



AWS Lake Formation

Introduction

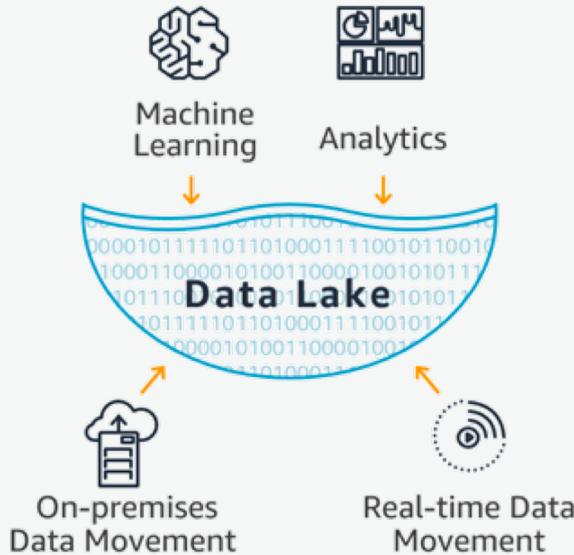
David Bayard, Solutions Architect

October 2019

Why Data Lakes?

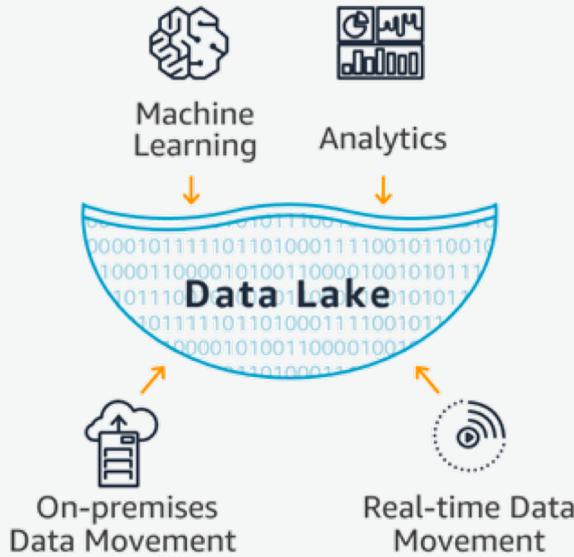
What is a Data Lake?

- A **centralized repository** for both **structured** and **unstructured data**
- Store data **as-is** in **open-source file formats** to enable **direct analytics**



Why a Data Lake?

- Decouple **storage** from **compute**, allowing you to **scale**
- Enable **advanced analytics** across all of your data sources
- Reduce **complexity** in ETL and operational overhead
- Future **extensibility** as new database and analytics technologies are invented



Building a Data Lake on AWS

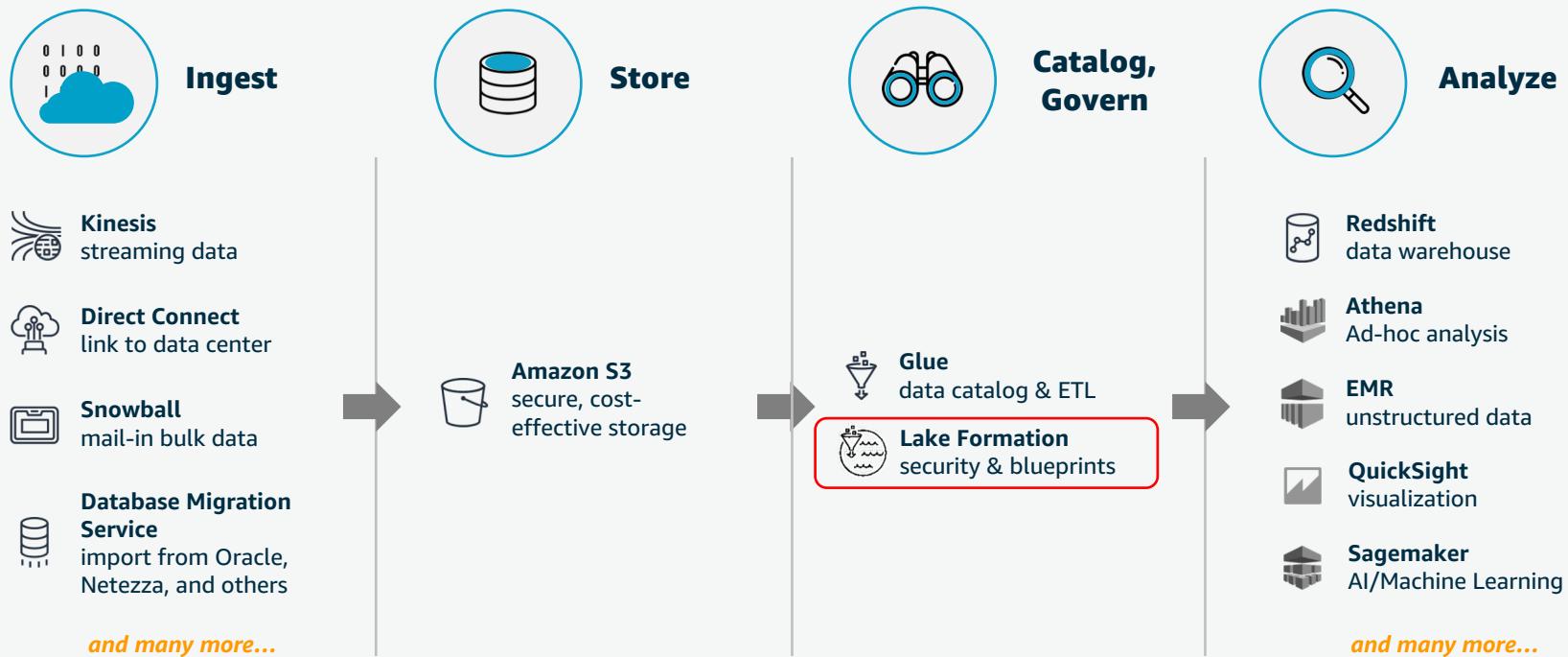
More data lakes & analytics on AWS than anywhere else



Realizing customer and operational insight from your data requires a robust Data Pipeline



Data Lake Reference Architecture

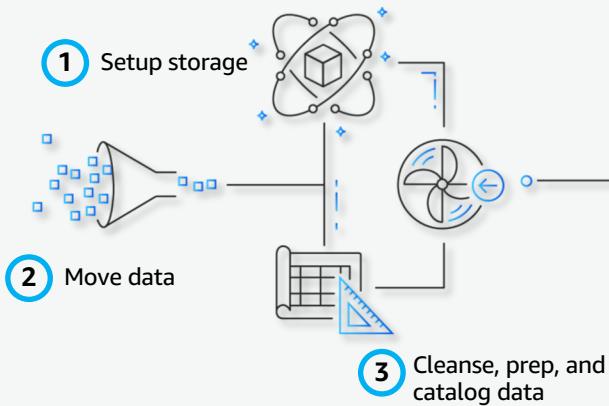


AWS Lake Formation

Manually building secure data lakes is **hard**

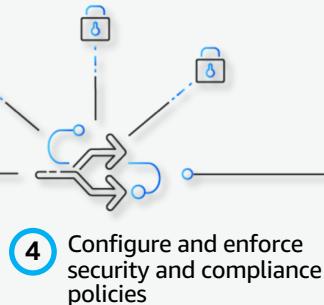
Typical steps of building a data lake

Ingestion & cleaning



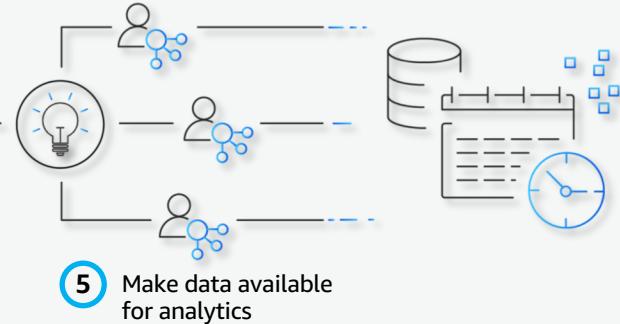
Data
Engineer

Security



Data Security
Officer

Analytics & ML



Data
Analyst

Sample of steps required

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](#)

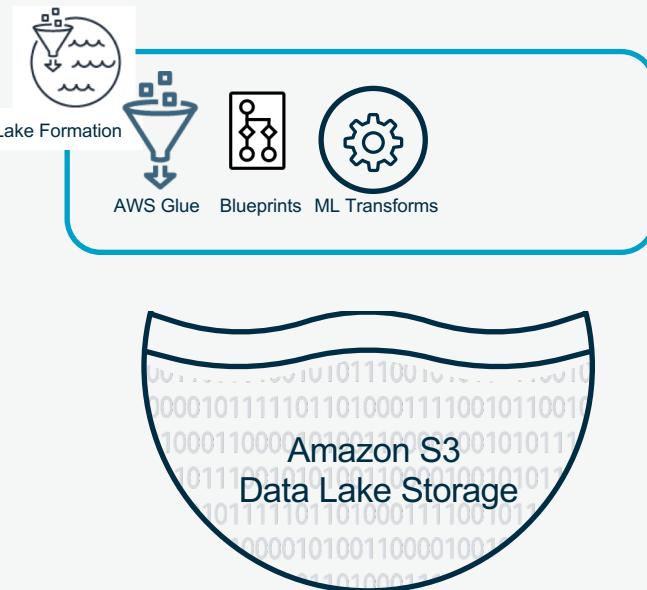
Lake Formation simplifies
building secure data lakes

Built on Amazon S3 a robust data lake infrastructure



Cost effective, durable storage with
global replication capabilities

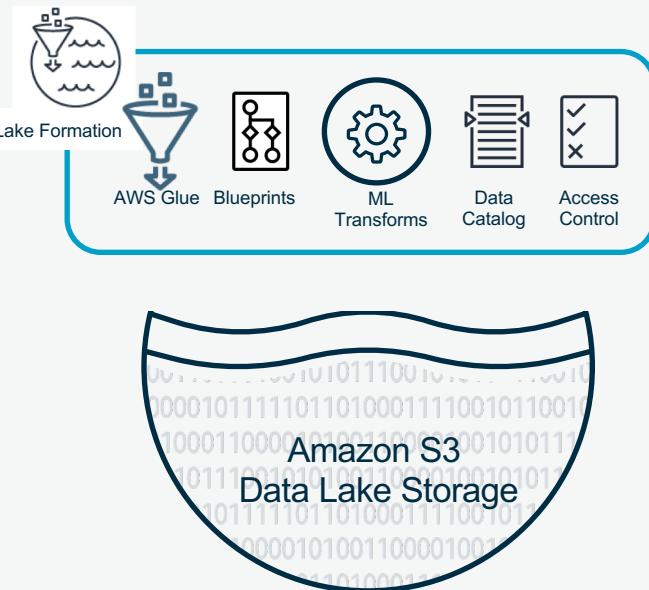
Automates manual, repetitive, low value tasks



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

Cost effective, durable storage with global replication capabilities

Provides a central locus of control

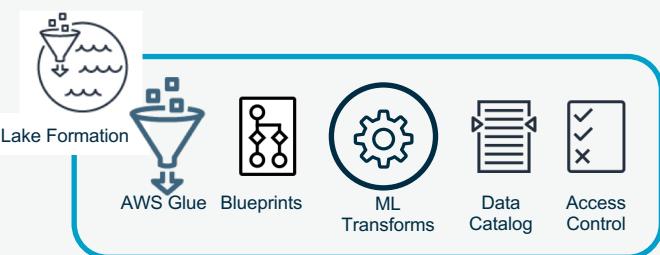


Centralized management of **fine grained permissions** empower security officers

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

AWS Lake Formation Pricing

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

Tools that enable **data engineers**, **security officers** & **data analysts**

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

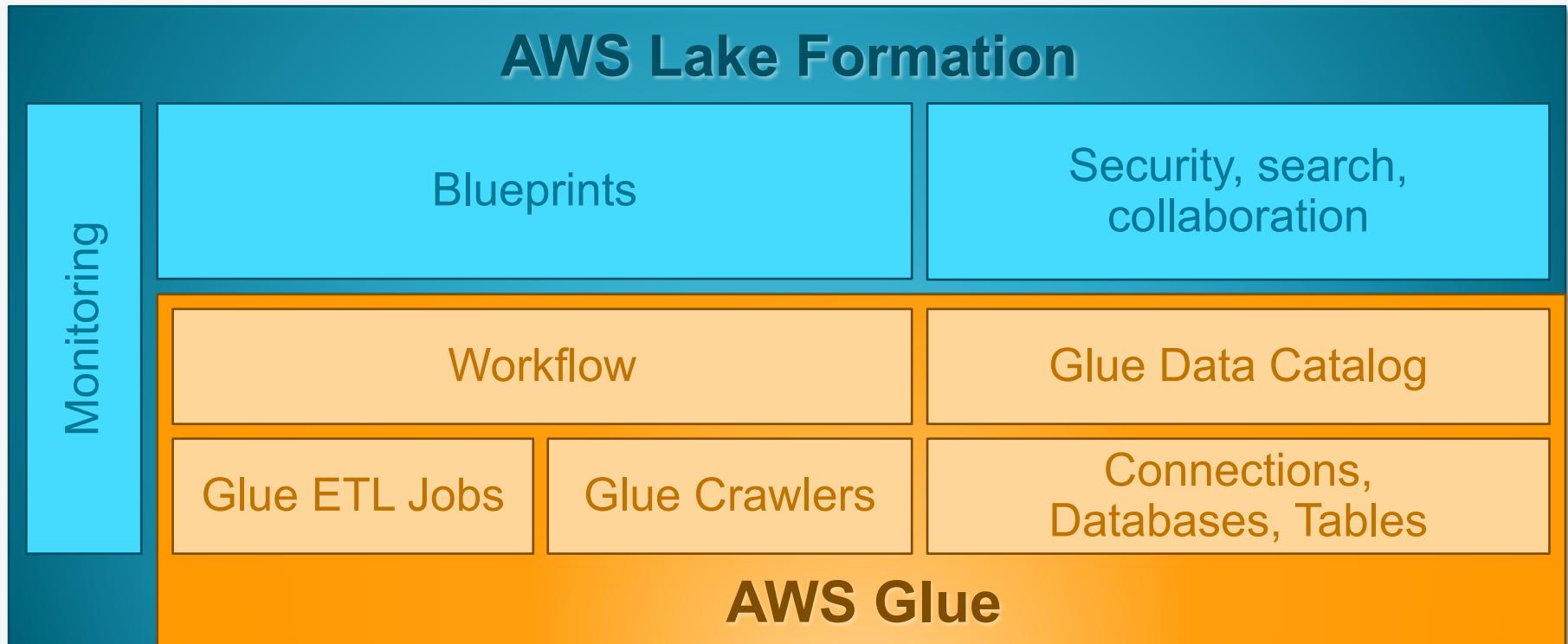


EMR

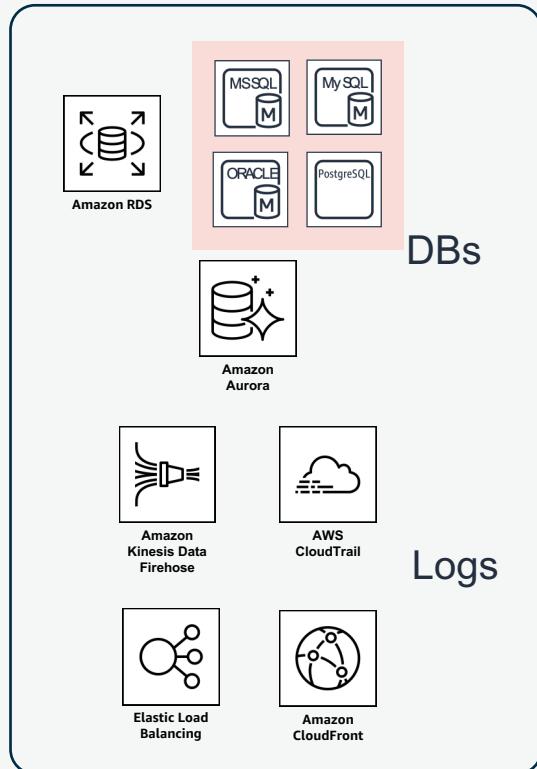


Athena

AWS Lake Formation is fully integrated w/ AWS Glue

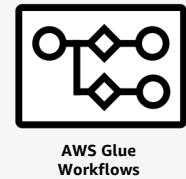


Easily load data into your data lake w/ blueprints



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**

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

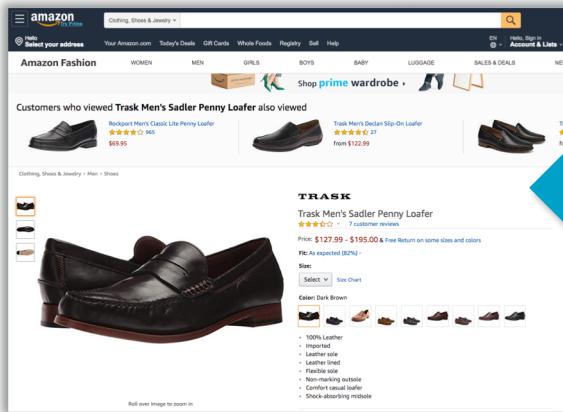
Track data that was already processed

Customize to your needs

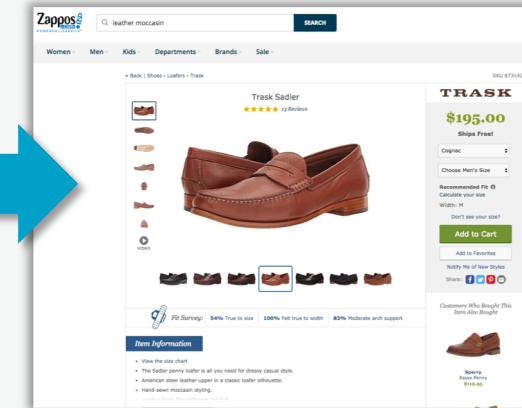
Glue ML Transformations

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*



ML FindMatches



Record matching

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

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

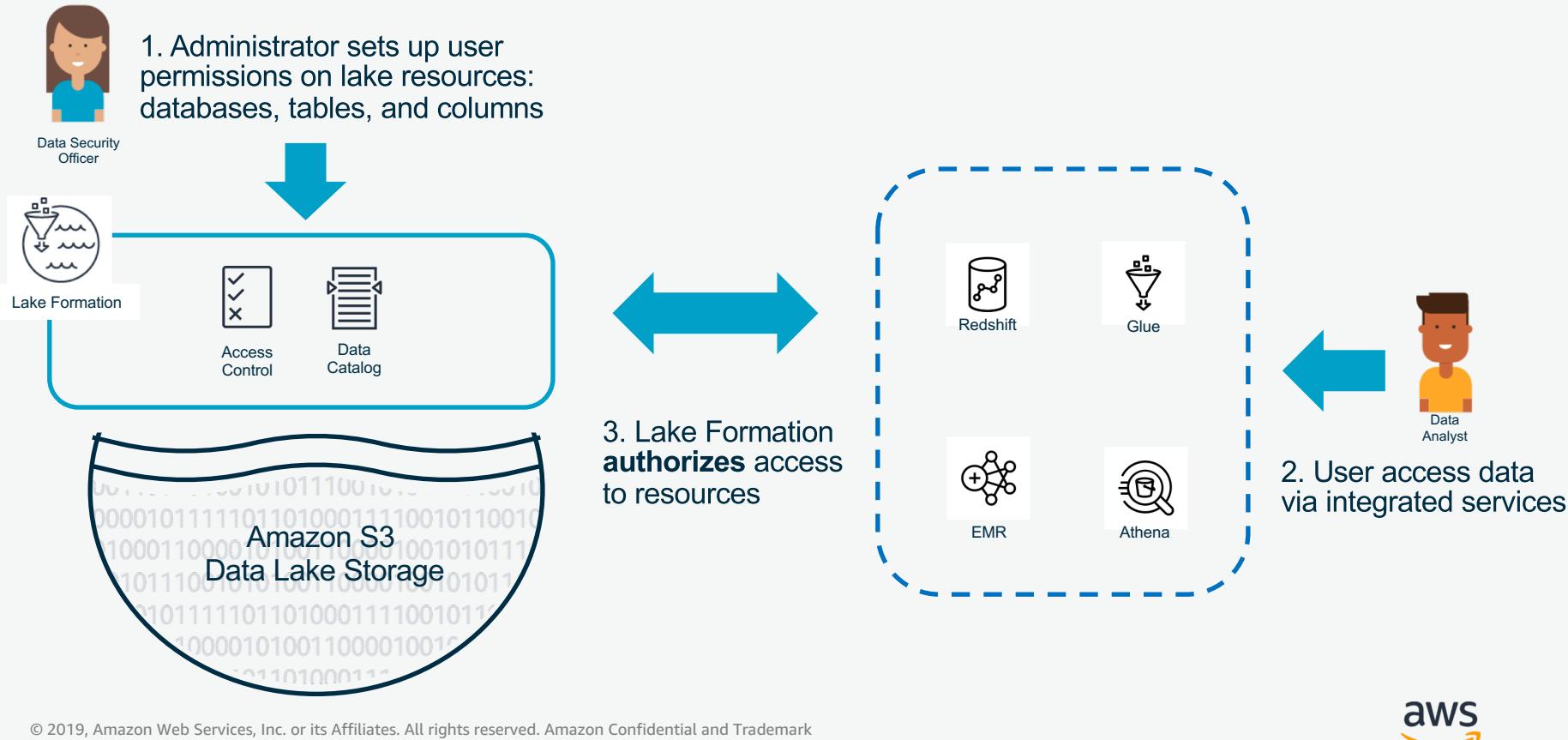


EMR



Athena

Centralized permissions



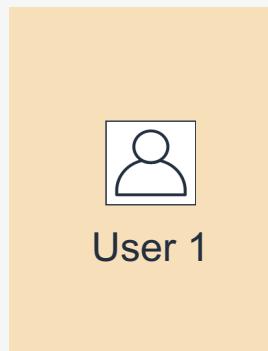
Security permissions in Lake Formation

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



User 1

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

A white square icon containing a black outline of a person's head and shoulders, representing User 2.

User 2

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

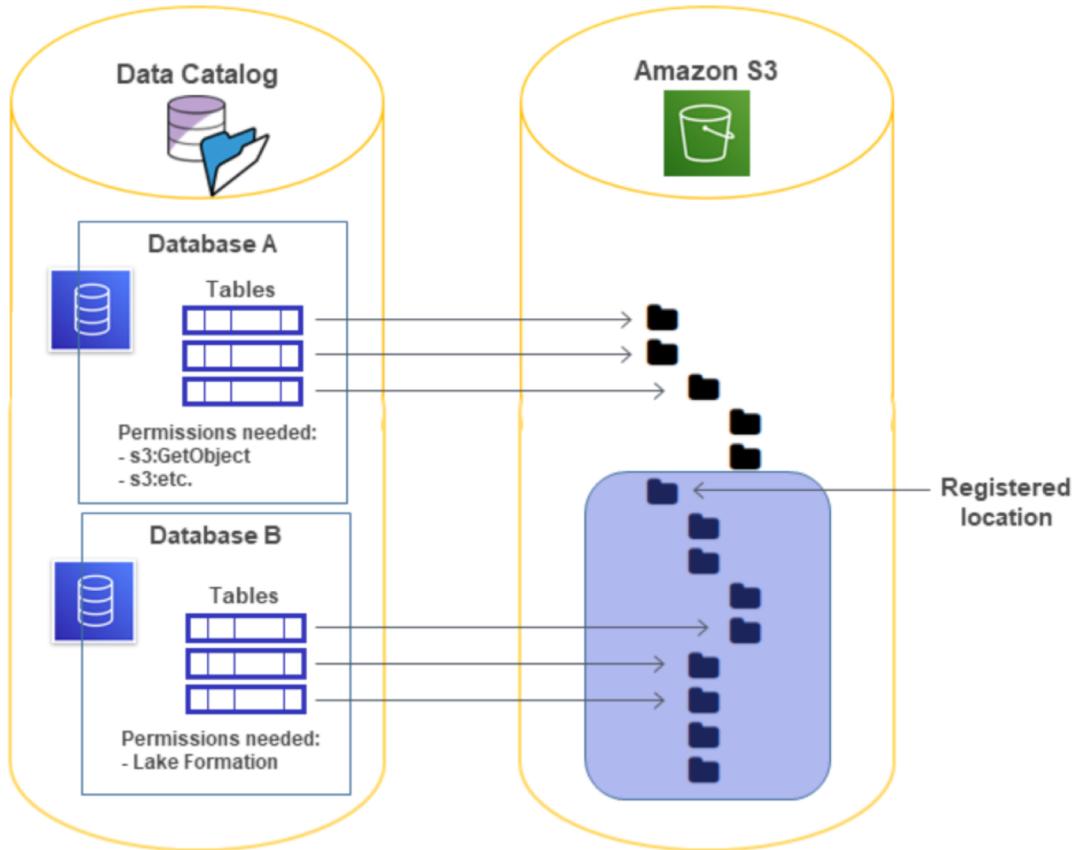
For backwards compatibility, the default settings enable the **Glue** permissions system

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



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





Database A is an example of the old permissions model.

Database B is an example of the new Lake Formation permissions model.

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 interface on the left and the Amazon Athena interface on the right, connected by a large blue arrow pointing from one to the other.

AWS Lake Formation: The sidebar shows navigation options like Dashboard, Data catalog, Databases, Tables (selected), Settings, Register and ingest, Permissions, Admins and database creators, Data permissions, and Data locations. The main area lists tables under the heading "Tables (52)". A search bar at the top of the table view contains the query "Database : amazoncloudtrail". One table entry, "amazoncloudtrail_cloudtrail", is highlighted with a blue box around its "Actions" dropdown menu, which includes View data, Edit, Drop, Location, Classification, Security, Grant, Revoke, Verify permissions, and View permissions. A tooltip "Query data in Amazon Athena" points to the "View data" option.

Amazon Athena: The interface shows a query editor with the following SQL code:

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

Below the query editor is a results table with columns: eventversion, eventid, eventtime, sharedeventid, and requestparameters.durationseconds. The results list 10 rows of data, each corresponding to a log entry from the "cloudtrail.parquetstrail" table.

Annotations:

- A large blue arrow points from the "View data" button in the AWS Lake Formation table list to the "Run Query" button in the Amazon Athena query editor.
- A blue box highlights the "View data" button in the AWS Lake Formation table list.
- A blue box highlights the "Run Query" button in the Amazon Athena query editor.
- A blue box highlights the "Results" table in the Amazon Athena interface.
- A blue box highlights the "Text-based search and filtering" text in the center of the slide.

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

The screenshot shows the AWS Lake Formation console. On the left, there's a sidebar with options like Dashboard, Data catalog, Databases, Tables, Settings, Register and ingest (Data lake locations, Blueprints, Crawlers, Jobs), and Permissions (Admins and database creators, Data permissions, Data locations). The main area is titled 'Data lake setup' with three stages: Stage 1 (Register your Amazon S3 storage), Stage 2 (Create a database), and Stage 3 (Grant permissions). Below these stages are 'Register location', 'Create database', and 'Grant permissions' buttons. At the bottom right, there's a section titled 'Recent access activity (0/50)' with a table:

Event name	Principal	Alert time
BatchGrantPermissions	lakeformationuser	Wed, Aug 7, 2019, 10:25 PM UTC
ListPermissions	lakeformationuser	Wed, Aug 7, 2019, 10:25 PM UTC
GetDataLakeSettings	lakeformationuser	Wed, Aug 7, 2019, 10:24 PM UTC
ListPermissions	lakeformationuser	Wed, Aug 7, 2019, 10:24 PM UTC
GetDataLakeSettings	lakeformationuser	Wed, Aug 7, 2019, 10:24 PM UTC

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

Comprehensive portfolio of integrated tools

Compliant services honor
Lake Formation permissions



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

All access is logged and
auditable

The screenshot shows the AWS Athena Query Editor interface. The left sidebar displays the Catalog (Lake formation), Database (amazoncloudtrail), and Tables (1) section, which lists 'amazoncloudtrail_clouptrail' as a Partitioned table. The main area shows a query editor with the following code:

```
1 SELECT * FROM "amazoncloudtrail"."amazoncloudtrail_clouptrail" limit 10;
```

Below the query editor, the results pane shows the first 10 rows of data from the CloudTrail table. The columns listed are eventversion and useridentity. The data consists of 10 rows, each containing a value for eventversion (e.g., 1, 2, 3, ..., 10) and a JSON object for useridentity.

eventversion	useridentity
1	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
2	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:Meta31", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
3	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
4	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
5	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
6	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:Meta31", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
7	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
8	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
9	{"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
10	... {"type": "AssumedRole", "principalId": "AROA3N5FRCFAFYZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...

How does Lake Formation work?

Dashboard**Data catalog**

Databases

Tables

Settings

Register and ingest

Data lake locations

Blueprints

Crawlers Jobs **Permissions****▼ Data lake setup**

Quickly set up your data lake in Lake Formation.

Stage 1**Register your Amazon S3 storage**

Lake Formation manages access to designated storage locations within Amazon S3. Register the storage locations that you want to be part of the data lake.

Register location**Stage 2****Create a database**

Lake Formation organizes data into a catalog of logical databases and tables. Create one or more databases and then automatically generate tables during data ingestion for common workflows.

Create database**Stage 3****Grant permissions**

Lake Formation manages access for IAM users, roles, and Active Directory users and groups via flexible database, table, and column permissions. Grant permissions to one or more resources for your selected users.

Grant permissions

Step 1: Register S3 path as data lake location



Data
Engineer

Add storage

Amazon S3 storage

Register Amazon S3 storage as your data lake.

Amazon S3 path

Choose an Amazon S3 location for your data lake.

Browse

Existing location permissions - *optional*

Registering the selected location may result in your users gaining access to data already at that location. Before registering the location, we recommend that you audit permissions on resources in the location.

Audit location permissions

IAM role

To add or update data, Lake Formation needs read/write access to the chosen Amazon S3 path. Choose a role that you know has permission to do this, or choose the **AWSServiceRoleForLakeFormationDataAccess** service-linked role. When you register the first Amazon S3 path, the service-linked role and a new inline policy are created on your behalf. Lake Formation adds the first path to the inline policy and attaches it to the service-linked role. When you register subsequent paths, Lake Formation adds the path to the existing policy.

CancelAdd storage

Step 2: Load data with blueprints



Data
Engineer

Blueprint type
Configure a blueprint to create a workflow.

Database snapshot
Bulk load data to your data lake from MySQL, PostgreSQL, Oracle, and Microsoft SQL Server databases.

Incremental database
Load new data to your data lake from MySQL, PostgreSQL, Oracle, and SQL Server databases.

AWS CloudTrail
Bulk load data from AWS CloudTrail sources.

AWS ELB logs

AWS ALB logs

Import source
Configure the workflow source.

CloudTrail name
Choose a CloudTrail source.

Start date
Choose a CloudTrail source start date.

Import target
Configure the target of the workflow.

Target database
Choose a database in the Data Catalog. [Create database](#)

Data lake location
Choose where to import from your data lake storage locations.

Data format
Choose the output data format.

Import frequency
Schedule the workflow.

Frequency
Choose how often to run the workflow.

Import options
Configure the workflow.

Workflow name

Names may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and must be less than 256 characters long.

IAM role

Table prefix
The table prefix that is used for catalog tables that are created.

Maximum capacity - optional

Step 3: Grant permissions to users



Grant permissions

Grant access permissions to specific users and roles.

IAM users and roles
Add one or more IAM users or roles.

datalake_user_redshift X
User

Database
Add one or more databases.

amazoncloudtrail X

Table - optional
Add one or more tables.

amazoncloudtrail_clouptrail X

Column - optional
Grant permissions to:

Include columns
Add one or more columns to include.

useridentity X string **eventsoure** X string **eventname** X string
sourceipaddress X string

Table permissions
Choose the specific access permissions to grant.

Select all Alter Insert Drop
 Delete Select
 Grant all
Enabling this permission grants full access to the specified resources while still logging access requests based on the individual permission settings above. It is typically used for debugging. Specific individual permissions set above will take effect when Grant all is later removed.

Grantable permissions
Choose the specific permissions that may be granted to others.

Select all Create table Alter Drop
 Grant all
Enabling this permission grants full access to the specified resources while still logging access requests based on the individual permission settings above. It is typically used for debugging. Specific individual permissions set above will take effect when Grant all is later removed.

Step 4: Query data with compatible services



Data Analyst

The screenshot shows the AWS Athena Query Editor interface. On the left, the Catalog and Database dropdowns are set to "Lake formation" and "amazoncloudtrail" respectively. Under "Tables (1)", the "amazoncloudtrail_clouptrail (Partitioned)" table is selected. The main pane displays a query editor with the following SQL:

```
1 SELECT * FROM "amazoncloudtrail"."amazoncloudtrail_clouptrail" limit 10;
```

Below the query, the status bar indicates "(Run time: 3.02 seconds, Data scanned: 101.28 KB)". The "Results" section at the bottom shows 10 rows of data from the CloudTrail table, each containing eventversion and useridentity fields.

eventversion	useridentity
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:palisade", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:Meta31", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:palisade", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:palisade", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:palisade", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:Meta31", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:palisade", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:palisade", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:palisade", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...
1.05	{"type": "AssumedRole", "principalid": "AROA3N5FRCFAYFZAY4O6R:palisade", "am": "arn:aws:sts::785789292865:assumed-role/AwsS...

Step 5: Audit and monitor in real time



DATA SECURITY
OFFICER

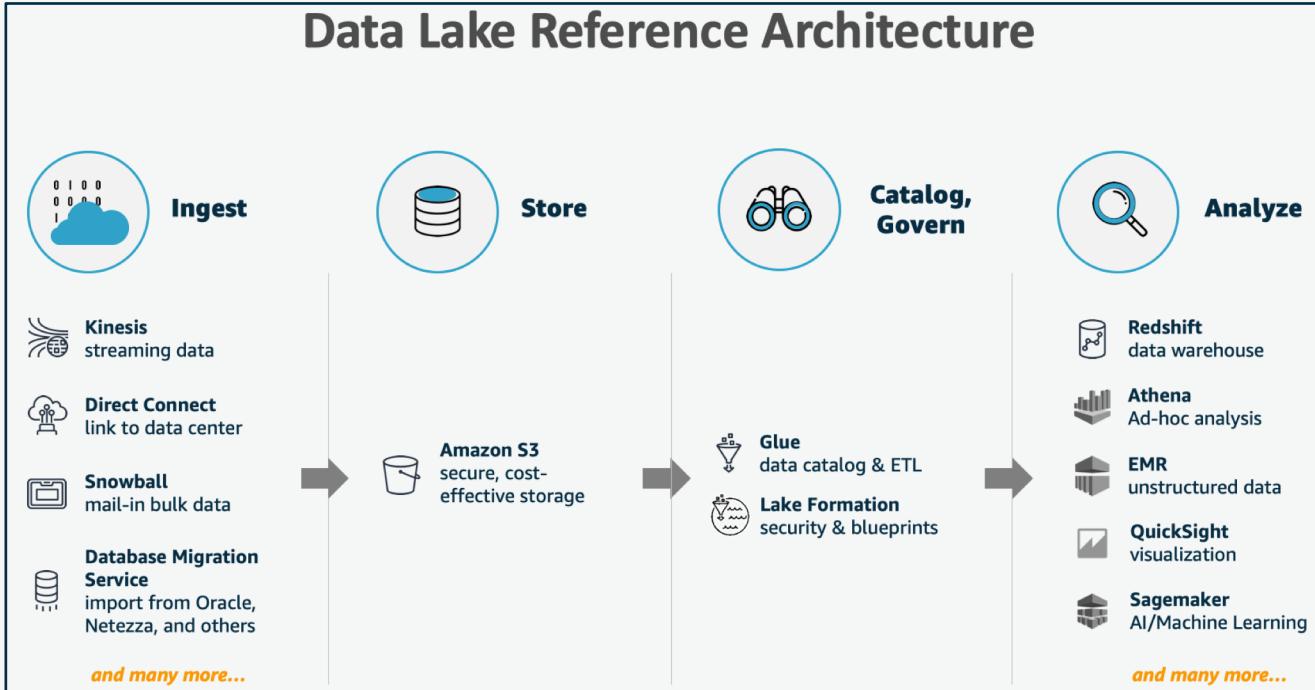
The screenshot shows the AWS Lake Formation Console dashboard. On the left, a sidebar menu includes options like Dashboard, Data catalog, Databases, Tables, Settings, Register and ingest (Data lake locations, Blueprints, Crawlers, Jobs), and Permissions (Admins and database creators, Data permissions, Data locations). The main content area features a "Recent access activity (50)" section with a table of log entries. The table has columns for Event name, Principal, and Alert time. The log entries show multiple calls from a user named "datalake_user" on June 30, 2019, and one call from "chessm" on June 30, 2019. The table includes a "View event" link and navigation controls.

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

AWS Lake Formation Pricing

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

Thank you!



Thank you!

Learn more:

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

