Immersion Day –
Building a Data Lake on AWS

Adebimpe (Bims) Daniells

Nov 6, 2019

Learning Objectives

- Learn the steps to build Data Lake on AWS
- Perform Data Cataloguing, Querying & Visualization
- Understand how AWS analytics fit together
- Machine Learning Integration with Data Lake

aws

Workshop Agenda

1. Overview of the Data Lake
2. Hydrating the Data Lake
3. Lab – Hydrating the Data Lake
4. Working Within the Data Lake
5. Lab - Querying the Data Lake with Amazon Athena and Amazon QuickSight
6. Consuming the Data Lake – Reporting, Analytics, Machine Learning
7. Demo - Machine Learning with Amazon SageMaker

aws

Data is a strategic asset for every organization

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

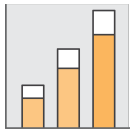
There is **more data** than people think

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

Data Scientists

Business Users



Analysts

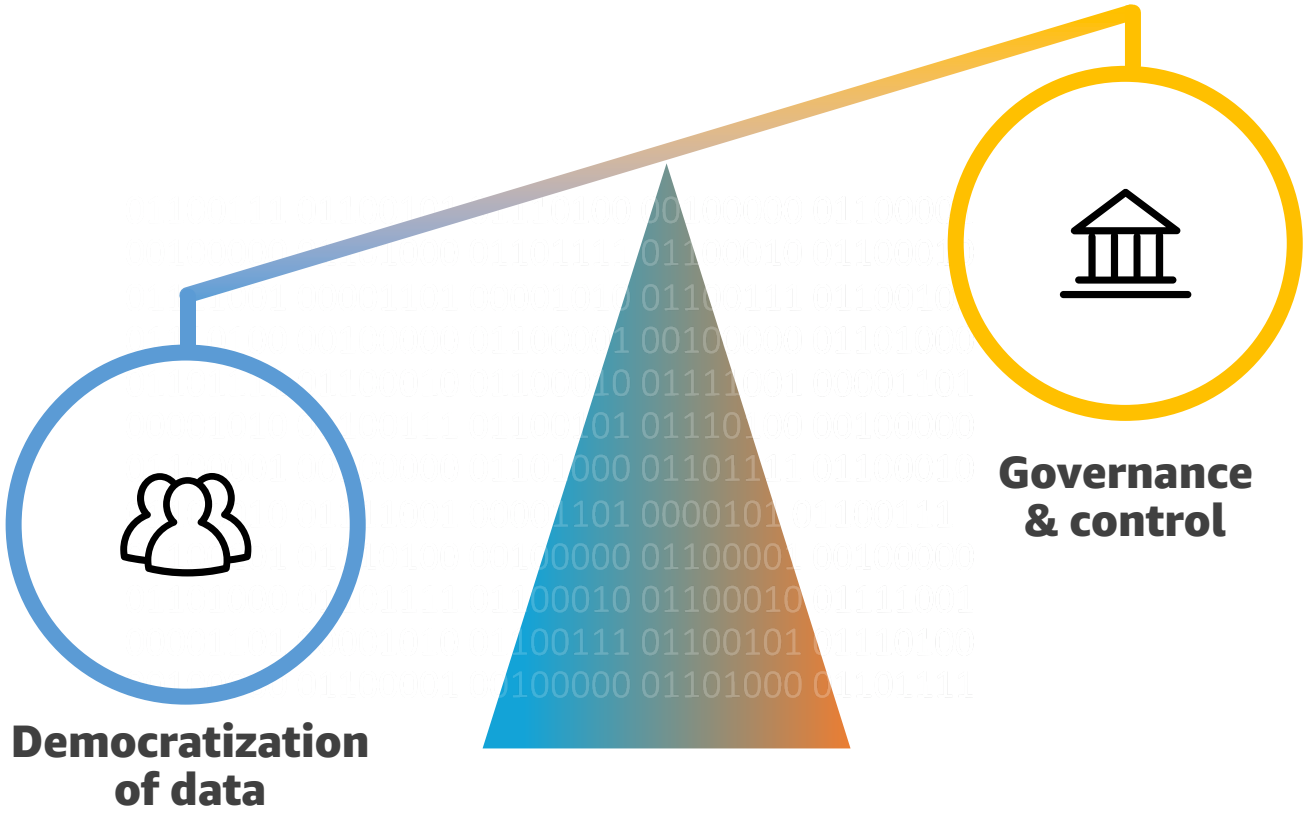
Applications

Secure **Real time**

Flexible **Scalable**

There are **more people** accessing data

And **more requirements** for making data available

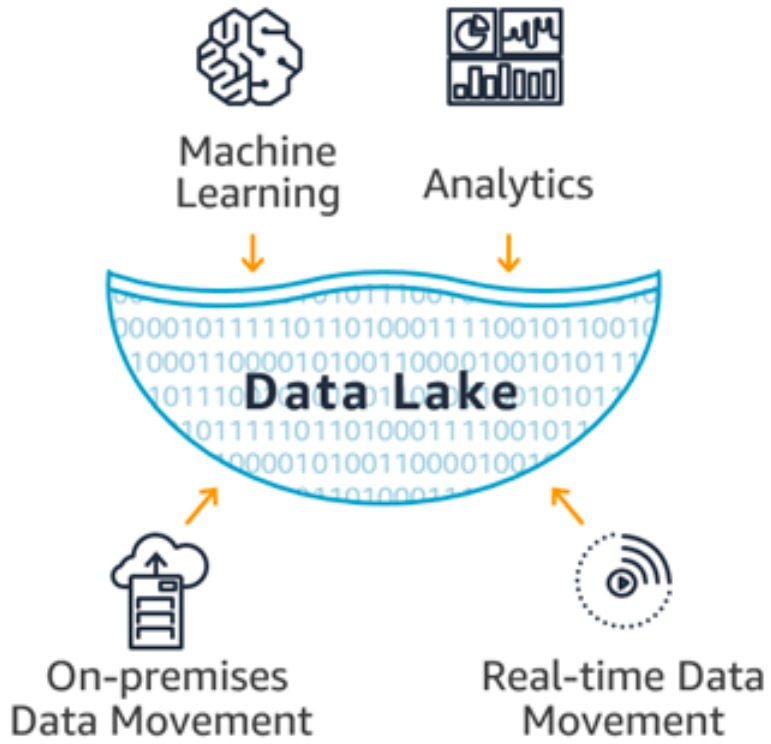


There are more people working with data than ever before

How do I provide democratized access to data to enable informed decisions while at the same time enforce data governance and prevent mismanagement of the data?

What is a Data Lake?

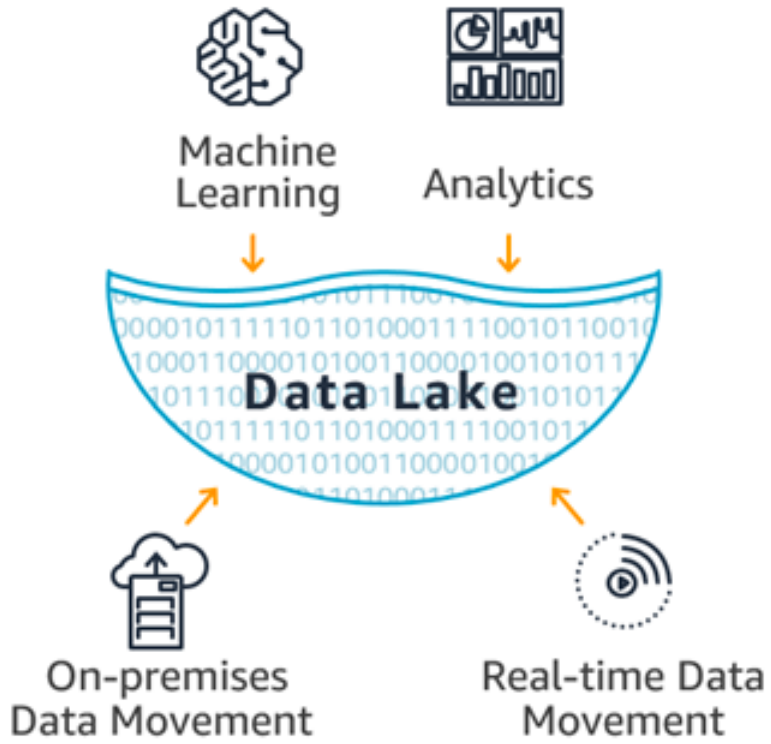
- A centralized repository for both structured and unstructured data at any scale
- Store data as-is in open-source file formats to enable direct analytics



aws

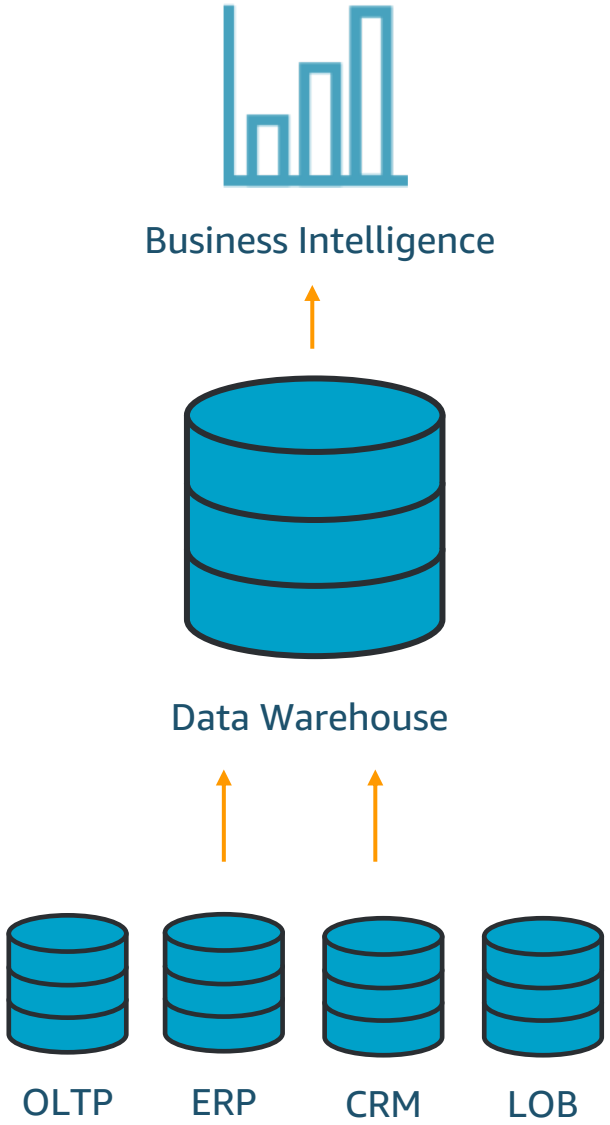
Why a Data Lake?

- 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



aws

Traditionally, Analytics Looked Like This



Relational Data

TBs-PBs Scale

Schema Defined Prior to Data Load

Operational and Ad Hoc Reporting

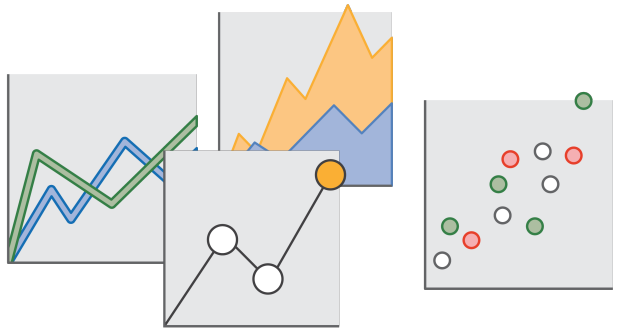
Large Initial Capex + $$K / TB/ Year

aws

Data lakes help you cost-effectively scale

- Extends or evolves their data warehouses
- Durable and available; exabyte scale
- Secure, compliant, auditable
- Run any type of analytics from DW to predictive
- Decoupling of compute and storage
- On-demand resources, tiering, cost choices

Benefits of a Data Lake – Schema on Read



"Is there a way I can apply multiple analytics and processing frameworks to the same data?"

A Data Lake enables ad-hoc analysis by applying schemas on read, not write.

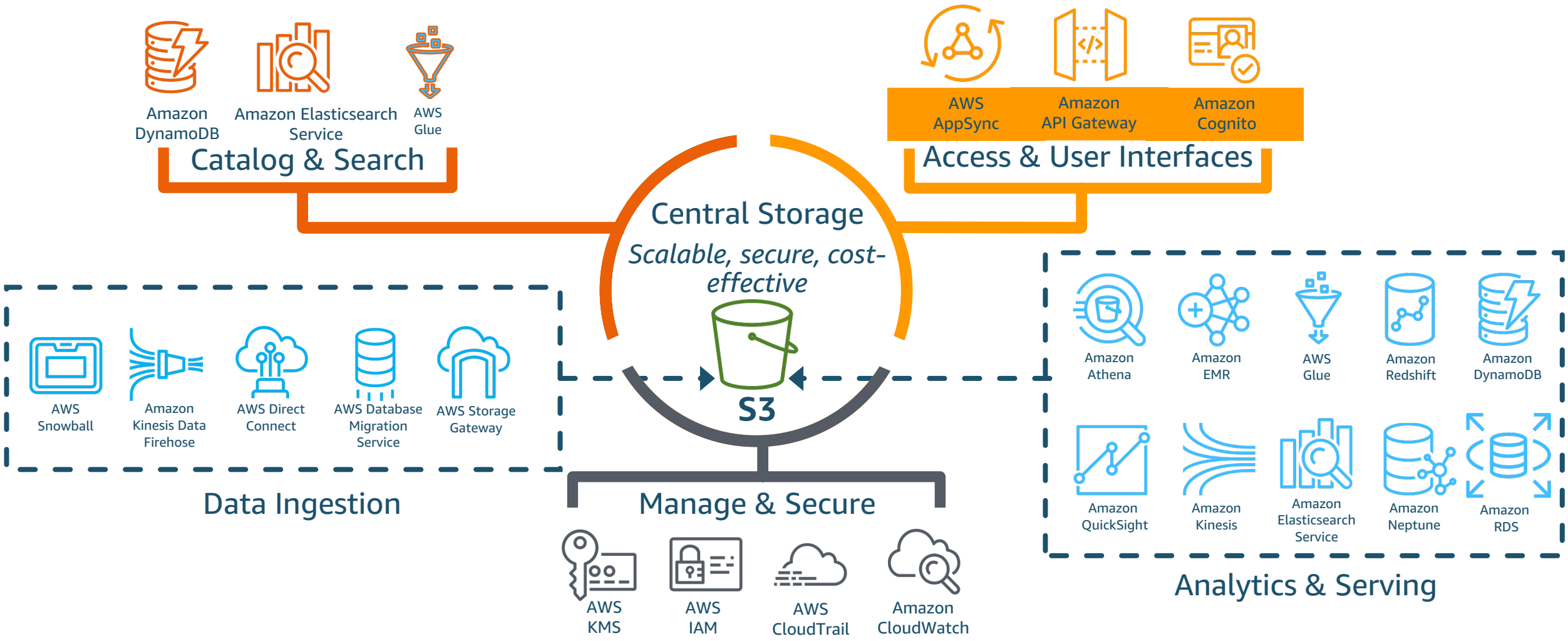
aws

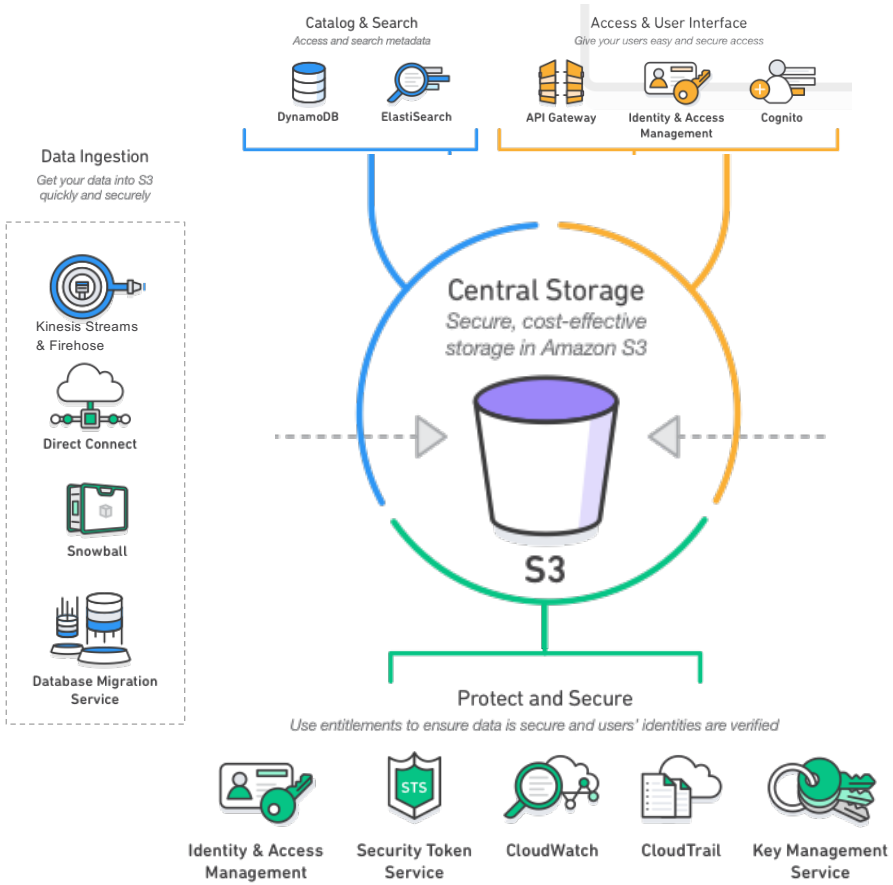
Building a Data Lake on AWS

Why AWS?

Implementing a Data Lake architecture requires a broad set of tools and technologies to serve an increasingly diverse set of applications and use cases.

Data Lake on AWS





Processing & Analytics

Real-time

- Elasticsearch Service
- Kinesis Data Analytics, Kinesis Data Streams
- Spark Streaming on EMR
- Apache Flink on EMR
- AWS Lambda
- Apache Storm on EMR

Analytics

- EMR Hadoop, Spark, Presto
- Athena Query Service
- Redshift Data Warehouse

AI & Predictive

- Amazon Lex Speech recognition
- Amazon Polly Text to speech
- Amazon Rekognition
- Machine Learning Predictive analytics
- SageMaker

Transactional & RDBMS

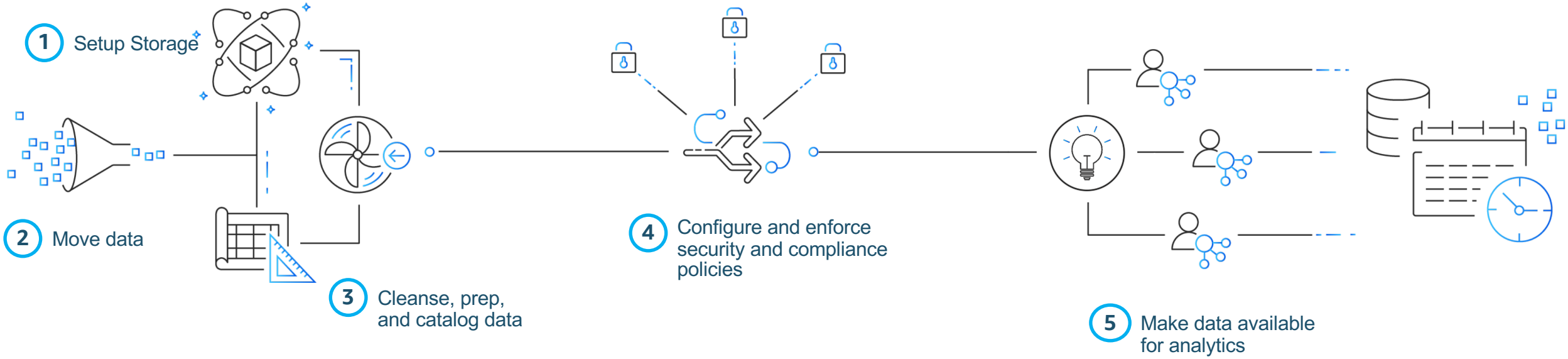
- DynamoDB NoSQL DB
- Aurora Relational Database

BI & Data Visualization

- Amazon QuickSight
- tableau
- MicroStrategy
- kibana

aws

Typical steps of building a data lake



Ways to move data into the Data Lake

Data movement from your datacenters

AWS Direct Connect
AWS Snowball
AWS Snowmobile
AWS Database Migration Service
AWS Storage Gateway

Data movement from real-time sources

AWS IoT Core
Amazon Kinesis Data Firehose
Amazon Kinesis Data Streams
Amazon Kinesis Video Streams

Data movement from on-premises datacenters

- Dedicated Network connection
- Secure appliances
- Ruggedized Shipping Container
- Database migration
- Gateway that lets applications write to the cloud

Data movement from real-time sources

- Connect devices to AWS
- Real-time data streams
- Real-time video streams

Why Amazon S3 for a Data Lake?

Durable

Designed for **11 9s** of durability

Available

Designed for **99.99**% availability

High performance

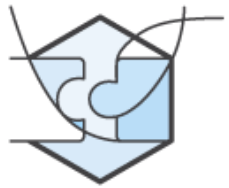
- Multiple upload
- Range GET

Easy to use

- Simple REST API
- AWS SDKs
- Read-after-create consistency
- Event notification
- Lifecycle policies

Scalable

- Store as much as you need
- Scale storage and compute independently
- No minimum usage commitments

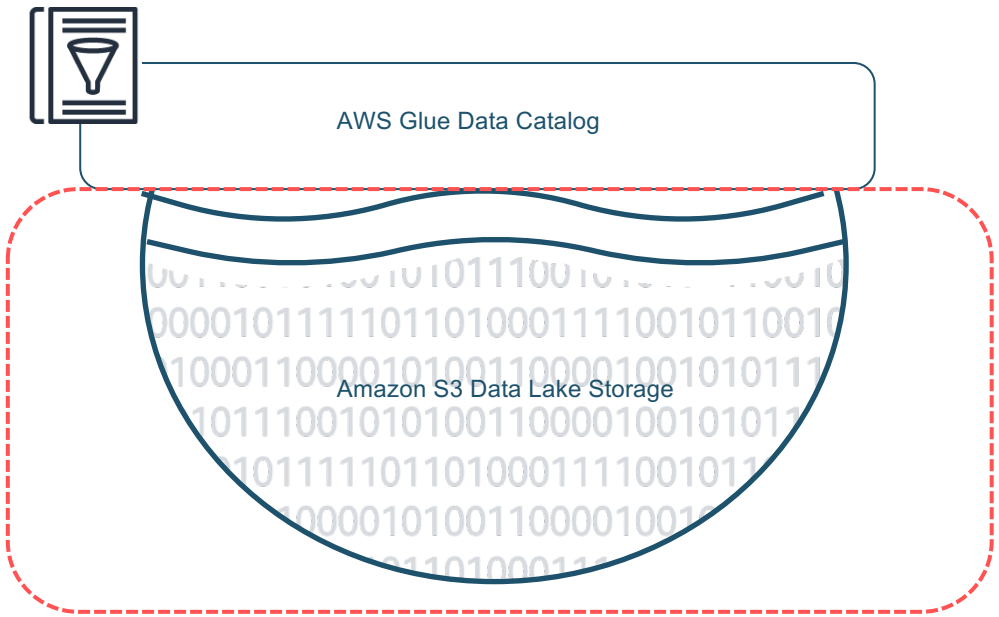


Integrated

- Amazon EMR
- Amazon Redshift
- Amazon DynamoDB
- Hive
- Spark
- Presto

aws

Secure data access



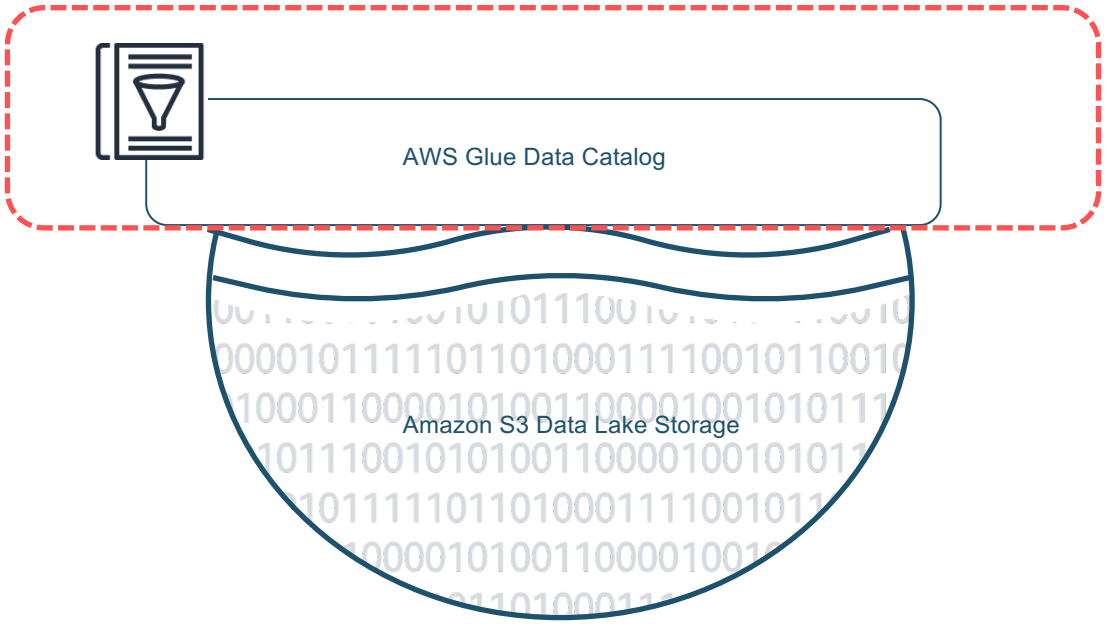
By default, all Amazon S3 buckets and objects are private

Secure data access with a combination of resource-based policies such as Amazon S3 bucket policies and IAM user policies

Use AWS KMS to enable client and server-side encryption for your data

Use object tagging (Classification = PHI) in conjunction with IAM

And secure the metadata access



Identity-based policies (AWS IAM policies)

- Managed by IAM
- Attached to IAM principals, e.g. IAM users, IAM roles

Resource-based policies

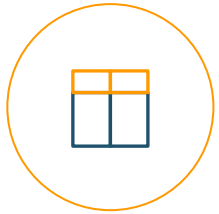
- Managed by AWS Glue
- One policy per account/catalog, similar to Amazon S3's bucket policies
- Required to allow cross account

Common data categories and use cases

Relational

Referential integrity, ACID transactions, schema-on-write

Lift and shift, ERP, CRM, finance



Key-value

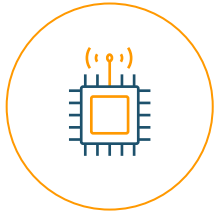
High throughput, low-latency reads and writes, endless scale

Real-time bidding, shopping cart, social, product catalog, customer preferences

Document

Store documents and quickly access querying on any attribute

Content management, personalization, mobile



In-memory

Query by key with microsecond latency

Leaderboards, real-time analytics, caching

Graph

Quickly and easily create and navigate relationships between data

Fraud detection, social networking, recommendation engine

Time-series

Collect, store, and process data sequenced by time

IoT applications, event tracking

Ledger

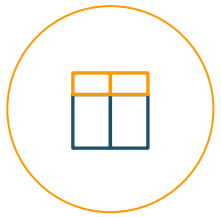
Complete, immutable, and verifiable history of all changes to application data

Systems of record, supply chain, health care, registrations, financial

Common use cases

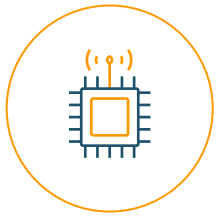
Common data categories and use cases

Relational



Key-value

Document



In-memory

Graph

Time-series

Ledger

RDS

DynamoDB

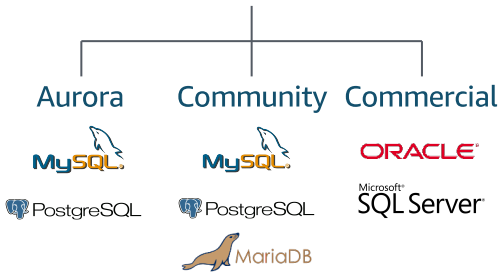
DocumentDB

ElastiCache

Neptune

Timestream

QLDB



aws

When to use which service?

Situation	Solution
Existing application	Use your existing engine on RDS • MySQL → Amazon Aurora, RDS for MySQL • PostgreSQL → Amazon Aurora, RDS for PostgreSQL • MariaDB → Amazon Aurora, RDS for MariaDB • Oracle → Use SCT to determine complexity → Amazon Aurora, RDS for Oracle • SQL Server → Use SCT to determine complexity → Amazon Aurora, RDS for SQL Server
New application	• If you can avoid relational features → DynamoDB • If you need relational features → Amazon Aurora
Fraud detection, Knowledge graphs, Network / IT Operations	• Amazon Neptune
In-memory store/cache	• Amazon ElastiCache
Time series data	• Amazon Timestream
Track every application change, crypto verifiable. Have a central trust authority	• Amazon Quantum Ledger Database (QLDB)
Don’t have a trusted central authority	• Amazon Managed Blockchain
Data Warehouse & BI	• Amazon Redshift, Amazon Redshift Spectrum, and Amazon QuickSight
Adhoc analysis of data in S3	• Amazon Athena and Amazon QuickSight
Apache Spark, Hadoop, HBase (needle in a haystack type queries)	• Amazon EMR

Building data lakes can still take months

Build a secure data lake in days with AWS Lake Formation

Move, store, catalog, and clean your data faster

Move, store, catalog, and clean your data faster with Machine Learning

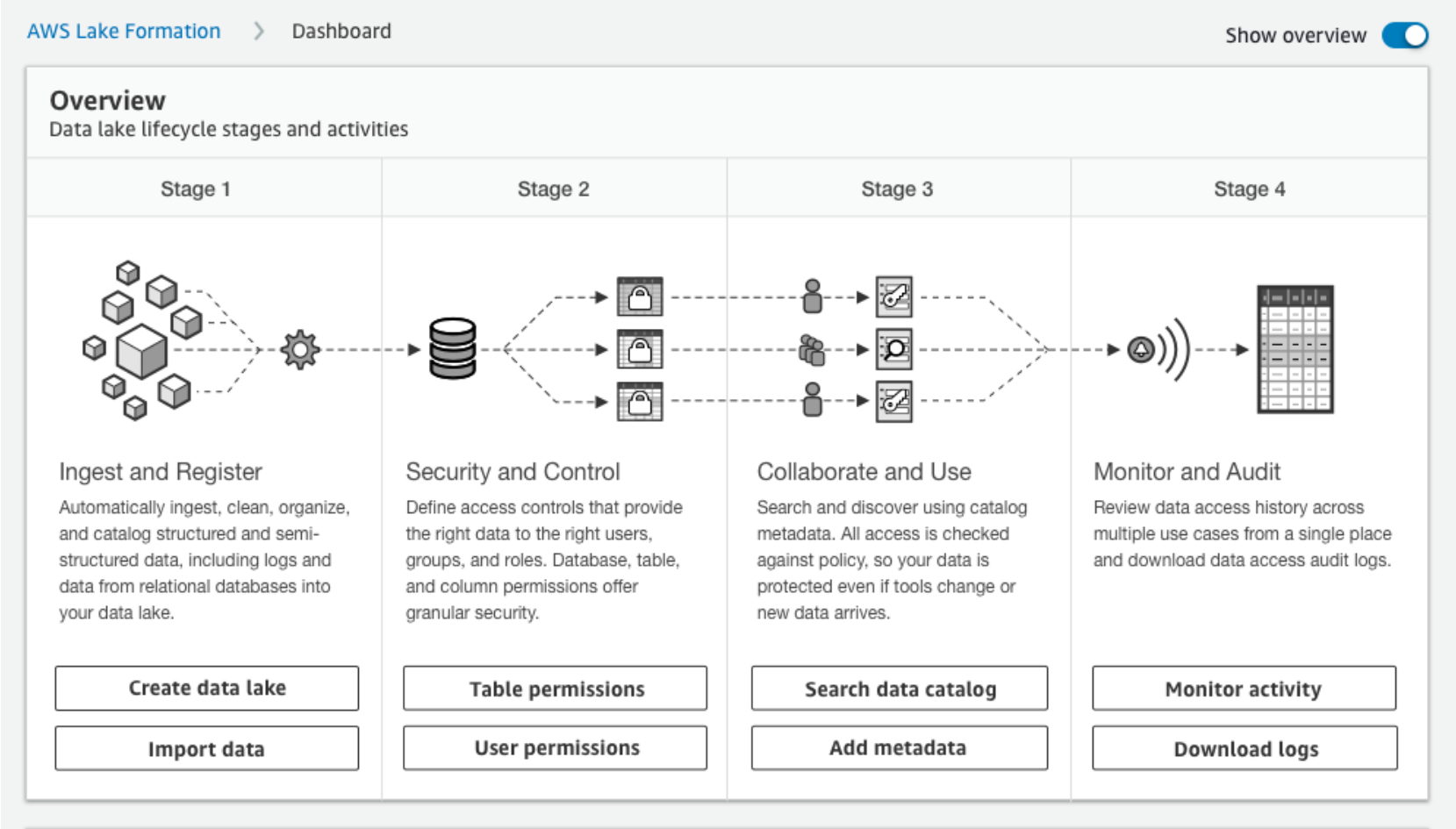
Enforce security policies across multiple services

Enforce security policies across multiple services

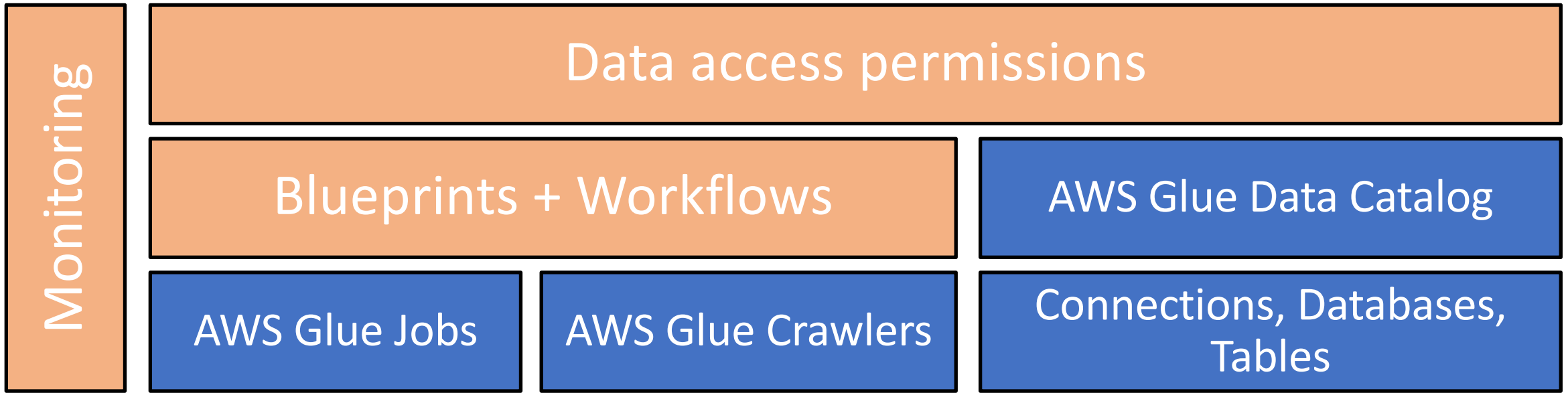
Gain and manage new insights

Empower analyst and data scientist to gain and manage new insights

How it works



AWS Lake Formation builds on AWS Glue capabilities



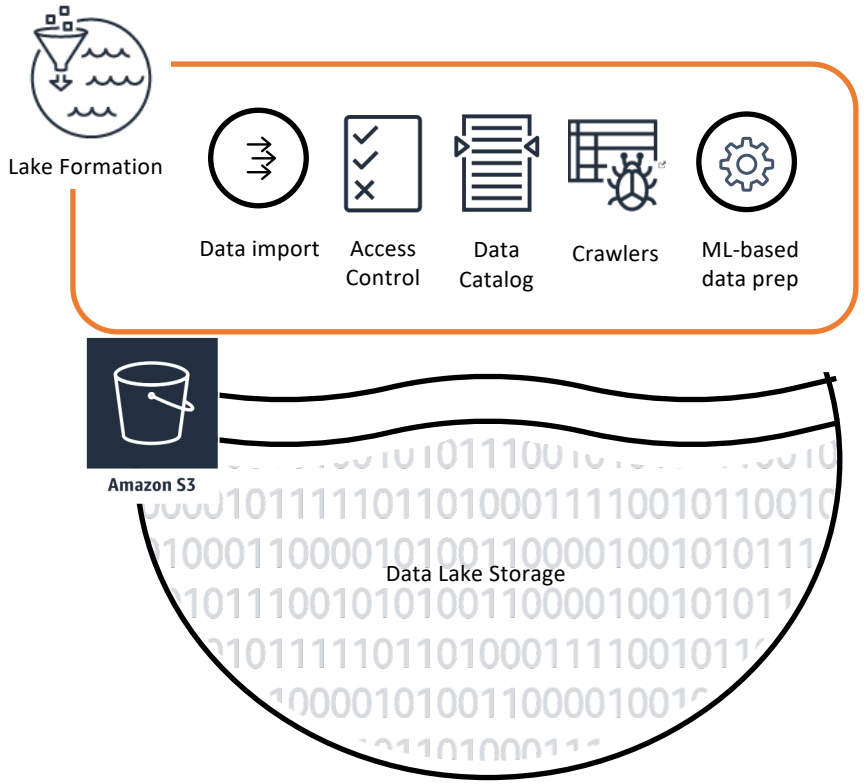
Register existing data or import new

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.



With blueprints

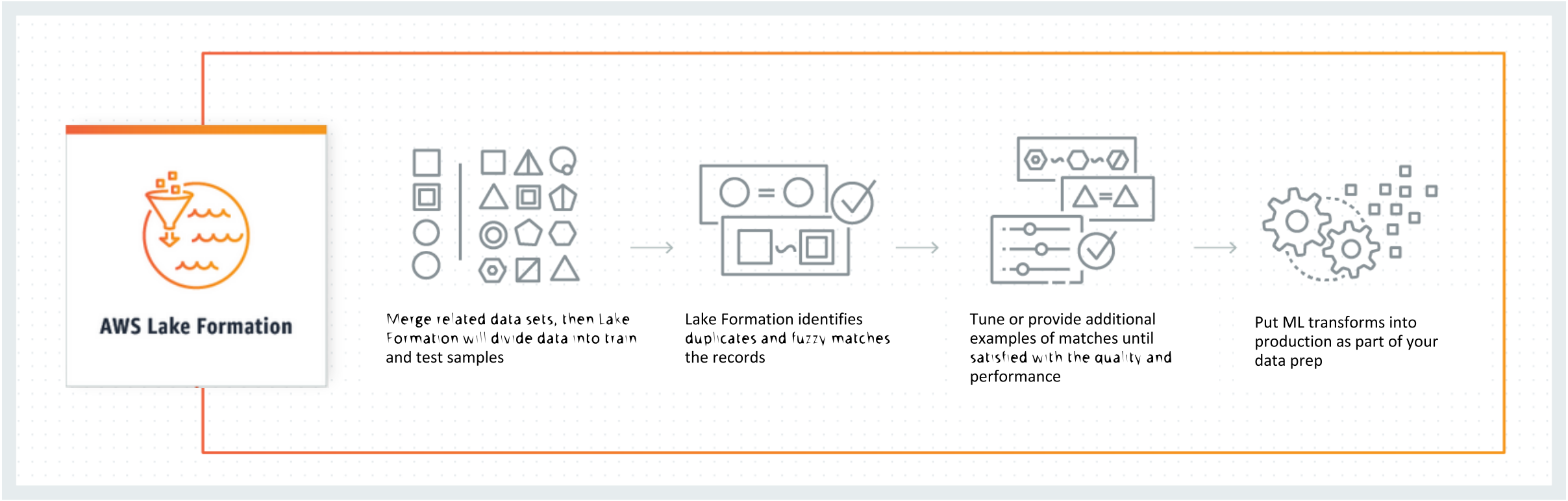
You

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

Easily de-duplicate your data with ML transforms



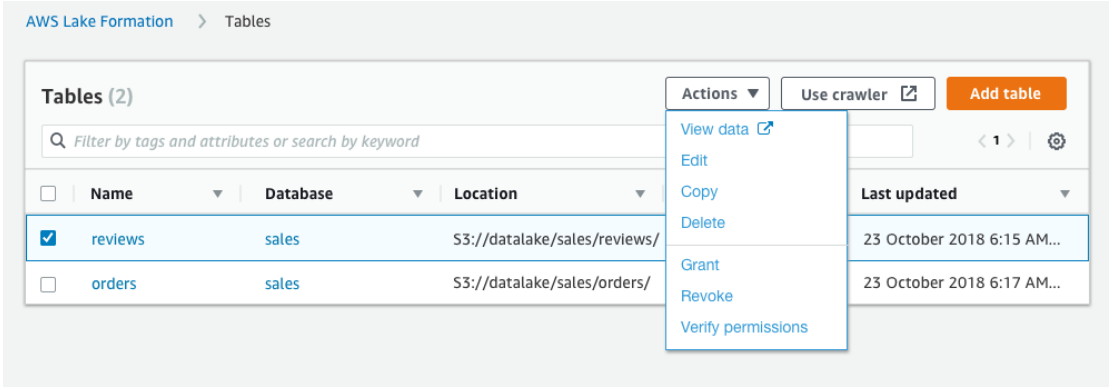
Security permissions in Lake Formation

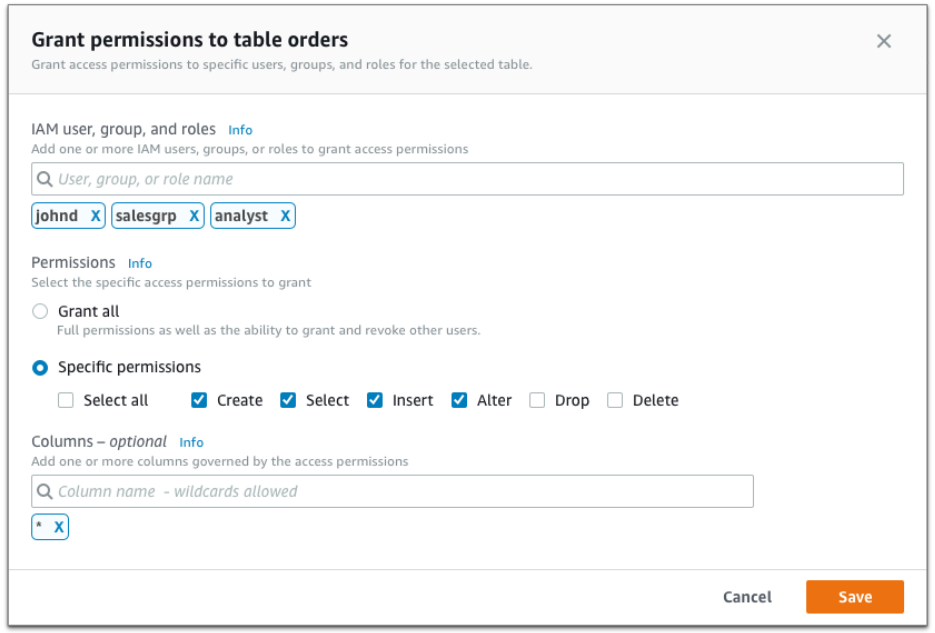
Control data access with simple grant and revoke permissions

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

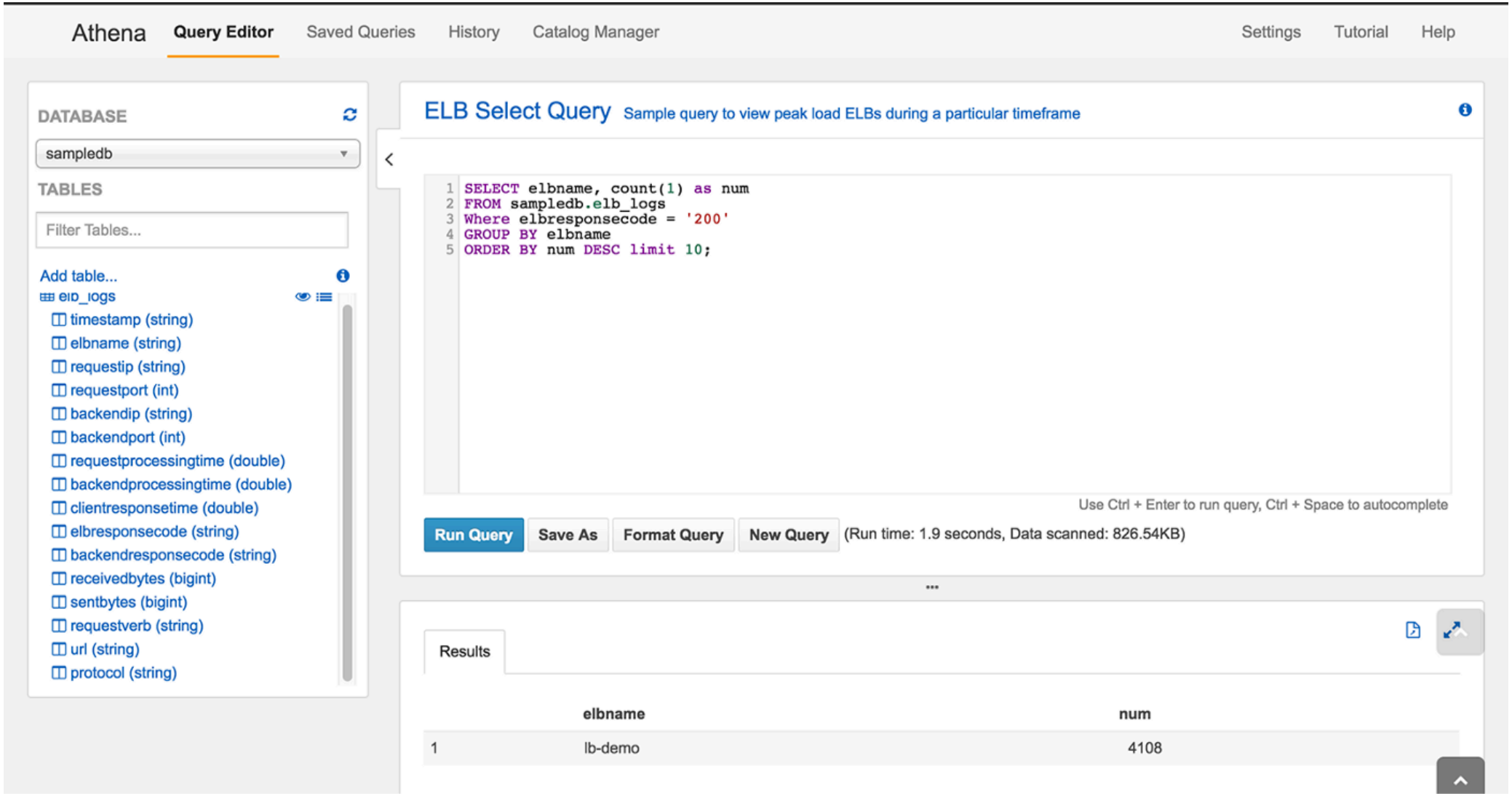




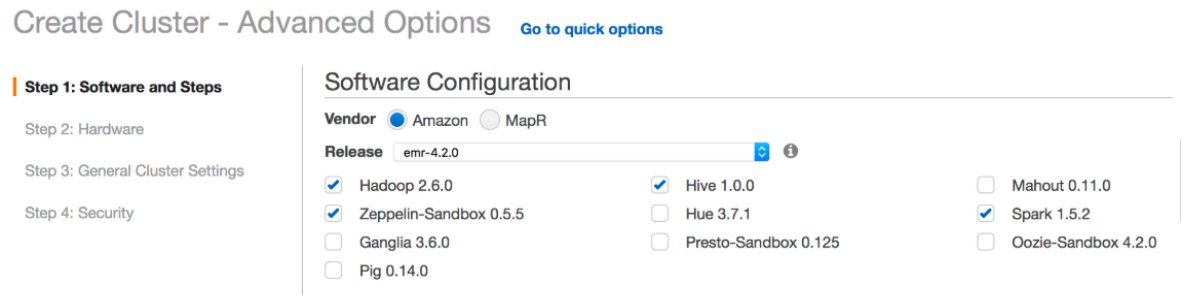
What can you do with a Data Lake?

aws

Query Directly with Amazon Athena



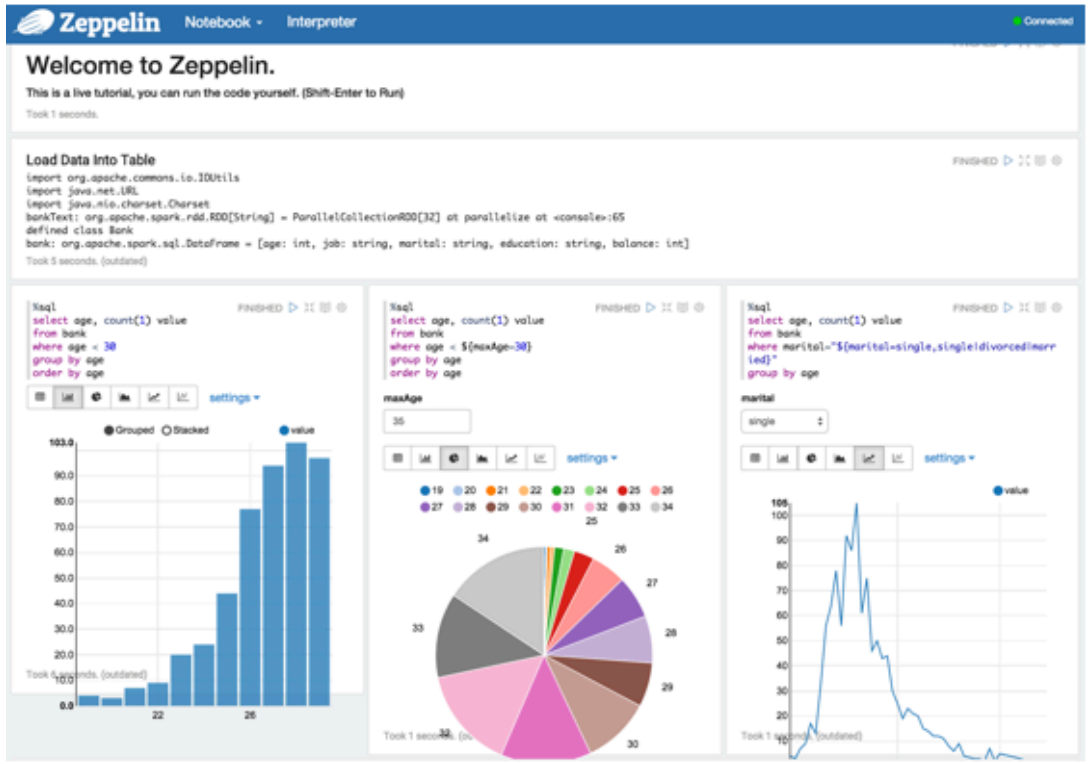
Analyze with Hadoop on Amazon EMR



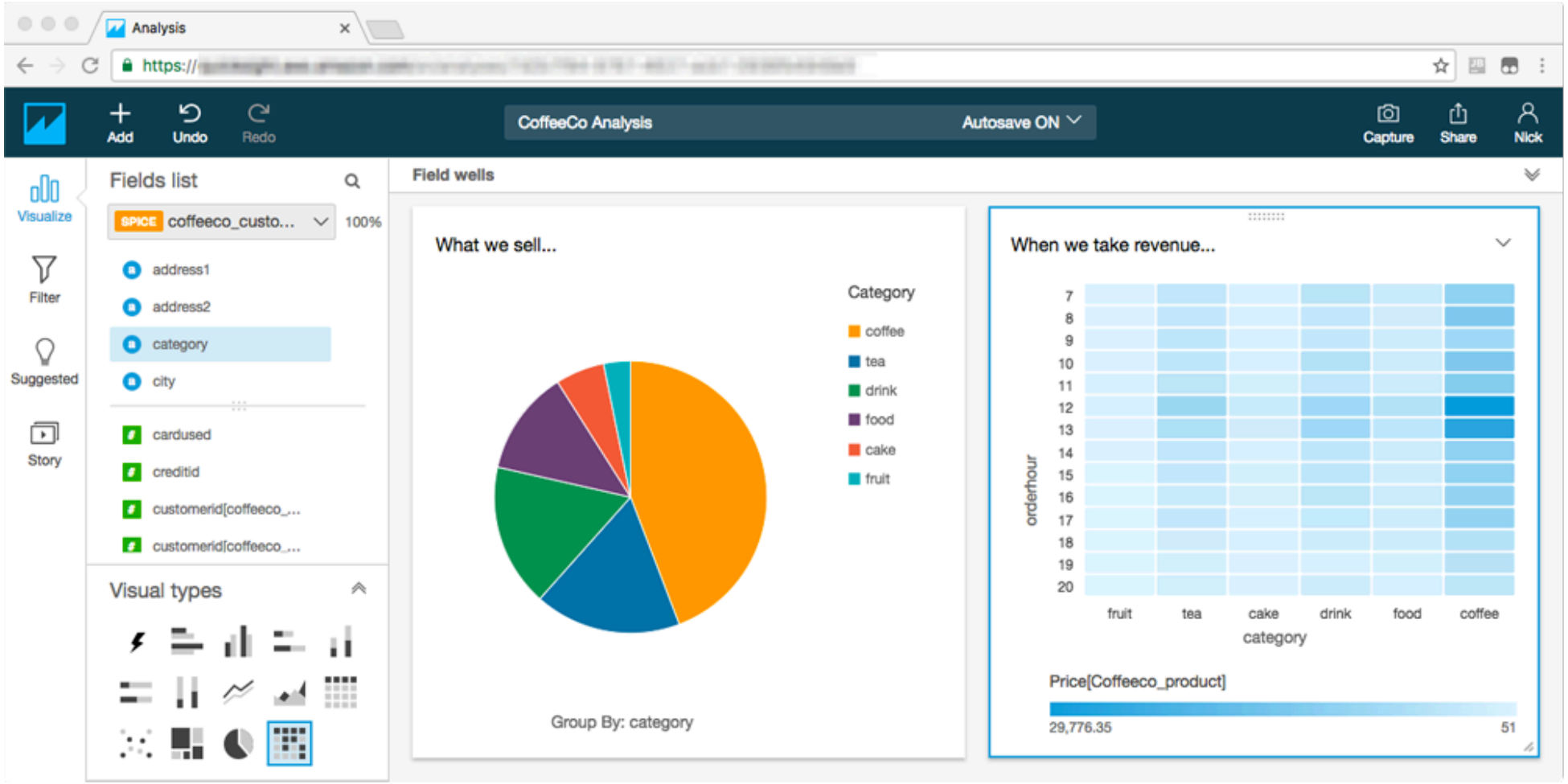

```scala
val movieLensHomeDir = "s3://emr.examples/movieLens/"

val movies = sc.textFile(movieLensHomeDir + "movies.dat").map { line =>
  val fields = line.split("::")
  // format: (movieId, movieName)
  (fields(0).toInt, fields(1))
}.collect.toMap

val ratings = sc.textFile(movieLensHomeDir + "ratings.dat").map { line =>
  val fields = line.split("::")
  // format: (timestamp % 10, Rating(userId, movieId, rating))
  (fields(3).toLong % 10, Rating(fields(0).toInt, fields(1).toInt, fields(2).toDouble))
}
```

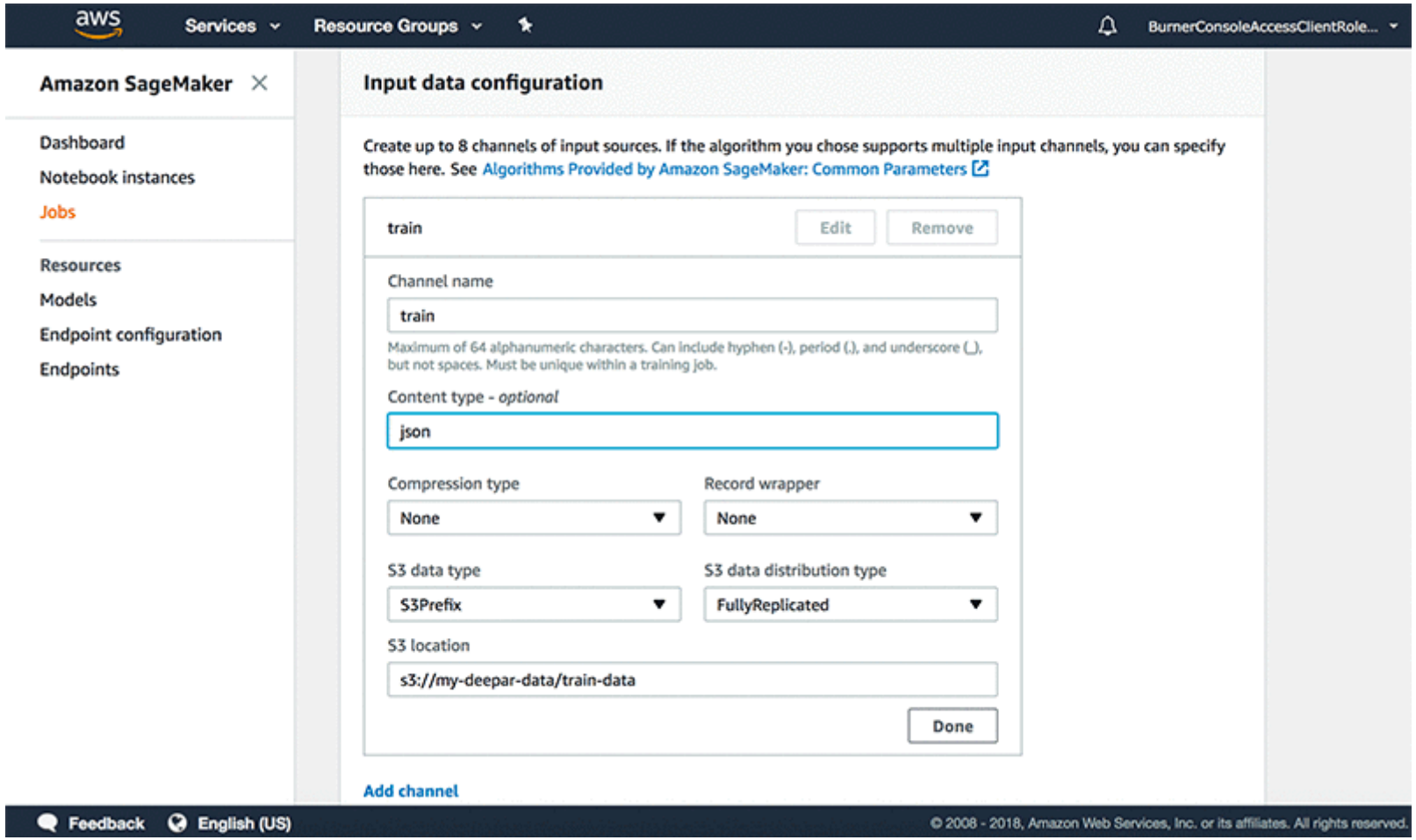


Create Visualizations with Amazon QuickSight

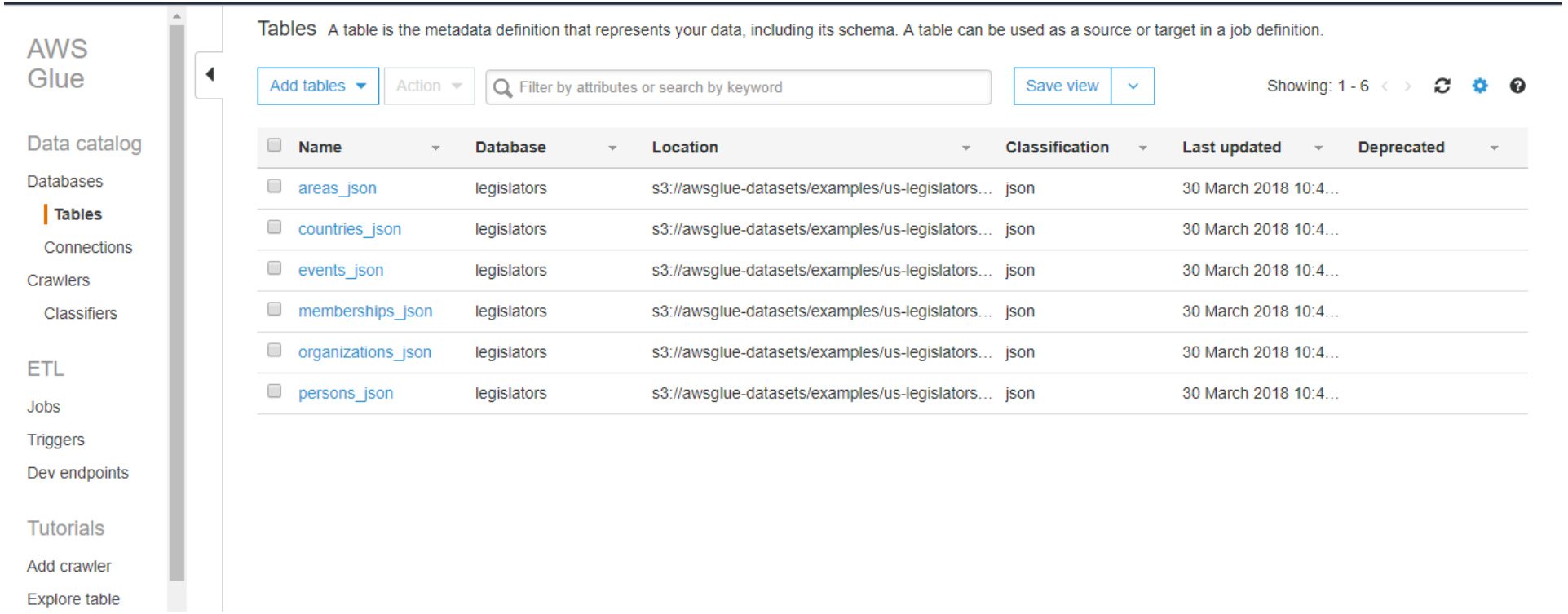


aws

Train ML Models with Amazon SageMaker



Create a Central Data Catalog with AWS Glue



AWS databases and analytics
Broad and deep portfolio, built for builders

Business Intelligence & Machine Learning

- **Amazon QuickSight**
- **Amazon SageMaker**
- **Amazon Comprehend**
- **Amazon Rekognition**
- **Amazon Lex**
- **Amazon Transcribe**
- **AWS DeepLens**

AWS Marketplace
250+ solutions

Databases

- **QLDB** NEW — Ledger Database
- **ElastiCache** — Redis, Memcached
- **Aurora** — MySQL, PostgreSQL
- **RDS** — MySQL, PostgreSQL, MariaDB, Oracle, SQL Server
- **Neptune** — Graph
- **DynamoDB** — Key value, Document
- **Timestream** NEW — Time Series
- **RDS on VMWare** NEW

Analytics

- **Amazon Redshift** — Data warehousing
- **Athena** — Interactive analytics
- **Amazon EMR** — Hadoop + Spark
- **Kinesis Analytics** — Real-time
- **Amazon Elasticsearch service** — Operational Analytics

Blockchain

- **Managed Blockchain** NEW
- **Blockchain Templates**

730+ Database solutions

600+ Analytics solutions

25+ Blockchain solutions

Data Lake

- **S3/Amazon Glacier**
- **Lake Formation** NEW — Data Lakes
- **AWS Glue** — ETL & Data Catalog

20+ Data lake solutions

Data Movement

Database Migration Service | Snowball | Snowmobile | Kinesis Data Firehose | Kinesis Data Streams | Data Pipeline | Direct Connect

30+ solutions

Kansas City, Missouri, Spurs Growth with Smart City Analytics on AWS

Challenge

With IoT devices generating data around the clock—plus more than 4,200 existing datasets—Kansas City, Missouri needed a solution to turn data into tangible outcomes.

Solution

The city engaged Xaqt, which provides an analytics and intelligence platform on AWS that helps policymakers better innovate how cities function.

Benefits

- Achieved expected road-maintenance cost savings of up to 50%
- Identified vacant buildings with 85% accuracy
- Ingests one million events per day and growing

> "The Xaqt solution built on AWS enables us to perform preventive roadwork. We expect to save up to 50% compared to emergency repair."
>
> **Bob Bennett**, Chief Innovation Officer, City of Kansas City, Missouri

Company:	City of Kansas City, Missouri
Industry:	Government & Public Services
Country:	United States
Employees:	4,800
Website:	www.kcmo.gov

About Kansas City, Missouri

With a population of 2.1 million, Kansas City, Missouri, is a fast-growing city located in the heart of the United States. In the downtown area, the city has seen more than $2.1 billion in development over the past 10 years.

Thank You!

adedan@amazon.com

In God we trust, every other must bring data….

aws