Immersion Day – Building a Data Lake on AWS

Rob Whelan, 2nd Watch, Practice Manager, Data Engineering
Ian Willoughby, 2nd Watch, Chief Architect
Hemant Javeri, 2nd Watch, Solutions Architect

Housekeeping Items:

Questions?	To ask a question, click the chat feature
Labs	You will receive an email with a link, login & password to use in today's & tomorrows labs
Uber Eats	During the event you will also receive your Uber Eats gift card via email
Survey	Don't forget to leave feedback using the survey link in the post webinar email
Presentation Slides	A copy of the presentation will be provided to all attendees

Breadth & Depth with Cloud

- Original AWS Premier Consulting Partner
- VMware Master Services Competency
- Managed Service Partner
- 10[redacted]nsul

RELEVANT COMPETENCIES

- Migration
- DevOps
- Windows Workloads
- Financial Services
- Managed Services

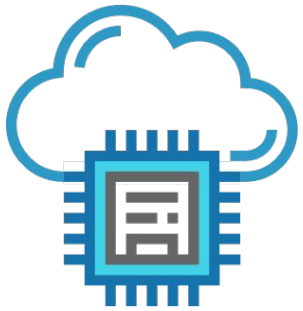
Your path to the cloud is different from everyone else's:

IT'S YOURS

We custom-tailor cloud strategies that:

Address people, process & technology

Leverage scalable processes & tools



Apply our unmatched enterprise experience to your challenges

Let you move with the agility of a start-up

Presenters

Rob Whelan
2nd Watch
Practice Manager - Data Engineering

Ian Willoughby
2nd Watch
Chief Architect

Hemant Javeri
2nd Watch
Solutions Architect

Workshop Agenda Day 1

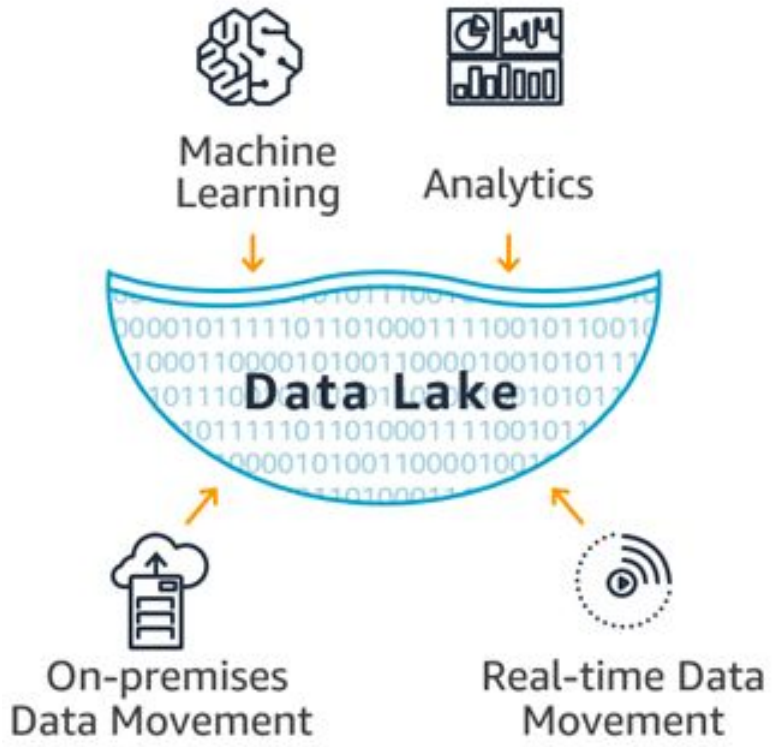
1. Overview of the Data Lake
2. Hydrating the Data Lake
3. Lab – Hydrating the Data Lake with Kinesis Lab
4. Lab – Hydrating the Data Lake with DMS Lab

Workshop Agenda Day 2

1. Working Within the Data Lake
2. Lab – Transforming Data in the Data Lake with Glue Lab
3. Lab - Querying the Data Lake with Amazon Athena and Amazon QuickSight
4. Consuming the Data Lake – Reporting, Analytics, Machine Learning
5. Demo - Machine Learning with Amazon SageMaker

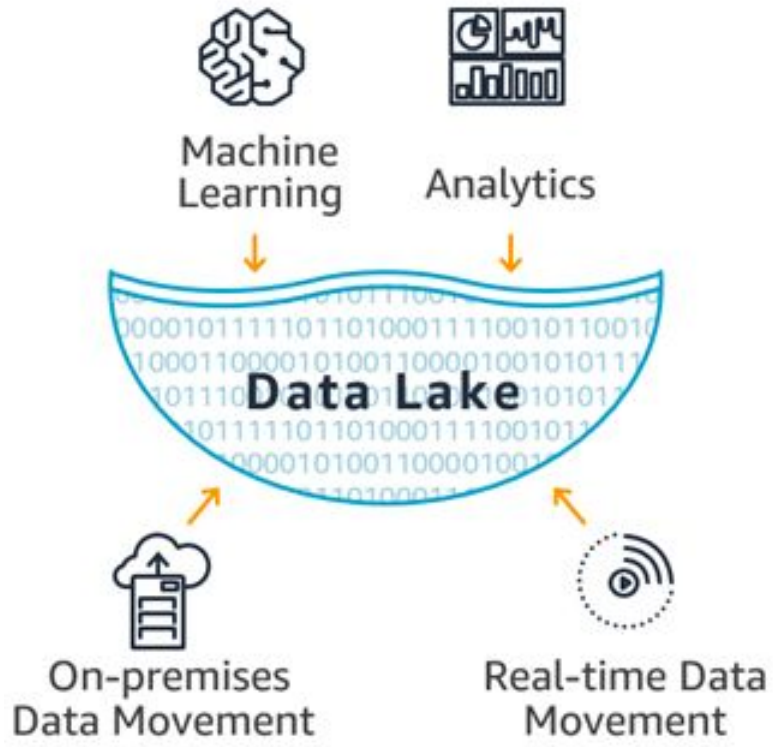
What is a Data Lake?

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

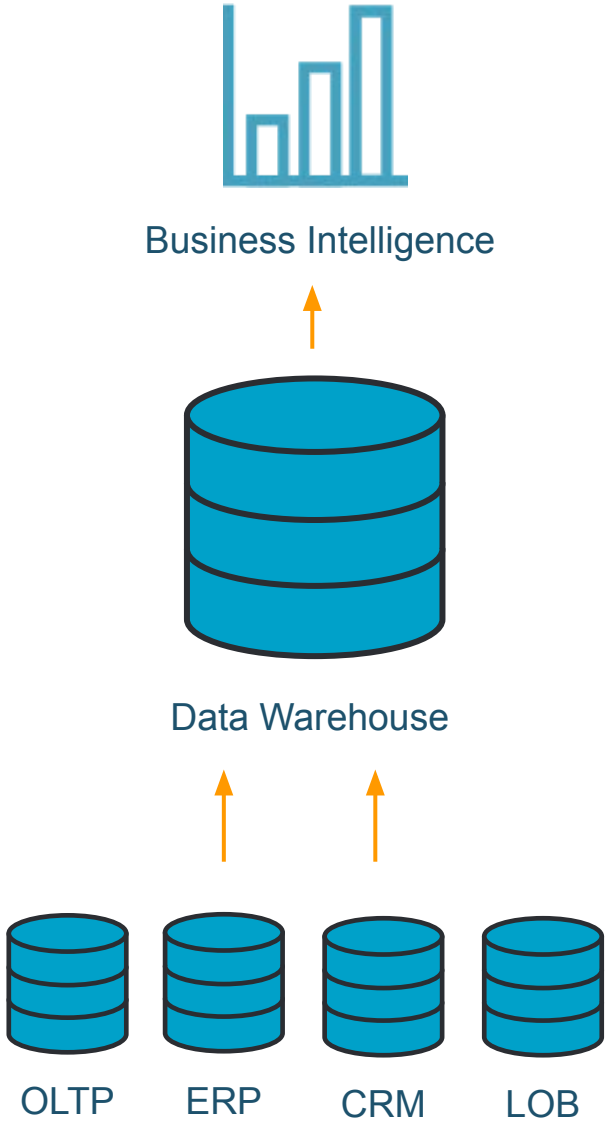


Why a Data Lake?

- 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



Traditionally, Analytics Looked Like This



Relational Data

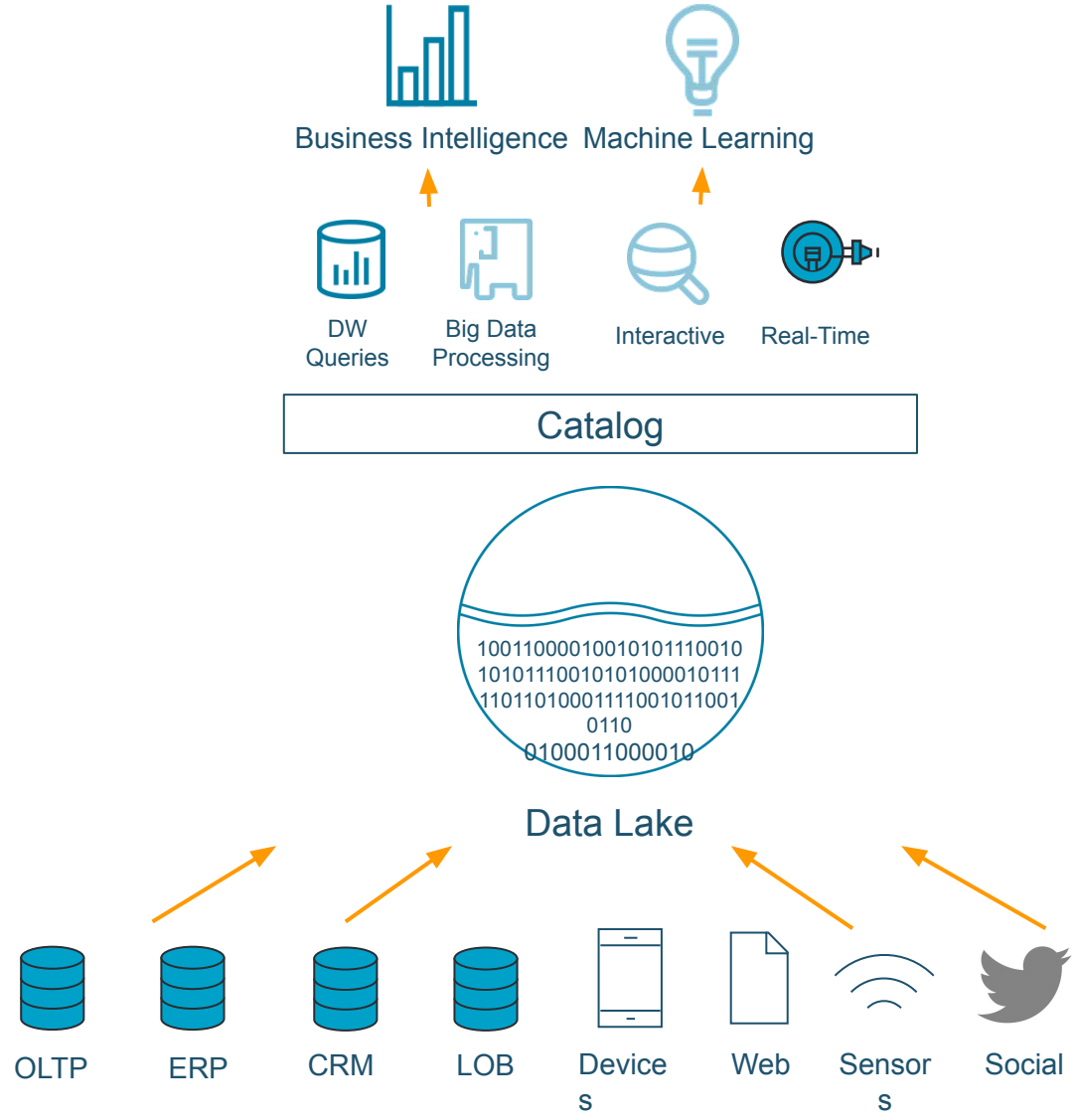
TBs-PBs Scale

Schema Defined Prior to Data Load

Operational and Ad Hoc Reporting

Large Initial Capex + $$K / TB/ Year

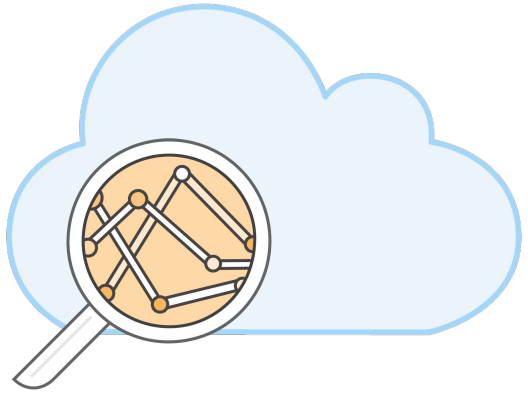
Data Lakes Extend the Traditional Approach



- TB-EBs Scale
- All Data in one place, a Single Source of Truth
- Relational and Non-Relational Data
- Decouples (low cost) Storage and Compute
- Schema on Read
- Diverse Analytical Engines

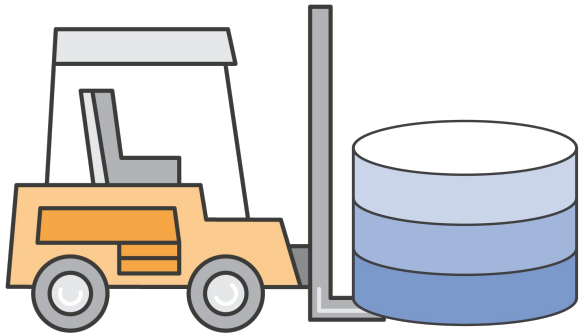
Benefits of a Data Lake – All Data in One Place

"Why is the data distributed in many locations? Where is the single source of truth ?"

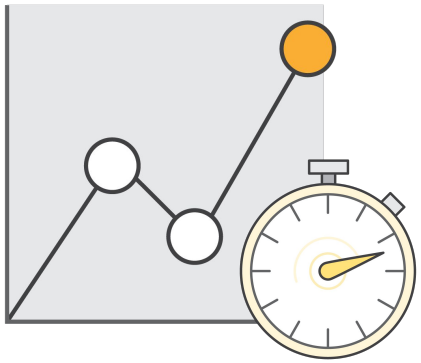


Store and analyze all of your data, from all of your sources, in one centralized location.

Benefits of a Data Lake – Quick Ingest

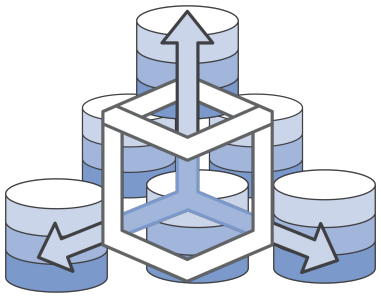


"How can I collect data quickly from various sources and store it efficiently?"

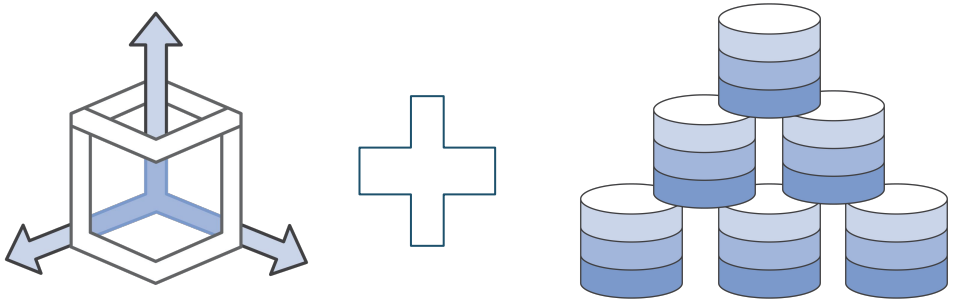


Quickly ingest data without needing to force it into a pre-defined schema.

Benefits of a Data Lake – Storage vs Compute

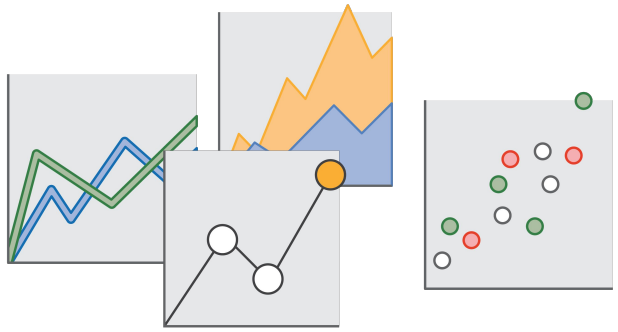


"How can I scale up with the volume of data being generated?"



Separating your storage and compute allows you to scale each component as required

Benefits of a Data Lake – Schema on Read



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

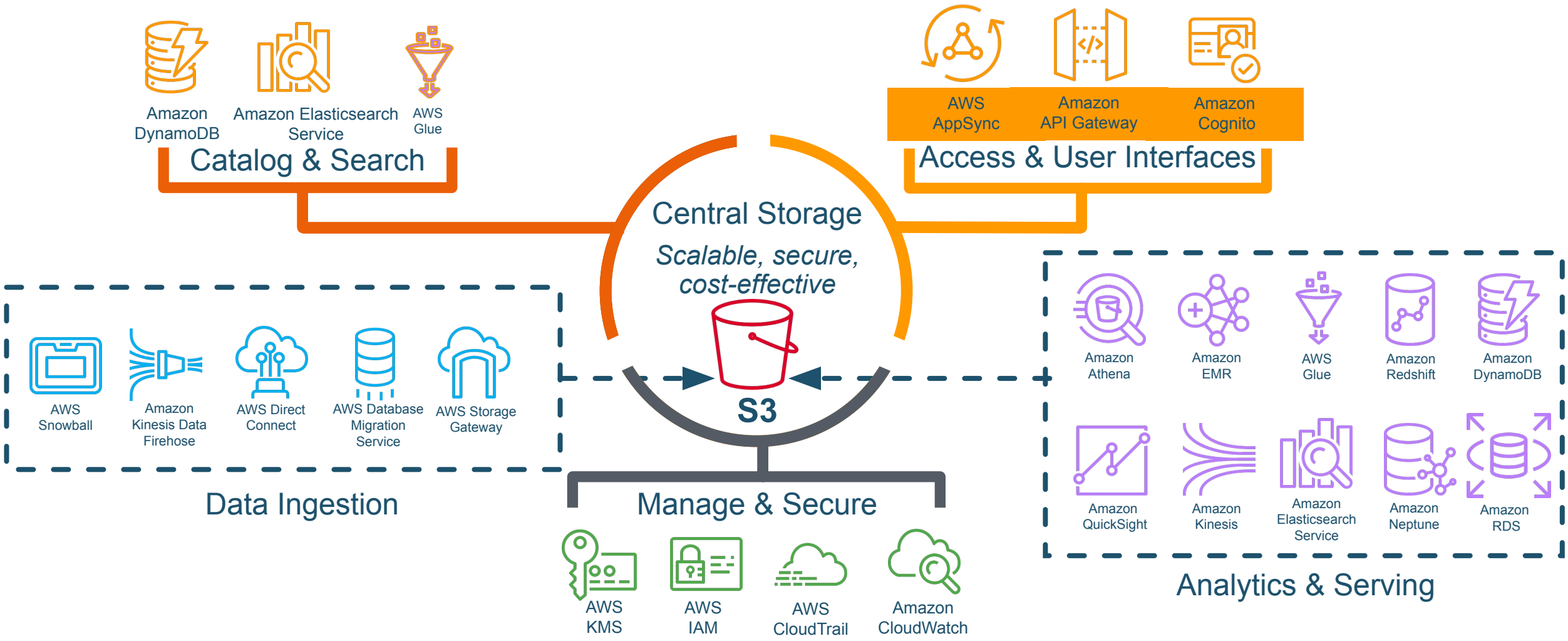
A Data Lake enables ad-hoc analysis by applying schemas on read, not write.

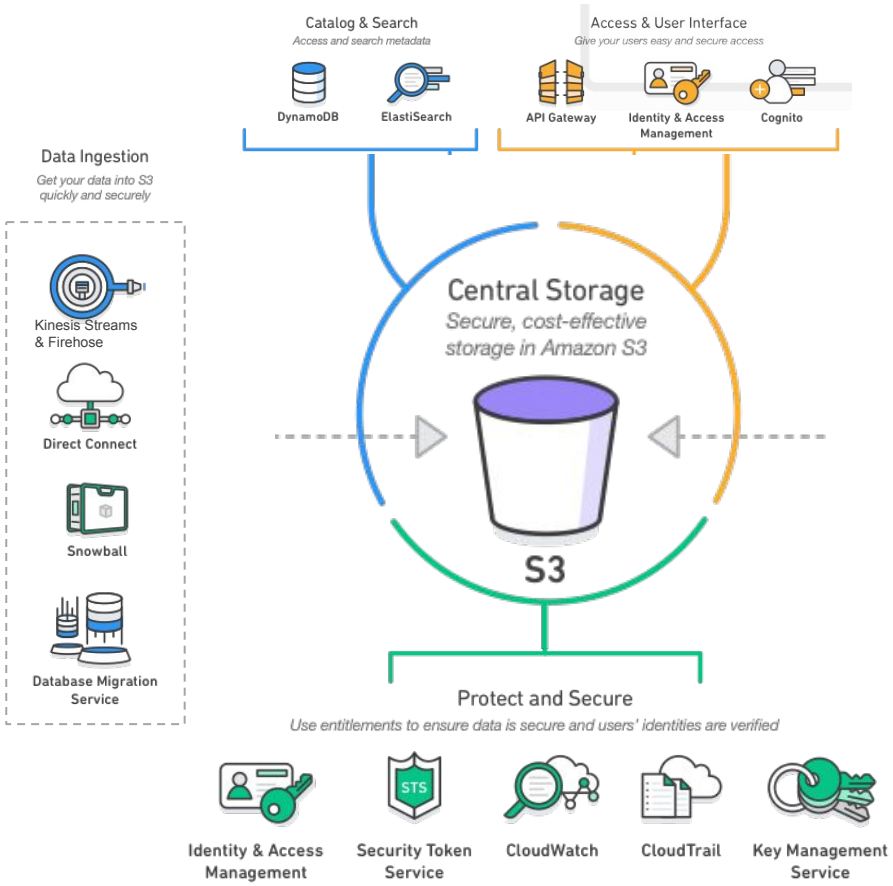
Building a Data Lake on AWS

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

Why Amazon S3 for a Data Lake?

Durable

Designed for **11 9s** of durability

Available

Designed for **99.99**% availability

High performance

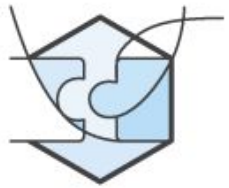
- Multiple upload
- Range GET

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

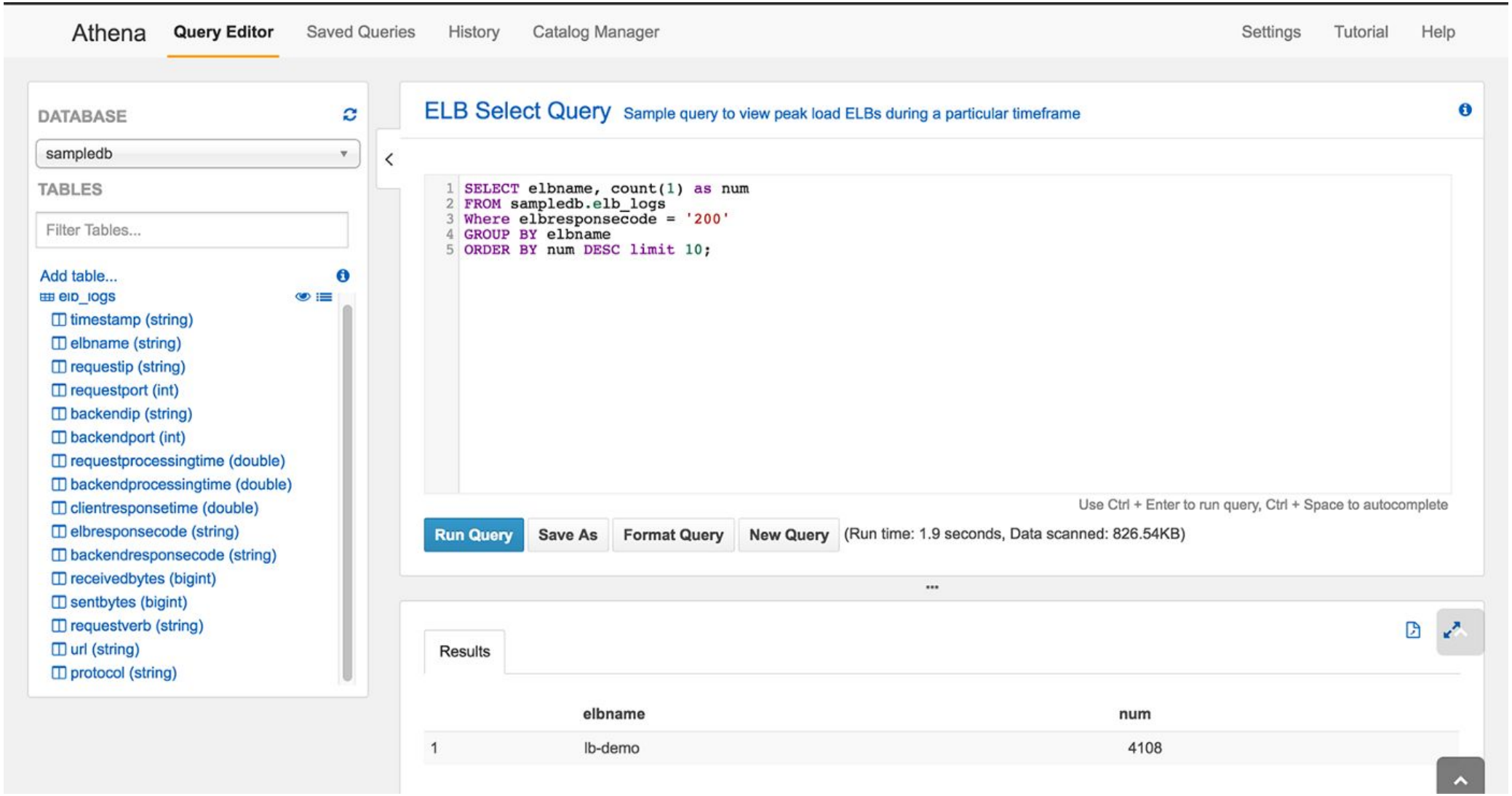


Integrated

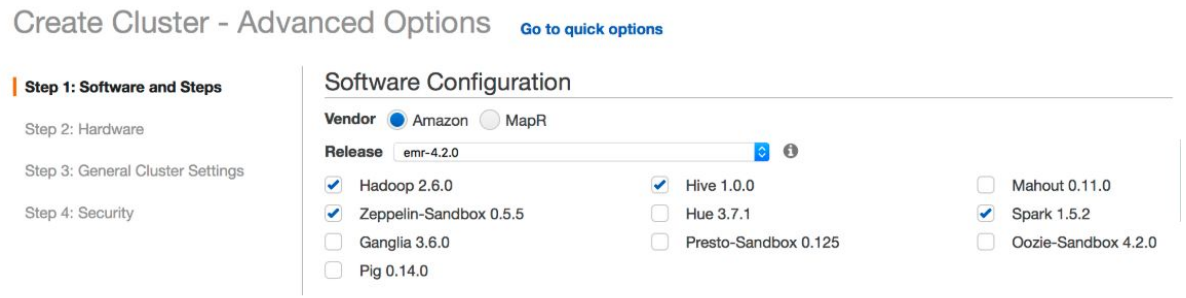
- Amazon EMR
- Amazon Redshift
- Amazon DynamoDB

What can you do with a Data Lake?

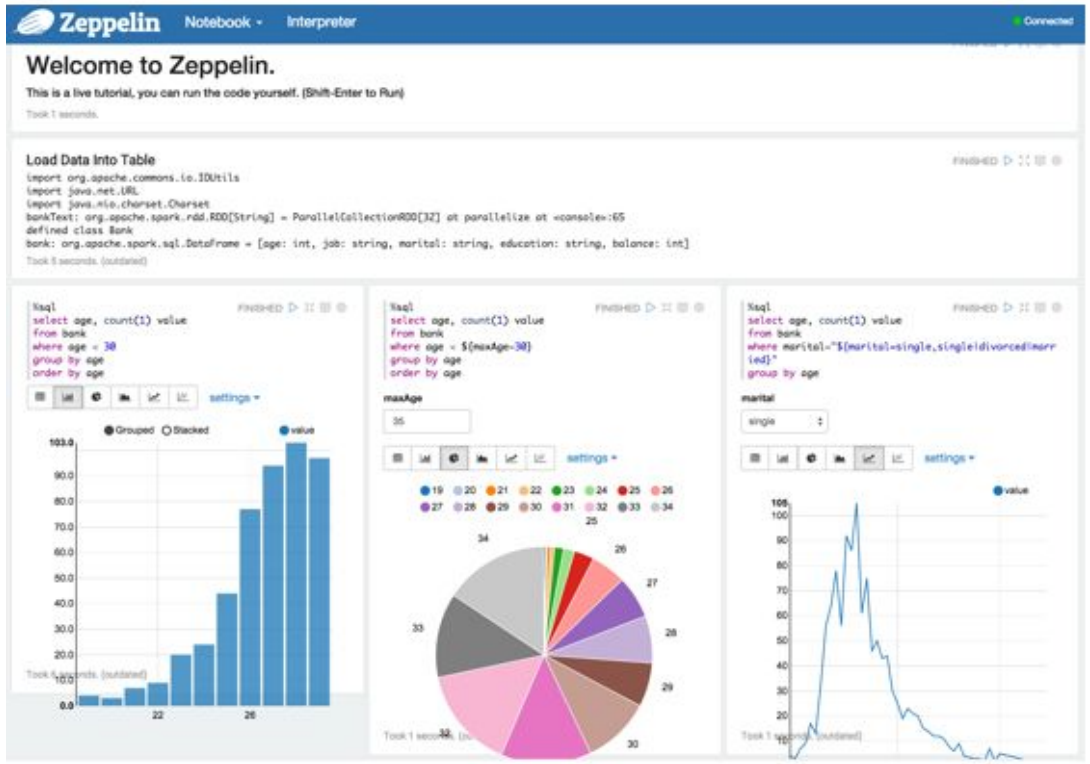
Query Directly with Amazon Athena



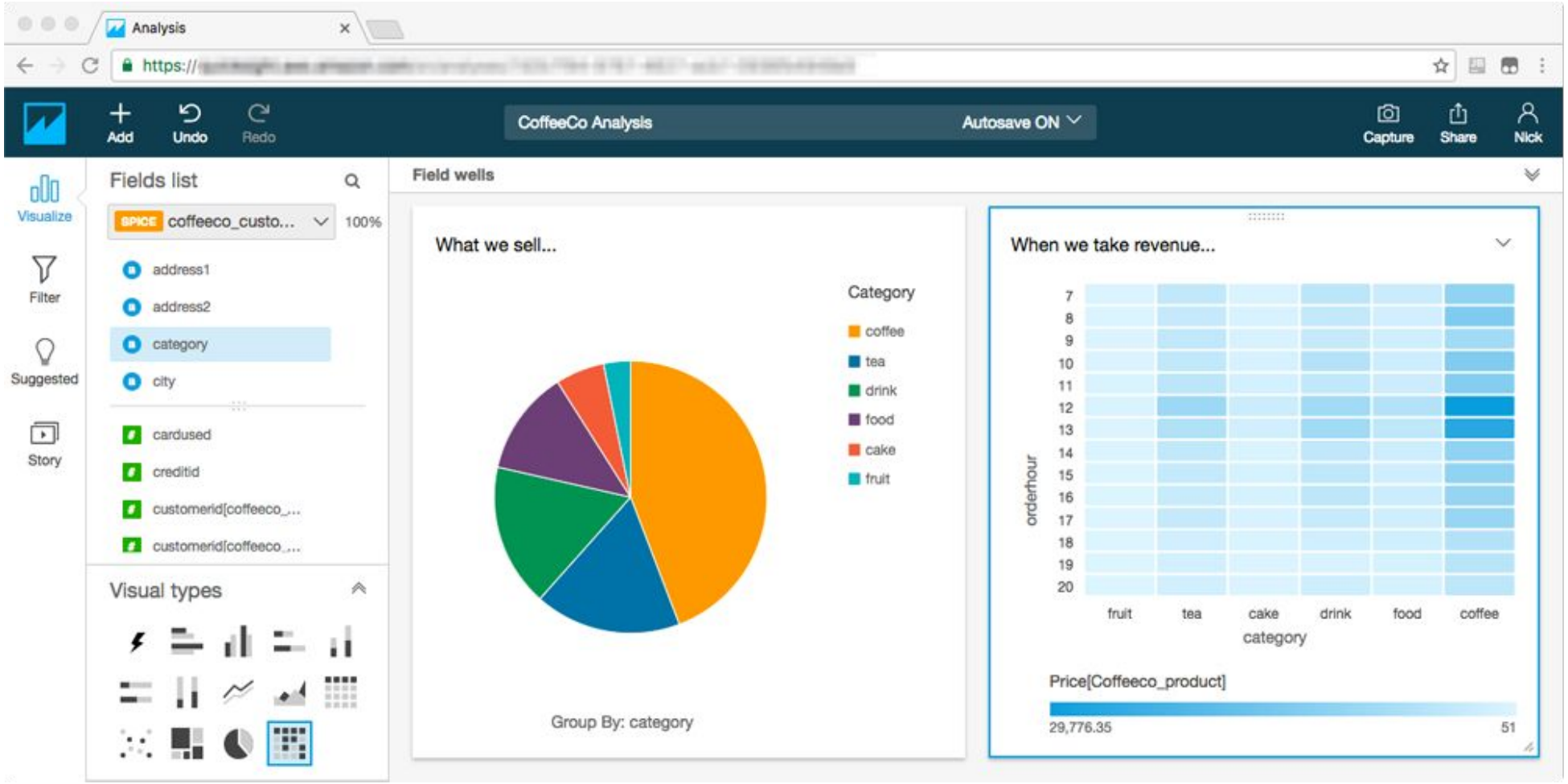
Analyze with Hadoop on Amazon EMR



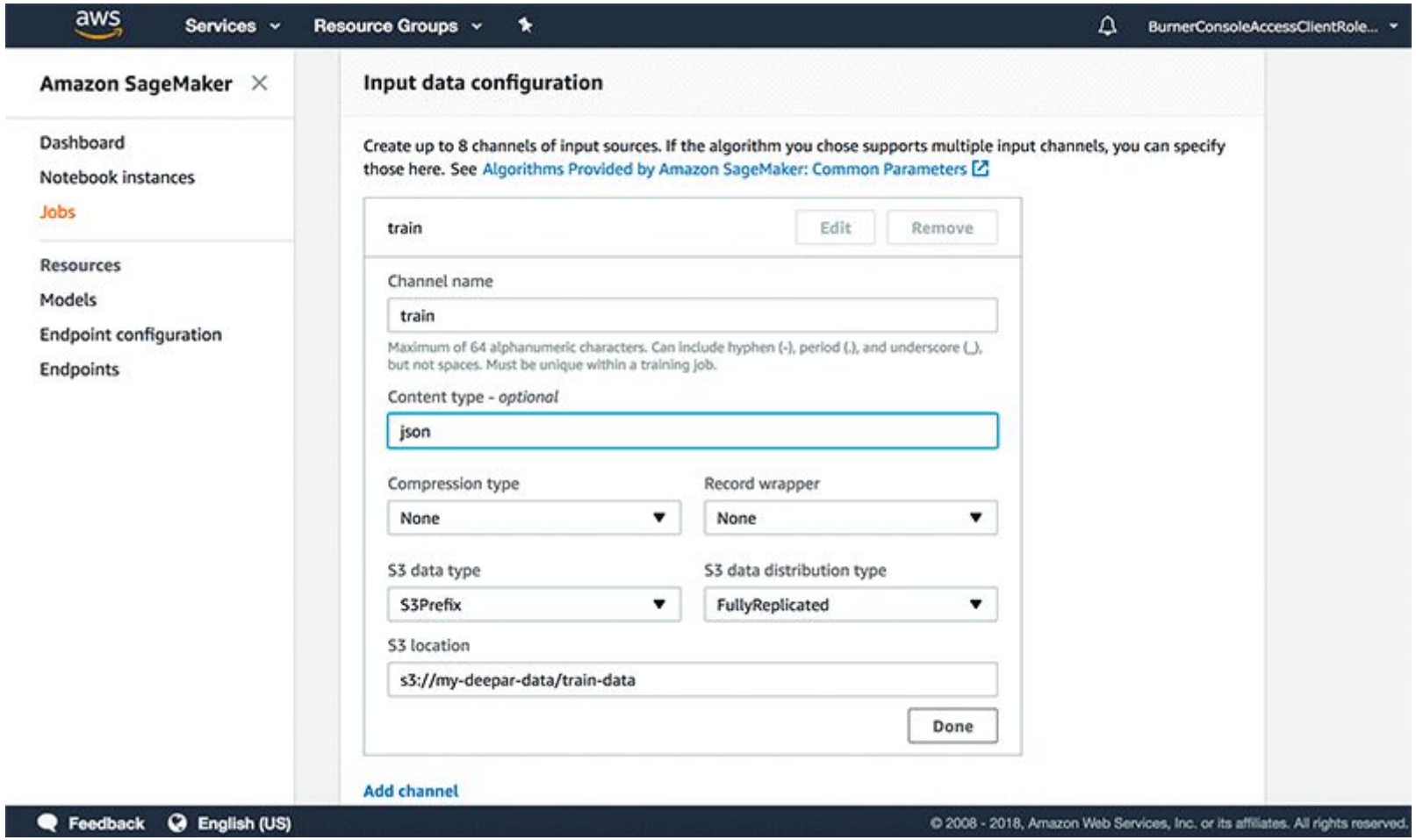

```scala
Scala
val movieLensHomeDir = "s3://emr.examples/movieLens/"

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))
}
```



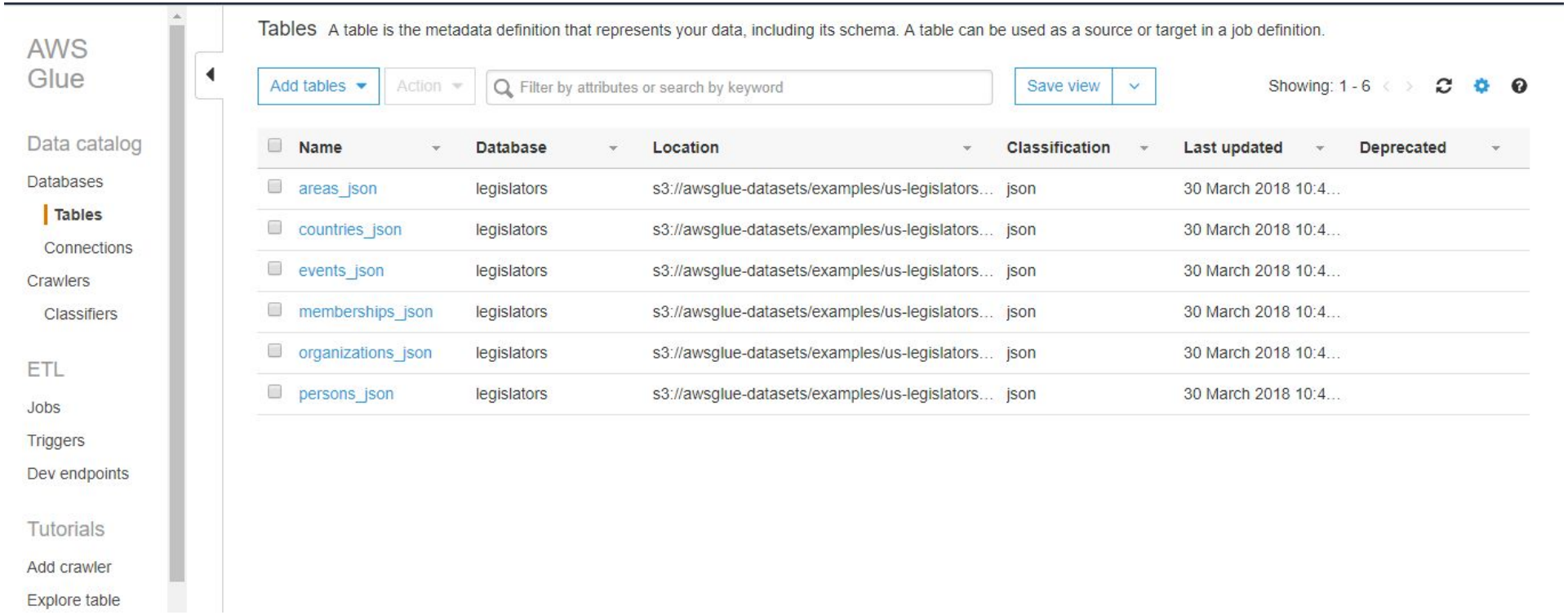
Create Visualizations with Amazon QuickSight



Train ML Models with Amazon SageMaker



Create a Central Data Catalog with AWS Glue



Load into Downstream Services

Amazon Redshift

Run complex analytic queries against petabytes of structured data

Amazon Aurora

A MySQL and PostgreSQL compatible relational database built for the cloud

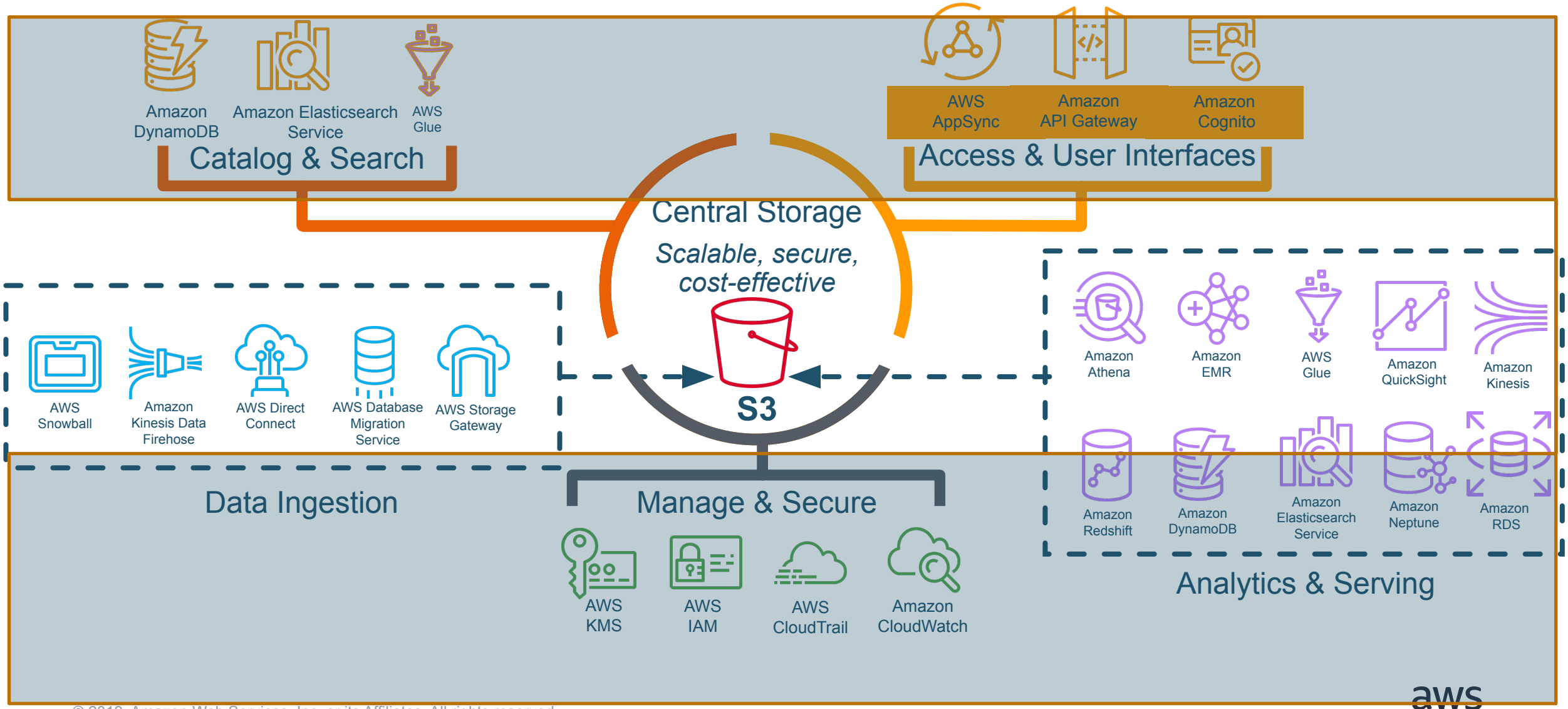
Amazon DynamoDB

A NoSQL database service that delivers consistent, single-digit millisecond latency at any scale.

Amazon Elasticsearch

Delivers Elasticsearch's real-time analytics capabilities alongside the availability, scalability, and security that production workloads require.

Data Lake on AWS



Thanks! You