Introduction to AWS Lake Formation

Prajakta Damle, Principal Product Manager, AWS Lake Formation & AWS Glue

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

Agenda

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

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

There is more data than people think

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

* 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.

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

aws

Data Scientists

Business Users

Analysts

Applications

Secure

Real time

Flexible

Scalable

There are **more people** accessing data

And **more requirements** for making data available

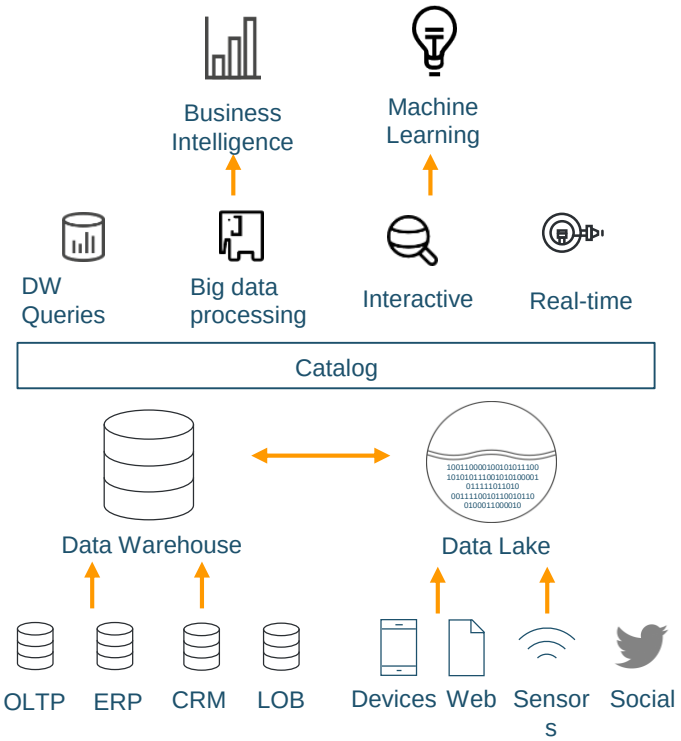
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

aws

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

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

Why data lakes?



Data Lakes provide:

Relational and non-relational data

Scale-out to EBs

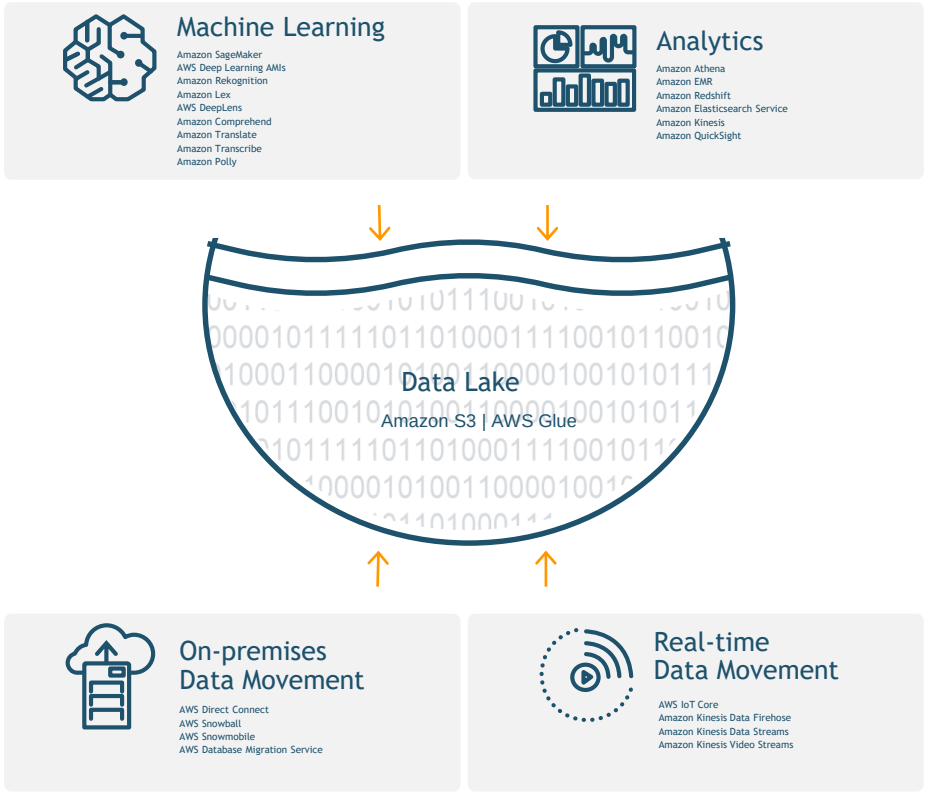
Diverse set of analytics and machine learning tools

Work on data without any data movement

Designed for low cost storage and analytics

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

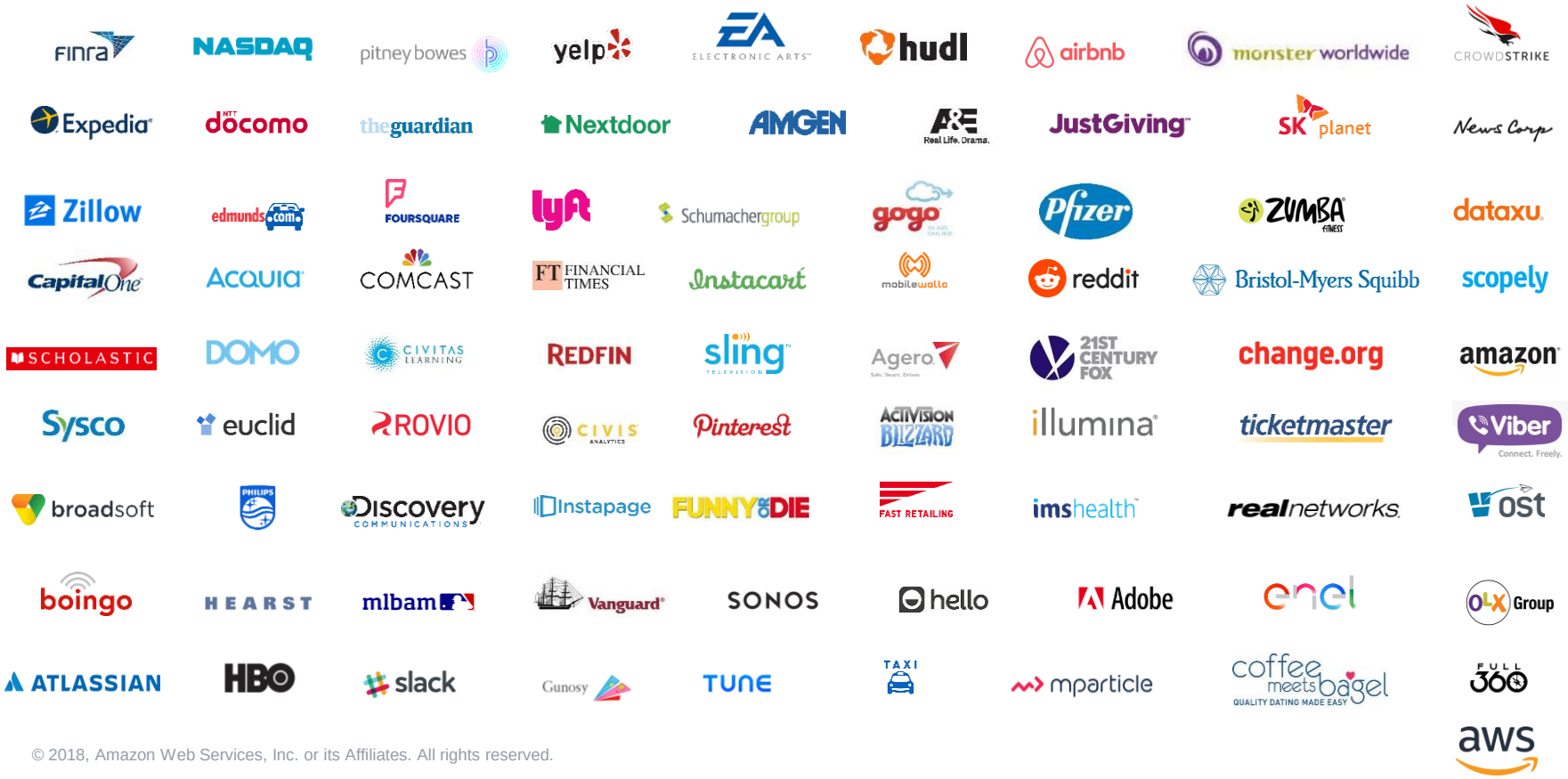
aws



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

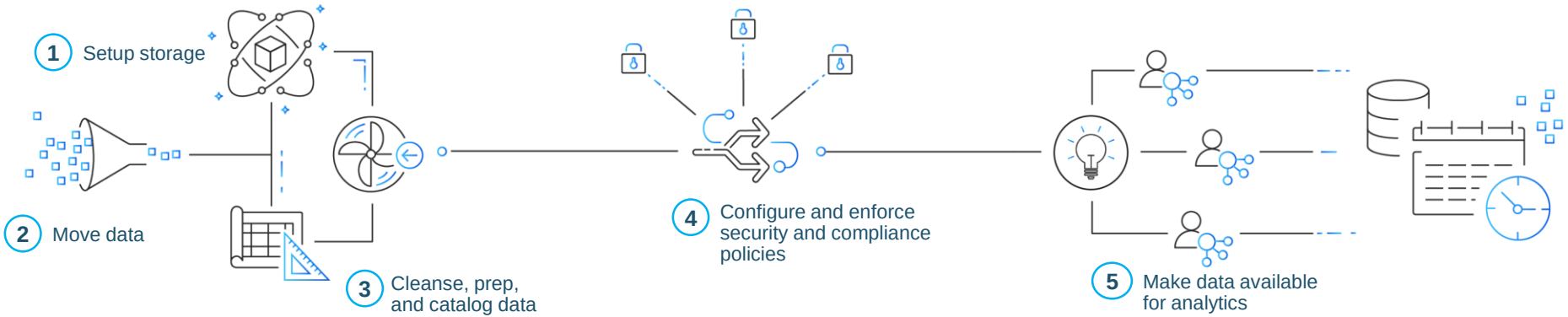
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

More data lakes & analytics on AWS than anywhere else



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

Typical steps of building a data lake

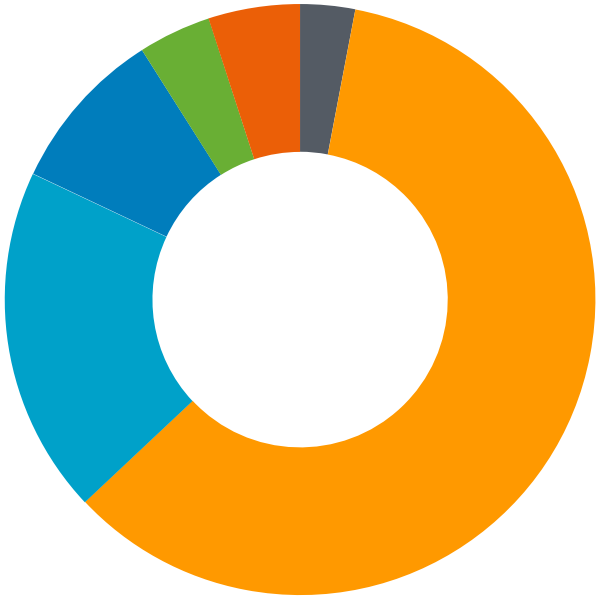


© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Building data lakes can still take months

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

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

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

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

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

AWS Lake Formation

Build a secure data lake in days

Identify, ingest, clean, and transform data

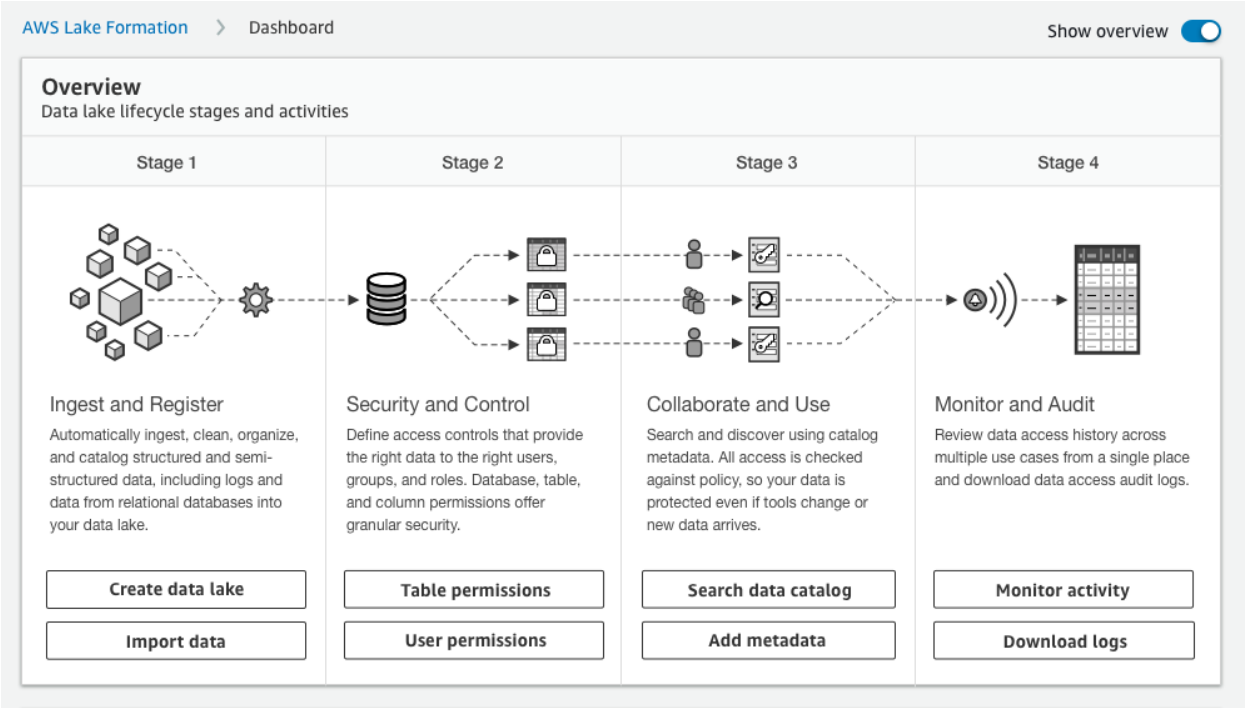
Enforce security policies across multiple services

Gain and manage new insights

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

aws

How it works



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

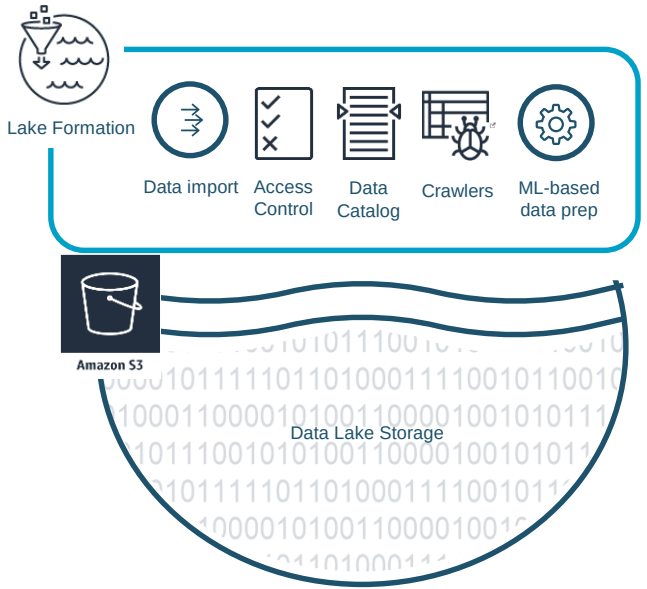
Register existing data or import new

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

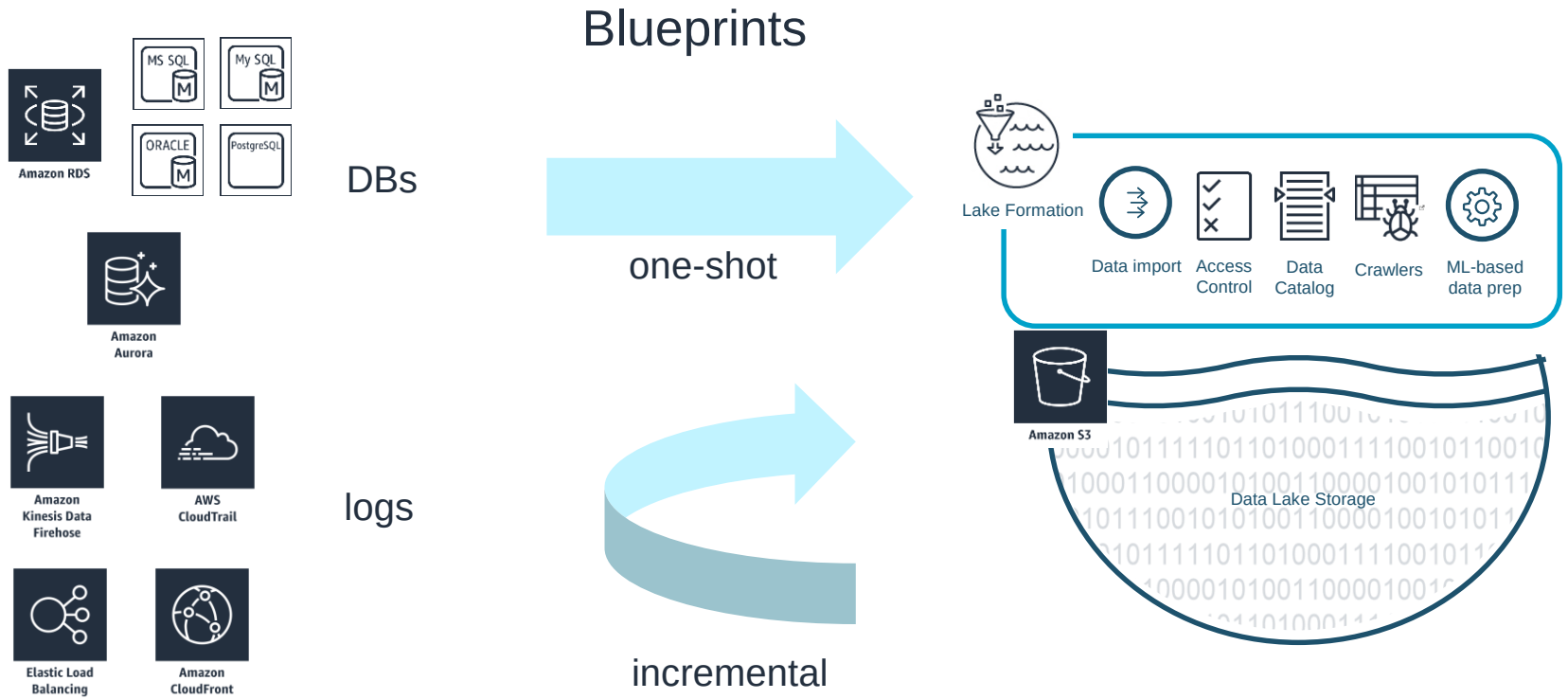
Data is stored in your account. You have direct access to it. No lock-in.



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

aws

Easily load data to your data lake



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

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

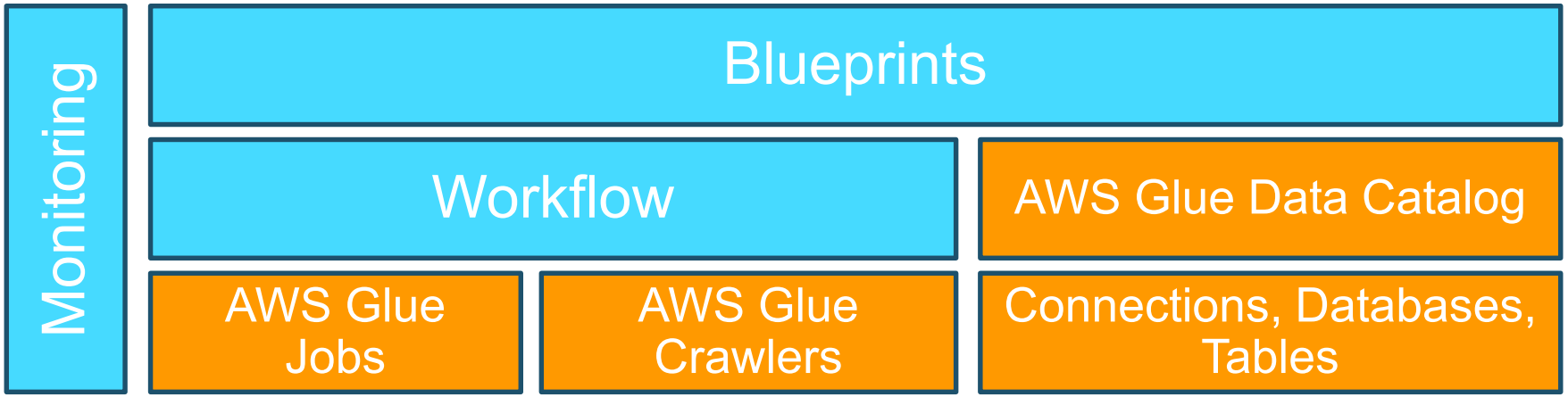
Blueprints

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

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

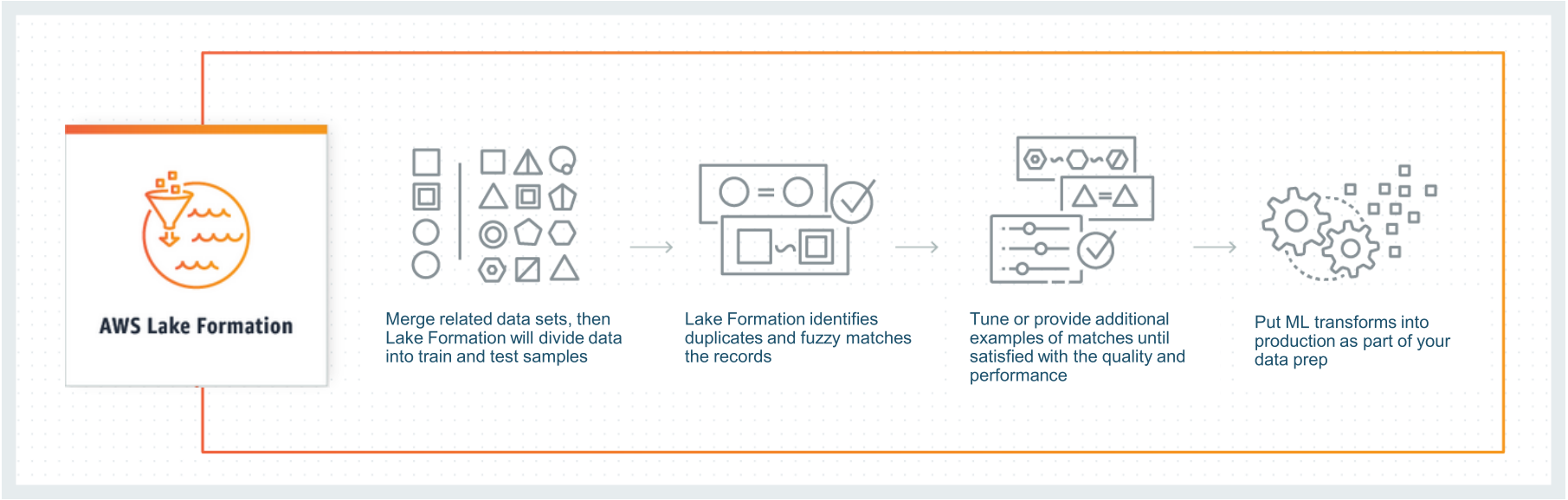
aws

Blueprints build on AWS Glue



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

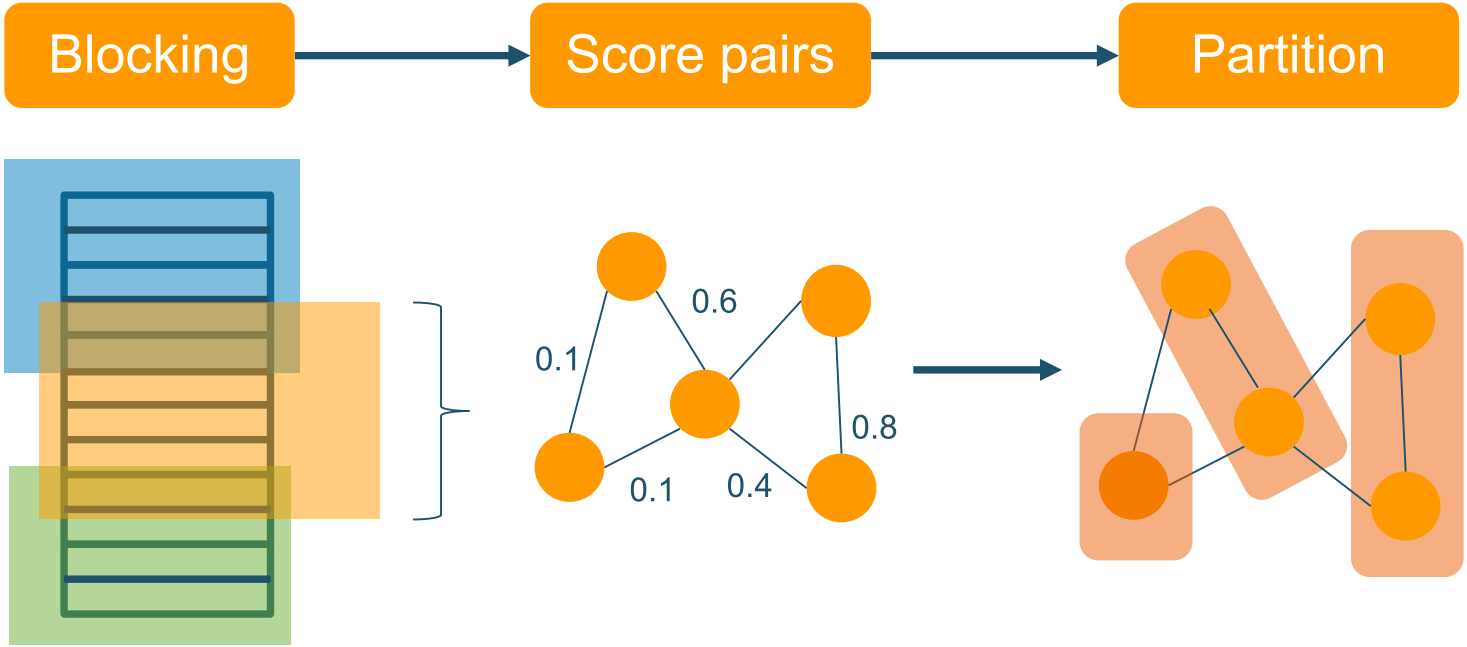
Easily de-duplicate your data with ML transforms



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

Fuzzy de-duplication – under the hood

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



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

aws

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

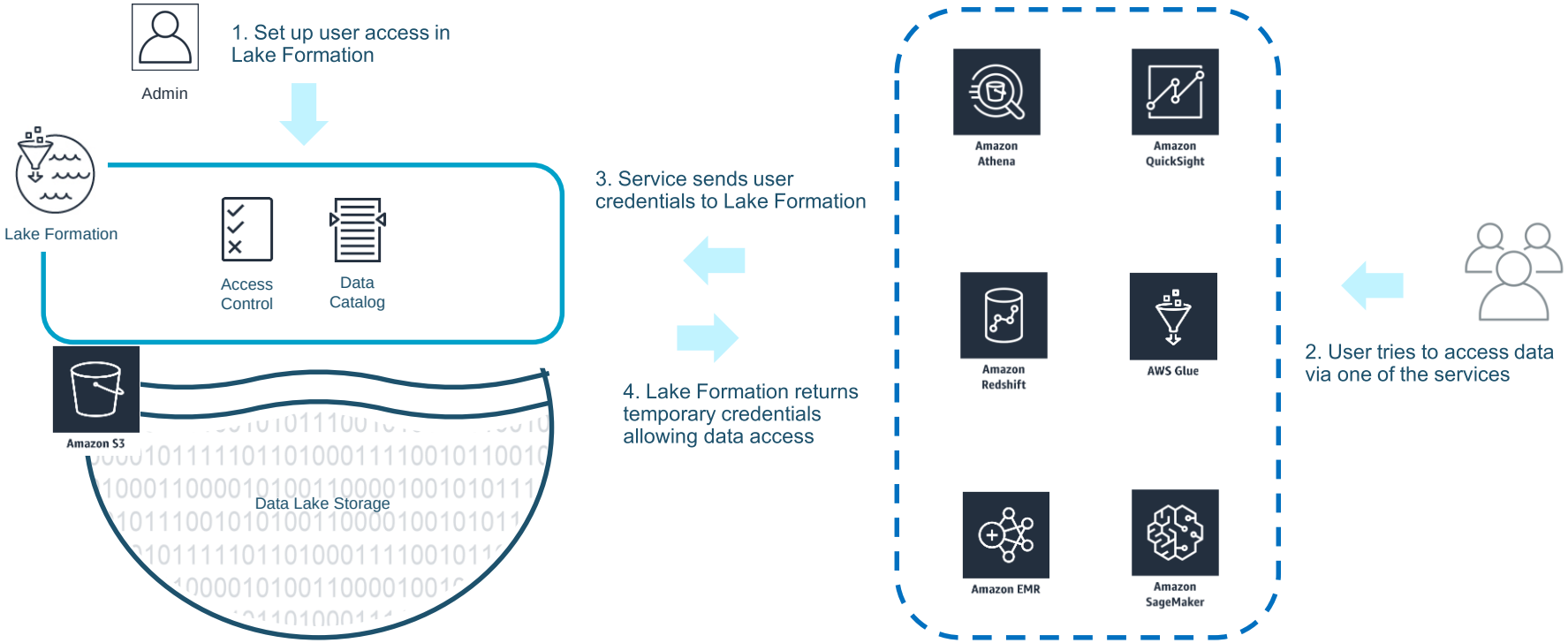
gives confidence of grouping

effective without tuning knobs

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

aws

Secure once, access in multiple ways



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

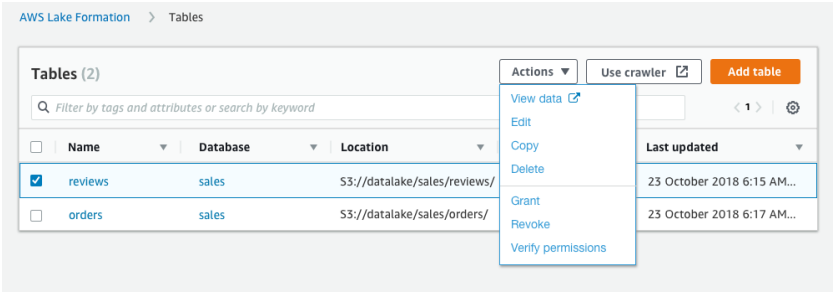
Security permissions in Lake Formation

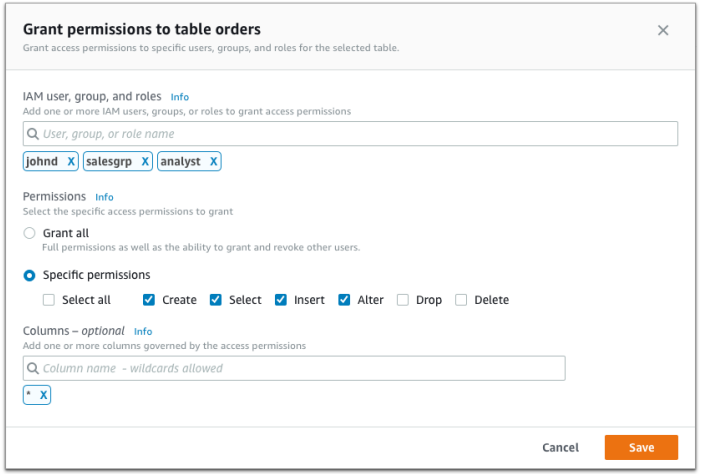
Control data access with simple grant and revoke permissions

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





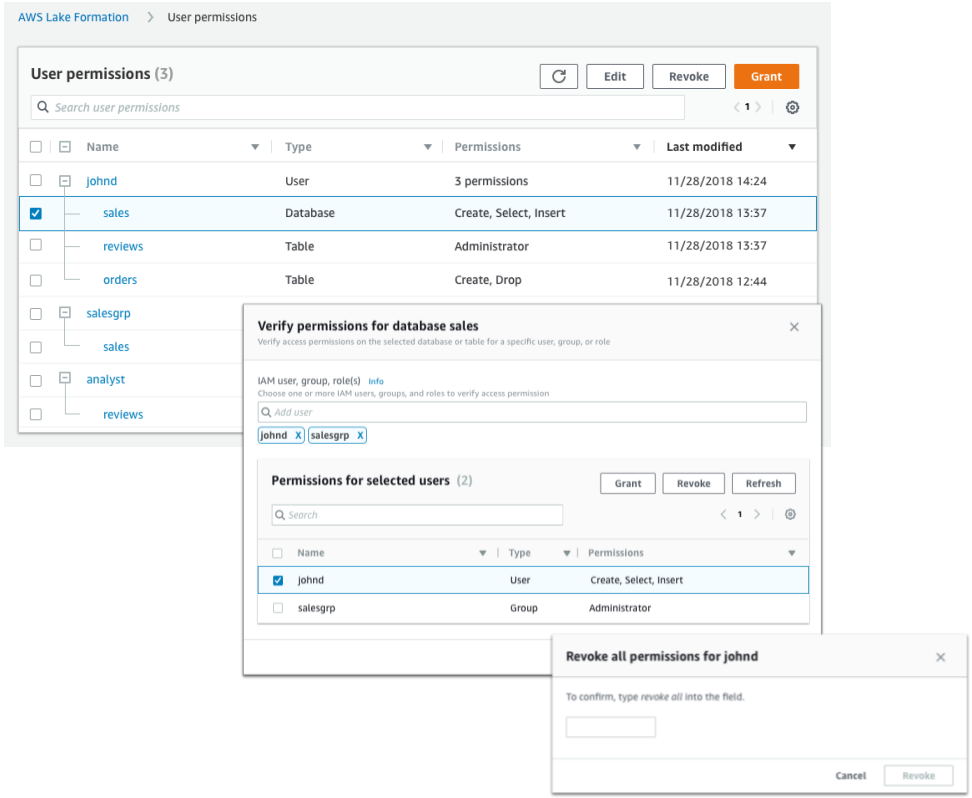
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Security permissions in Lake Formation

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

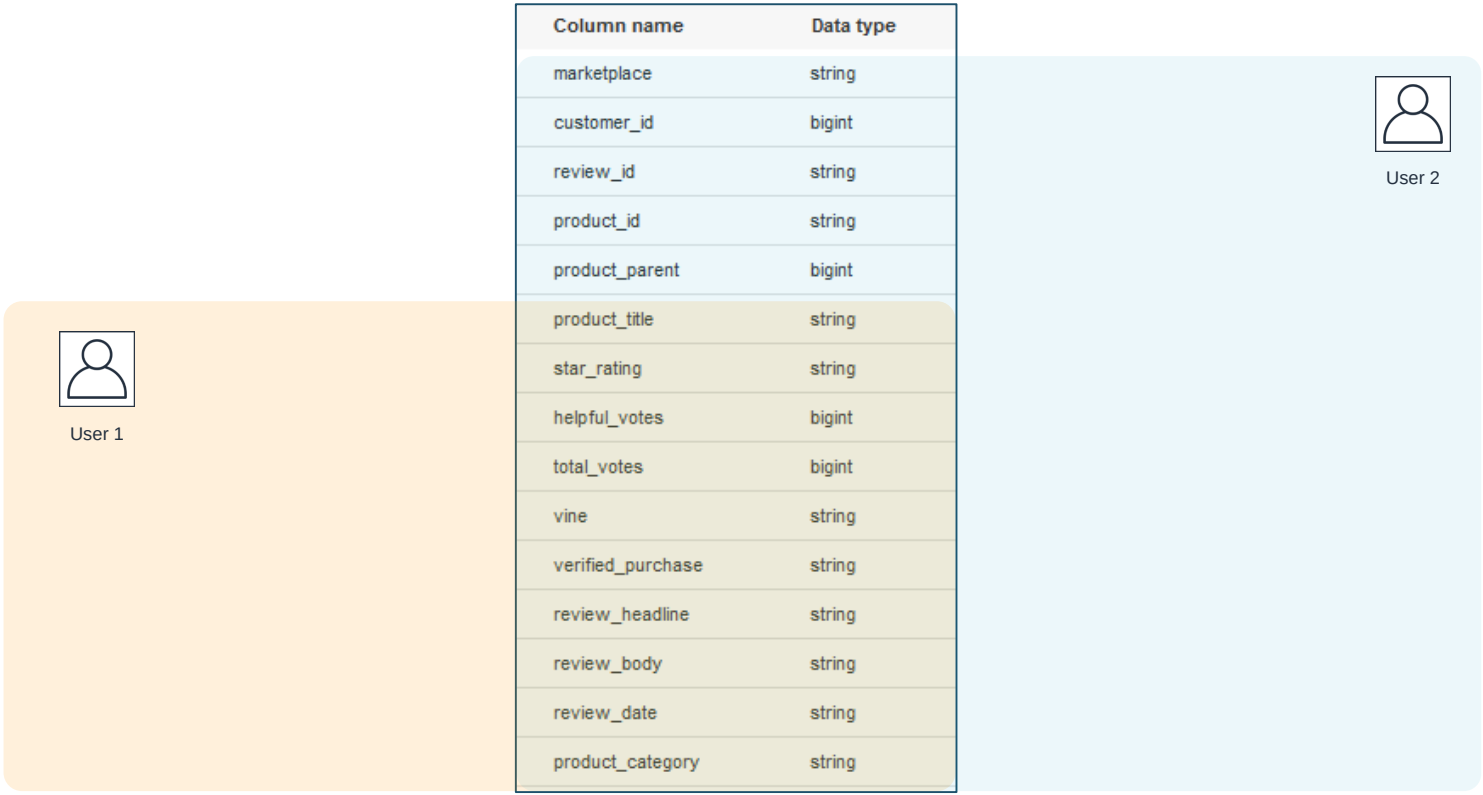
Verify permissions granted to a user

Easily revoke policies for a user



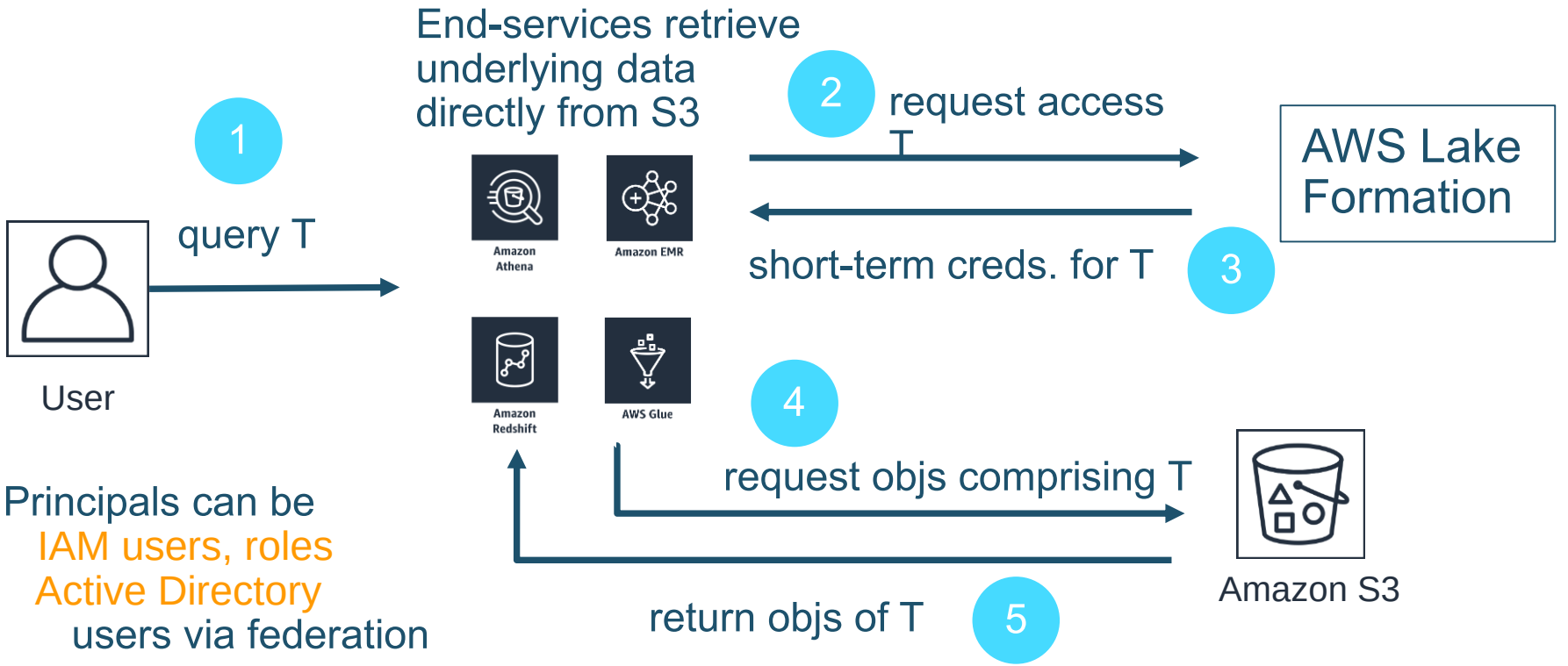
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Grant table and column-level permissions



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

Security – deep dive



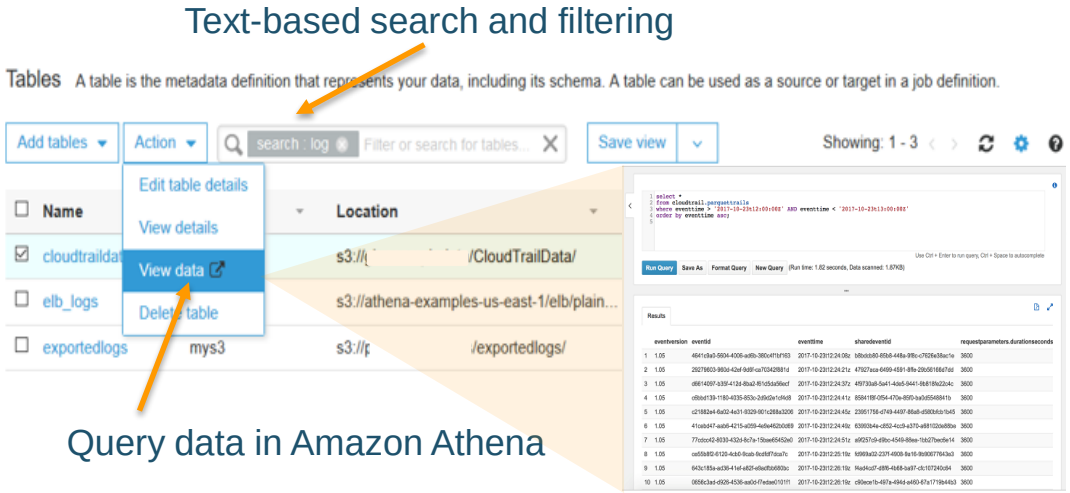
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

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



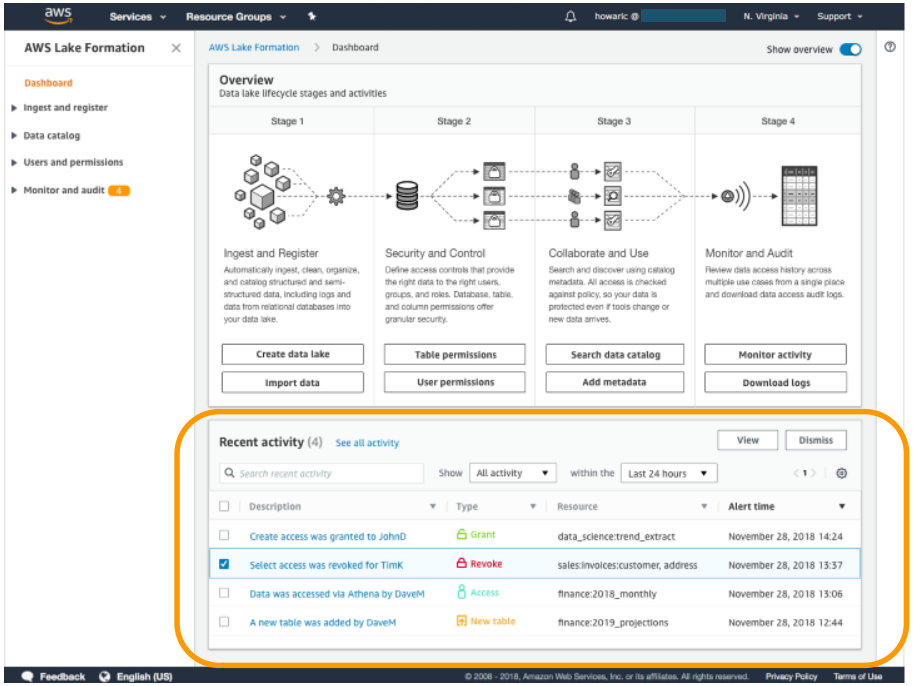
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

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



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

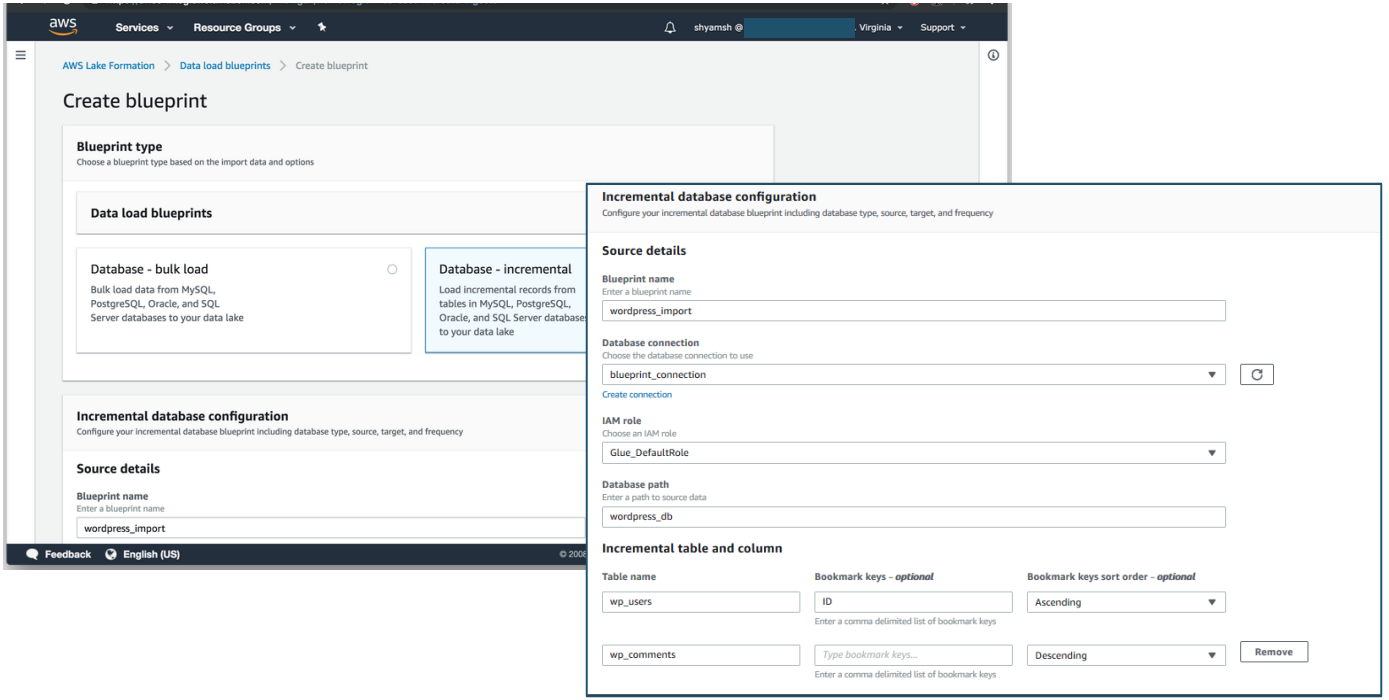
Example: a data lake in 3 easy steps

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

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

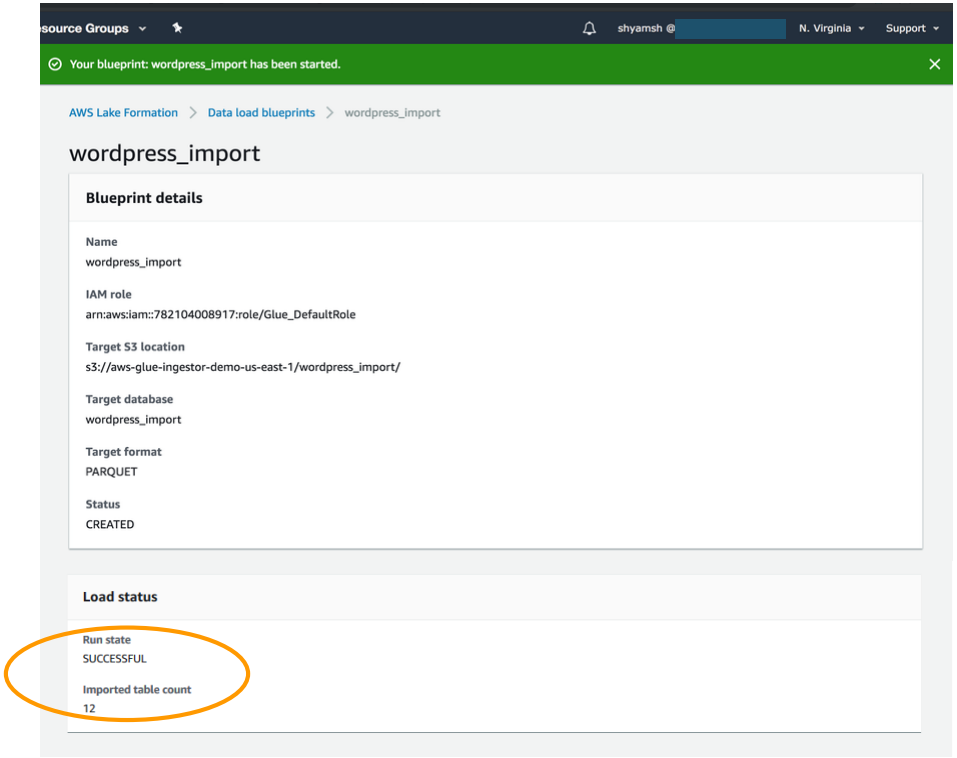
aws

Step 1: Blueprints to ingest data



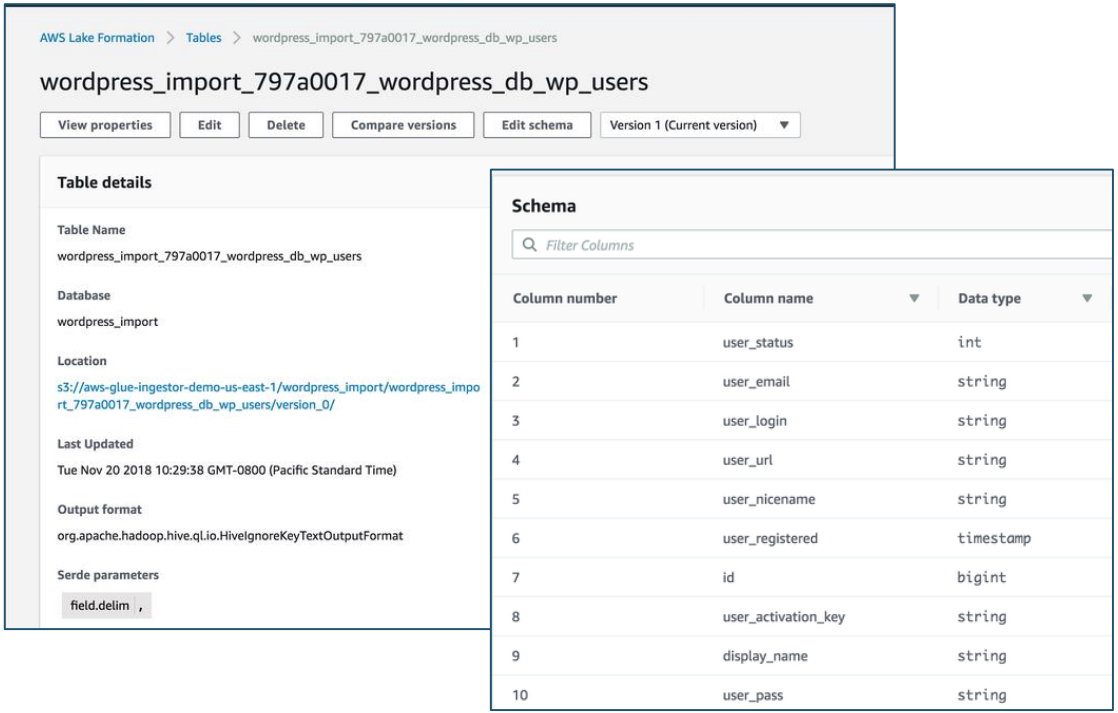
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Monitor the import



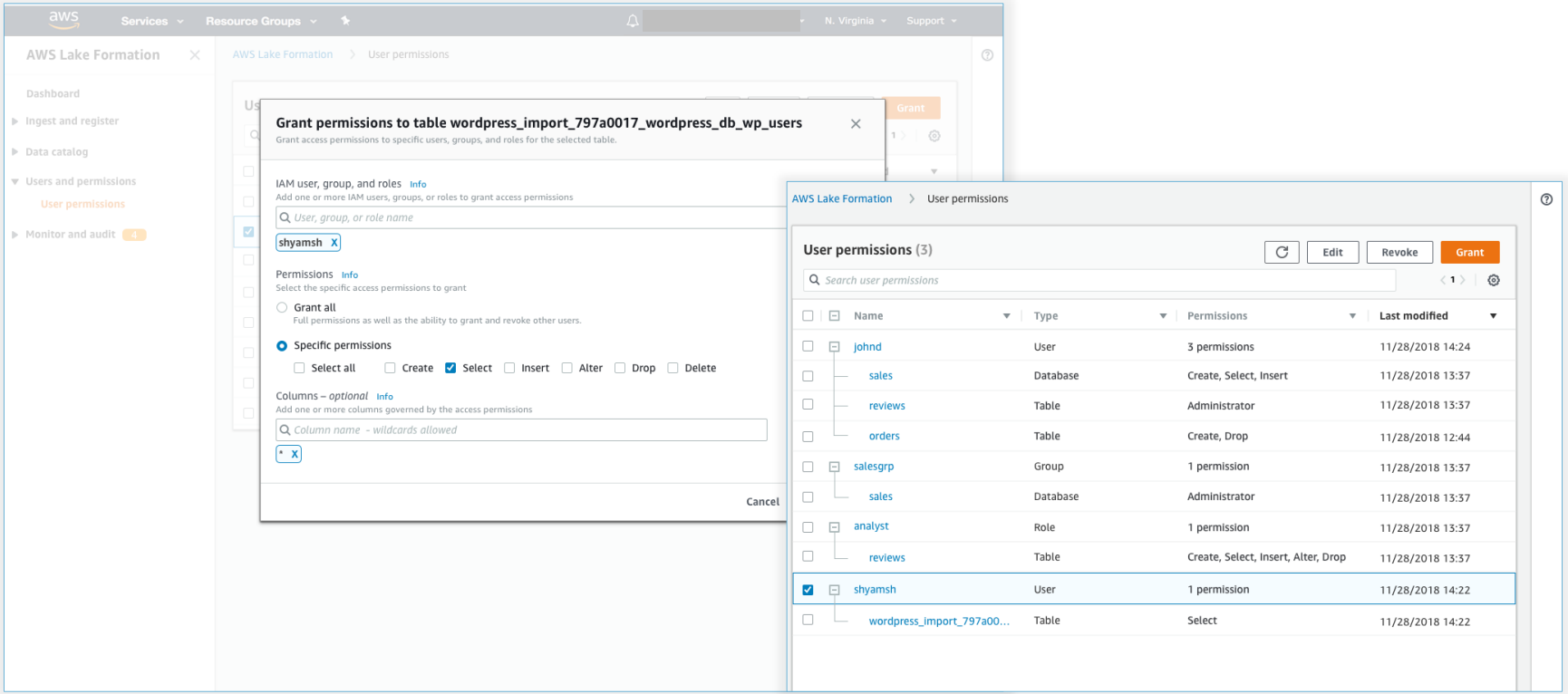
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Imported data as table in the data lake



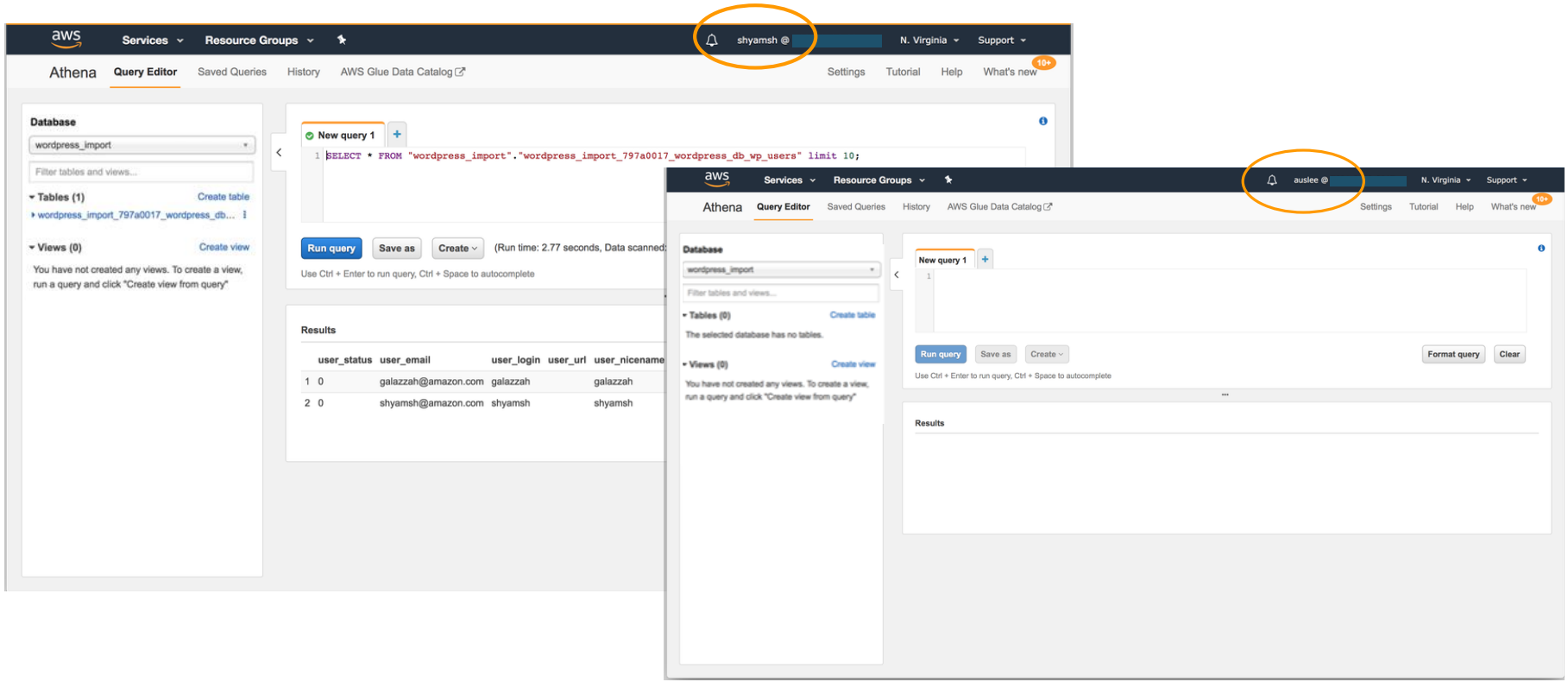
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

Step 2: Grant permissions to securely share data



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

Step 3: Run query in Amazon Athena



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

AWS Lake Formation Pricing

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

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

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

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

aws

Thank you!

Join the preview: https://pages.awscloud.com/lake-formation-preview.html

lakeformation-pm@amazon.com

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