

KEM # 7

Introducing AWS Lake Formation

Rahul Pathak
@rahulpathak
GM, Big Data, Data Lakes, Blockchain

aws
re:Invent

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



Setting up and managing data lakes today involves a lot of complicated and time-consuming tasks. **AWS Lake Formation** is a new service (coming soon) that will make it easy to set up a secure data lake in days. You will be able to **ingest, catalog, cleanse, transform, and secure** your data. Explore how AWS Lake Formation will make it easier to combine analytic tools, like **Amazon EMR, Redshift, Athena, Sagemaker, and QuickSight**, on data in your data lake.

Agenda

- Why did we build AWS Lake Formation?
- What is AWS Lake Formation?
- How does AWS Lake Formation help you?



There is **more data** than people think

Data

grows
>10x
every 5 years

Data platforms need to

live for
15
years

scale
1,000x

* IDC, Data Age 2021S: The Evolution of Data to Life-Critical Don't Focus on Big Data, Focus on the Data That's Big, April 2017.

aws
re:Invent

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



Data Scientists



Business Users



Analysts



Applications

Secure

Real time

Flexible

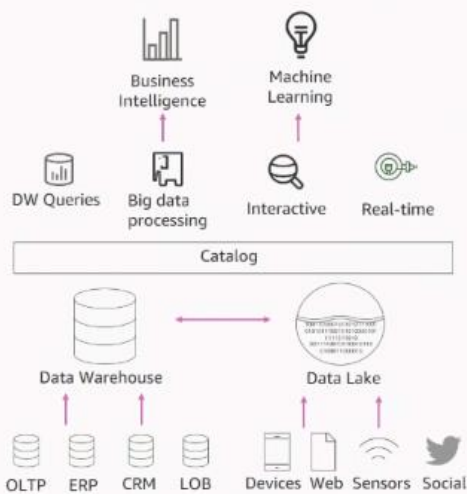
Scalable

There are **more people** accessing data

And **more requirements** for making data available

A data lake is a **centralized repository** that allows you to store all your **structured and unstructured data** at any scale

Why data lakes?



Data Lakes provide:

Relational and non-relational data

Scale-out to EBs

Diverse set of analytics and machine learning tools

Work on data without any data movement

Designed for low cost storage and analytics

aws
re:Invent

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



Any **analytic workload**, any scale,
at the lowest possible
cost

aws
re:Invent

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



More data lakes & analytics on AWS than anywhere else

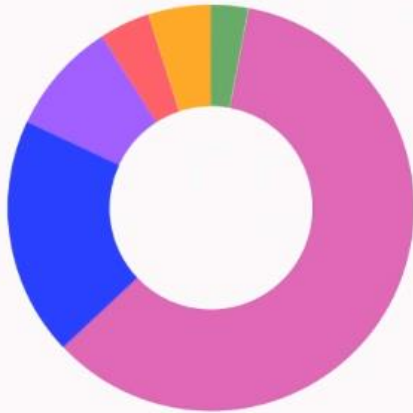


Typical steps of building a data lake



Building data lakes can still take **months**

Data preparation accounts for ~80% of the work



- Building training sets
- Cleaning and organizing data
- Collecting data sets
- Mining data for patterns
- Refining algorithms
- Other

Sample of steps required

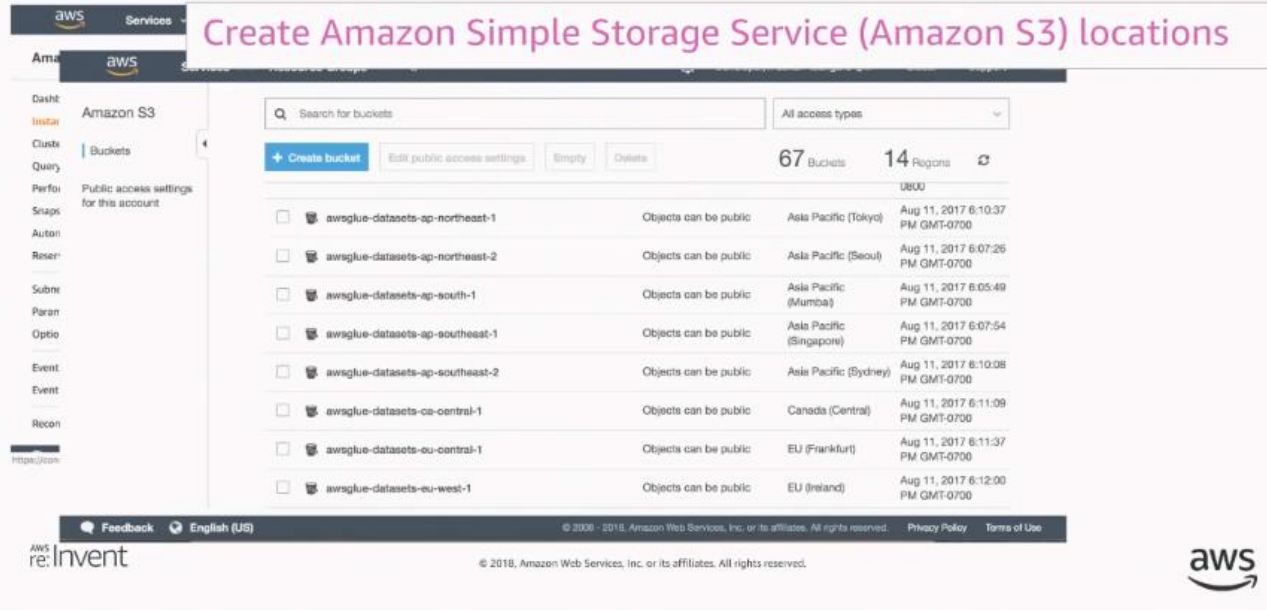
Find sources

The screenshot shows the Amazon RDS console interface. On the left is a navigation menu with options like Dashboard, Instances, Clusters, Query Editor, Performance Insights, Snapshots, Automated backups, Reserved instances, Subnet groups, Parameter groups, Option groups, Events, Event subscriptions, and Recommendations. The main area displays a table of database instances. At the top of the table, there are buttons for 'Instances (7)', 'Filter instances', 'Instance actions', 'Restore from S3', and 'Create database'. The table has columns for DB Instance, Engine, Status, and CPU. Below is a table with 7 rows of database instances.

DB Instance	Engine	Status	CPU
blueprint-source-db-instance	MySQL	available	0.20%
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%

Sample of steps required

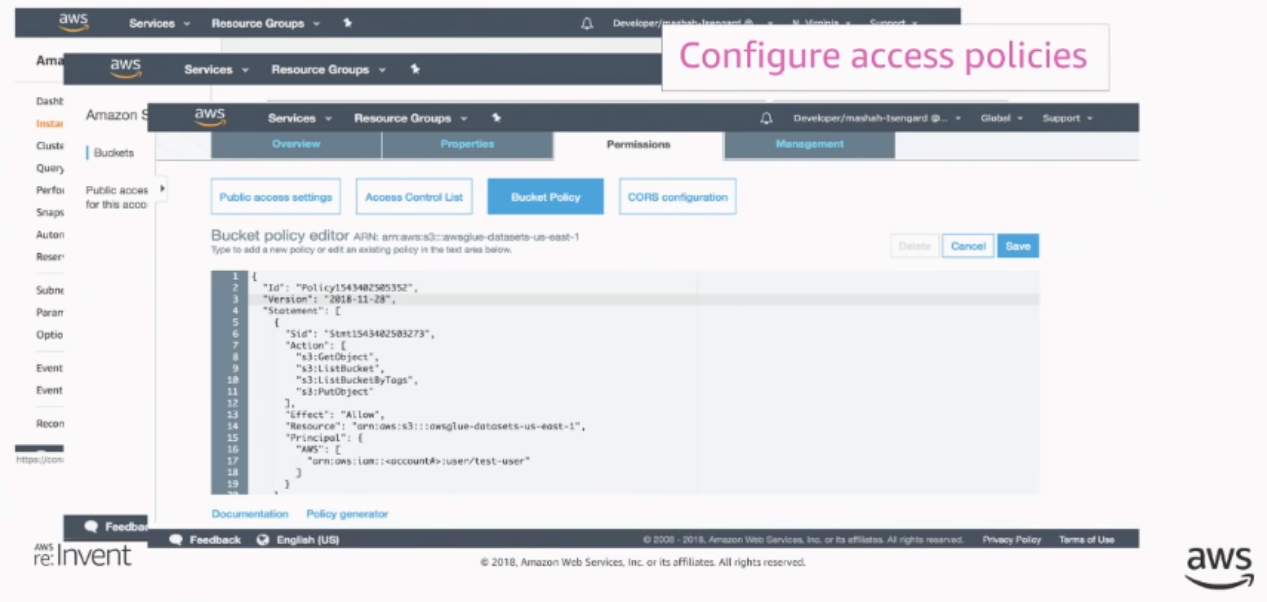
Create Amazon Simple Storage Service (Amazon S3) locations



Name	Access	Region	Creation Date
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

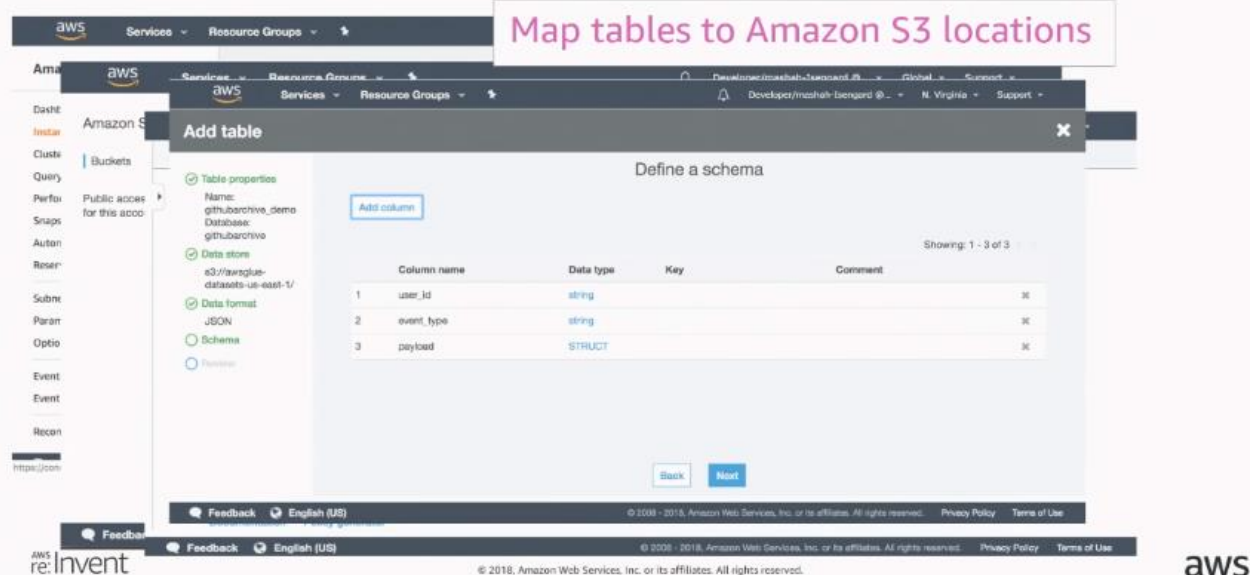
Sample of steps required

Configure access policies

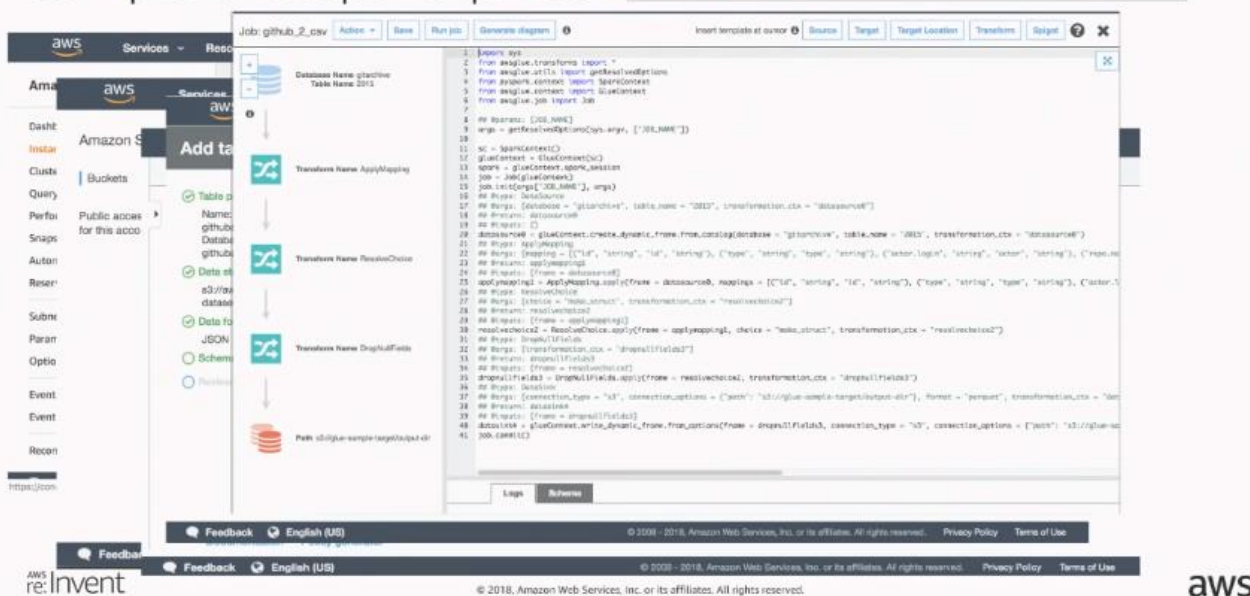


```
1 {
2   "Id": "Policy1543402505352",
3   "Version": "2018-11-28",
4   "Statement": [
5     {
6       "Sid": "Stmt1543402505273",
7       "Action": [
8         "s3:GetObject",
9         "s3:ListBucket",
10        "s3:ListBucketByTags",
11        "s3:PutObject"
12      ],
13      "Effect": "Allow",
14      "Resource": "arn:aws:s3:::awsglue-datasets-us-east-1",
15      "Principal": {
16        "AWS": [
17          "arn:aws:iam::<account-id>:user/test-user"
18        ]
19      }
20    }
21  ]
22 }
```


Sample of steps required



Sample of steps required



Sample of steps required

Job: github_2.csv Action Save Run job Generate diagram

Create policy

Create metadata access policies

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](#)

Visual editor JSON Import managed policy

```
{
  "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"
  ]
}
```

Cancel Review policy

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

Sample of steps required Configure access from analytics services

Job: github_2.csv Action Save Run job Generate diagram

Create policy

Configure access from analytics services

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](#)

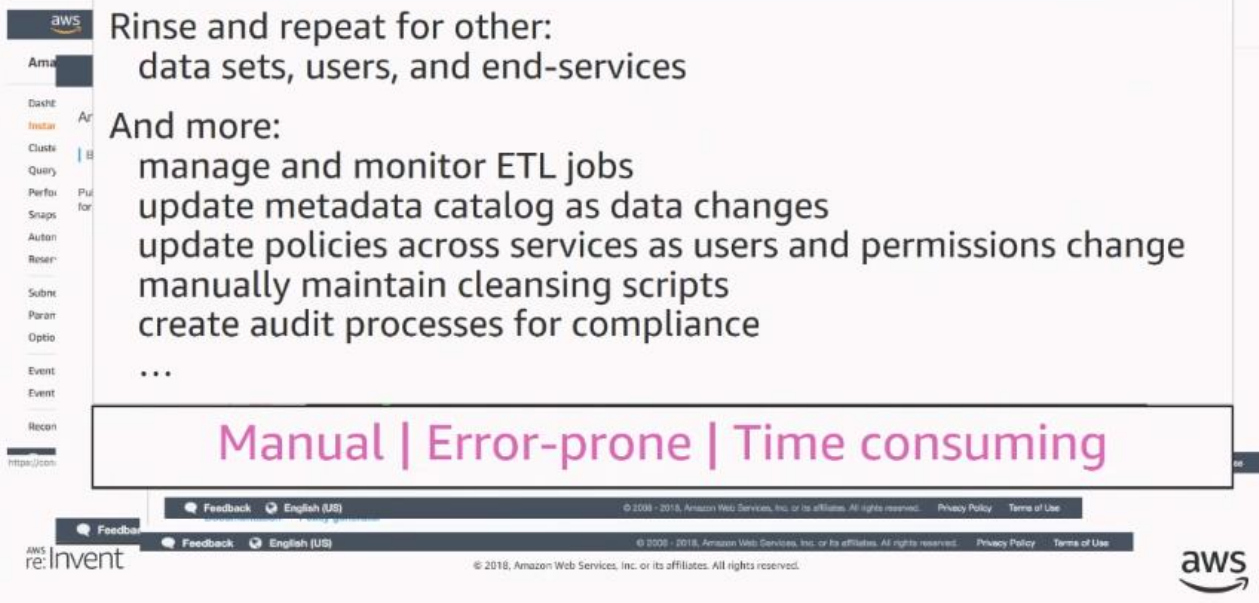
Visual editor JSON Import managed policy

```
{
  "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"
  ]
}
```

Cancel Review policy

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

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

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

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

re:Invent

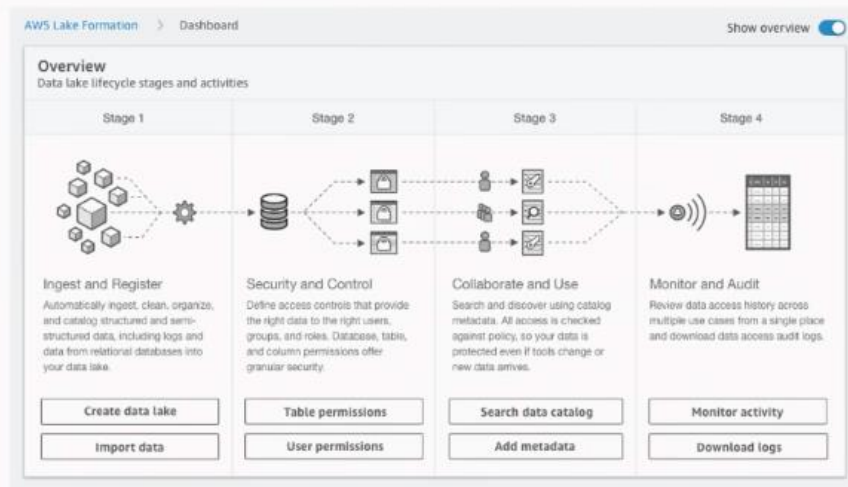
aws

AWS Lake Formation

Build a secure data lake in days

-  Identify, ingest, clean, and transform data
-  Enforce security policies across multiple services
-  Gain and manage new insights

How it works



aws
re:Invent

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



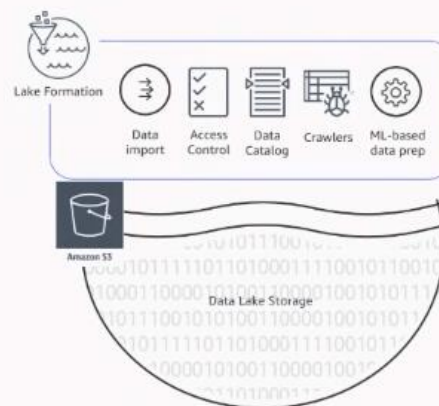
Register existing data or import new

Amazon S3 forms the storage layer for Lake Formation

Register existing S3 buckets that contain your data

Ask Lake Formation to create required S3 buckets and import data into them

Data is stored in your account. You have direct access to it. No lock-in.

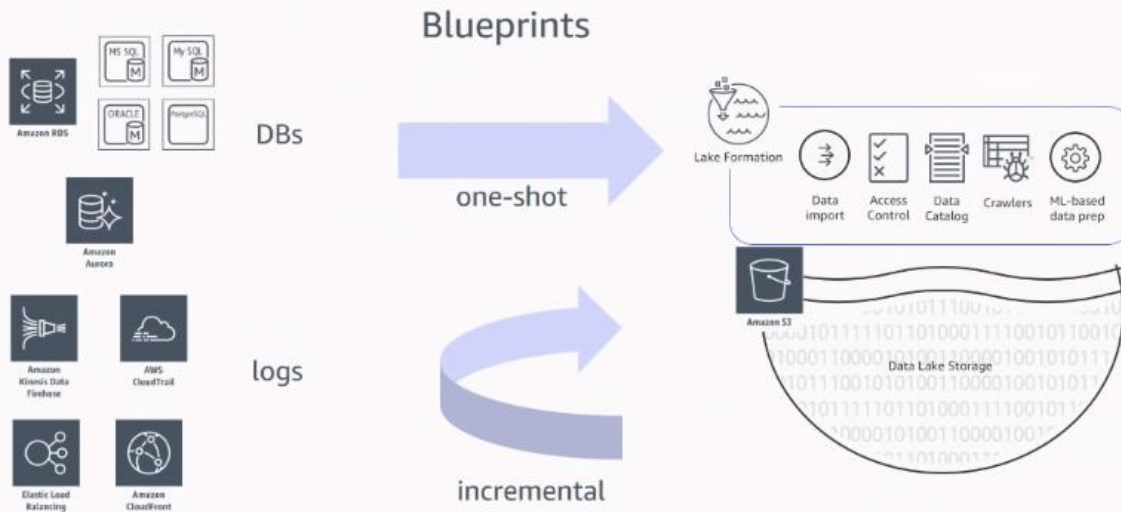


aws
re:Invent

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



Easily load data to your data lake



With blueprints

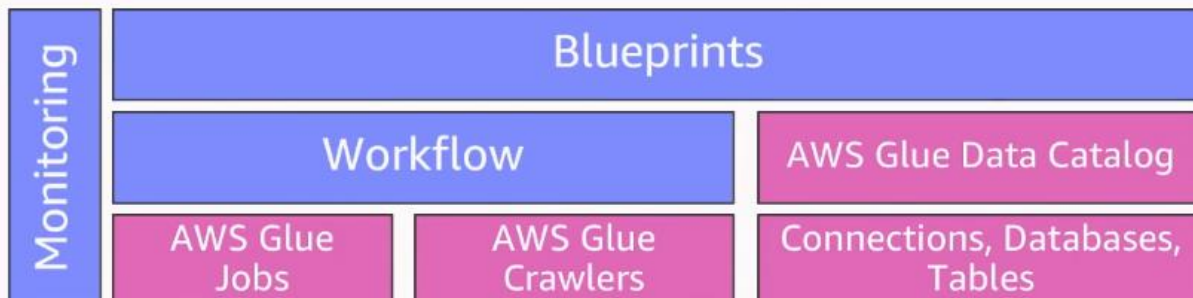
You

1. Point us to the source
2. Tell us the location to load to in your data lake
3. Specify how often you want to load the data

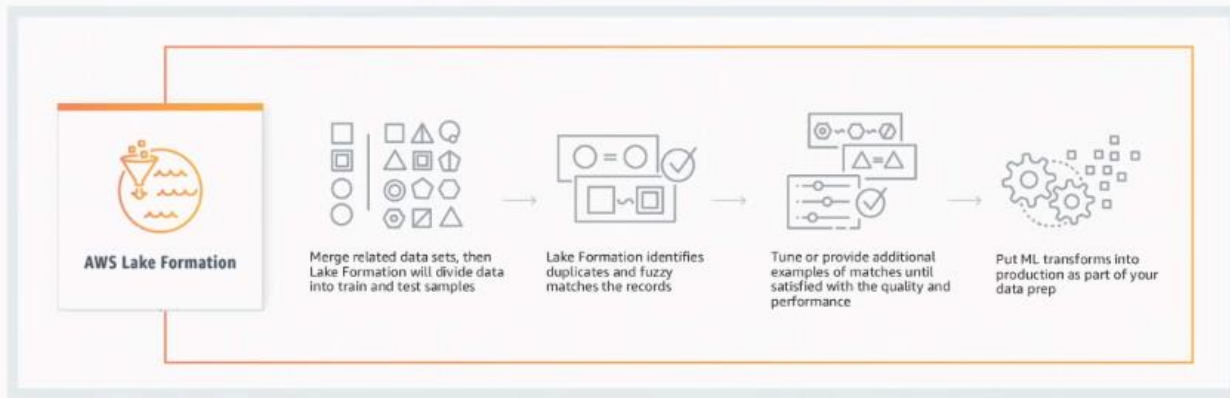
Blueprints

1. Discover the source table(s) schema
2. Automatically convert to the target data format
3. Automatically partition the data based on the partitioning schema
4. Keep track of data that was already processed
5. You can customize any of the above

Blueprints build on AWS Glue

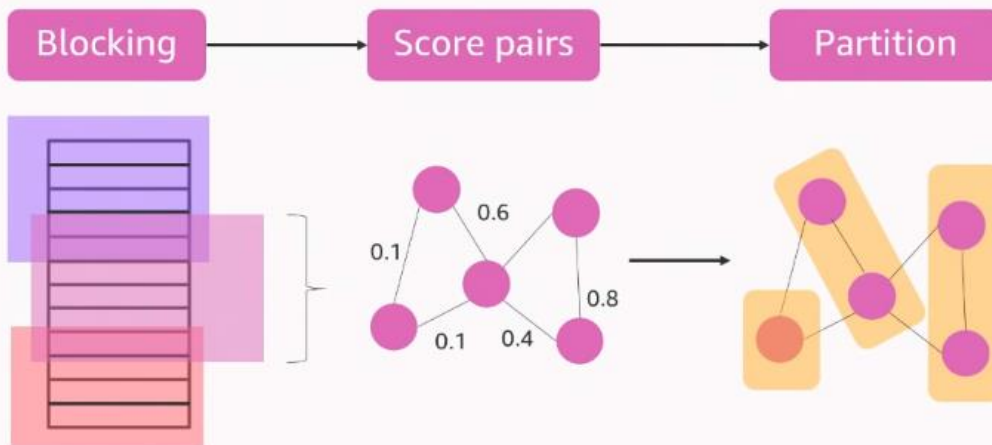


Easily de-duplicate your data with ML transforms



Fuzzy de-duplication – under the hood

Naïve: look at all pairs, N^2 – state-of-the-art:



Fuzzy de-duplication – Innovations

Intersection Dynamic Blocking (VLDB 2008)

parallelizable & performant
blocks on dynamic mix of columns

400M+ rows
7.5B+ candidate pairs
2.5 hours

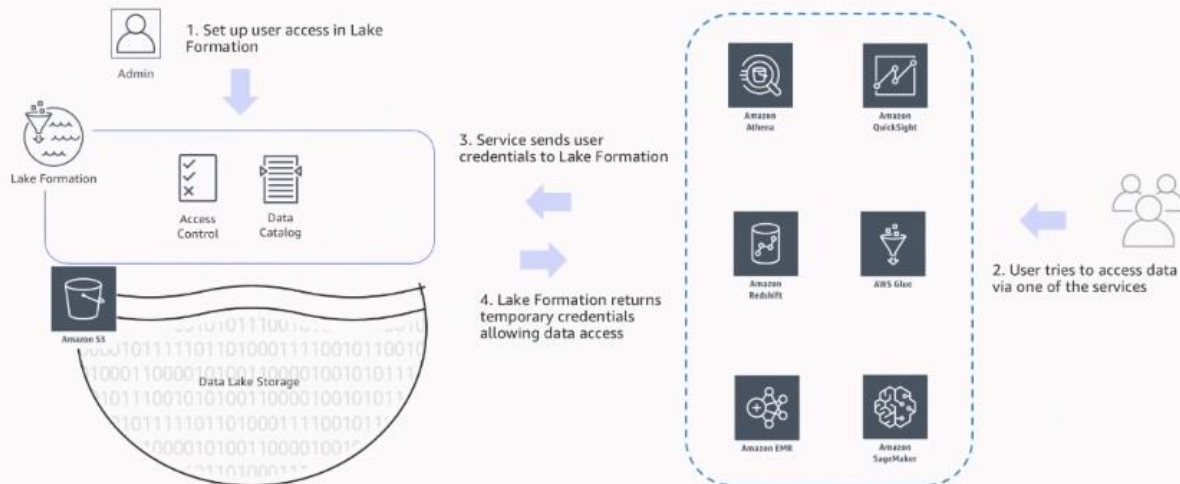
SuperPart

partitions based on customer-provided ground-truth

gives confidence of grouping

effective without tuning knobs

Secure once, access in multiple ways



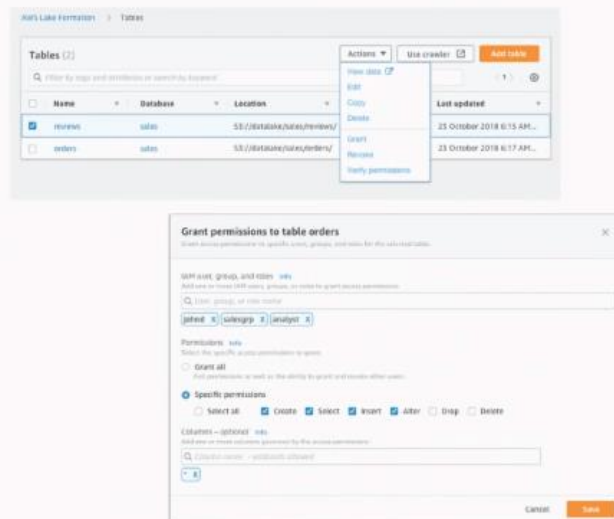
Security permissions in Lake Formation

Control data access with simple grant and revoke permissions

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

Easily view policies granted to a particular user

Audit all data access at one place

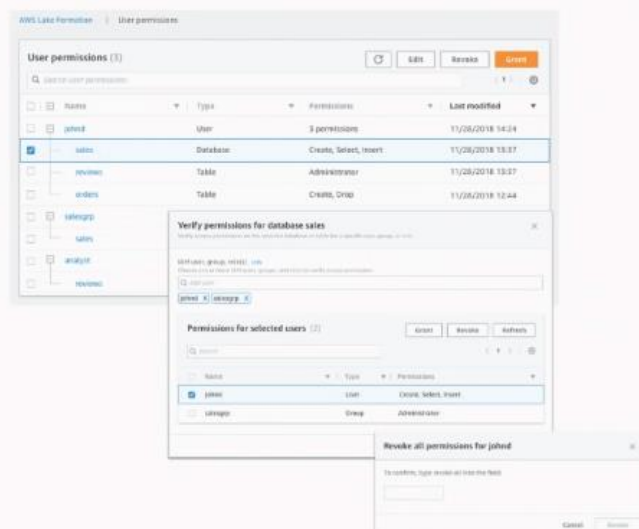


Security permissions in Lake Formation

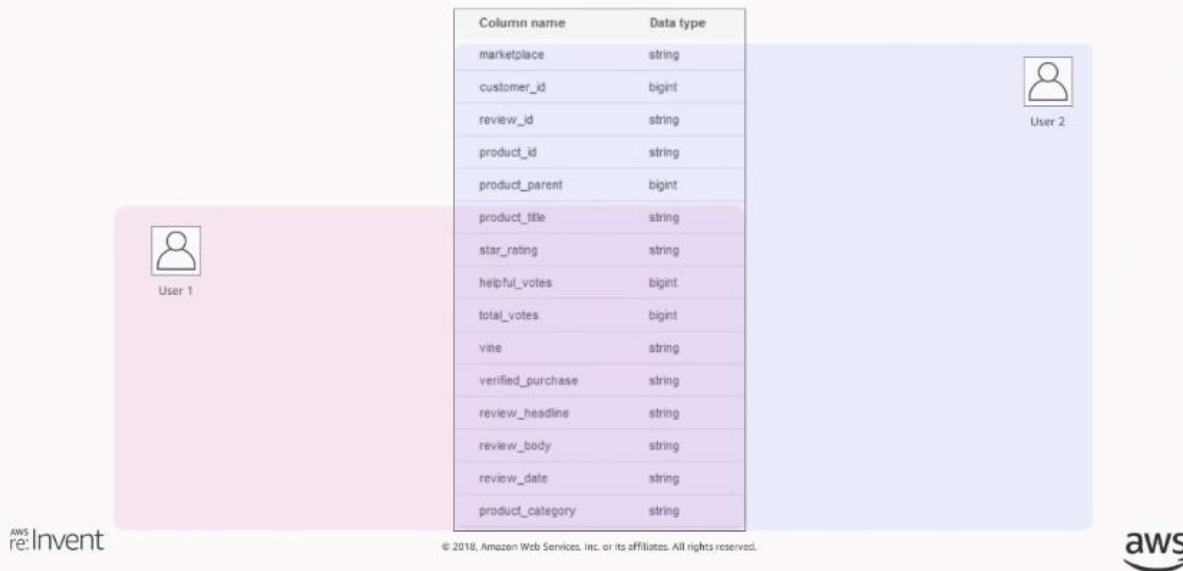
Search and view permissions granted to a user, role, or group in one place

Verify permissions granted to a user

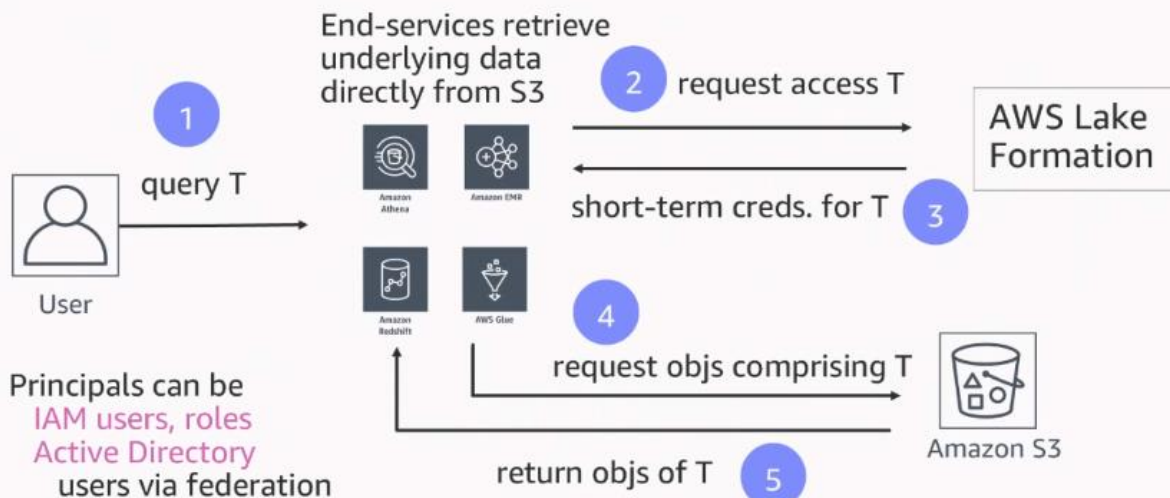
Easily revoke policies for a user



Grant table and column-level permissions



Security – deep dive



Search and collaborate across multiple users

Text-based, faceted search across all metadata

Add attributes like Data owners, stewards, and other as table properties

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

Text-based search and filtering

Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

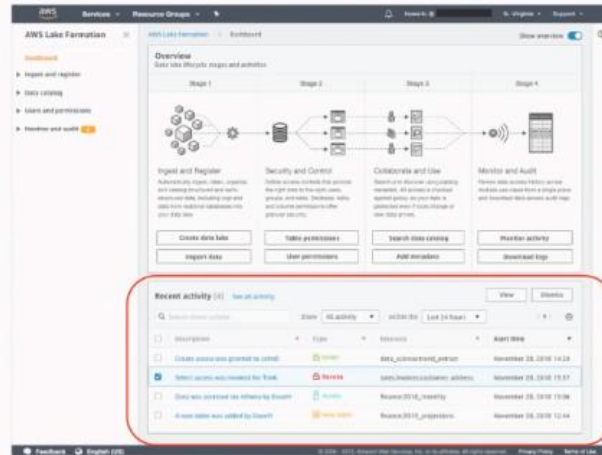
Query data in Amazon Athena

Audit and monitor in real time

See detailed alerts in the console

Download audit logs for further analytics

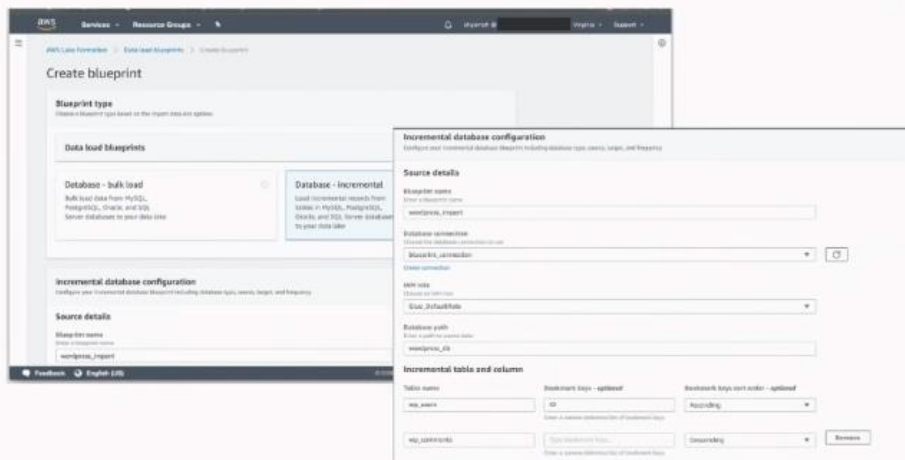
Data ingest and catalog notifications also published to Amazon CloudWatch events



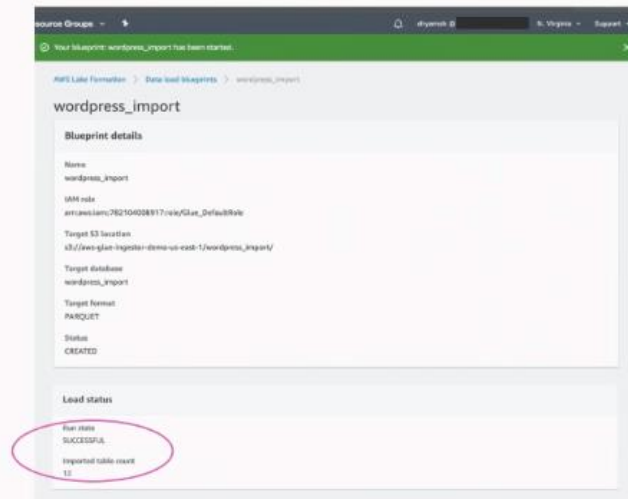
Example: a data lake in 3 easy steps

1. Use blueprints to ingest data
2. Grant permissions to securely share data
3. Query the data (Amazon Athena)

Step 1: Blueprints to ingest data



Monitor the import

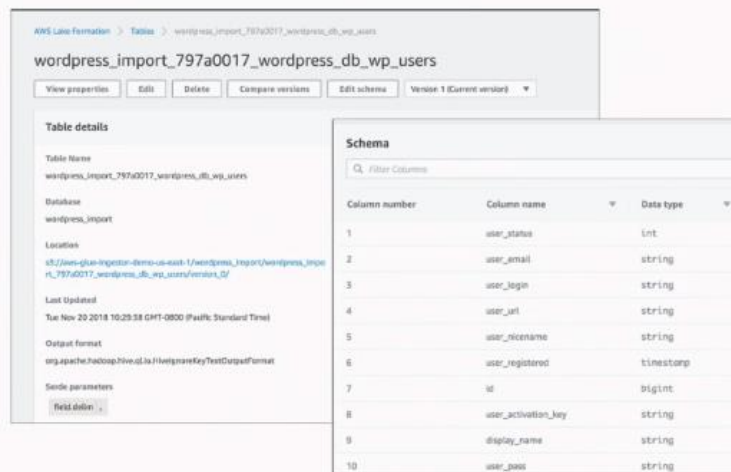


aws
re:Invent

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



Imported data as table in the data lake

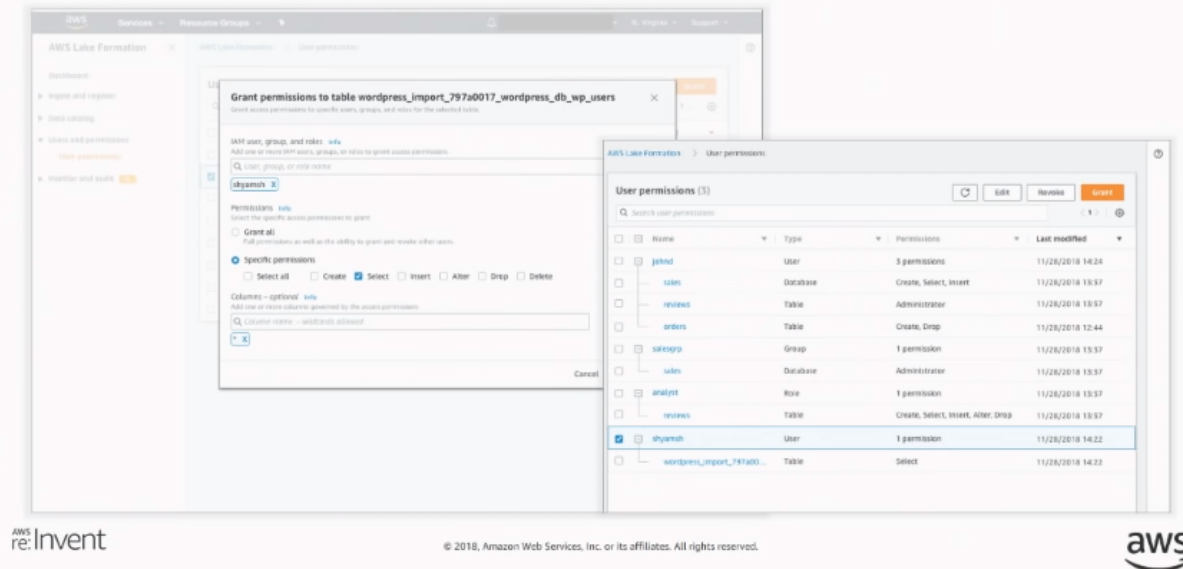


aws
re:Invent

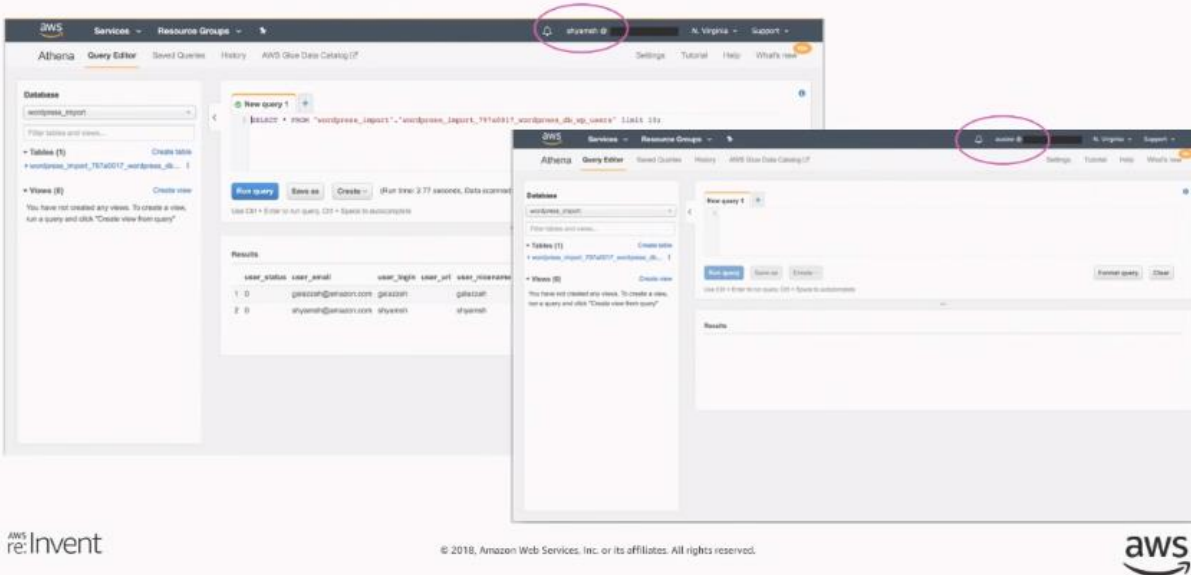
© 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Step 2: Grant permissions to securely share data



Step 3: Run query in Amazon Athena



AWS Lake Formation Pricing

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

Customer interest



"We are very excited about the launch of AWS Lake Formation, which provides a **central point of control** to easily load, clean, secure, and catalog data from thousands of clients to our AWS-based data lake, **dramatically reducing our operational load**. ... Additionally, AWS Lake Formation will be **HIPAA compliant** from day one ..."

Aaron Symanski, CTO, Change Healthcare



"I can't wait for my team to get our hands on AWS Lake Formation. With an enterprise-ready option like Lake Formation, we will be able to **spend more time deriving value from our data** rather than doing the heavy lifting involved in manually setting up and managing our data lake."

Joshua Couch, VP Engineering, Fender Digital

aws
re:Invent

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



Thank you!

Rahul Pathak
rapathak@amazon.com
lakeformation-pm@amazon.com

aws
re:Invent

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

