-demand, pay-as-you-go

Amazon S3: ubiquitous storage allows you to centralize data sets

Want a **single point of control**

सत्यमेव जयते

Ministry of Housing and Urban Affairs

Government of India

The Smart Cities Mission wanted to set up a state-of-the-art India Urban Observatory to plug into the various sources of data from cities, both from real-time and archival sources. The time taken from idea to execution was 5 weeks. Platform has scaled from 500 data sets to 500,000 data sets.

CSIRO

Commonwealth Scientific and Industrial Research Organisation, Australia Making sense of the human genome by using serverless AWS architecture.

“The system avoids data silos while preserving data ownership and patient privacy,”- CSIRO

ventures

Future Health designed+delivered

Ventures Health, New Zealand achieves 90% cost reduction using AWS data lake with just 2 data engineer in few weeks. The cost savings are being reinvested in its core mission. They serve around half a million patients on the North Island of New Zealand and specializes in electronic health, data services, digital communications, online and market places.

More data lakes & analytics on AWS than anywhere else

FINRA · NASDAQ · pitney bowes · yelp · EA ELECTRONIC ARTS · hudl · airbnb · monster worldwide · CROWDSTRIKE

Expedia · NTT docomo · theguardian · Nextdoor · AMGEN · A&E Real Life. Drama. · JustGiving · SK planet · News Corp

Zillow · edmunds.com · FOURSQUARE · lyft · Schumachergroup · gogo IN AIR. ONLINE. · Pfizer · ZUMBA FITNESS · dataxu

Capital One · ACQUIA · COMCAST · FT FINANCIAL TIMES · Instacart · mobilewalla · reddit · Bristol-Myers Squibb · scopely

SCHOLASTIC · DOMO · CIVITAS LEARNING · REDFIN · sling TELEVISION · Agero Safe. Smart. Driven. · 21ST CENTURY FOX · change.org · amazon

Sysco · euclid · ROVIO · CIVIS ANALYTICS · Pinterest · ACTIVISION BLIZZARD · illumina · ticketmaster · Viber Connect. Freely.

broadsoft · PHILIPS · Discovery COMMUNICATIONS · Instapage · FUNNY OR DIE · FAST RETAILING · imshealth · realnetworks · ost

StateFarm · HEARST · mlbam · Vanguard · SONOS · hello · Adobe · enel · OLX Group

ATLASSIAN · HBO · slack · Gunosy · TUNE · TAXI · mparticle · coffee meets bagel QUALITY DATING MADE EASY · FULL 360

Agenda

Trends driving the revolution

What is a data lake?

Why AWS for data lake?

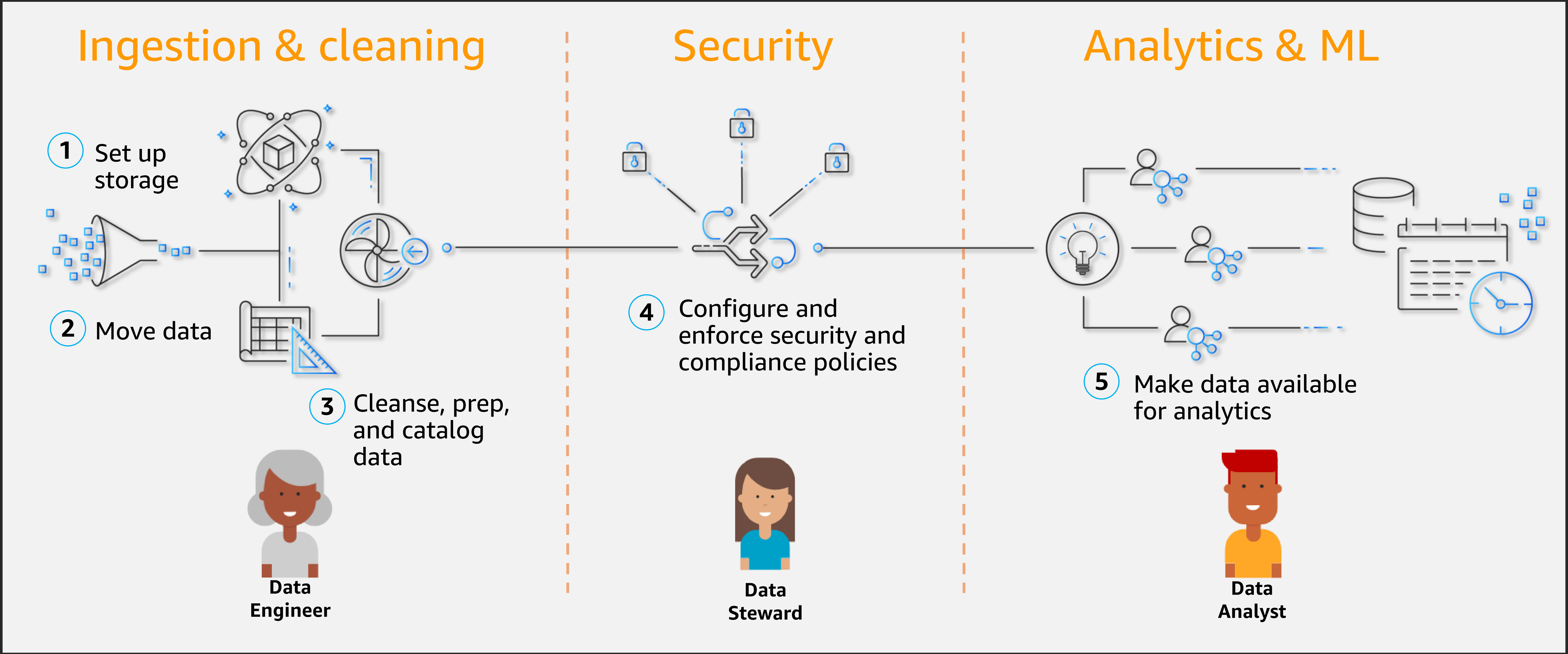
What is hard about building data lakes?

AWS Lake Formation and its blueprints makes data lakes easy.

Demo!

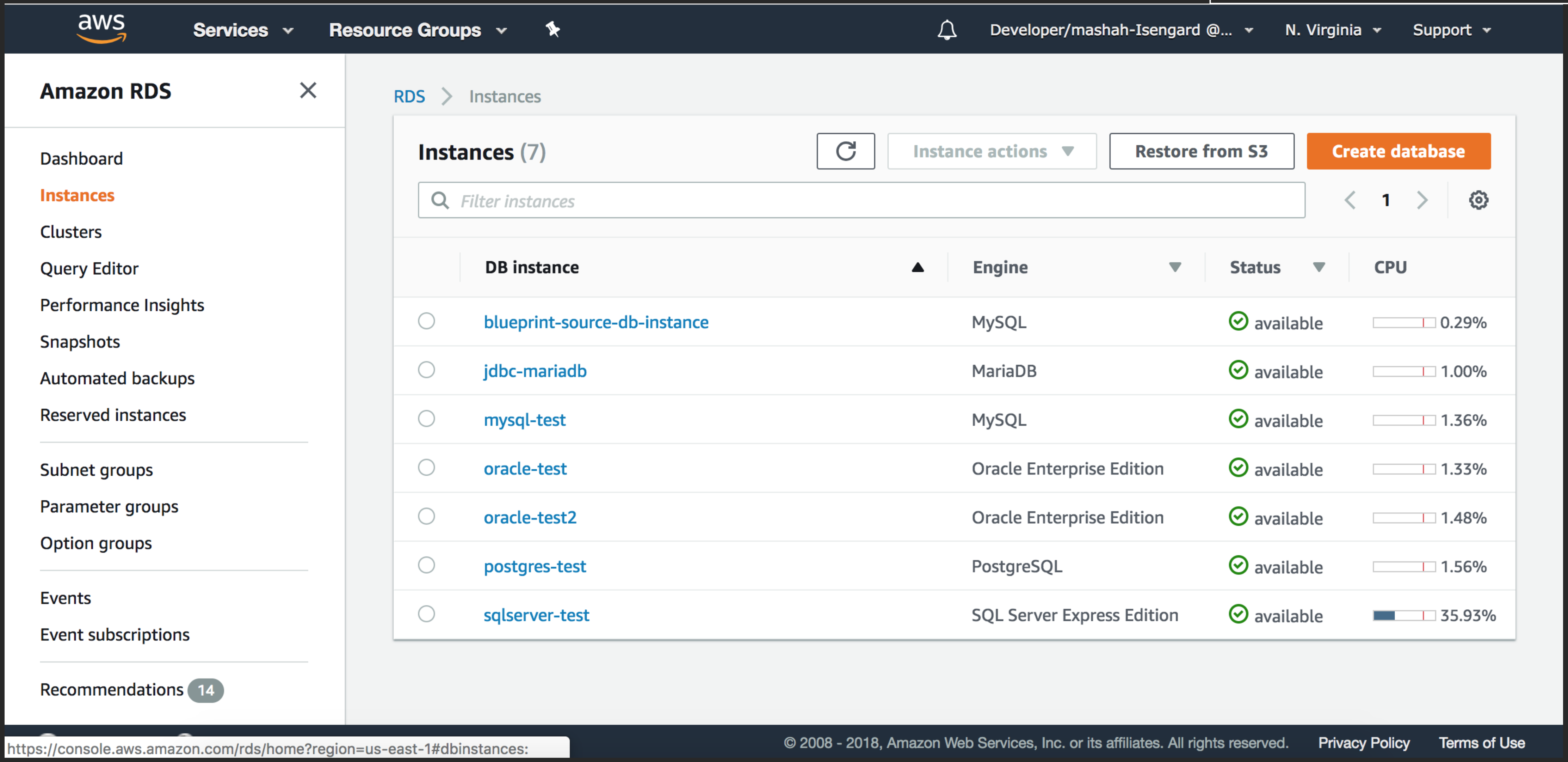
Building clean and secure data lakes can take months

Typical steps of building a data lake


Ingestion & cleaning
Security
Analytics & ML
1 Set up storage
2 Move data
3 Cleanse, prep, and catalog data
Data Engineer
4 Configure and enforce security and compliance policies
Data Steward
5 Make data available for analytics
Data Analyst

Sample of steps required

Find sources



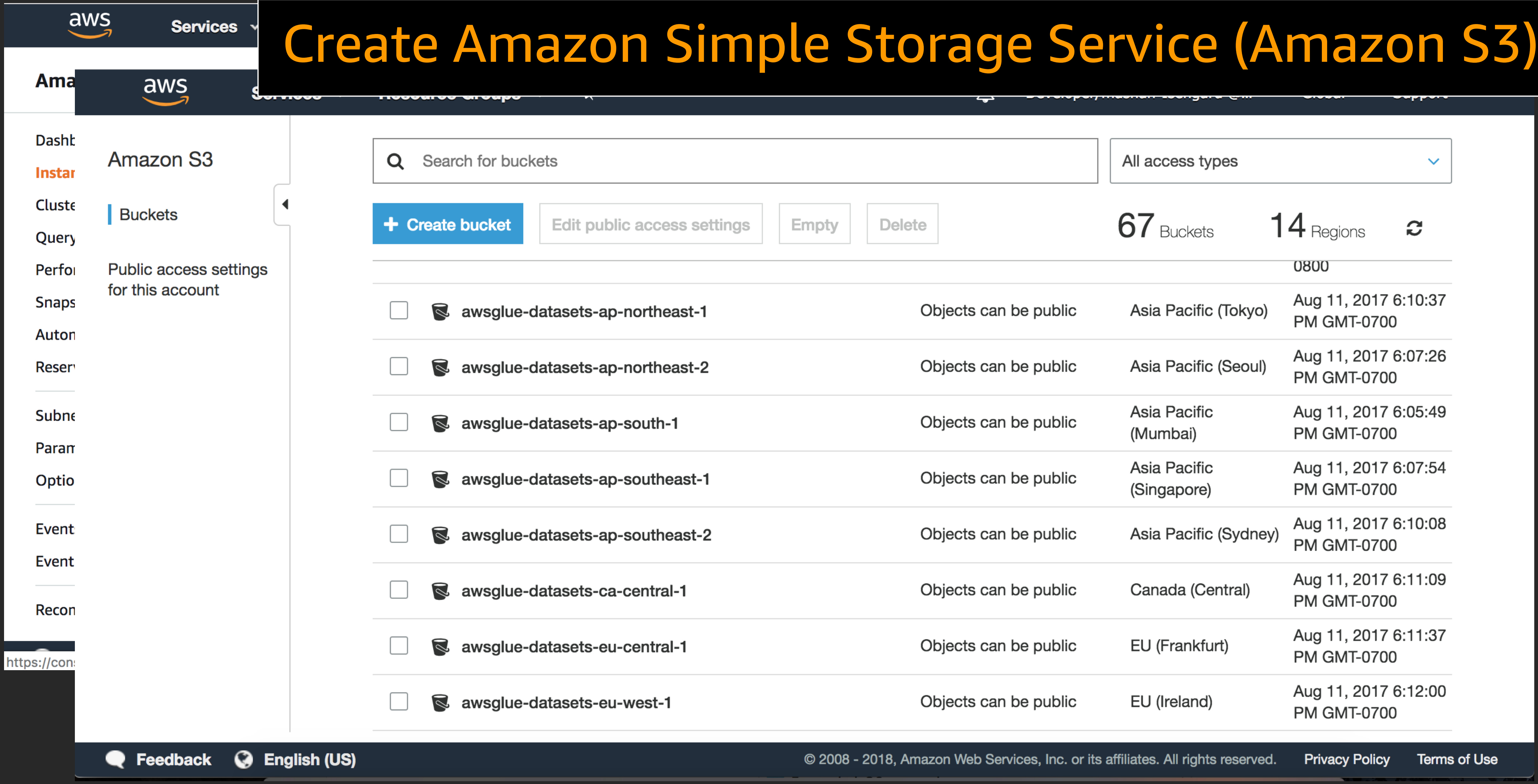
DB instance	Engine	Status	CPU
blueprint-source-db-instance	MySQL	available	0.29%
jdbc-mariadb	MariaDB	available	1.00%
mysql-test	MySQL	available	1.36%
oracle-test	Oracle Enterprise Edition	available	1.33%
oracle-test2	Oracle Enterprise Edition	available	1.48%
postgres-test	PostgreSQL	available	1.56%
sqlserver-test	SQL Server Express Edition	available	35.93%

https://console.aws.amazon.com/rds/home?region=us-east-1#dbinstances:

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

Sample of steps required

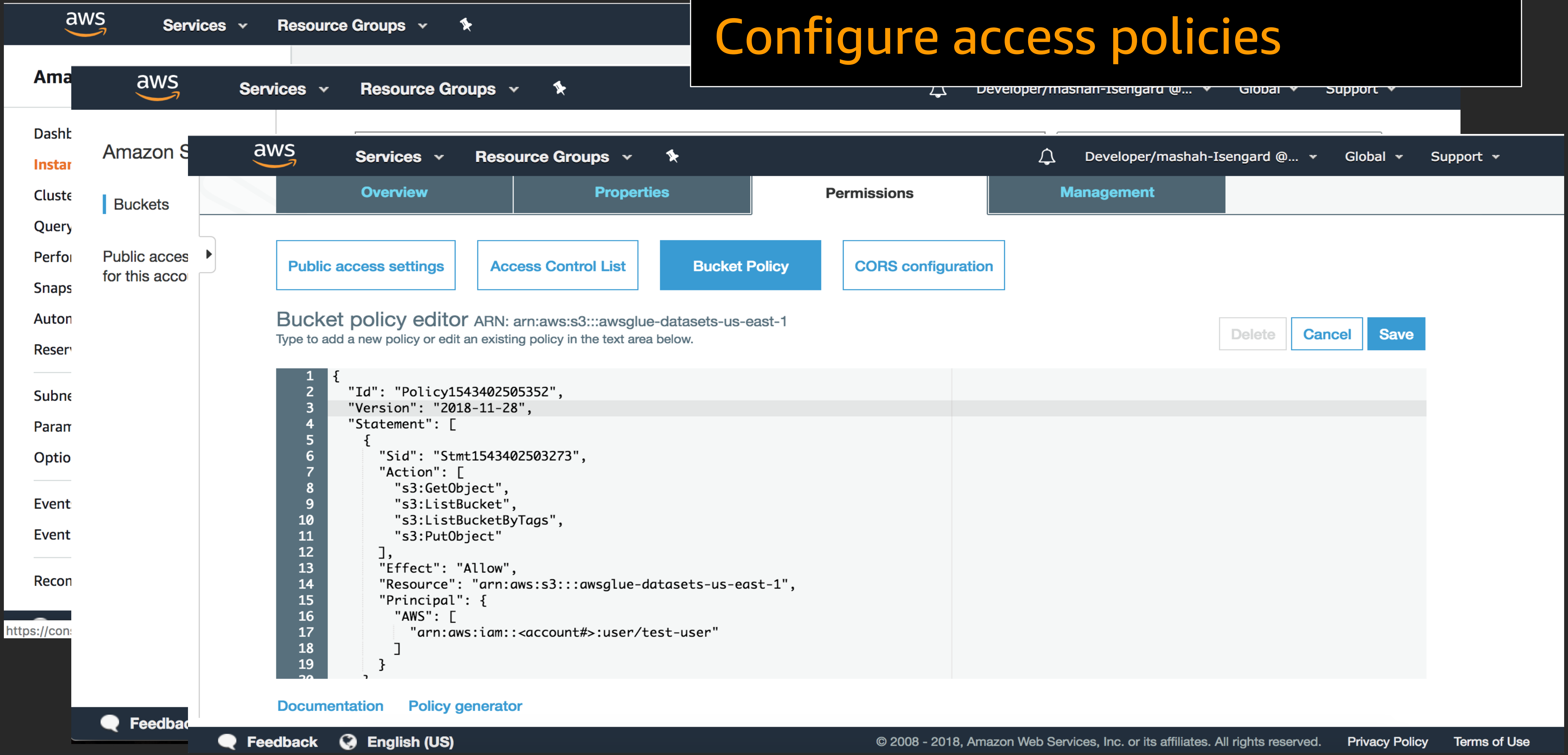
Create Amazon Simple Storage Service (Amazon S3) locations



			0800
awsglue-datasets-ap-northeast-1	Objects can be public	Asia Pacific (Tokyo)	Aug 11, 2017 6:10:37 PM GMT-0700
awsglue-datasets-ap-northeast-2	Objects can be public	Asia Pacific (Seoul)	Aug 11, 2017 6:07:26 PM GMT-0700
awsglue-datasets-ap-south-1	Objects can be public	Asia Pacific (Mumbai)	Aug 11, 2017 6:05:49 PM GMT-0700
awsglue-datasets-ap-southeast-1	Objects can be public	Asia Pacific (Singapore)	Aug 11, 2017 6:07:54 PM GMT-0700
awsglue-datasets-ap-southeast-2	Objects can be public	Asia Pacific (Sydney)	Aug 11, 2017 6:10:08 PM GMT-0700
awsglue-datasets-ca-central-1	Objects can be public	Canada (Central)	Aug 11, 2017 6:11:09 PM GMT-0700
awsglue-datasets-eu-central-1	Objects can be public	EU (Frankfurt)	Aug 11, 2017 6:11:37 PM GMT-0700
awsglue-datasets-eu-west-1	Objects can be public	EU (Ireland)	Aug 11, 2017 6:12:00 PM GMT-0700

Sample of steps required

Configure access policies

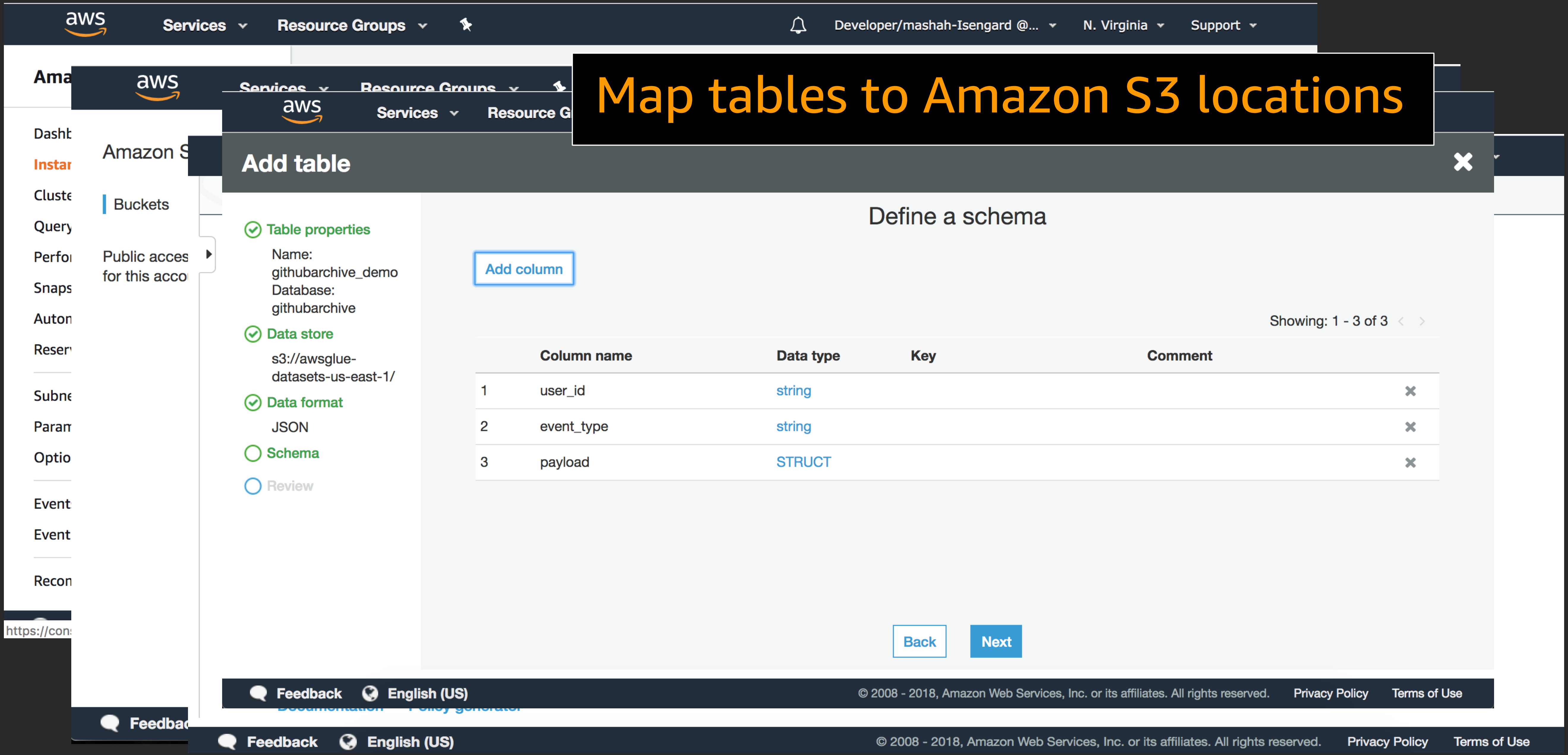


Bucket policy editor ARN: arn:aws:s3:::awsglue-datasets-us-east-1
Type to add a new policy or edit an existing policy in the text area below.

```
{
  "Id": "Policy1543402505352",
  "Version": "2018-11-28",
  "Statement": [
    {
      "Sid": "Stmt1543402503273",
      "Action": [
        "s3:GetObject",
        "s3:ListBucket",
        "s3:ListBucketByTags",
        "s3:PutObject"
      ],
      "Effect": "Allow",
      "Resource": "arn:aws:s3:::awsglue-datasets-us-east-1",
      "Principal": {
        "AWS": [
          "arn:aws:iam::<account#>:user/test-user"
        ]
      }
```

© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sample of steps required



Map tables to Amazon S3 locations

Define a schema

Table properties
Name: githubarchive_demo
Database: githubarchive

Data store
s3://awsglue-datasets-us-east-1/

Data format
JSON

Schema

Review

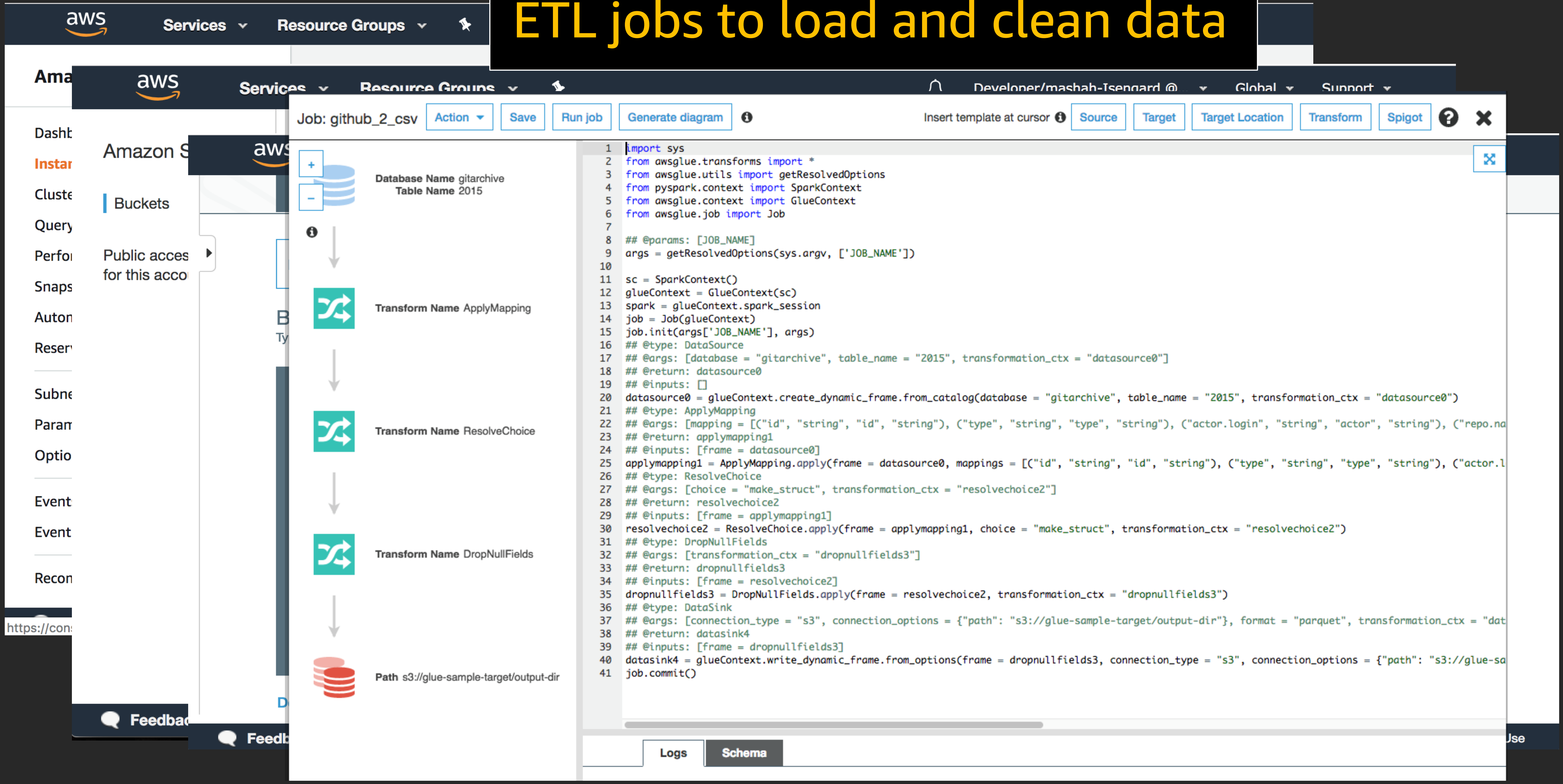
Showing: 1 - 3 of 3

	Column name	Data type	Key	Comment
1	user_id	string		
2	event_type	string		
3	payload	STRUCT		

© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Sample of steps required

ETL jobs to load and clean data



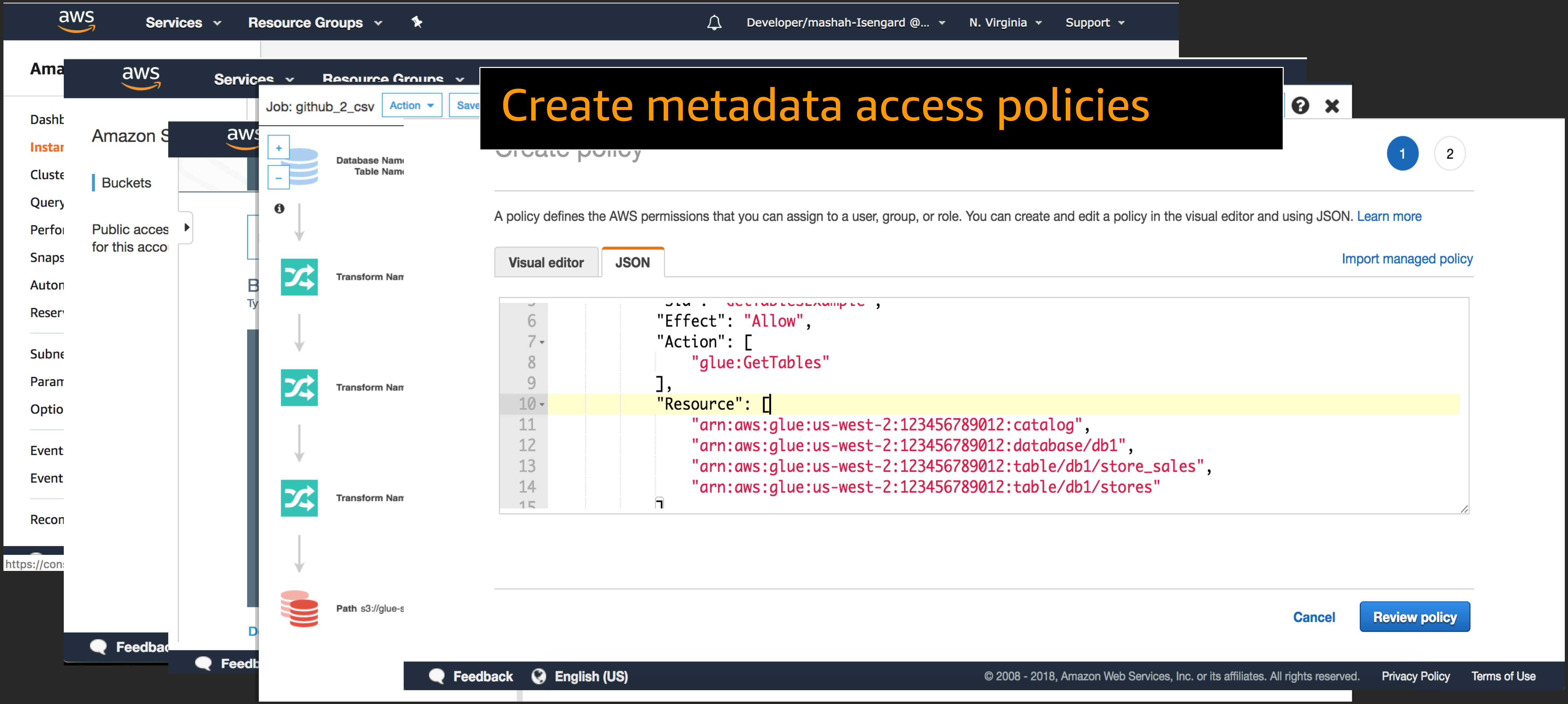
```python
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
## @type: DataSource
## @args: [database = "gitarchive", table_name = "2015", transformation_ctx = "datasource0"]
## @return: datasource0
## @inputs: []
datasource0 = glueContext.create_dynamic_frame.from_catalog(database = "gitarchive", table_name = "2015", transformation_ctx = "datasource0")
## @type: ApplyMapping
## @args: [mapping = [("id", "string", "id", "string"), ("type", "string", "type", "string"), ("actor.login", "string", "actor", "string"), ("repo.na
## @return: applymapping1
## @inputs: [frame = datasource0]
applymapping1 = ApplyMapping.apply(frame = datasource0, mappings = [("id", "string", "id", "string"), ("type", "string", "type", "string"), ("actor.l
## @type: ResolveChoice
## @args: [choice = "make_struct", transformation_ctx = "resolvechoice2"]
## @return: resolvechoice2
## @inputs: [frame = applymapping1]
resolvechoice2 = ResolveChoice.apply(frame = applymapping1, choice = "make_struct", transformation_ctx = "resolvechoice2")
## @type: DropNullFields
## @args: [transformation_ctx = "dropnullfields3"]
## @return: dropnullfields3
## @inputs: [frame = resolvechoice2]
dropnullfields3 = DropNullFields.apply(frame = resolvechoice2, transformation_ctx = "dropnullfields3")
## @type: DataSink
## @args: [connection_type = "s3", connection_options = {"path": "s3://glue-sample-target/output-dir"}, format = "parquet", transformation_ctx = "dat
## @return: datasink4
## @inputs: [frame = dropnullfields3]
datasink4 = glueContext.write_dynamic_frame.from_options(frame = dropnullfields3, connection_type = "s3", connection_options = {"path": "s3://glue-sa
job.commit()
```

https://con

Sample of steps required



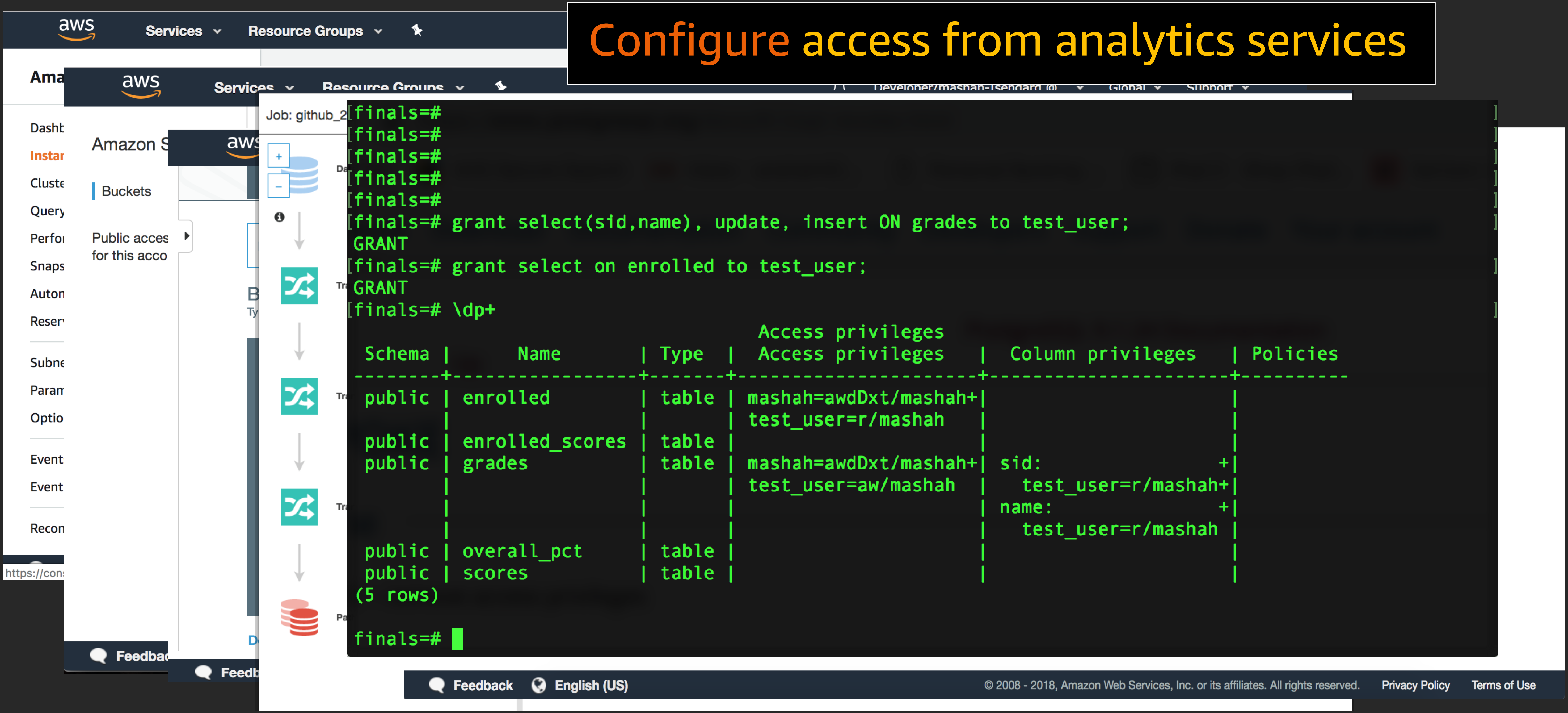
Create metadata access policies

Create policy

A policy defines the AWS permissions that you can assign to a user, group, or role. You can create and edit a policy in the visual editor and using JSON. Learn more

```
            "Effect": "Allow",
            "Action": [
                "glue:GetTables"
            ],
            "Resource": [
                "arn:aws:glue:us-west-2:123456789012:catalog",
                "arn:aws:glue:us-west-2:123456789012:database/db1",
                "arn:aws:glue:us-west-2:123456789012:table/db1/store_sales",
                "arn:aws:glue:us-west-2:123456789012:table/db1/stores"
            ]
```

Sample of steps required

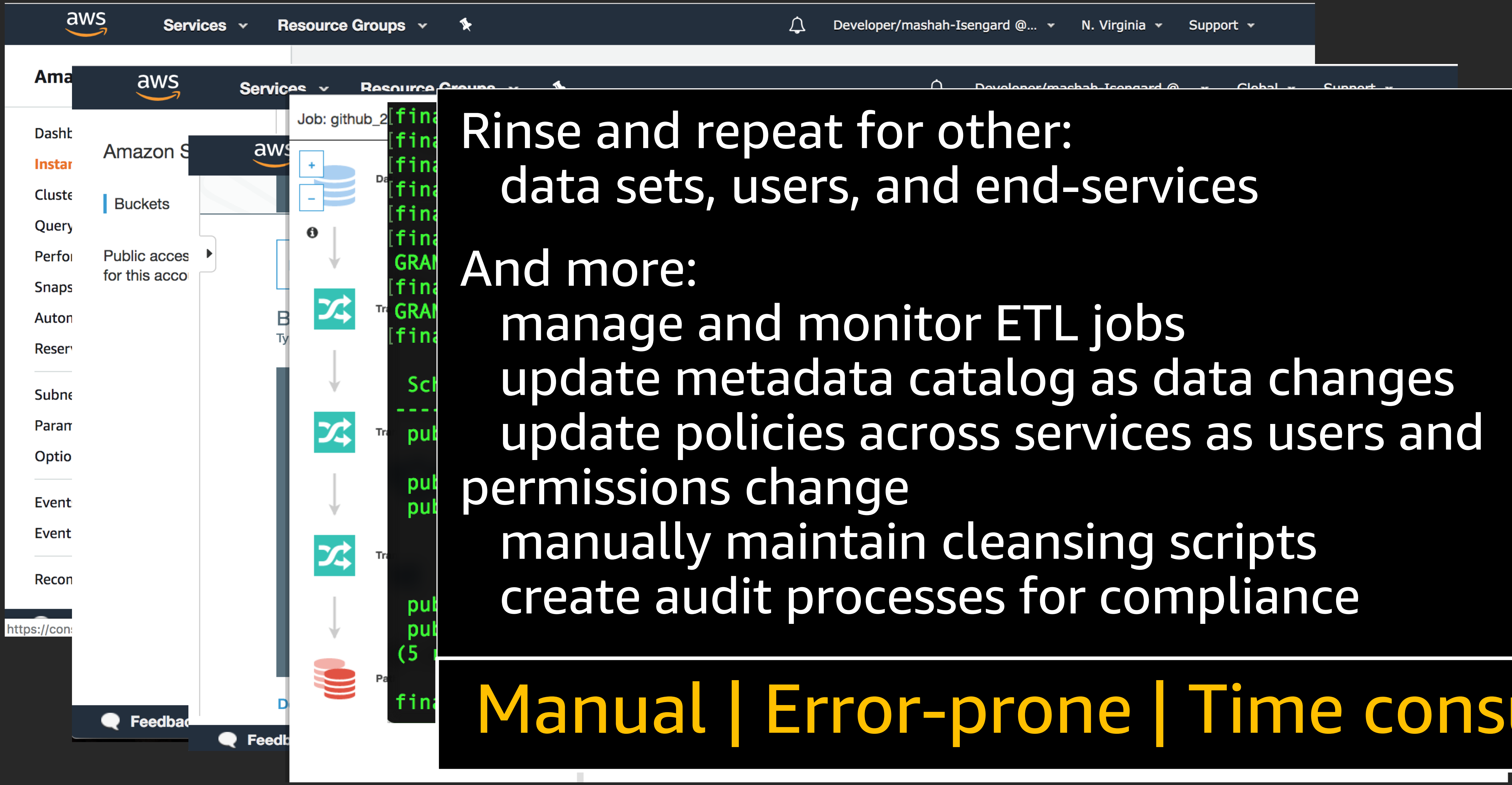


Configure access from analytics services

```
[finals=#
[finals=#
[finals=#
[finals=#
[finals=#
[finals=# grant select(sid,name), update, insert ON grades to test_user;
GRANT
[finals=# grant select on enrolled to test_user;
GRANT
[finals=# \dp+
                                    Access privileges
 Schema |      Name       | Type  |   Access privileges    |  Column privileges   | Policies
--------+-----------------+-------+------------------------+----------------------+----------
 public | enrolled        | table | mashah=awdDxt/mashah+|                      |
        |                 |       | test_user=r/mashah   |                      |
 public | enrolled_scores | table |                        |                      |
 public | grades          | table | mashah=awdDxt/mashah+| sid:                +|
        |                 |       | test_user=aw/mashah  |   test_user=r/mashah+|
        |                 |       |                        | name:               +|
        |                 |       |                        |   test_user=r/mashah |
 public | overall_pct     | table |                        |                      |
 public | scores          | table |                        |                      |
(5 rows)

finals=# 
```

Sample of steps required



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

Agenda

Trends driving the revolution

What is a data lake?

Why AWS for data lake?

What is hard about building data lakes?

AWS Lake Formation and its blueprints makes data lakes easy.

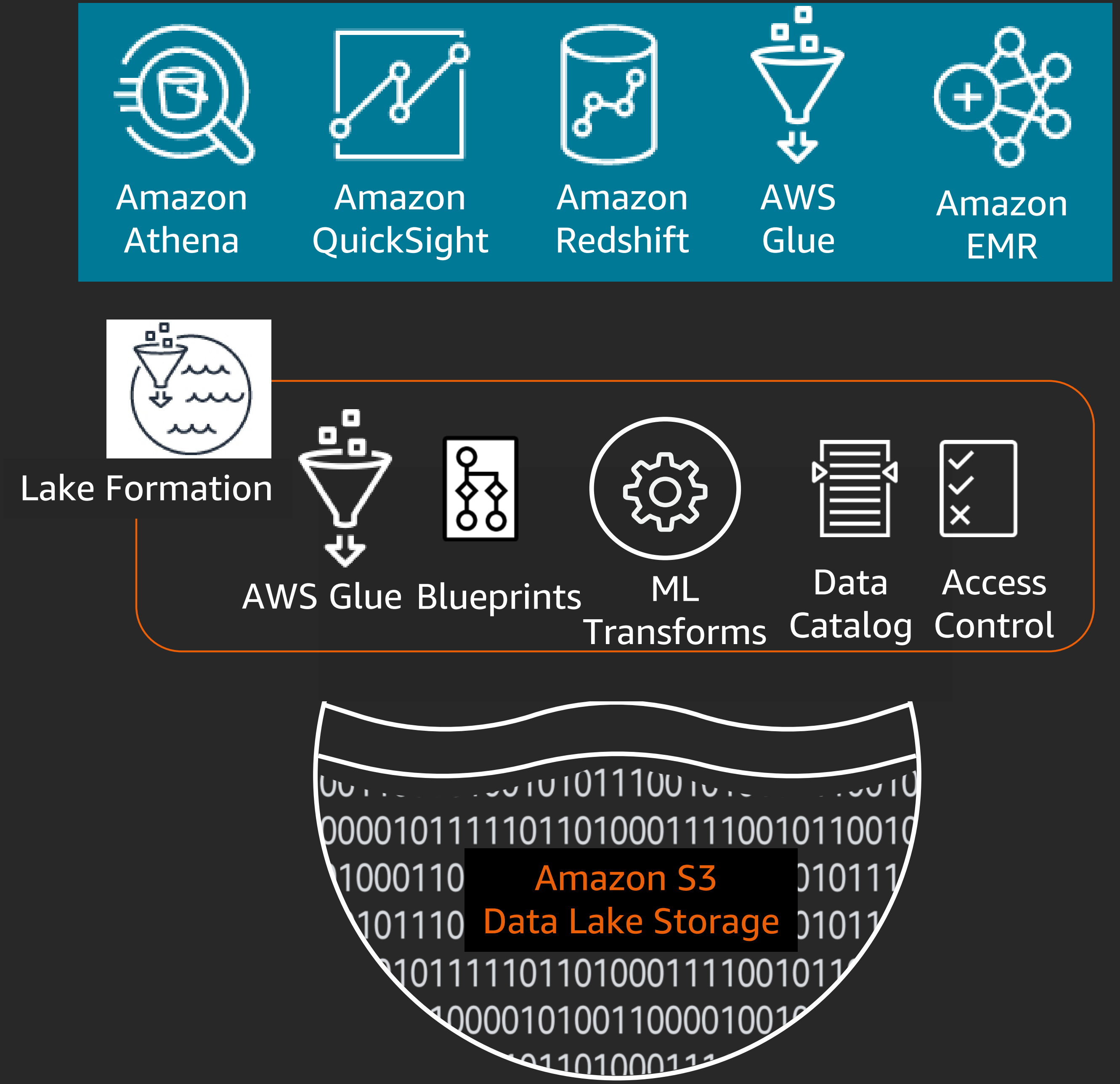
Demo!

Fully managed service that enables

data engineers | data stewards | data analysts

to build clean and secure data lakes in days

AWS Lake Formation solution stack



Discovery, **sharing**, and **integrated tools** to enable every user

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

Building data lakes with AWS Lake Formation

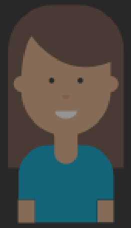
Ingestion & cleaning

Data Engineer

Blueprints simplify ingest

ML transforms for data cleaning

Security



Data Steward

Centralized permissions

Real time monitoring & integrated auditing

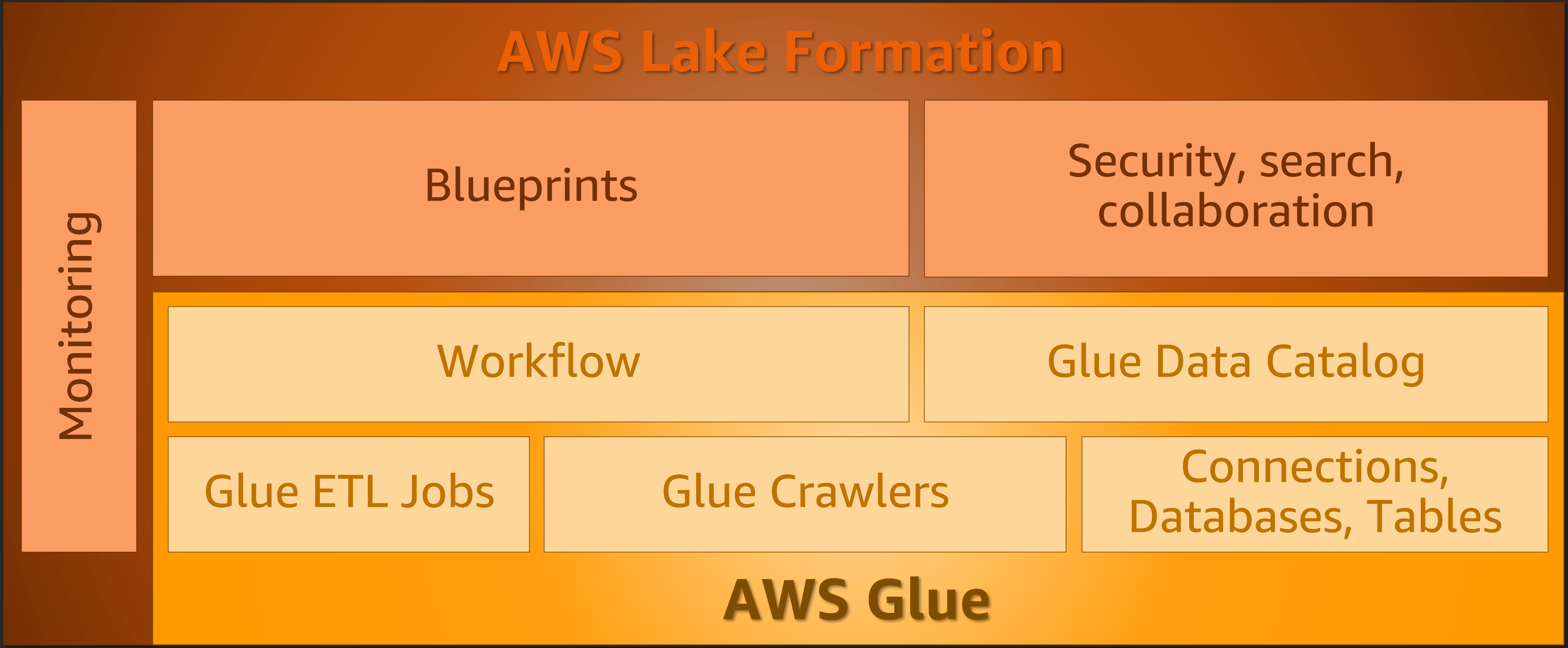
Analytics & ML

Data Analyst

Right tool for the job

Comprehensive portfolio of integrated tools

AWS Lake Formation builds on AWS Glue



AWS Glue provides scalable serverless components

Data Catalog

Apache Hive Metastore compatible

Integrated with AWS analytic services

Crawlers

Automatically infer schemas

Populate data catalog

Serverless ETL

Interactive development

Apache Spark / Python shell jobs

Serverless execution

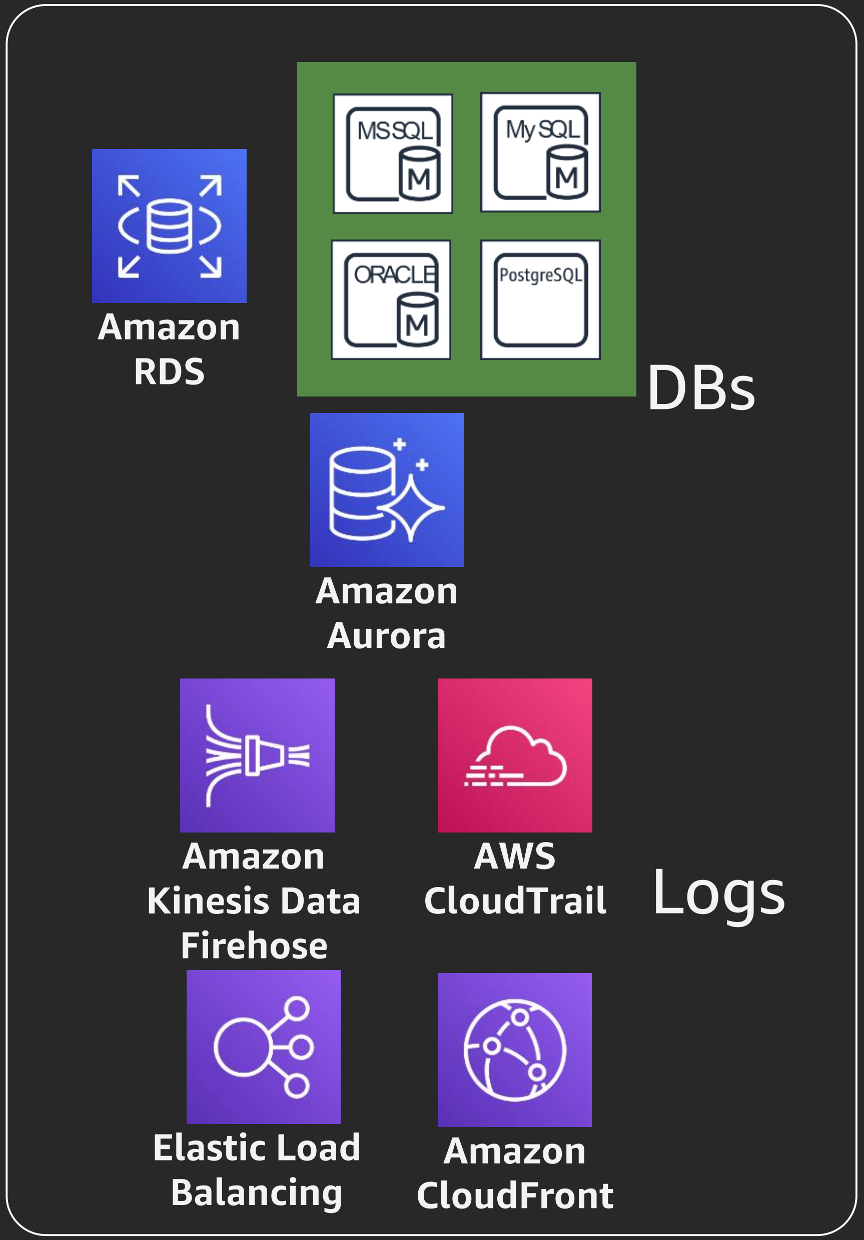
Flexible Workflows

Orchestrate triggers, crawlers & jobs

Author & monitor entire flows

Integrated alerting

Easily load data into your data lake w/ blueprints



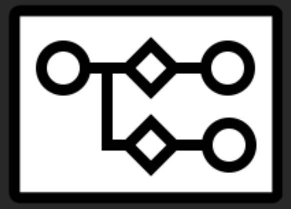
Prebuilt templates to serve common ingestion use cases

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 centralized permissions

Ingestion & cleaning

Blueprints simplify ingest

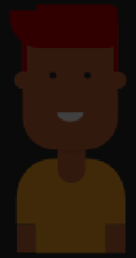
ML transforms for data cleaning

Security

Centralized permissions

Real-time monitoring & integrated auditing

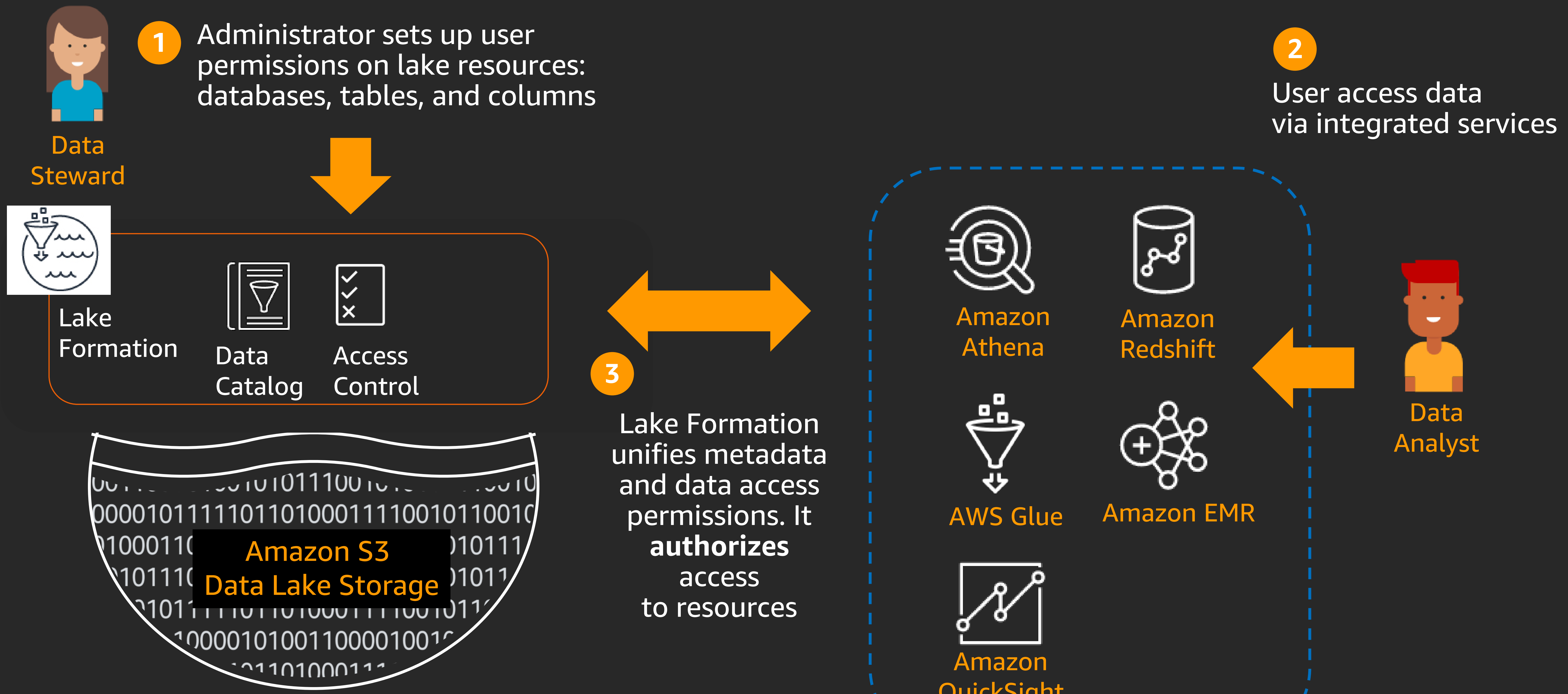
Analytics & ML



Right tool for the job

Comprehensive portfolio of integrated tools

Centralized permissions


1
Administrator sets up user permissions on lake resources: databases, tables, and columns
Data Steward
Lake Formation
Data Catalog
Access Control
Amazon S3
Data Lake Storage
2
User access data via integrated services
Amazon Athena
Amazon Redshift
AWS Glue
Amazon EMR
Amazon QuickSight
Data Analyst
3
Lake Formation unifies metadata and data access permissions. It **authorizes** access to resources

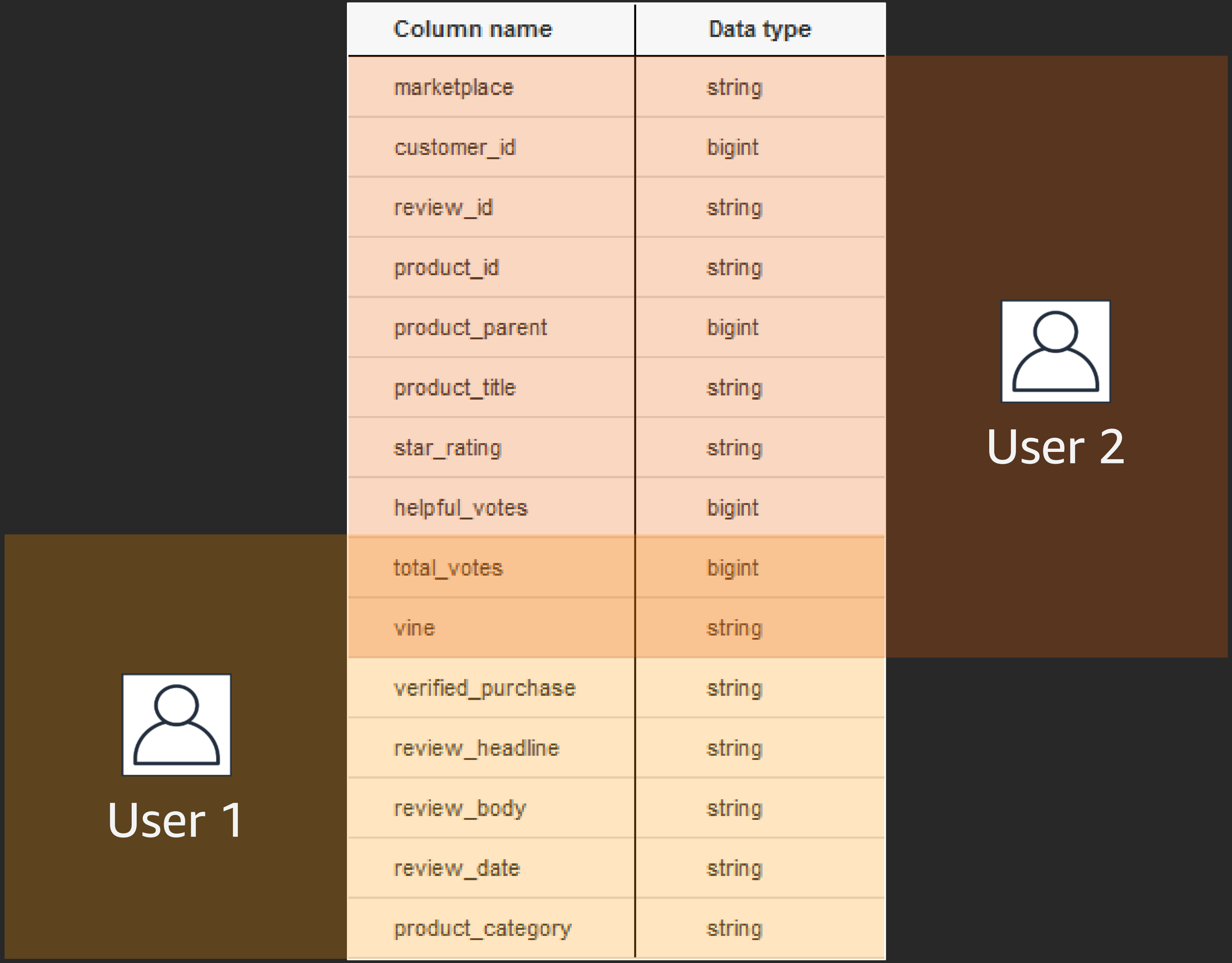
Security permissions in AWS Lake Formation

Control data access with simple grant and revoke permissions

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

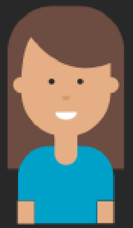
Easily view permissions granted to a particular user

Audit all data access in one place



Column name	Data type
marketplace	string
customer_id	bigint
review_id	string
product_id	string
product_parent	bigint
product_title	string
star_rating	string
helpful_votes	bigint
total_votes	bigint
vine	string
verified_purchase	string
review_headline	string
review_body	string
review_date	string
product_category	string

Audit data access

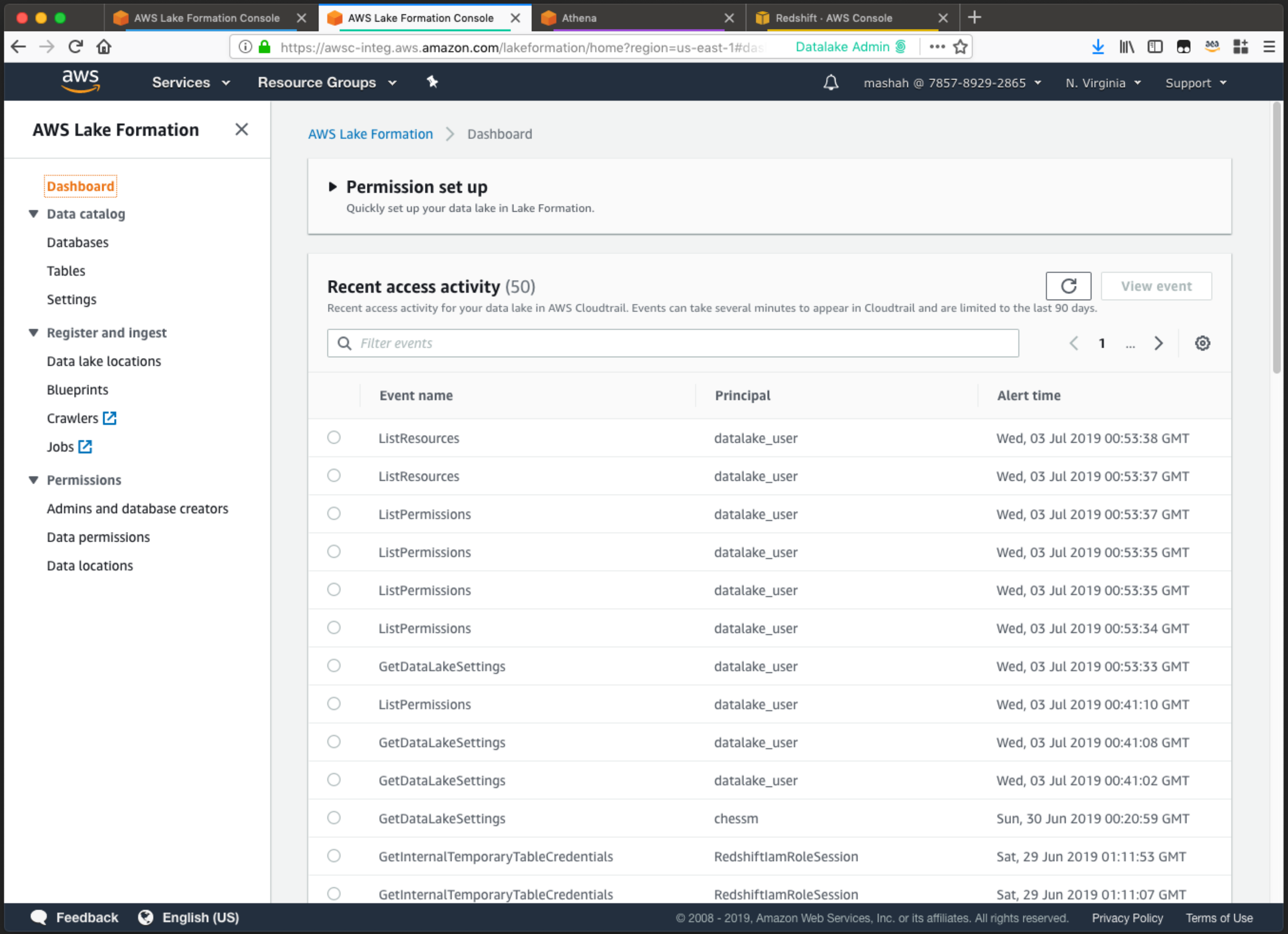


Data Steward

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



Permission set up

Quickly set up your data lake in Lake Formation.

Recent access activity (50)

View event

Recent access activity for your data lake in AWS Cloudtrail. Events can take several minutes to appear in Cloudtrail and are limited to the last 90 days.

Filter events

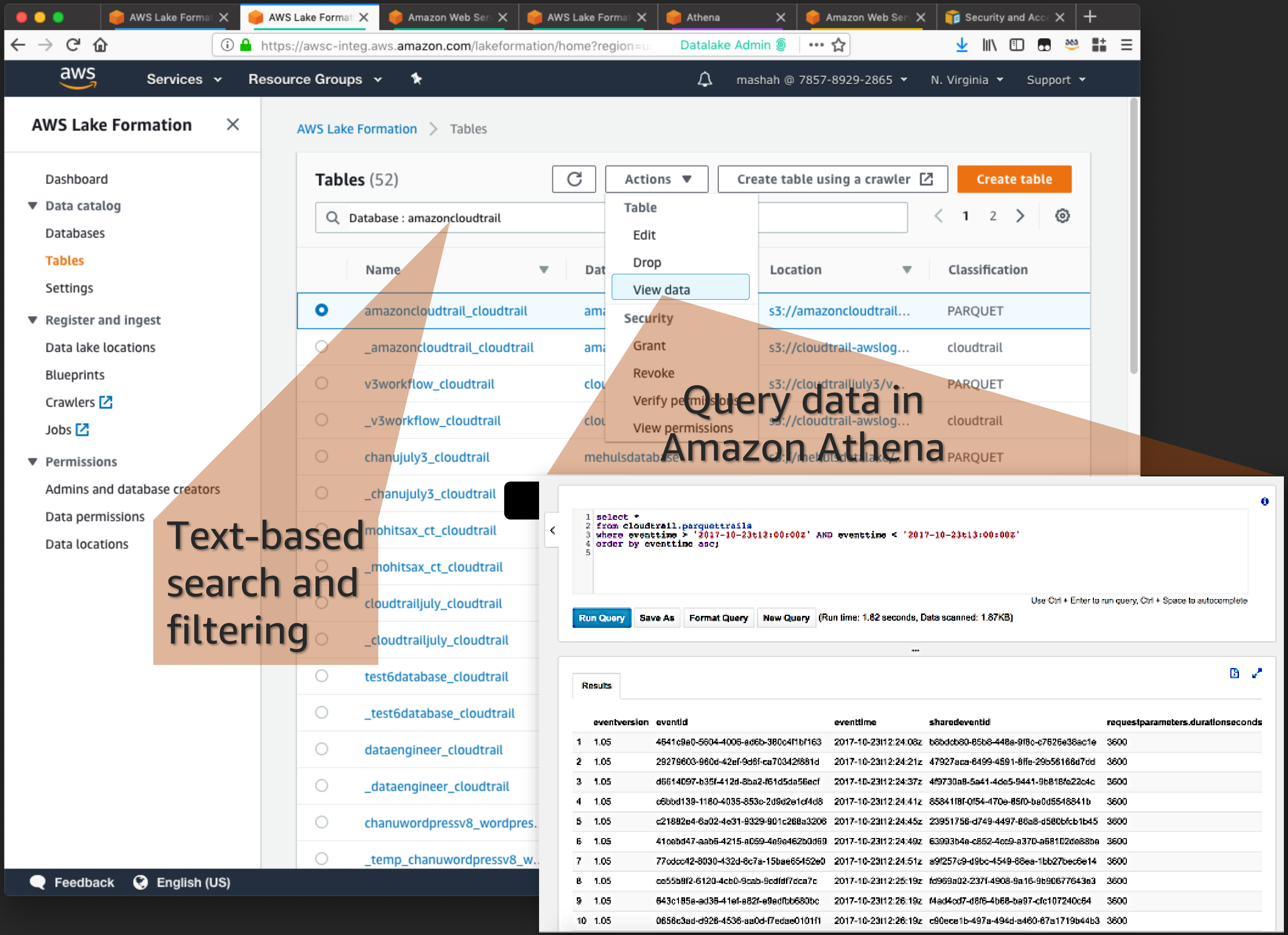
Event name	Principal	Alert time
ListResources	datalake_user	Wed, 03 Jul 2019 00:53:38 GMT
ListResources	datalake_user	Wed, 03 Jul 2019 00:53:37 GMT
ListPermissions	datalake_user	Wed, 03 Jul 2019 00:53:37 GMT
ListPermissions	datalake_user	Wed, 03 Jul 2019 00:53:35 GMT
ListPermissions	datalake_user	Wed, 03 Jul 2019 00:53:35 GMT
ListPermissions	datalake_user	Wed, 03 Jul 2019 00:53:34 GMT
GetDataLakeSettings	datalake_user	Wed, 03 Jul 2019 00:53:33 GMT
ListPermissions	datalake_user	Wed, 03 Jul 2019 00:41:10 GMT
GetDataLakeSettings	datalake_user	Wed, 03 Jul 2019 00:41:08 GMT
GetDataLakeSettings	datalake_user	Wed, 03 Jul 2019 00:41:02 GMT
GetDataLakeSettings	chessm	Sun, 30 Jun 2019 00:20:59 GMT
GetInternalTemporaryTableCredentials	RedshiftIamRoleSession	Sat, 29 Jun 2019 01:11:53 GMT
GetInternalTemporaryTableCredentials	RedshiftIamRoleSession	Sat, 29 Jun 2019 01:11:07 GMT

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



Name	Dat	Location	Classification
amazoncloudtrail_cloudtrail	ama	s3://amazoncloudtrail...	PARQUET
_amazoncloudtrail_cloudtrail	ama	s3://cloudtrail-awslog...	cloudtrail
v3workflow_cloudtrail	clou	s3://cloudtrailjuly3/v...	PARQUET
_v3workflow_cloudtrail	clou	s3://cloudtrail-awslog...	cloudtrail
chanujuly3_cloudtrail	mehulsdatabase	s3://mehulsdatalake/...	PARQUET
_chanujuly3_cloudtrail			
mohitsax_ct_cloudtrail			
_mohitsax_ct_cloudtrail			
cloudtrailjuly_cloudtrail			
_cloudtrailjuly_cloudtrail			
test6database_cloudtrail			
_test6database_cloudtrail			
dataengineer_cloudtrail			
_dataengineer_cloudtrail			
chanuwordpressv8_wordpres...			
_temp_chanuwordpressv8_w...			

Text-based search and filtering

Query data in Amazon Athena

```
select *
from cloudtrail.parquettrails
where eventtime > '2017-10-23t12:00:00z' AND eventtime < '2017-10-23t13:00:00z'
order by eventtime asc;
```

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

Run Query | Save As | Format Query | New Query (Run time: 1.82 seconds, Data scanned: 1.87KB)

	eventversion	eventid	eventtime	sharedeventid	requestparameters.durationseconds
1	1.05	4541c9a0-5604-4006-ad6b-380c4f1bf163	2017-10-23t12:24:08z	b8bdcb80-85b8-448a-9f8c-c7526e38ac1e	3600
2	1.05	29279600-960d-42ef-9d6f-ca70342f881d	2017-10-23t12:24:21z	47927aca-6499-4591-8ffe-29b56166d7dd	3600
3	1.05	d6614097-b35f-412d-8ba2-f61d5da56ecf	2017-10-23t12:24:37z	4f9730a8-5a41-4de5-9441-9b818fe22c4c	3600
4	1.05	c6bbd139-1160-4035-853c-2d9d2e1cf4d8	2017-10-23t12:24:41z	85841f8f-0f54-470e-85f0-ba0d5548841b	3600
5	1.05	c21892e4-6a02-4e31-9329-901c268a3206	2017-10-23t12:24:45z	23951756-d749-4497-86a8-d580bfcb1b45	3600
6	1.05	41cebd47-aab6-4215-a059-4e9e462b0d59	2017-10-23t12:24:49z	63993b4e-c852-4cc9-a370-a68102de88be	3600
7	1.05	77cdcc42-8030-432d-8c7a-15bae65452e0	2017-10-23t12:24:51z	a9f257c9-d9bc-4549-88ea-1bb27bec6e14	3600
8	1.05	ce55b9f2-6120-4cb0-9cab-9cdfdf7dca7c	2017-10-23t12:25:19z	fd969a02-237f-4908-9a16-9b90677643e3	3600
9	1.05	643c185a-ad35-41ef-a82f-e9edfbb680bc	2017-10-23t12:26:19z	f4ad4cd7-d8f6-4b68-ba97-cfc107240c64	3600
10	1.05	0656c3ad-d926-4536-aa0d-f7edae0101f1	2017-10-23t12:26:19z	c90ece1b-497a-494d-a460-67a1719b44b3	3600

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

Amazon Athena

Amazon QuickSight

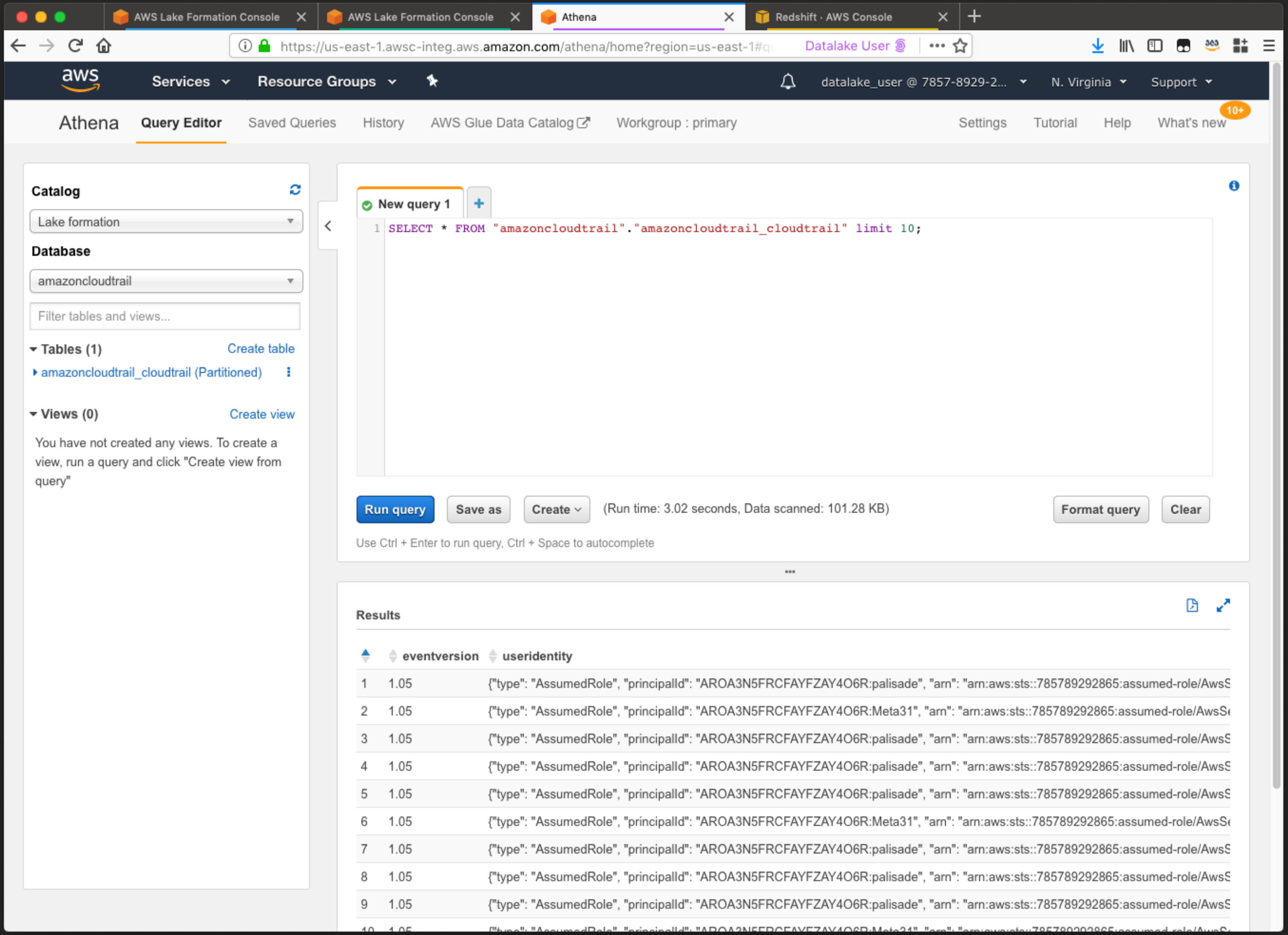
Comprehensive portfolio of integrated tools

Compliant services honor
Lake Formation permissions

Amazon QuickSight | Amazon Redshift | Amazon EMR | AWS Glue | AWS Athena

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

All access is logged and auditable



You have not created any views. To create a view, run a query and click "Create view from query"

```
SELECT * FROM "amazoncloudtrail"."amazoncloudtrail_cloudtrail" limit 10;
```

(Run time: 3.02 seconds, Data scanned: 101.28 KB)

Use Ctrl + Enter to run query, Ctrl + Space to autocomplete

	eventversion	useridentity
1	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS
2	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:Meta31", "arn": "arn:aws:sts::785789292865:assumed-role/AwsSe
3	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS
4	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS
5	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS
6	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:Meta31", "arn": "arn:aws:sts::785789292865:assumed-role/AwsSe
7	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS
8	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS
9	1.05	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS

Demo

Thank you!

Mona Mona

https://www.linkedin.com/in/mona-08b73837/