

Deep Dive into AWS Lake Formation

Roy Hasson – Principal BDM – Analytics and Data Lakes

Agenda

Why data lakes?

What is hard about building data lakes?

Why AWS Lake Formation?

What is Lake Formation?

How it works - Demo

Decision making used to...

...revolve around the
Enterprise Data Warehouse



OLTP



LOB



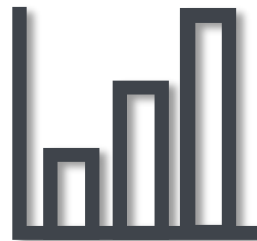
ERP



CRM



Enterprise Data Warehouse



Business Intelligence

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, April 2017.

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

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

Why data lakes?

What is hard about building data lakes?

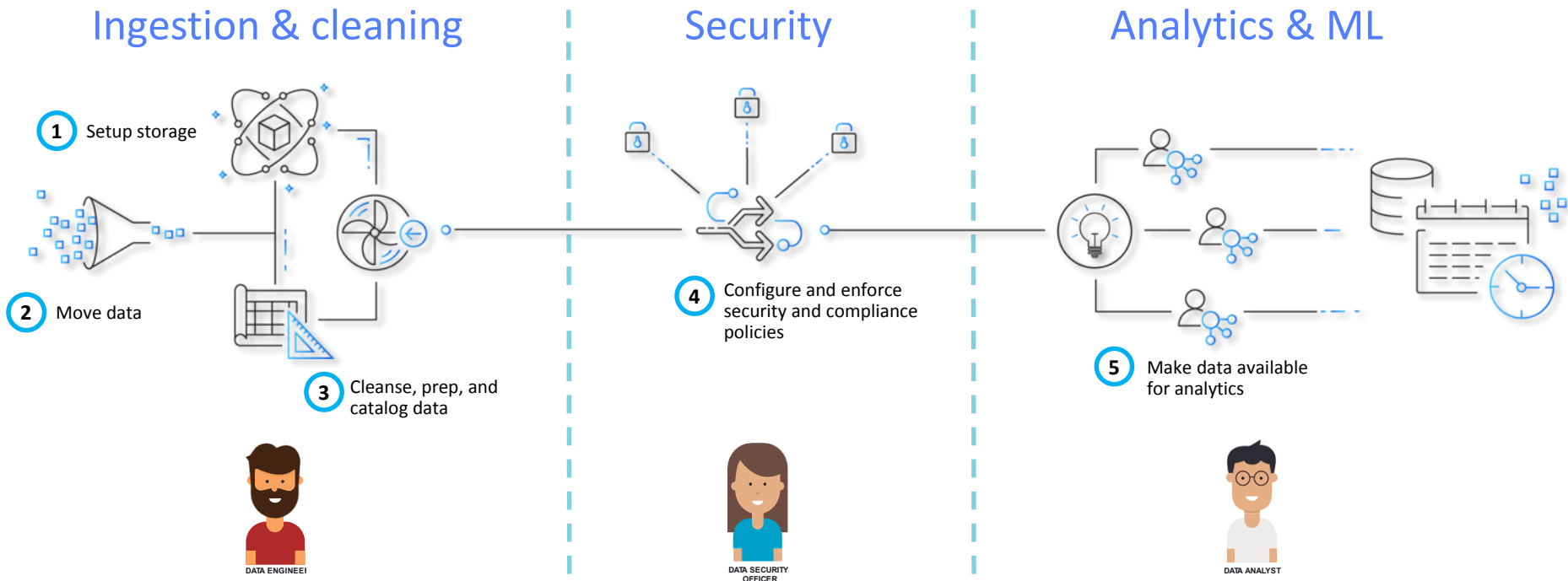
Why Lake Formation for data lakes?

What is Lake Formation?

How it works!

Manually building secure data lakes is **hard**

Typical steps of building a data lake



Sample of steps required to **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

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

Agenda

Why data lakes?

What is hard about building data lakes?

Why Lake Formation for data lakes?

What is Lake Formation?

How it works!

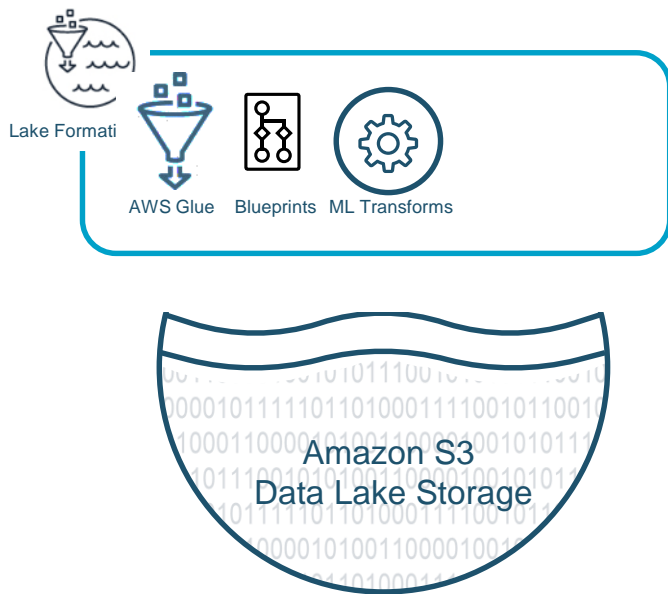
Lake Formation lets you
build secure data lakes in **days**

Built on Amazon S3 as a robust data lake infrastructure



Cost effective, durable storage with global replication capabilities

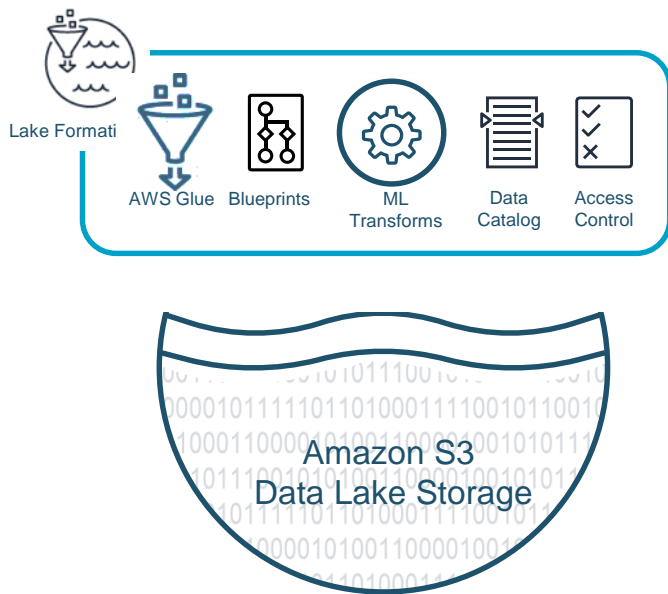
Automates manual, repetitive, low value tasks



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

Cost effective, durable storage with global replication capabilities

Provides a central locus of control

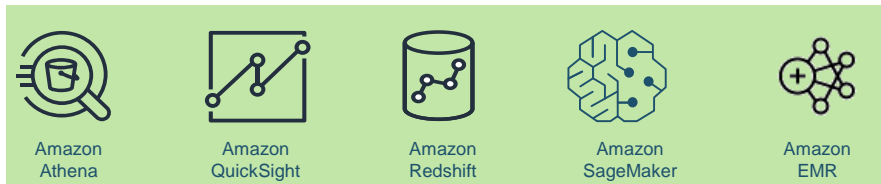


Centralized management of **fine grained permissions** empower security officers

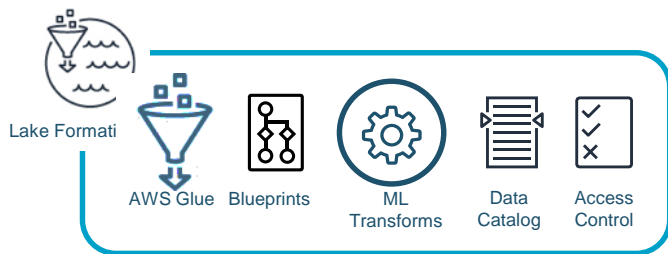
Simplified **ingest & cleaning** enables data engineers to build faster

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 & cleaning** enables data engineers to build faster



Cost effective, durable storage with global replication capabilities

Agenda

Why data lakes?

What is hard about building data lakes?

Why Lake Formation for data lakes?

What is Lake Formation?

How it works!

Tools that enable data engineers, security officers
& data analysts
to build, manage and use your data lake

Building data lakes with Lake Formation

Ingestion & cleaning



AWS Glue

Serverless Spark

Blueprints

ML Transforms

Security



Data catalog

Centralized permissions

Real time monitoring

Auditing

Analytics & ML



Comprehensive portfolio
of integrated tools



Redshift



Glue



EMR



Athena

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

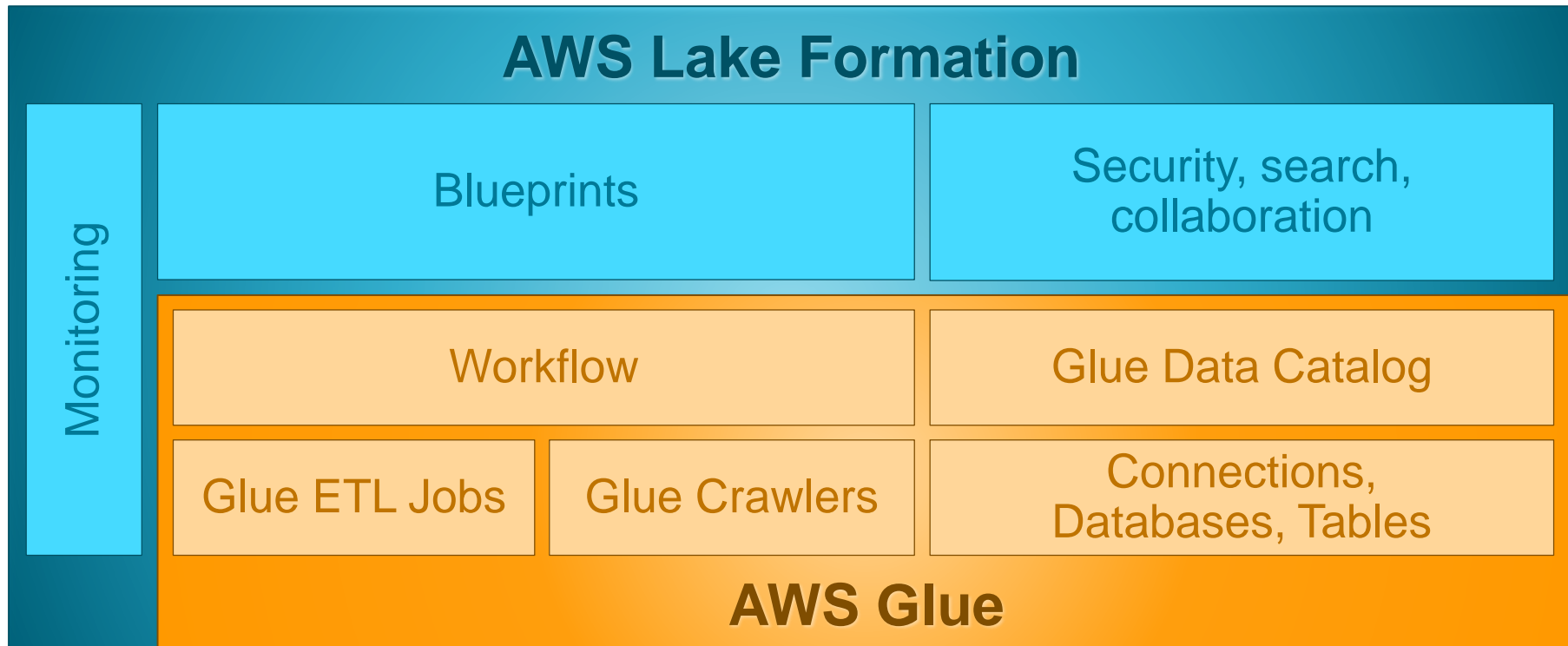


Orchestration

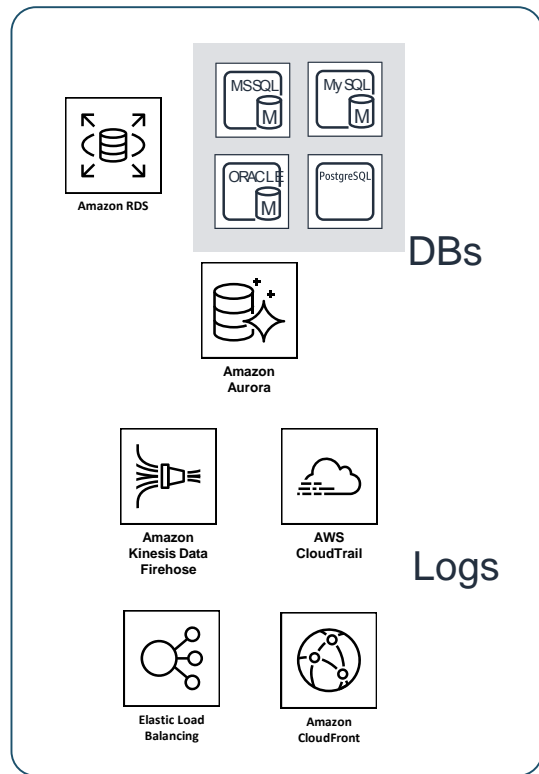
Deploy

- Flexible workflows
- Monitoring and alerting
- External integrations

AWS Lake Formation is fully integrated w/ AWS Glue



Easily load data into your data lake w/ blueprints



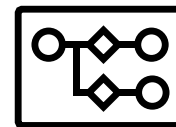
Prebuilt templates to serve common ingestion use cases

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**

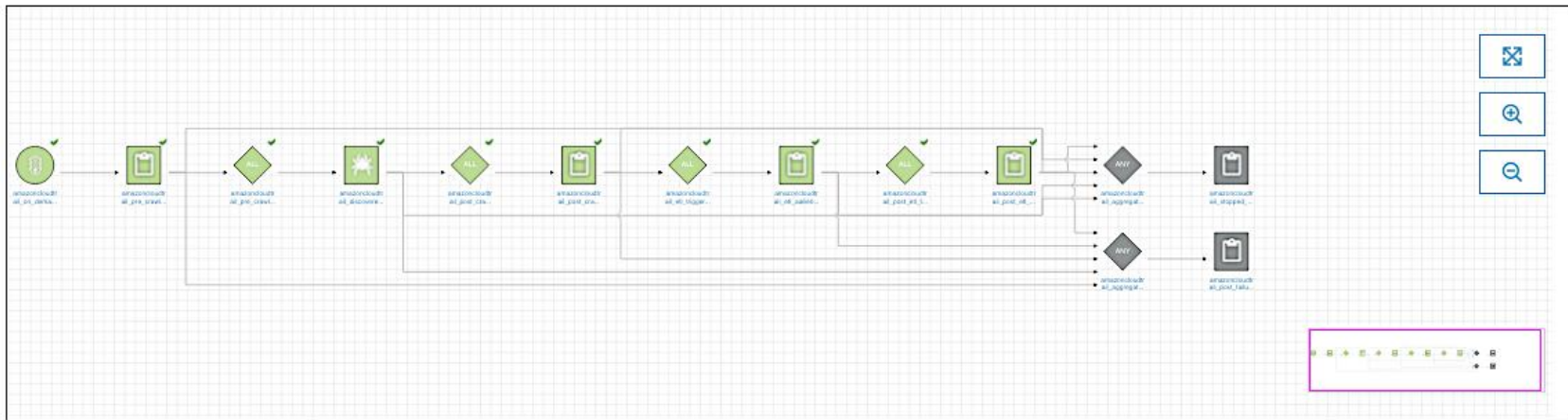


AWS Glue Workflows

Blueprints create AWS Glue workflows

Graph

Legend:  Completed  Running  Warning  Error  Deleting



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

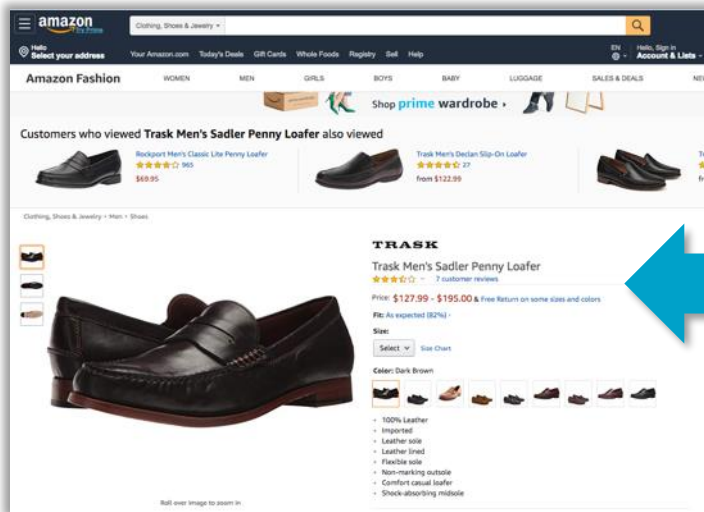
Leverage machine learning to solve hard problems

Deduplication

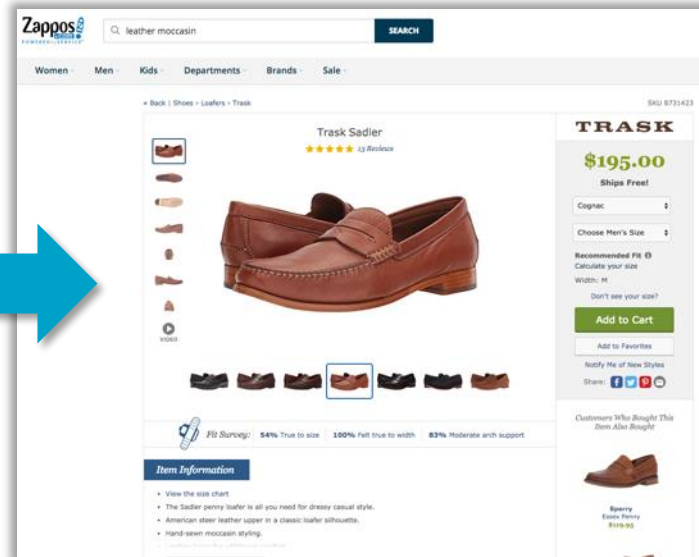
Transforming a dataset that has multiple rows referring to the *same actual thing* into a dataset where no two rows refer to the *same actual thing*

Record matching

Finding the relationships between multiple datasets, even when those datasets do not share an identifier (or when their identifier is unreliable)



ML FindMatches



Securing data lakes with Lake Formation

Ingestion & cleaning



Serverless Spark

AWS Glue

Glue ML transformations

Blueprints

Security



Data catalog

Centralized permissions

Real time monitoring

Integrated auditing

Analytics & ML



Comprehensive portfolio
of integrated tools



Redshift



Glue



EMR



Athena

Data Catalog & Permissions

Permissions are set on data catalog objects

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



Choice of using the **Glue** or the **Lake Formation** permissions system

For backwards compatibility, the default settings enable the **Glue** permissions system

Existing **Glue** crawlers, jobs, triggers and workflows will not change



Crawlers



ETL Jobs



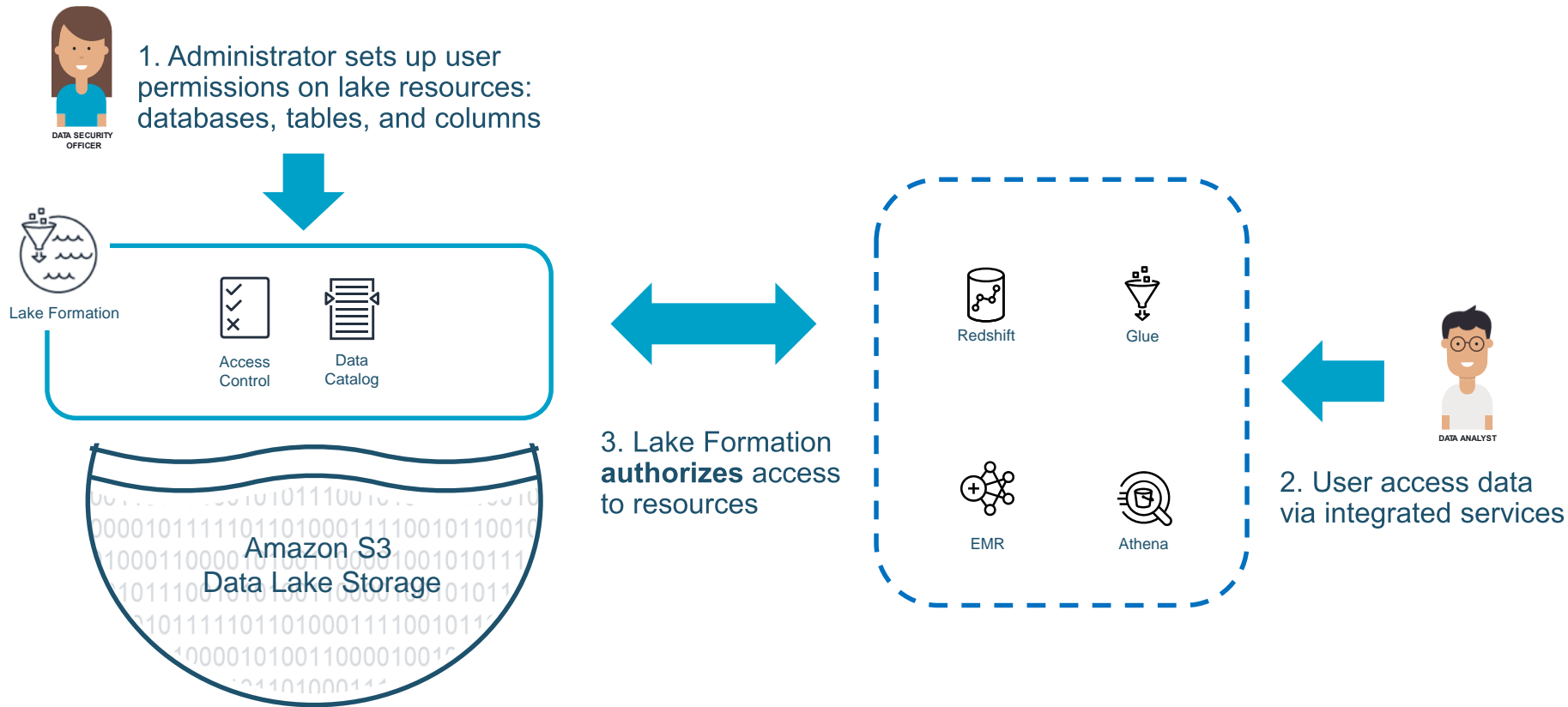
Workflows

Existing access to **Glue** resources will still be governed by **IAM & S3 policies**



Access
Control

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

		Column name	Data type
 User 1	 User 2	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
 User 1	 User 2	verified_purchase	string
		review_headline	string
		review_body	string
		review_date	string
		product_category	string

Upgrading to the Lake Formation permissions model

Not using the Glue Catalog?

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

Using the Glue 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

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**

The screenshot displays the AWS Lake Formation console interface. On the left, a navigation sidebar includes links to Dashboard, Data catalog, Databases, Tables, Settings, Register and ingest, Data lake locations, Blueprints, Crawlers, Jobs, Permissions, Admins and database creators, Data permissions, and Data locations. The main content area shows a list of tables under the 'amazoncloudtrail' database. A search bar at the top of the table list contains the text 'Database : amazoncloudtrail'. A table with columns 'Name', 'Data', 'Location', and 'Classification' is visible, listing various cloudtrail tables. A 'View data' button is highlighted in the context menu for the first table. Overlaid on the right is a screenshot of the Amazon Athena console, showing a SQL query and its results. The query filters for events from a specific time range. The results table has columns: eventversion, eventid, eventtime, sharedeventid, and requestparameters.durationseconds. The results show 10 rows of event data.

Text-based search and filtering

Query data in Amazon Athena

```
1 select *
2 from cloudtrail.parcetrails
3 where eventtime > '2017-10-23T12:00:00Z' AND eventtime < '2017-10-23T13:00:00Z'
4 order by eventtime asc
5
```

	eventversion	eventid	eventtime	sharedeventid	requestparameters.durationseconds
1	1.05	4841c8a0-6004-4006-a0b5-380c41f1f163	2017-10-23T12:24:00z	b5b3d800-89b8-448a-08bc-c7826a35ac1e	3600
2	1.05	29279603-9606-42af-9a8f-ca703a20881d	2017-10-23T12:24:21z	47927aca-6499-4591-0ffe-29056168a70d	3600
3	1.05	d8814c97-a359-4126-8ba2-461d5da56efc	2017-10-23T12:24:37z	409730a8-5a41-40a5-9441-908180a2204c	3600
4	1.05	c8b0d139-1180-4935-8530-26f02e1d4f08	2017-10-23T12:24:41z	8584118f-0594-470a-8500-ba05548041b	3600
5	1.05	c21882e4-6a02-4e31-8329-861c288a3206	2017-10-23T12:24:45z	23951756-d749-4497-86a8-d5802fcb1b45	3600
6	1.05	410ebd47-aab6-4215-a059-4e8e4620a889	2017-10-23T12:24:49z	63093b4e-c852-4cc5-a370-a581f20e08ba	3600
7	1.05	77d0dc42-8030-432d-6c7e-150ae65452e0	2017-10-23T12:24:51z	a92527c9-49bc-4549-86ea-1bb27bce0e14	3600
8	1.05	ce050892-6123-4c80-9a9b-9cdf0f7cca7c	2017-10-23T12:25:19z	bd90ba02-237f-4908-ba16-9b80677643e3	3600
9	1.05	643c185a-ac34-41af-a82f-8a0f8a66803c	2017-10-23T12:26:19z	Ma4d0d7-d086-4068-ba07-c1c107240d64	3600
10	1.05	0856c3ad-6928-4536-aa05-f7edae01011f	2017-10-23T12:26:19z	c80ccc1b-494a-496a-a600-87a1719a4463	3600

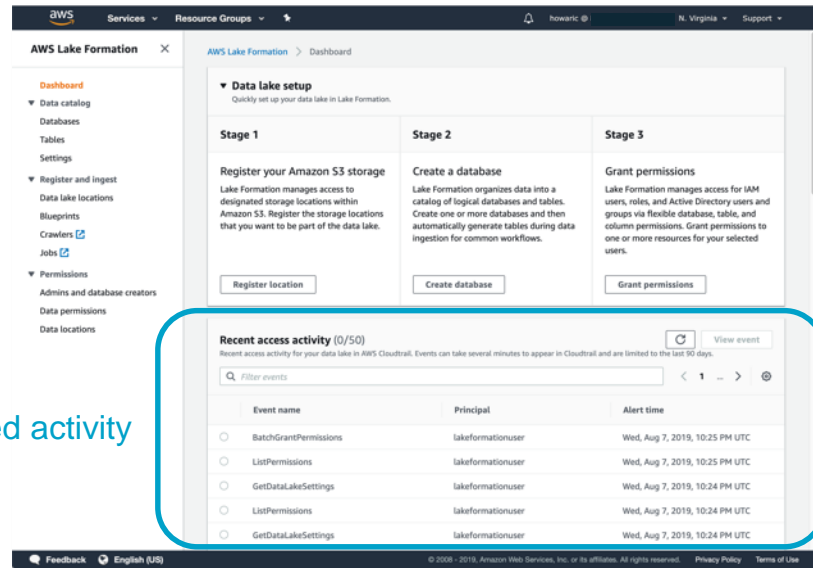
Audit and monitor in real time

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



Serverless Spark

AWS Glue

Glue ML transformations

Blueprints

Security



Data catalog

Centralized permissions

Real time monitoring

Auditing

Analytics & ML



Comprehensive portfolio
of integrated tools



Redshift



Glue



EMR



Athena

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**

The screenshot displays the AWS Athena Query Editor interface. On the left, the 'Catalog' pane shows the 'Lake formation' database and the 'amazoncloudtrail' database. Under 'Tables (1)', 'amazoncloudtrail_cloudtrail' is listed as a partitioned table. The main editor area shows a SQL query: `SELECT * FROM "amazoncloudtrail"."amazoncloudtrail_cloudtrail" limit 10;`. Below the query, buttons for 'Run query', 'Save as', and 'Create' are visible, along with a status bar indicating a run time of 3.02 seconds and 101.28 KB of data scanned. The 'Results' pane at the bottom shows a table with two columns: 'eventversion' and 'useridentity'. The results are paginated, showing 10 rows of data.

	eventversion	useridentity
1	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts:785789292865:assumed-role/AwsE
2	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:Meta311", "arn": "arn:aws:sts:785789292865:assumed-role/AwsSt
3	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts:785789292865:assumed-role/AwsE
4	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts:785789292865:assumed-role/AwsE
5	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts:785789292865:assumed-role/AwsE
6	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:Meta311", "arn": "arn:aws:sts:785789292865:assumed-role/AwsSt
7	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts:785789292865:assumed-role/AwsE
8	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts:785789292865:assumed-role/AwsE
9	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts:785789292865:assumed-role/AwsE
10	1.05	("type": "AssumedRole", "principalId": "AROA3NSFRCFAYFZAY4O6R:Meta311", "arn": "arn:aws:sts:785789292865:assumed-role/AwsSt

Agenda

Why data lakes?

Why choose AWS for data lakes?

Why Lake Formation for data lakes?

What is Lake Formation?

Lets look at a demo!

AWS Lake Formation Pricing

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

Thank you!

Learn more: <https://aws.amazon.com/lake-formation/>

Contact us: lakeformation-feedback@amazon.com