

Big Data on AWS

Gustavo Veloso

Senior Solutions Architect, AWS
velosog@amazon.com

December, 2019

Data is a strategic asset for every organization

66 The world's most valuable resource is no longer oil, but data.* 99



*Copyright: The Economist, 2017, David Parkins

We need to
rethink what we
mean by data
and analytics





This is data

This is data



This is data

Skip the trip.
one-hour delivery

Exclusively for Amazon Prime Members



Data no longer fits



There is **more data** than people think

Data is **more diverse**

Data

grows
>10x
every 5 years

Data platforms need to

live for
15
years

scale
1,000x

* IDC, Data Age 2025: The Evolution of Data to Life-Critical. Don't Focus on Big Data. Focus on the Data That's Big. Apr. 2017

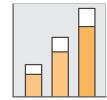
Broader workloads



Data Scientists



Business Users



Analysts



Applications

machine
learning

scientific

SQL analytics

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

Decision making used to...

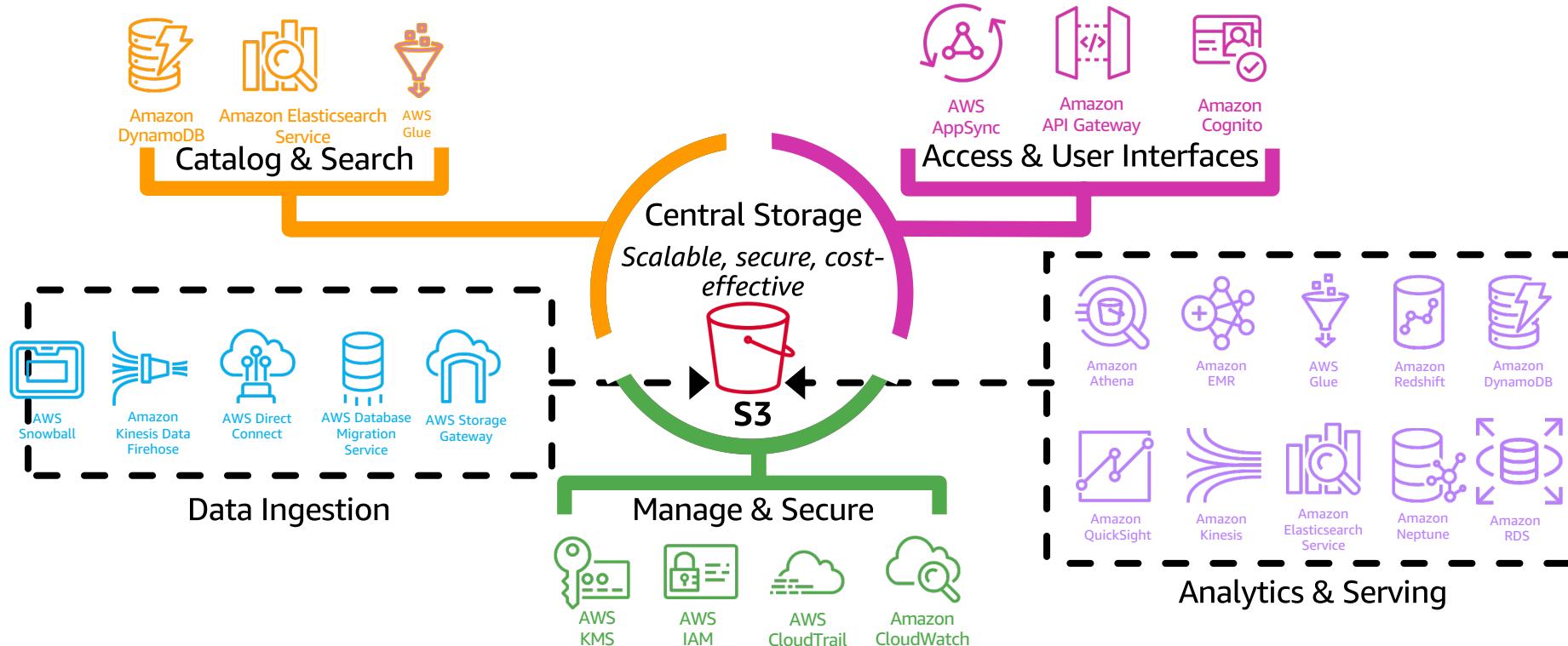
...revolve around the
Enterprise Data Warehouse



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

Data lake on AWS



Tier 1 Data Lake: Ingestion



Amazon S3

Single source of truth for raw data

Use least transformations

Use lifecycle policies to Amazon Simple Storage Service (Amazon S3) IA or Amazon Glacier

Tier 2 Data Lake: Analytics



Amazon S3

Use columnar formats – Parquet/ORC
Organized into partitions
Coalescing to larger partitions over time
Optimized for analytics

Tier 3 Data Lake: Analytics



Amazon S3

Domain level DataMart

Organized by use cases

Optimized for specialized analysis



Amazon Redshift

Data Warehouse:

Fast speeds over structured schemas

Serves dashboards and reports

Fine-grained access controls

Supports joining native and external tables

Lifecycle back to S3 Data Lake

EQUINOX

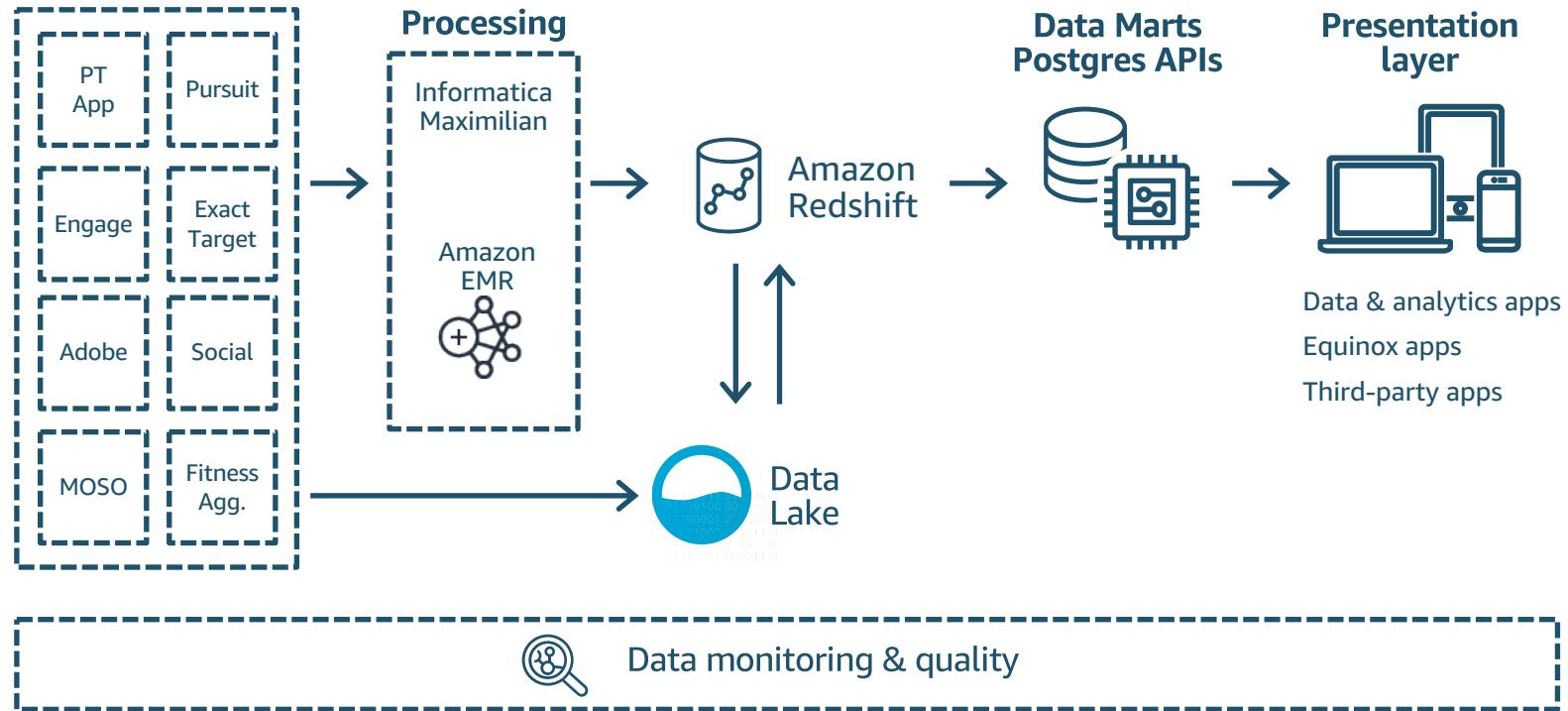
Equinox Fitness Clubs is a company with integrated luxury and lifestyle offerings centered on movement, nutrition and regeneration. Equinox built connected experiences using applications that connect to Apple Health and built data collection in their exercise equipment.



More than **200 locations** within every major city across the U.S., London, and Canada



Data lake architecture





Fortnite | 125+ million players

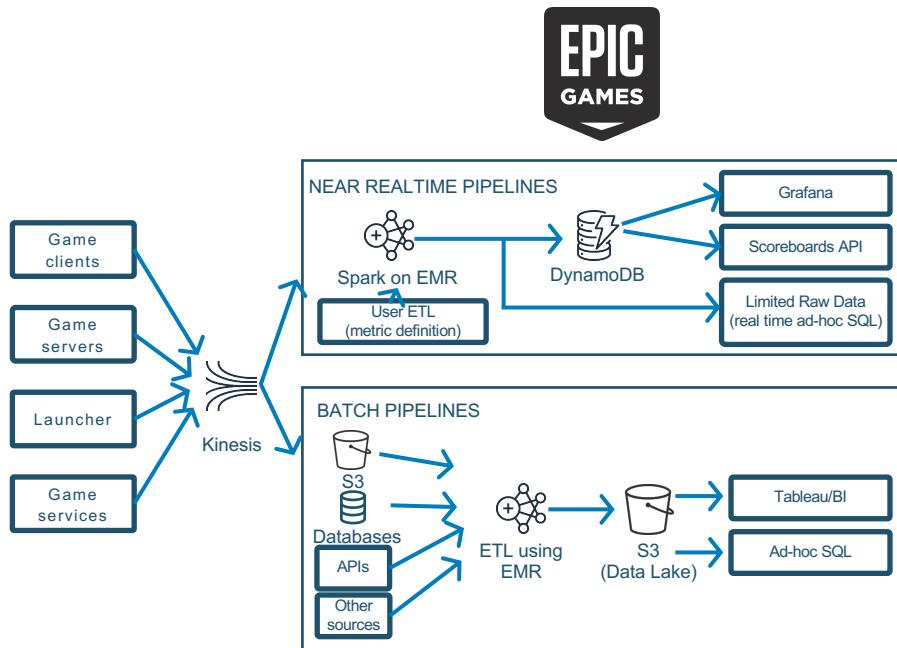
CHALLENGE

Need to create constant feedback loop
for designers

Gain up-to-the-minute understanding
of gamer satisfaction to guarantee
gamers are engaged, thus resulting in
the most popular game played in the
world



Epic Games uses Data Lakes and analytics



Entire analytics platform running on AWS

S3 leveraged as a Data Lake

All telemetry data is collected with Kinesis

Real-time analytics done through Spark on EMR, DynamoDB to create scoreboards and real-time queries

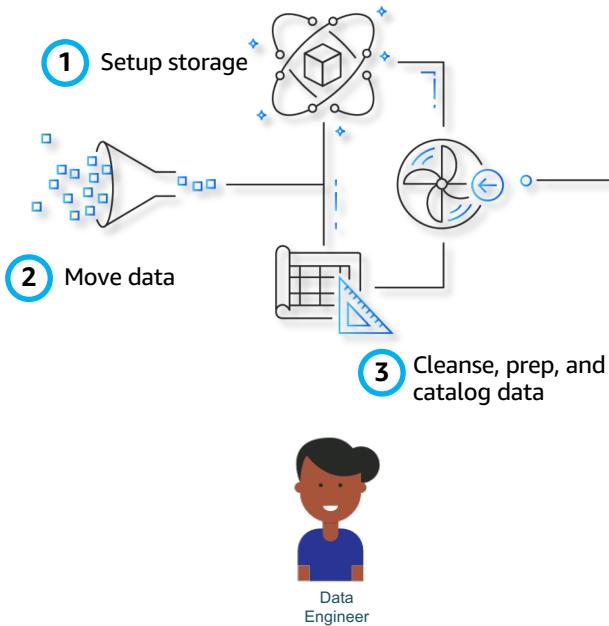
Use Amazon EMR for large batch data processing

Game designers use data to inform their decisions

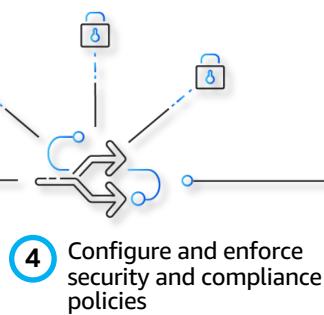
Manually building secure data lakes is **hard**

Typical steps of building a data lake

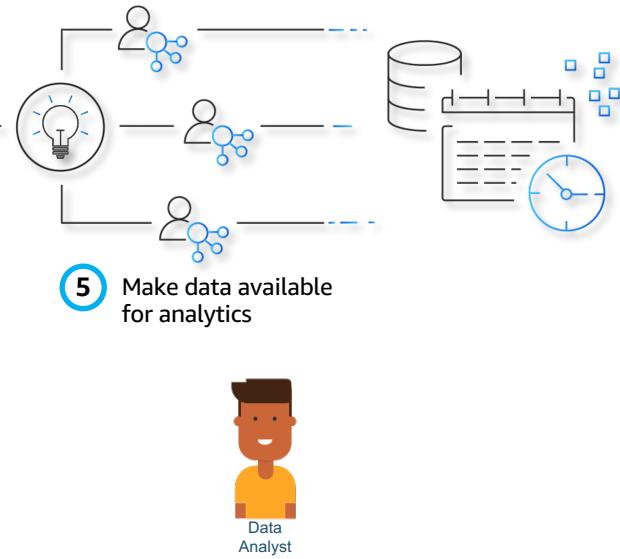
Ingestion & cleaning



Security



Analytics & ML



Sample of steps required

Configure access from analytics services



Ama

Dasht

Instar

Cluste

Query

Perfor

Snaps

Auton

Reserv

Subne

Param

Optio

Event

Event

Recon

<https://con>

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](#)

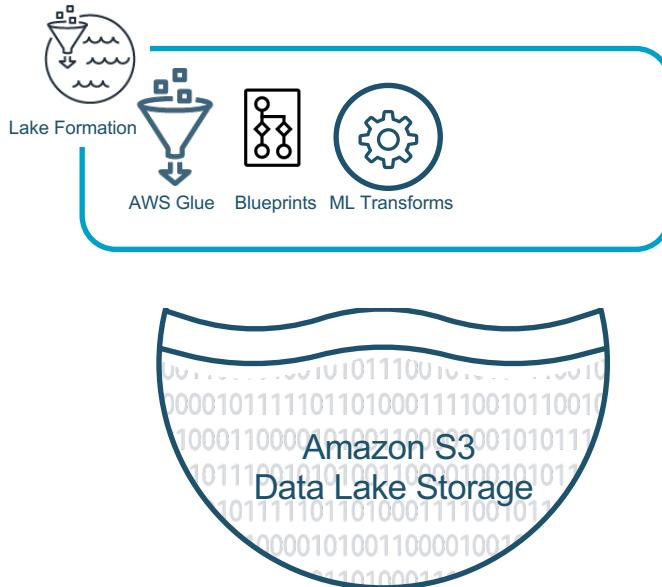
Lake Formation lets you
build secure data lakes in **days**

Built on Amazon S3 a robust data lake infrastructure



Cost effective, durable storage with global replication capabilities

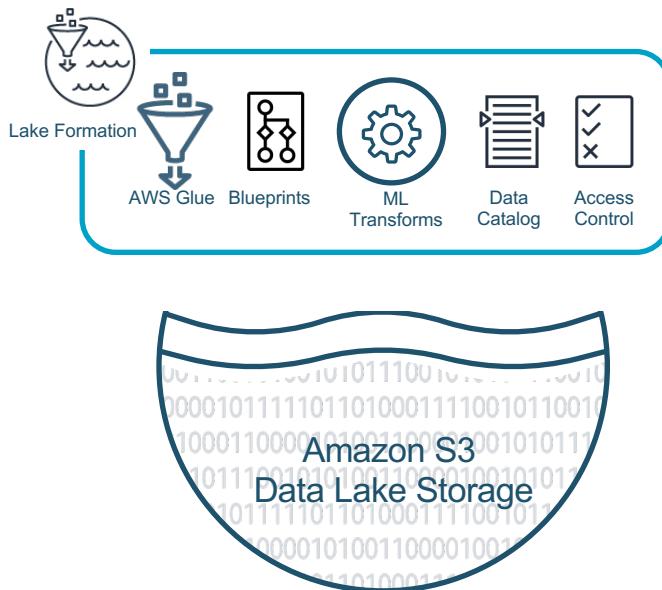
Automates manual, repetitive, low value tasks



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

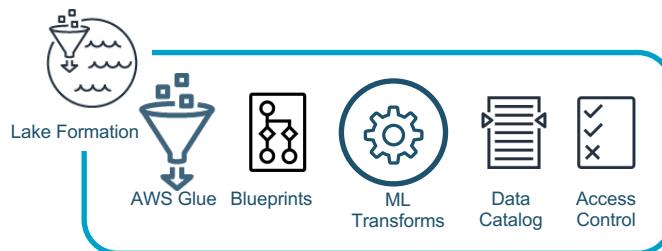
Simplified **ingest & cleaning** enables data engineers to build faster

Cost effective, durable storage with global replication capabilities

Enables all your data users



Comprehensive set of **integrated tools** enable every user equally



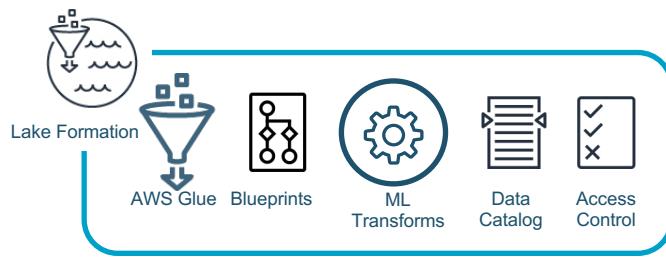
Centralized management of **fine grained permissions** empower security officers



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

Cost effective, durable storage with global replication capabilities

Fastest way to build secure data lakes



Enables all your users to run any **analytics workload**, at any scale, in a secure and cost-effective manner

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

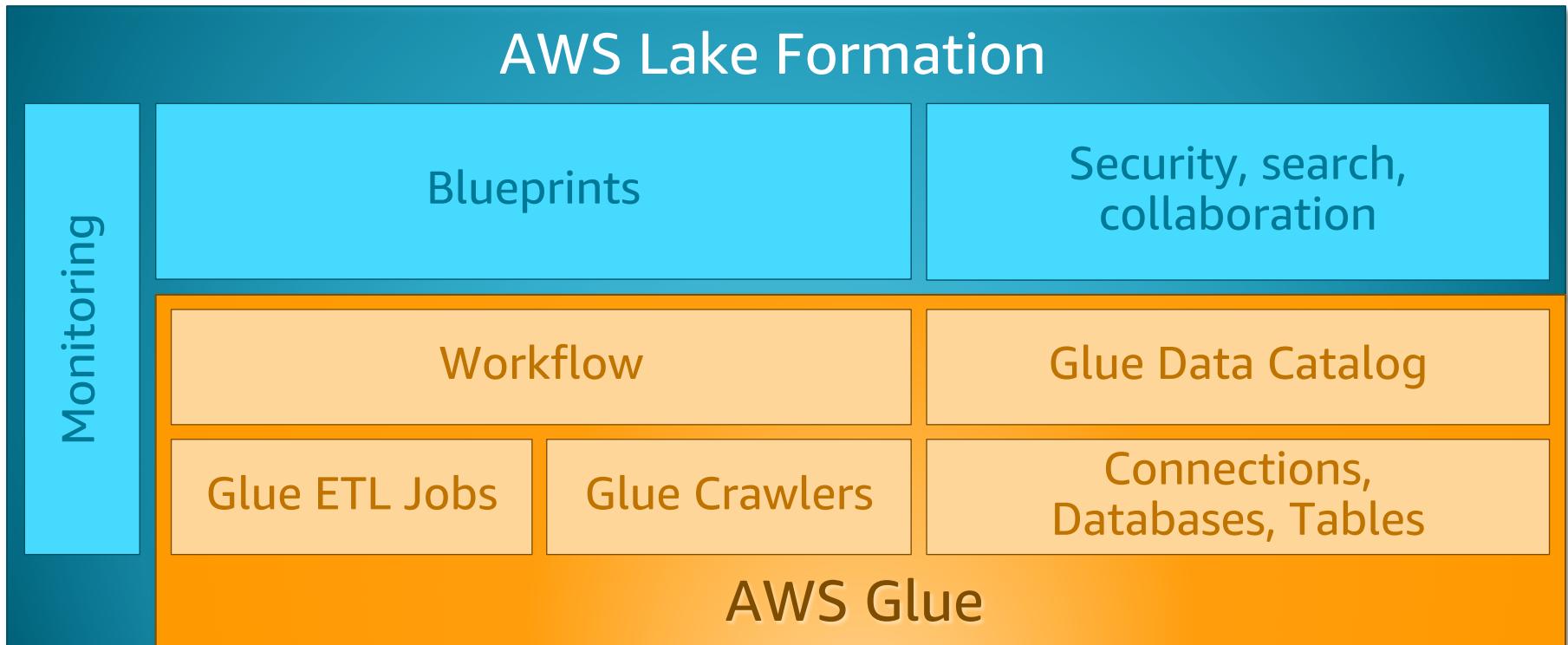


EMR



Athena

AWS Lake Formation is fully integrated w/ AWS Glue



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

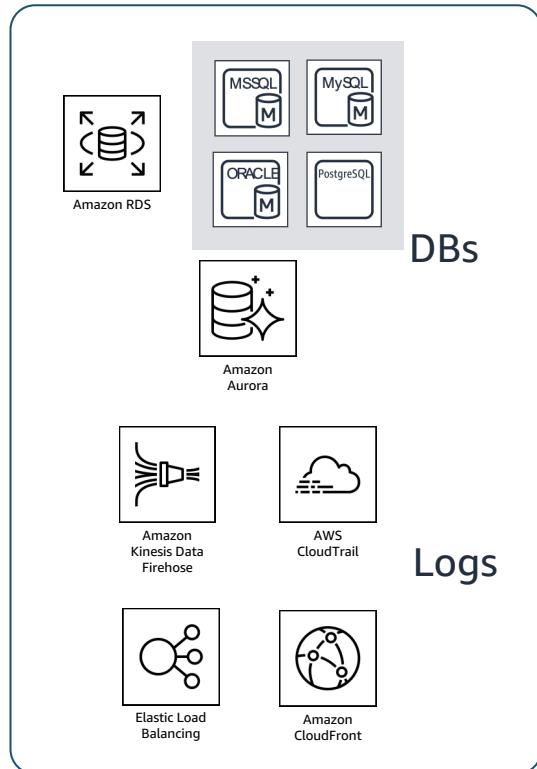
Deploy

Flexible workflows

Monitoring and alerting

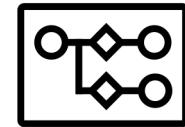
External integrations

Easily load data into your data lake w/ blueprints



Prebuilt templates to serve common ingestion use cases

Automatically build AWS Glue workflows



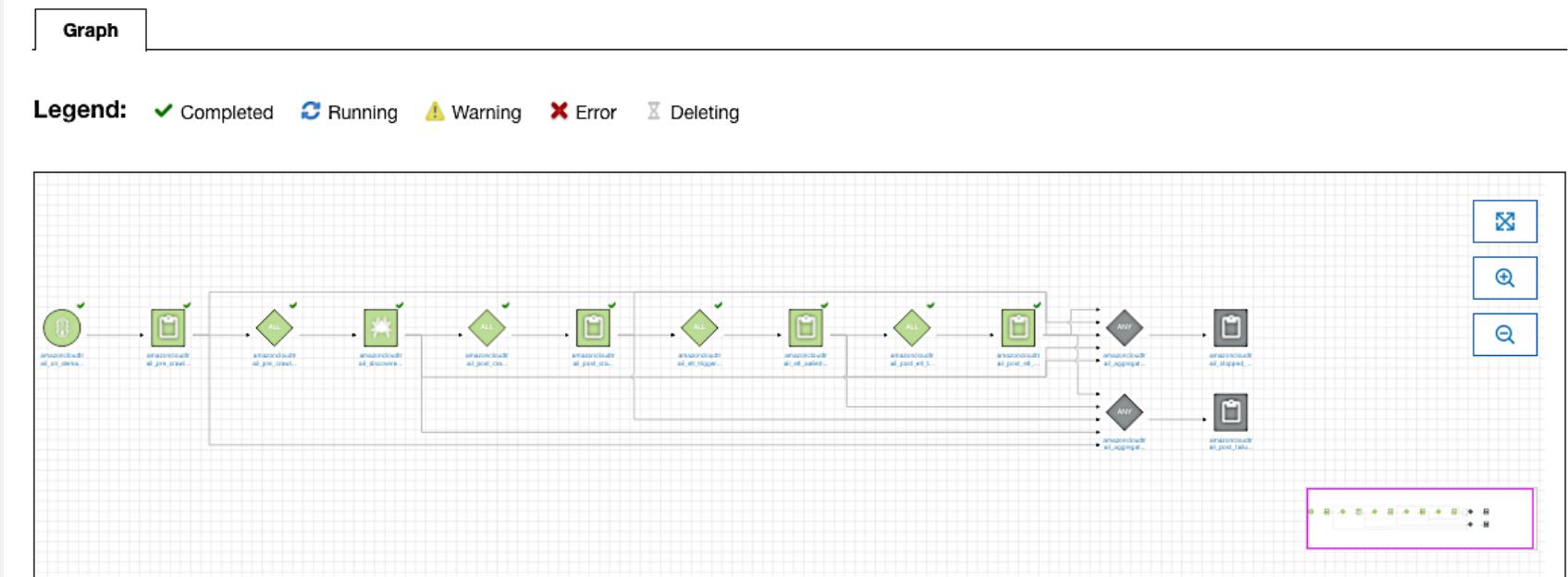
AWS Glue Workflows

AWS Glue jobs and crawlers discover, transform and structure data

Automatically populate the Data Catalog

Load data incrementally or in full

Blueprints create AWS Glue workflows



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

Securing data lakes with Lake Formation

Ingestion & cleaning



Data
Engineer

Security



Data Security
Officer

Analytics & ML



Data
Analyst

Serverless Spark

Data catalog

Comprehensive portfolio
of integrated tools

AWS Glue

Centralized permissions

Glue ML transformations

Blueprints

Real time monitoring

Integrated auditing



Redshift



Glue

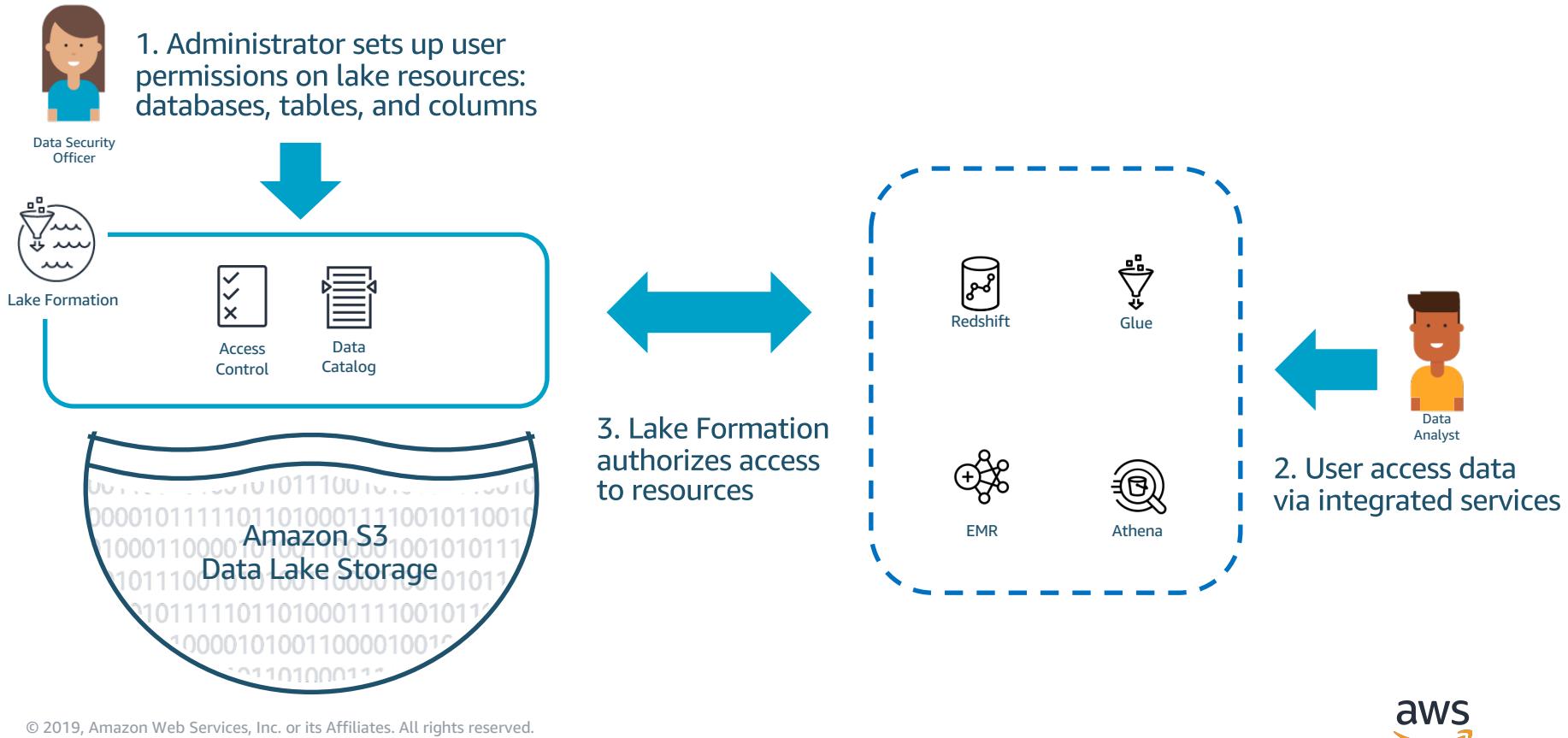


EMR



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



User 2

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 shows the AWS Lake Formation console with a search result for the database 'amazoncloudtrail'. A large blue callout box highlights the search bar and the results table, with the text 'Text-based search and filtering' overlaid. Another blue callout box highlights the 'View data' button in the Actions menu for one of the listed tables. A third blue callout box highlights the 'Run Query' button at the bottom of the page, with the text 'Query data in Amazon Athena' overlaid. The results table lists various tables with their names, locations, and classification. One row is selected, showing its details in a modal.

Name	Location	Classification
amazoncloudtrail_cloudtrail	s3://amazoncloudtrai...	PARQUET
_amazoncloudtrail_cloudtrail	s3://cloudtrail-awslog...	cloudtrail
v3workflow_cloudtrail	s3://cloudtrailjuly3/v...	PARQUET
_v3workflow_cloudtrail	s3://cloudtrail-awslog...	cloudtrail
chanujuly3_cloudtrail	mehulsdatabase...	PARQUET
_chanujuly3_cloudtrail	mehulsdatabase...	PARQUET
mhohsax_ct_cloudtrail	mehulsdatabase...	PARQUET
mhohsax_ct_clouptrail	mehulsdatabase...	PARQUET
cloudtrailjuly_cloudtrail	mehulsdatabase...	PARQUET
_cloudtrailjuly_cloudtrail	mehulsdatabase...	PARQUET
test6database_cloudtrail	mehulsdatabase...	PARQUET
_test6database_cloudtrail	mehulsdatabase...	PARQUET
dataengineer_cloudtrail	mehulsdatabase...	PARQUET
_dataengineer_cloudtrail	mehulsdatabase...	PARQUET
chanuwordpressv8_wordpres...	mehulsdatabase...	PARQUET
_temp_chanuwordpressv8_w...	mehulsdatabase...	PARQUET

Run Query Save As Format Query New Query (Run time: 1.82 seconds, Data scanned: 1.87KB)

Feedback English (US)

Results

eventversion	eventid	eventtime	sharedeventid	requestparameters.durationseconds
1	1.05	45419e0-5694-400b-90d8-3804f1bf163	b80d3b50-852b-44ba-9fb8-c752e38ac1e	3600
2	1.05	29278003-900c-42f9-9d51-70342981d	47927a6c-6499-4591-9ff6-25551687bd5	3600
3	1.05	d8614097-33f5-412a-8ba2-f515dd55bed	9373ab-5a41-4d65-9441-18818e224c	3600
4	1.05	6560d139-1180-4035-b329-2092e214d48	8584118f-0544-70e8-850a-0e055458341b	3600
5	1.05	c21882e4-6a02-4e31-8329-90c1b8a3206	23951758-d749-4497-8e68-d5e0bcb1b45	3600
6	1.05	410ebd7-a9b5-4215-9059-4e69e426989	6393b0e4-e852-4c0b-a370-a9f1025683e	3600
7	1.05	772dc02-0330-432d-8c7e-15baee5452e0	92925749-d9bc-1459-88ee-18d70e5e514	3600
8	1.05	c055ab9f-6123-4c2b-92d5-ecdf1cc47c	f2693a31-23714908-9a16-9a8d6776423	3600
9	1.05	643c185a-a335-41ef-a822-9ae8fb6500c	3add40f-d986-4668-ba97-4c10724064	3600
10	1.05	08563ad-d926-4593-aa0d-7edea101f1	c90cece1-497a-494d-a460-67a1719a463	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

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

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

Below the table, there are links for "Feedback", "English (US)", and copyright information: "© 2008 - 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use". A blue rounded rectangle highlights the "Recent access activity" section.

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



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 shows the AWS Athena Query Editor interface. The top navigation bar includes tabs for 'Athena' (selected), 'Query Editor', 'Saved Queries', 'History', 'AWS Glue Data Catalog', 'Workgroup: primary', and links for 'Settings', 'Tutorial', 'Help', and 'What's new'. There are four tabs open in the browser: 'AWS Lake Formation Console', 'AWS Lake Formation Console', 'Athena' (active), and 'Redshift - AWS Console'. The main area has a sidebar with 'Catalog' (Lake formation), 'Database' (amazoncloudtrail), and sections for 'Tables (1)' (amazoncloudtrail_cloudtrail) and 'Views (0)'. A query editor window titled 'New query 1' contains the SQL command: 'SELECT * FROM "amazoncloudtrail"."amazoncloudtrail_cloudtrail" limit 10;'. Below the editor are buttons for 'Run query', 'Save as', 'Create', and 'Format query'. The results section displays 10 rows of data from the 'amazoncloudtrail_cloudtrail' table, with columns labeled 'eventversion' and 'useridentity'. The data consists of 10 entries, each with a value of 1.05 for both columns. At the bottom of the results table, there is a truncated row starting with '10 1.05'.

Thanks

velosog@amazon.com

