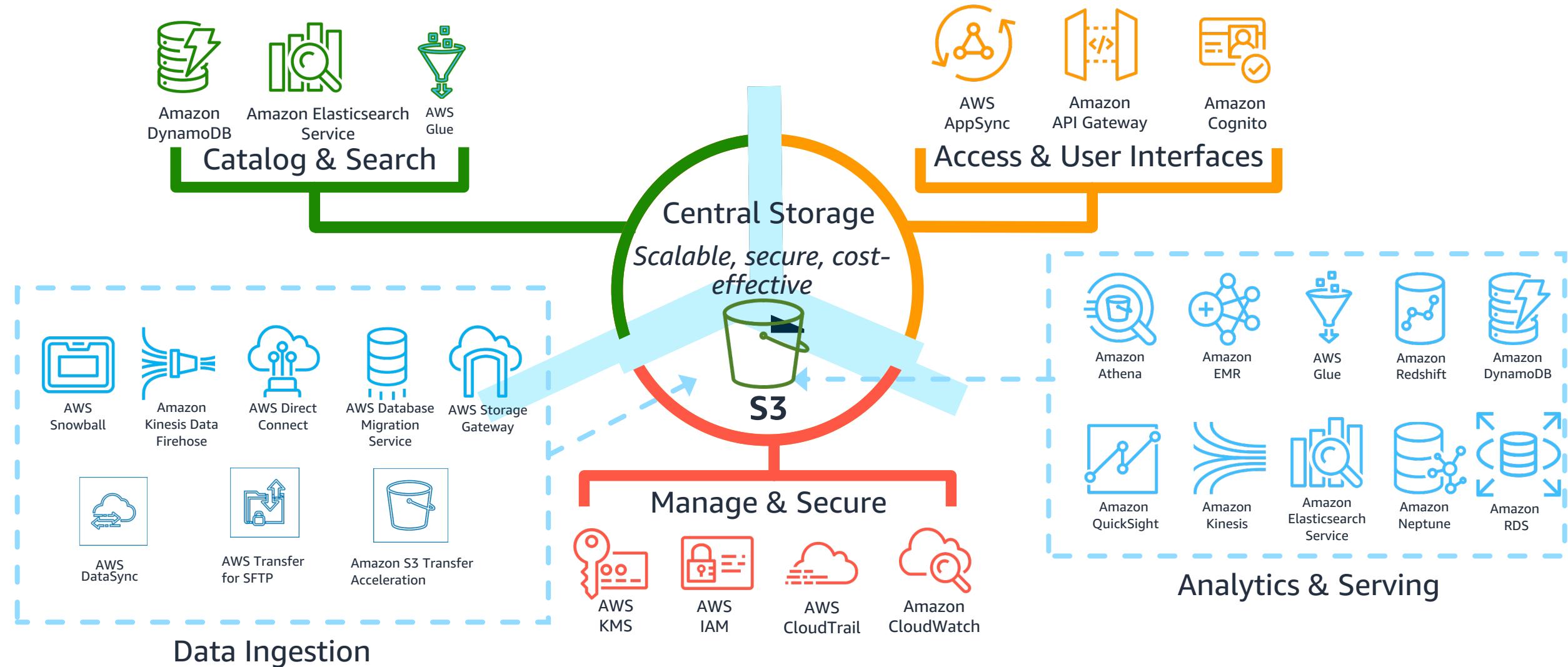




# Automate Data Lake with AWS Lake Formation



# Covering It All With Lake Formation



# Agenda

- Lake Formation Value Proposition
- AWS Lake Formation
  - Steps to build Data Lake
  - Reference Architecture
- Lake Formation Console Walkthrough

# What is hard about building data lakes?

Manually building secure data lakes is **hard**

# Sample of steps required

Configure access from analytics services

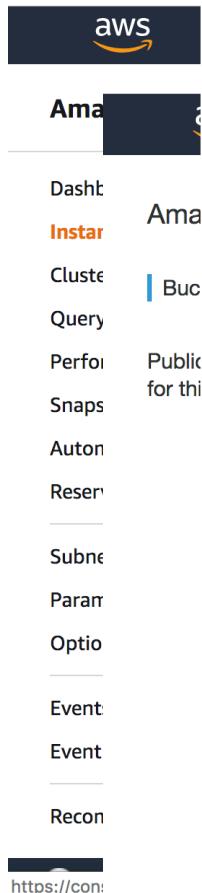
Rinse and repeat for other:  
data sets, users, and end-services

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 \(US\)](#)

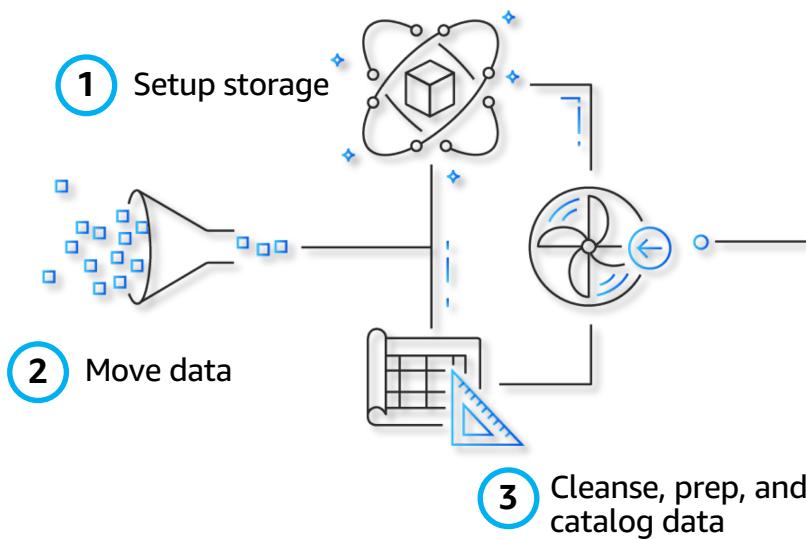
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#)

[Feedback](#) [English \(US\)](#)

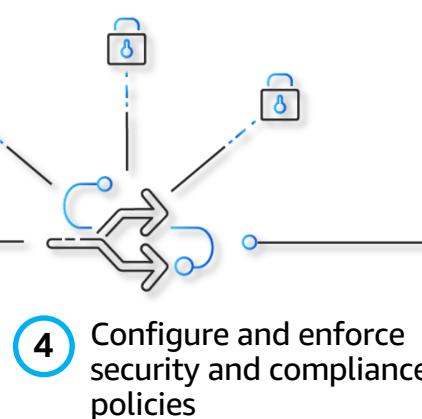
© 2008 - 2018, Amazon Web Services, Inc. or its affiliates. All rights reserved. [Privacy Policy](#) [Terms of Use](#)

# Typical steps of building a data lake

## Ingestion & cleaning

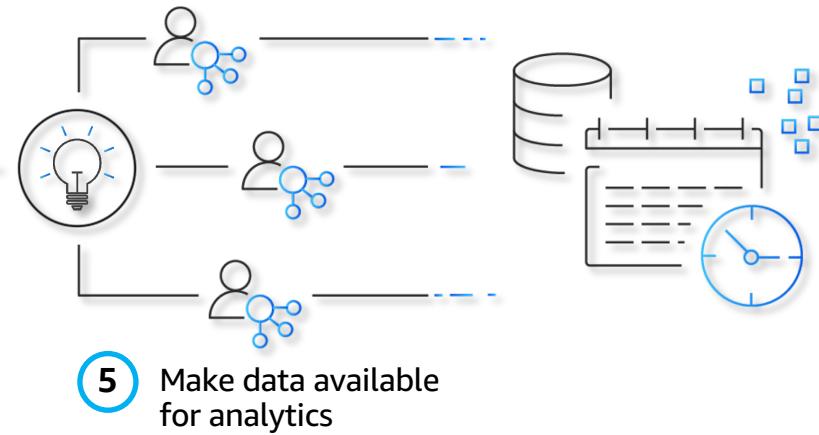


## Security



Data Security Officer

## Analytics & ML



Data Analyst

# What is Lake Formation?

Automating Data Lake



# Features and Benefits of Lake Formation



Amazon  
Athena



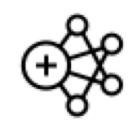
Amazon  
QuickSight



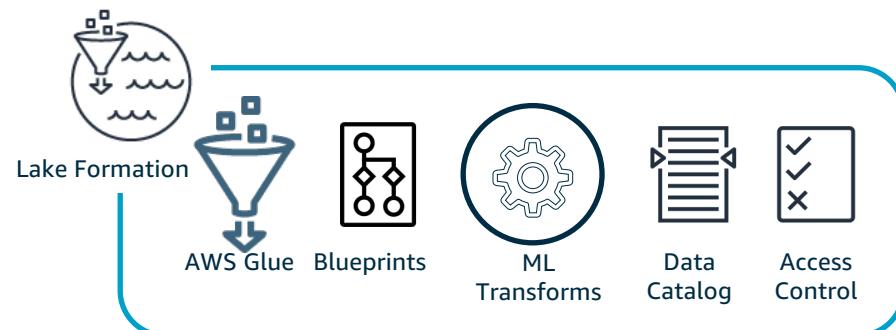
Amazon  
Redshift



Amazon  
SageMaker



Amazon  
EMR



00...1000101110010...0010  
0000101111011010001110010110010  
100011000010100110001001010111  
10111001010011000100101011  
1011111011010001111001011  
10000101001100001001  
10110100011

Amazon S3  
Data Lake Storage

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

# Building data lakes with Lake Formation

## Ingestion & cleaning



AWS Glue  
Serverless Spark  
Blueprints  
ML Transforms

## Security



Data catalog  
Centralized permissions  
Real time monitoring  
Auditing

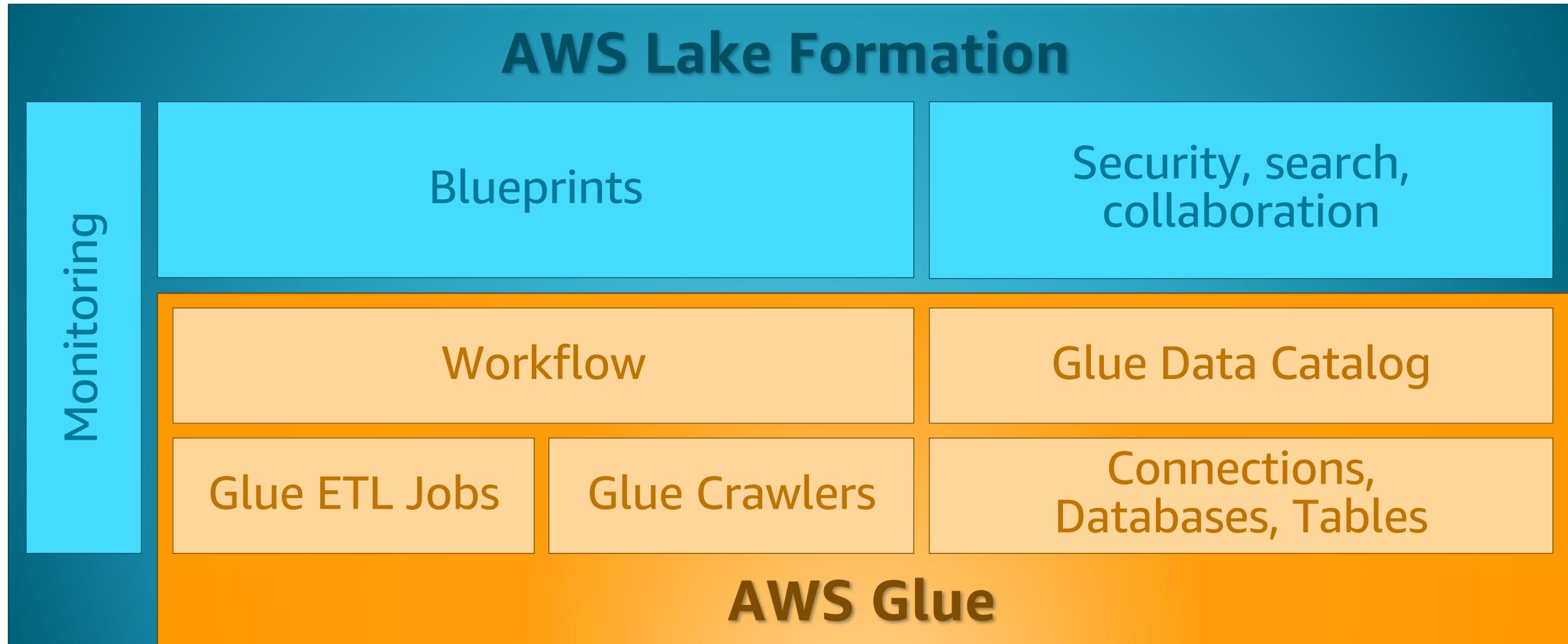
## Analytics & ML



Comprehensive portfolio  
of integrated tools



# AWS Lake Formation is fully integrated w/ AWS Glue



# AWS Glue Components (Recap)

---



## 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

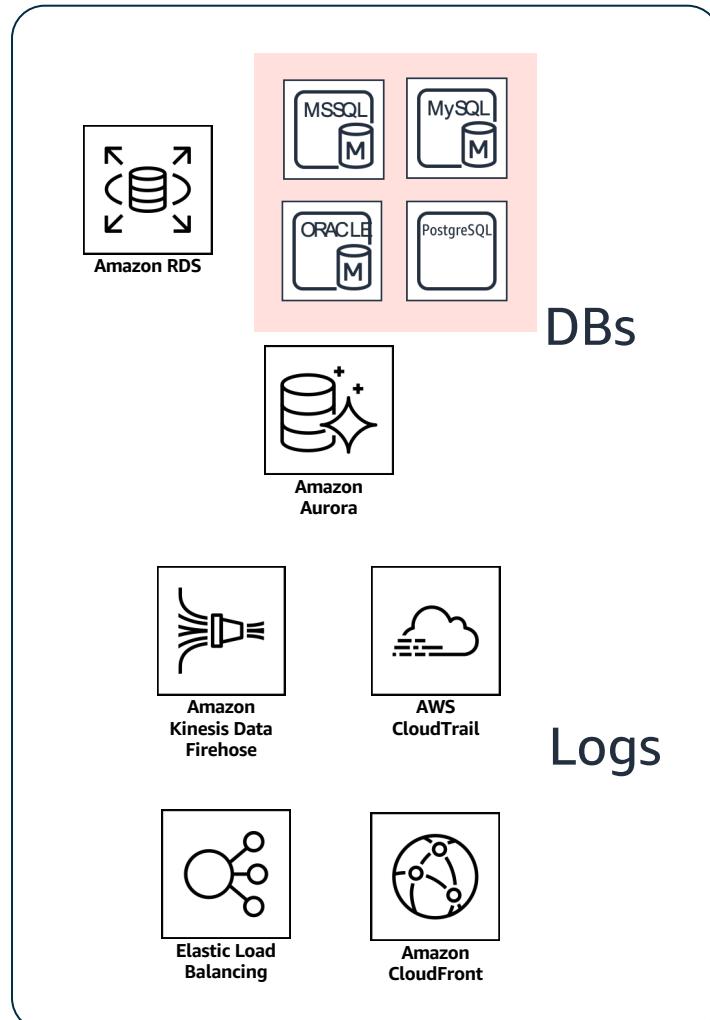
### Deploy

Flexible workflows

Monitoring and alerting

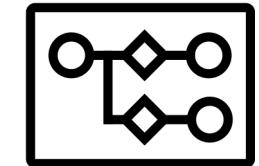
External integrations

# Easily load data into your data lake w/ blueprints



Prebuilt templates to serve common ingestion use cases

Automatically build **AWS Glue workflows**



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

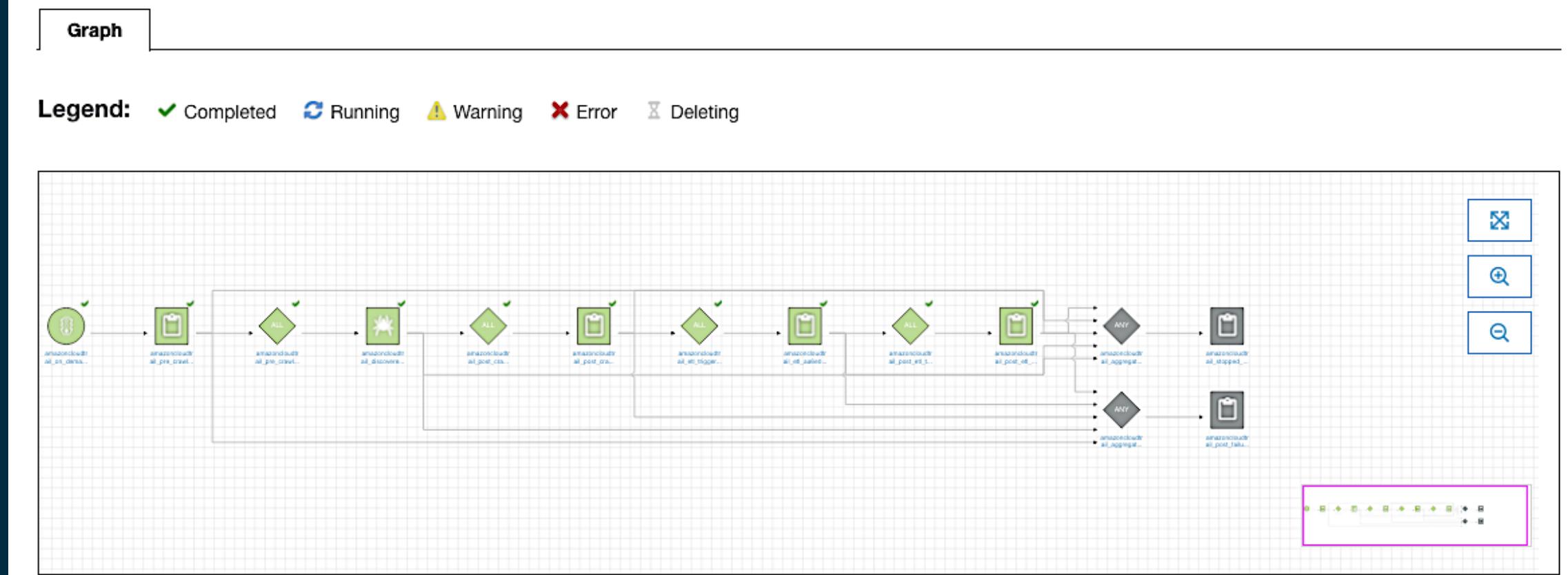
**Convert** to target data format

**Partition** data automatically

**Track** data that was already processed

**Customize** to your needs

# Blueprints create AWS Glue workflows



# 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

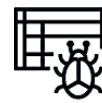


# Data Catalog & Permissions

- 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**



Data Catalog



Crawlers



Workflows



Access Control

# Upgrading to the Lake Formation permissions model

## Not using the Glue Catalog?

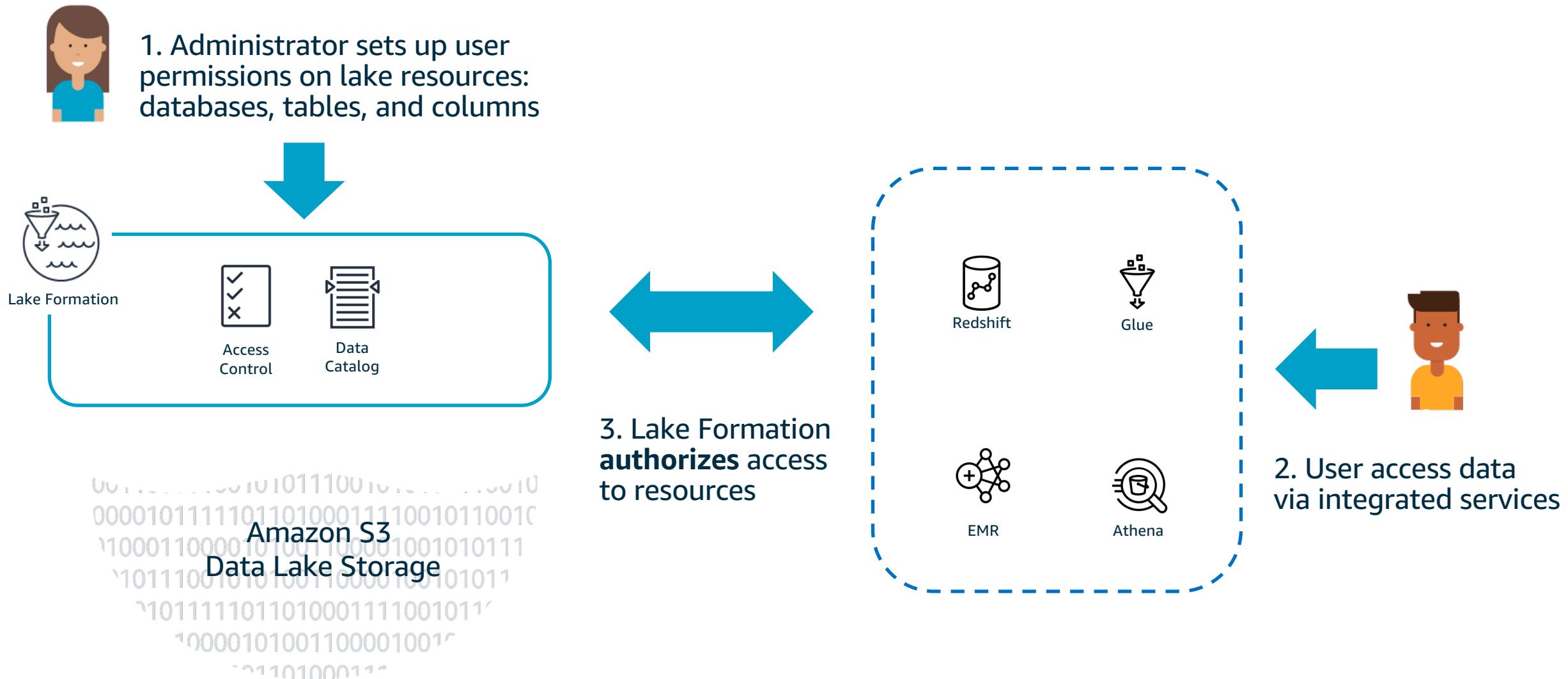
Change the default settings to start using the [Lake Formation permissions system](#)

## Using the Glue Catalog?

Explicitly upgrade each data [location](#), [database](#) and [table](#) when ready

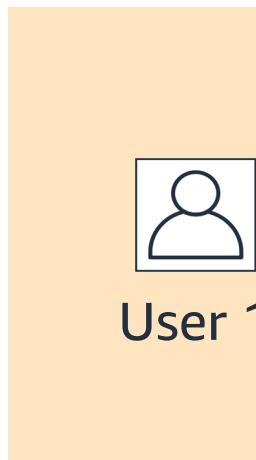
- 1) Understand [existing policies / access / usage](#)
- 2) Configure corresponding [Lake Formation policies](#)
- 3) Remove the [Glue permissions system](#) by changing the default settings
- 4) Turn on the [Lake Formation permissions system](#) by registering the location

# 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
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



User 2

# Data catalog and metadata management

- Text-based **search** 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 shows the AWS Lake Formation console. On the left, the navigation pane includes options like Dashboard, Data catalog (Tables selected), Databases, Settings, Register and ingest, Permissions, and Admins and database creators. The main area displays a list of tables under the 'amazoncloudtrail' database, with 52 tables listed. A context menu is open over one of the tables, showing options like Actions (Edit, Drop, View data, Security, Grant, Revoke, Verify permissions, View permissions), Create table using a crawler, and Create table. A large blue callout bubble highlights the 'View data' option. Another blue callout bubble highlights the 'Text-based search and filtering' feature. In the bottom right, there's a preview of an Amazon Athena query window with a sample SQL query and results table.

Text-based search and filtering

Query data in Amazon Athena

eventversion	eventid	eventtime	shareeventid	requestparameters.durationseconds	
1	1.05	4641c9e0-5604-4008-ad6b-380c4ff1bf163	b6bdcb80-85b8-448a-9f8c-c782e38ac1e	3600	
2	1.05	29279603-98d0-42ef-9d8f-ca70342881d	47927aca-8499-4591-8ff8-29b5f188d7dd	3600	
3	1.05	d6614097-b35f-412d-8ba2-f615d5d56cef	2017-10-23T12:24:37z	49730db-5a41-4de5-9441-9fb1b622c4c	3600
4	1.05	c66bd139-1160-4035-853c-2d9d2e1cf4d8	2017-10-23T12:24:41z	65841ff0-0f54-470a-8f50-be0d5546841b	3600
5	1.05	c21882e4-6a02-4e31-9329-901c268a3206	2017-10-23T12:24:45z	23951756-d749-4497-8ea8-d580fcbb145	3600
6	1.05	410ebd47-aab5-4215-a059-4e9e462b9d9f	2017-10-23T12:24:49z	63993b34-e852-4c99-a370-6f81d26a838be	3600
7	1.05	770dc42-8030-432d-8c7a-15uae65452e0	2017-10-23T12:24:51z	e9f257e9-9bc4-4549-8ea8-1bb27bec6e14	3600
8	1.05	cc55a8f2-6120-4ccb-9cab-8cdff7c7ca7c	2017-10-23T12:25:19z	fd969a02-2371-4908-9a16-9b9c6776a3e3	3600
9	1.05	643c185a-e33b-41ef-a82f-e9edfb6890be	2017-10-23T12:26:19z	f4ad4c07-d8f6-ab97-cf107240b54	3600
10	1.05	0656c3ad-d928-4536-aa0d-f7ede0101f1	2017-10-23T12:26:19z	c90ece1b-497a-494d-a460-67a1719b44b3	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

The screenshot shows the AWS Lake Formation Dashboard. On the left, there's a navigation sidebar with options like Dashboard, Data catalog, Register and ingest, and Permissions. The main area has sections for Data lake setup, Stage 1 (Register your Amazon S3 storage), Stage 2 (Create a database), and Stage 3 (Grant permissions). At the bottom, there's a table titled "Recent access activity (0/50)" with columns for Event name, Principal, and Alert time. A blue oval highlights this table, with the text "Detailed activity" written above it.

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



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

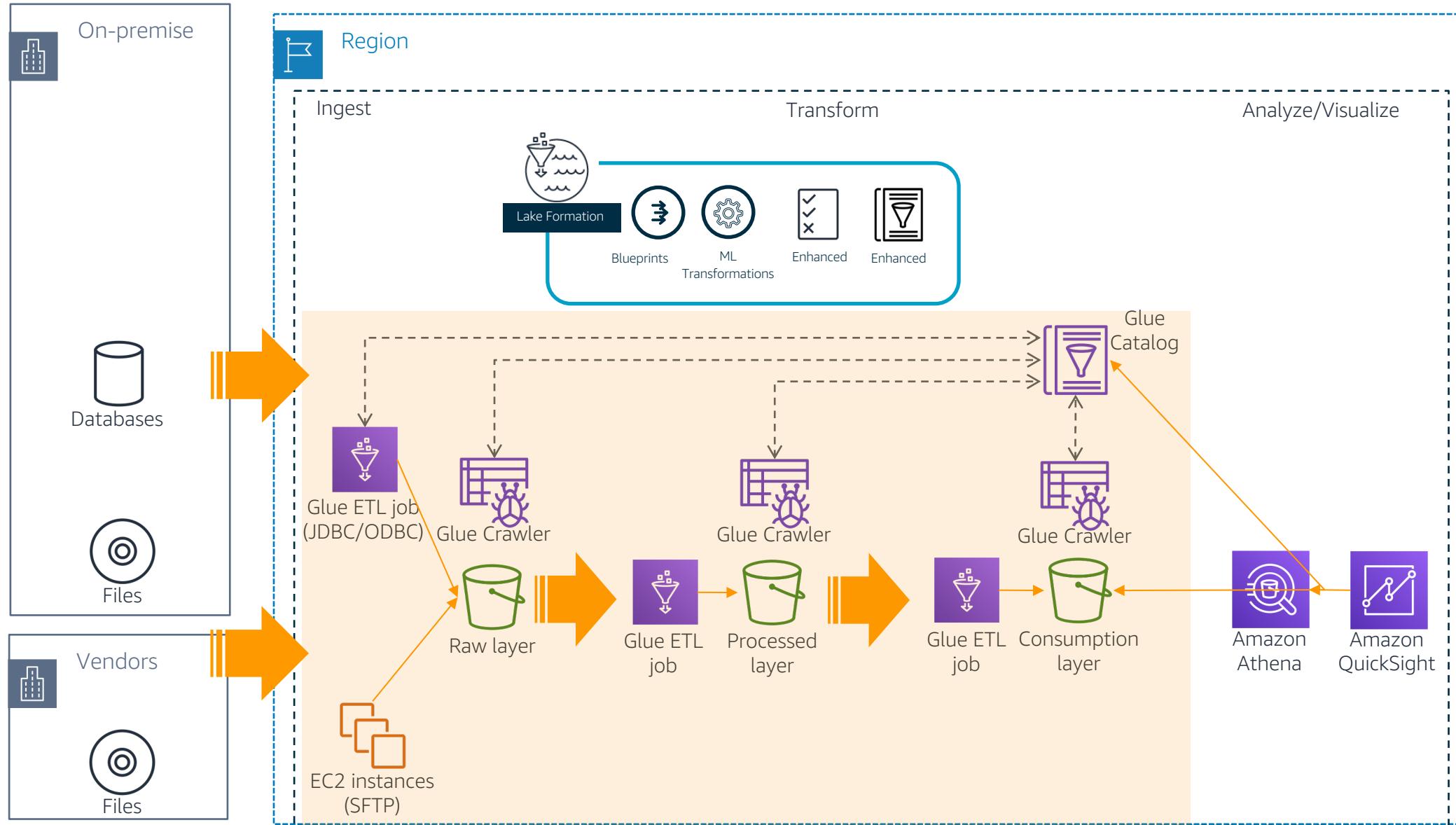


- They guarantee that users only see tables & columns
- All access is logged and auditable

The screenshot shows the AWS Athena Query Editor interface. The left sidebar displays the Catalog (Lake formation) and Database (amazoncloudtrail). Under Tables, there is one entry: amazoncloudtrail\_cldtrail (Partitioned). The main area shows a query editor with a single query: "SELECT \* FROM "amazoncloudtrail"."amazoncloudtrail\_cldtrail" limit 10;". Below the query, buttons for Run query, Save as, Create, Format query, and Clear are visible. The status bar indicates a run time of 3.02 seconds and 101.28 KB scanned. At the bottom, a results table displays 10 rows of JSON data, each representing an event record from the CloudTrail log.

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

# Serverless Lake Formation for Financial Services



# How It Works

Console Walkthrough



# Step 1: Create a data lake administrator

A user who can create databases and tables in the data catalog and grant fine-grained permissions for other users

The screenshot shows the AWS Lake Formation console interface. On the left, there is a user icon of a person with dark skin and short hair, wearing a blue shirt. The main area has two main sections:

- Data lake administrators (0/2)**: Administrators can view all metadata in the Data Catalog. They can also grant and revoke permissions on data resources to principals, including themselves.
  - Filter administrators: A search bar with a magnifying glass icon.
  - Actions: Buttons for "C" (Create), "Grant", and a gear icon for settings.
- Database creators**: Choose IAM principals permitted to create databases in your Data Catalog.
  - Filter database creators: A search bar with a magnifying glass icon.
  - Actions: Buttons for "C" (Create), "Revoke", and "Grant".
  - Pagination: Navigation arrows and a page number "1".
  - Settings: A gear icon.
  - Table Headers: Principal, Principal type, Permissions, Grantable.
  - Data: No permissions listed.

# Step 2: Register S3 path as data lake location



## 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.

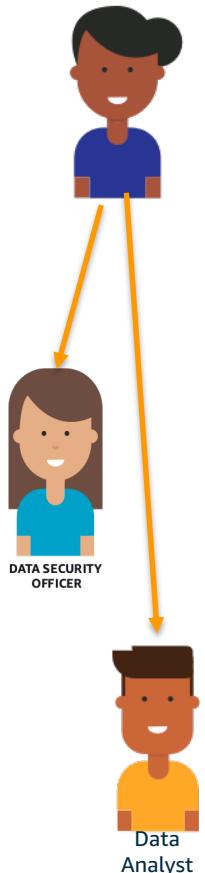
Audit location permissions

### 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 3: Grant permissions to users at tables and columnar level



### 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\_cloudtrail** 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.

**Cancel** **Grant**

# Step 4: Load data with blueprints



## 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.

IsengardTrail-DO-NOT-DELETE

### Start date

Choose a CloudTrail source start date.

2019/06/17



## Import target

Configure the target of the workflow.

### Target database

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

v3\_chanu\_clouptrail



### Data lake location

Choose where to import from your data lake storage locations.

s3://v3-datalake-clouptrail



### Data format

Choose the output data format.

Parquet



## Import frequency

Schedule the workflow.

### Frequency

Choose how often to run the workflow.

Run on demand



## Import options

Configure the workflow.

### Workflow name

v3chanucloudtrailtest

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

### IAM role

v3\_datalake\_service\_role



### Table prefix

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

v3chanucloudtrailtest

### Maximum capacity - optional

Enter a maximum capacity

Cancel

Create

# Step 5: Query data with compatible services



Data Analyst

The screenshot shows the AWS Athena Query Editor interface. On the left, the Catalog pane displays 'Lake formation' as the Catalog and 'amazoncloudtrail' as the Database. Under 'Tables (1)', there is a single entry: 'amazoncloudtrail\_cloudtrail (Partitioned)'. The main workspace shows a query titled 'New query 1' with the following SQL code:

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

Below the query, status information indicates: '(Run time: 3.02 seconds, Data scanned: 101.28 KB)'. At the bottom of the workspace are buttons for 'Run query', 'Save as', 'Create', 'Format query', and 'Clear'. A note below the workspace says: 'Use Ctrl + Enter to run query, Ctrl + Space to autocomplete'.

The 'Results' section at the bottom displays 10 rows of JSON data, each representing an event record from the CloudTrail log. The columns shown are 'eventversion' and 'useridentity'. The data is truncated for brevity.

	eventversion	useridentity
1	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
2	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:Meta31", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
3	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
4	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
5	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
6	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:Meta31", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
7	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
8	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
9	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:palisade", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...
10	1.05	{"type": "AssumedRole", "principalId": "ARO3N5FRCFAYFZAY4O6R:Meta31", "arn": "arn:aws:sts::785789292865:assumed-role/AwsS...

# Step 6: Audit and monitor in real time



DATA SECURITY  
OFFICER

The screenshot shows the AWS Lake Formation Dashboard. On the left, there is a sidebar with the following navigation options:

- Dashboard (selected)
- Data catalog
- Databases
- Tables
- Settings
- Register and ingest
  - Data lake locations
  - Blueprints
  - Crawlers (with a question mark icon)
  - Jobs (with a question mark icon)
- Permissions
  - Admins and database creators
  - Data permissions
  - Data locations

The main content area displays a section titled "Recent access activity (50)". It includes a "View event" button and a "Filter events" search bar. Below is a table showing the following data:

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

# Thank you!