



AWS  
re:Invent

**A R C 3 4 5 - R**

# Architecting data lakes with AWS data and analytics services

**Shree Kenghe**

Solutions Architect  
Amazon Web Services

**Prahlad Rao**

Solutions Architect  
Amazon Web Services

# Agenda

Data lake architecture

Ingest

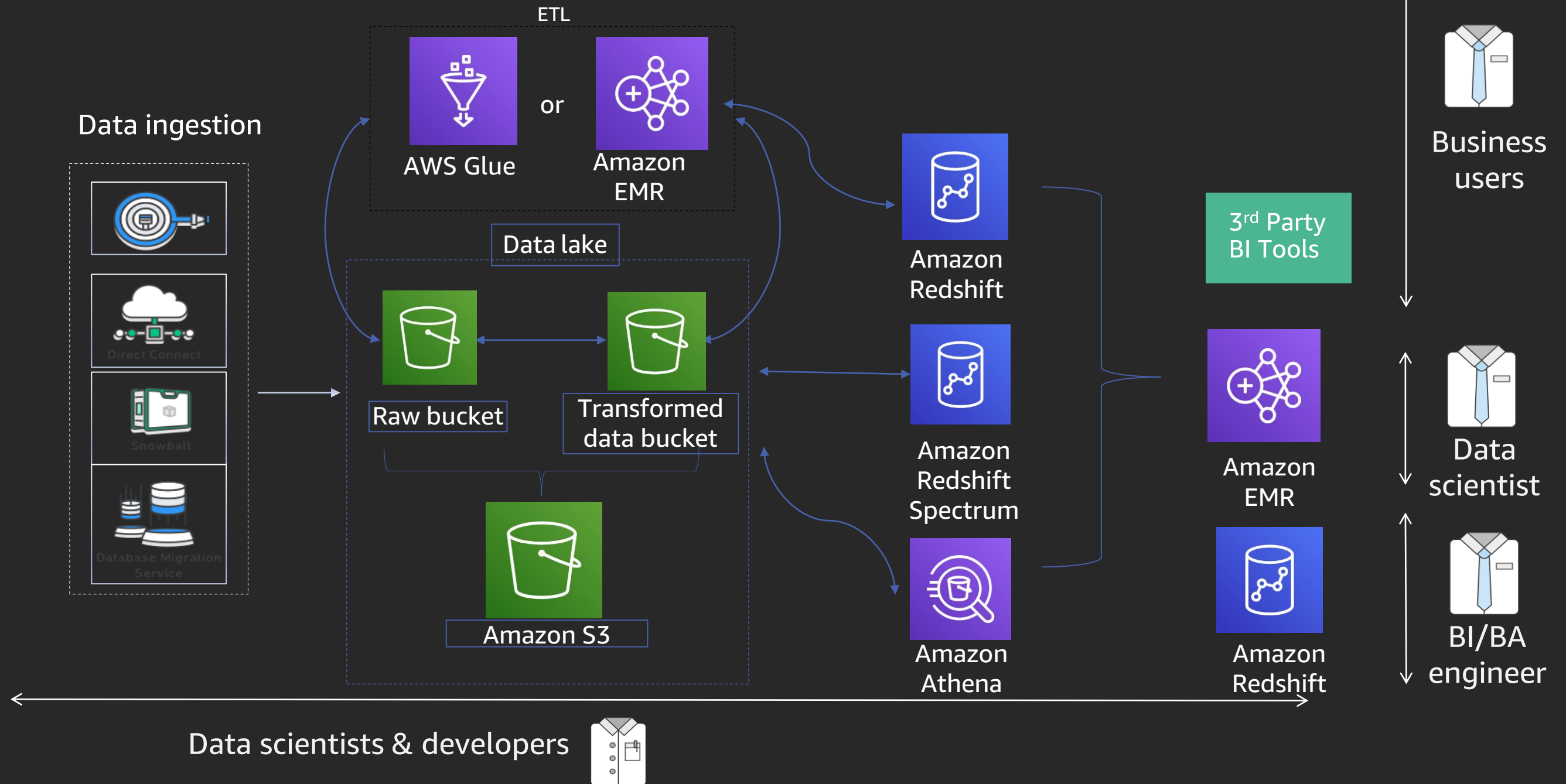
Store

Secure

Analyze

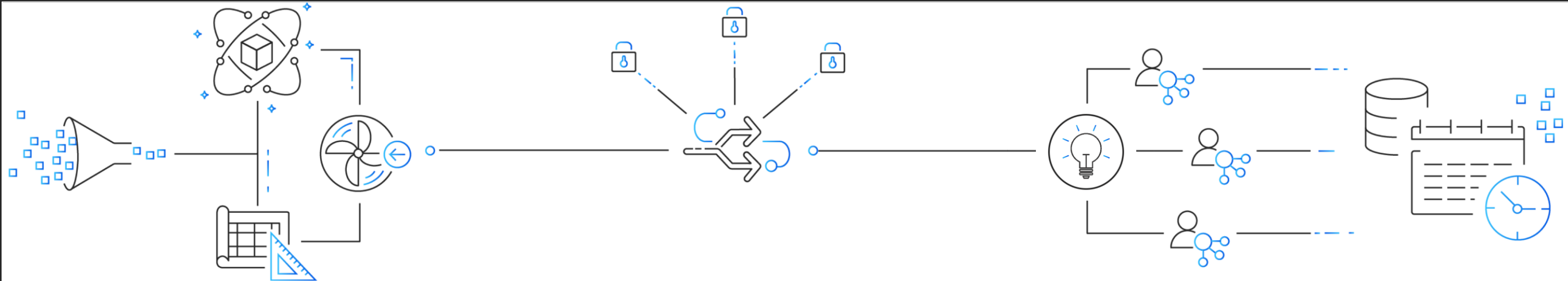
# Data lake architecture

# Data lake reference architecture



# Typical steps of building a data lake

## 1 Set up storage



## 2 Move data

## 3 Cleanse, prep, and catalog data

## 4 Configure and enforce security and compliance policies

## 5 Make data available for analytics

# Ingest

# Data ingest options



AWS Database  
Migration Service  
(AWS DMS)



Amazon  
Kinesis Data  
Firehose



Amazon Kinesis  
Data Streams



Amazon Kinesis  
Video Streams



Amazon S3  
Transfer  
Acceleration



AWS  
Storage  
Gateway



AWS  
Snowball



AWS  
Snowball Edge



AWS  
Snowmobile



AWS  
DataSync



AWS  
Transfer  
for SFTP



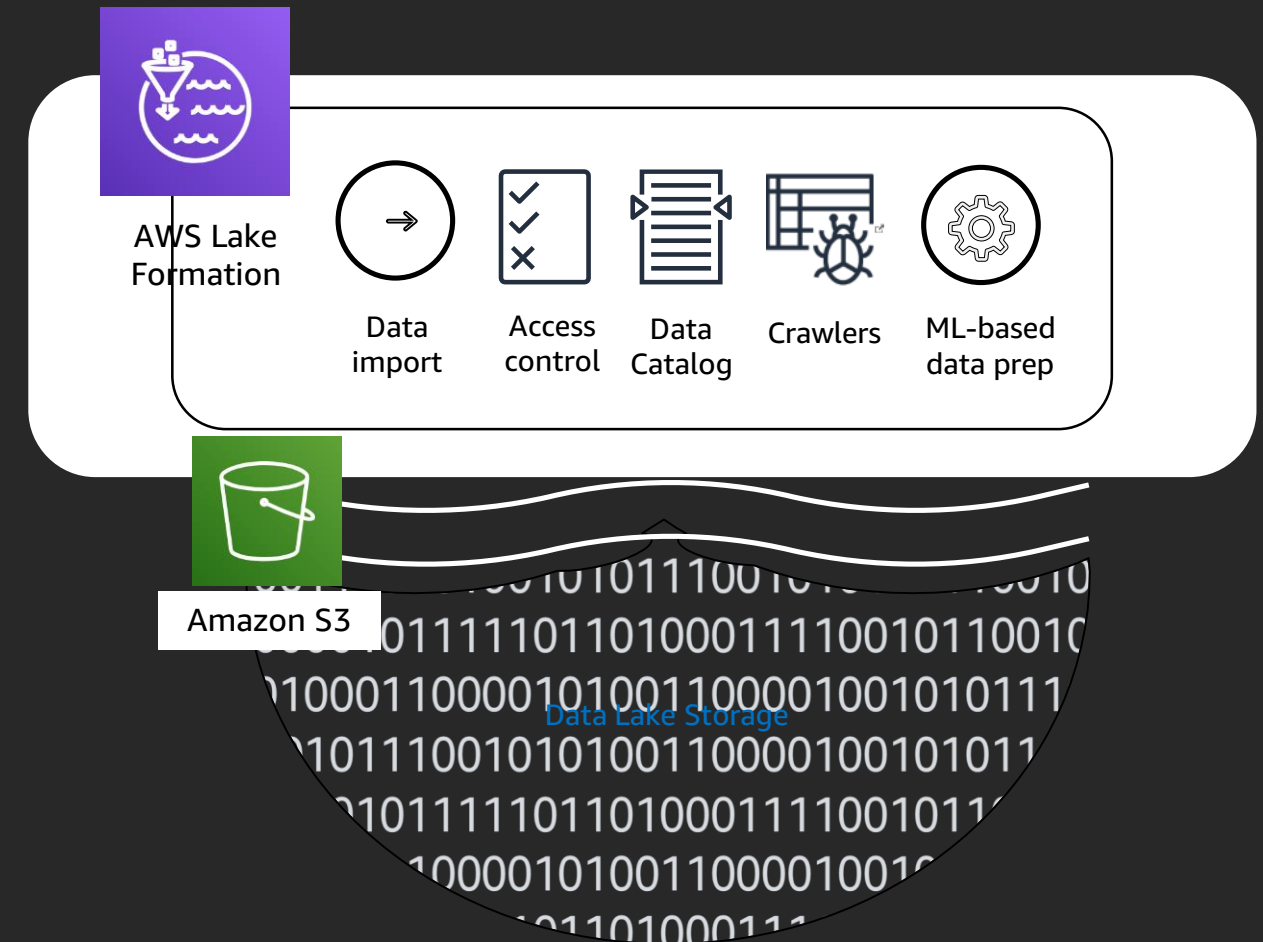
# Register existing data or import new

Amazon Simple Storage Service (Amazon S3) forms the storage layer for AWS 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



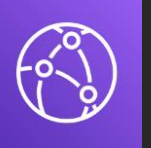
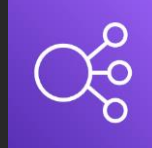


# Blueprints: Easily load data to your data lake



Amazon RDS

DBs

Amazon Aurora



Amazon Kinesis  
Data Firehose

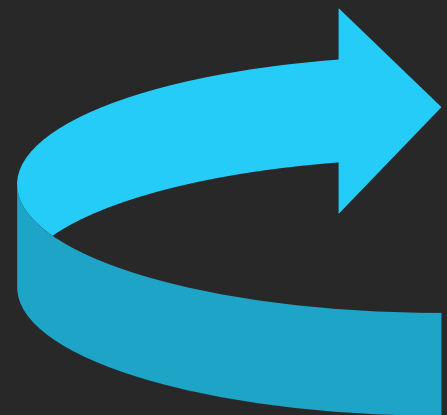
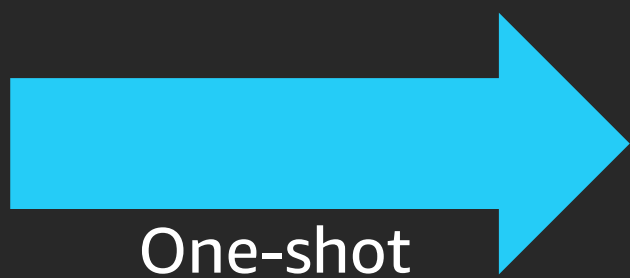
Logs





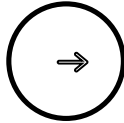
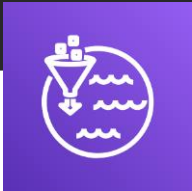
AWS CloudTrail

Elastic Load  
Balancing

Amazon  
CloudFront

## Blueprints





AWS Lake  
Formation

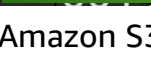

Data  
import

Access  
control

Data  
Catalog

Crawlers

ML-based  
data prep



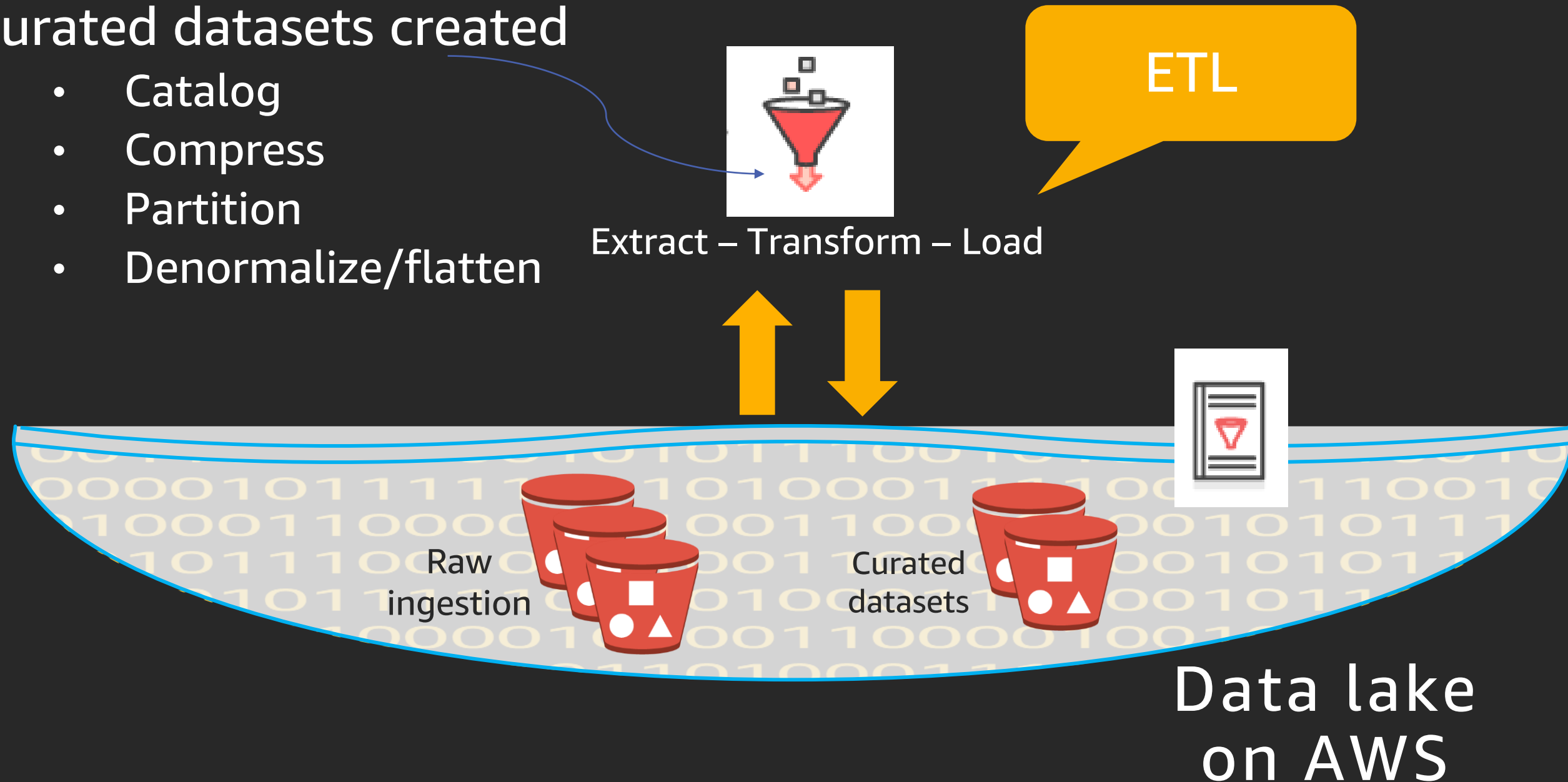
Amazon S3

Data Lake Storage

# Store

# Preparing raw data for optimized analytics

- Raw data stored in data lake
- Curated datasets created
  - Catalog
  - Compress
  - Partition
  - Denormalize/flatten



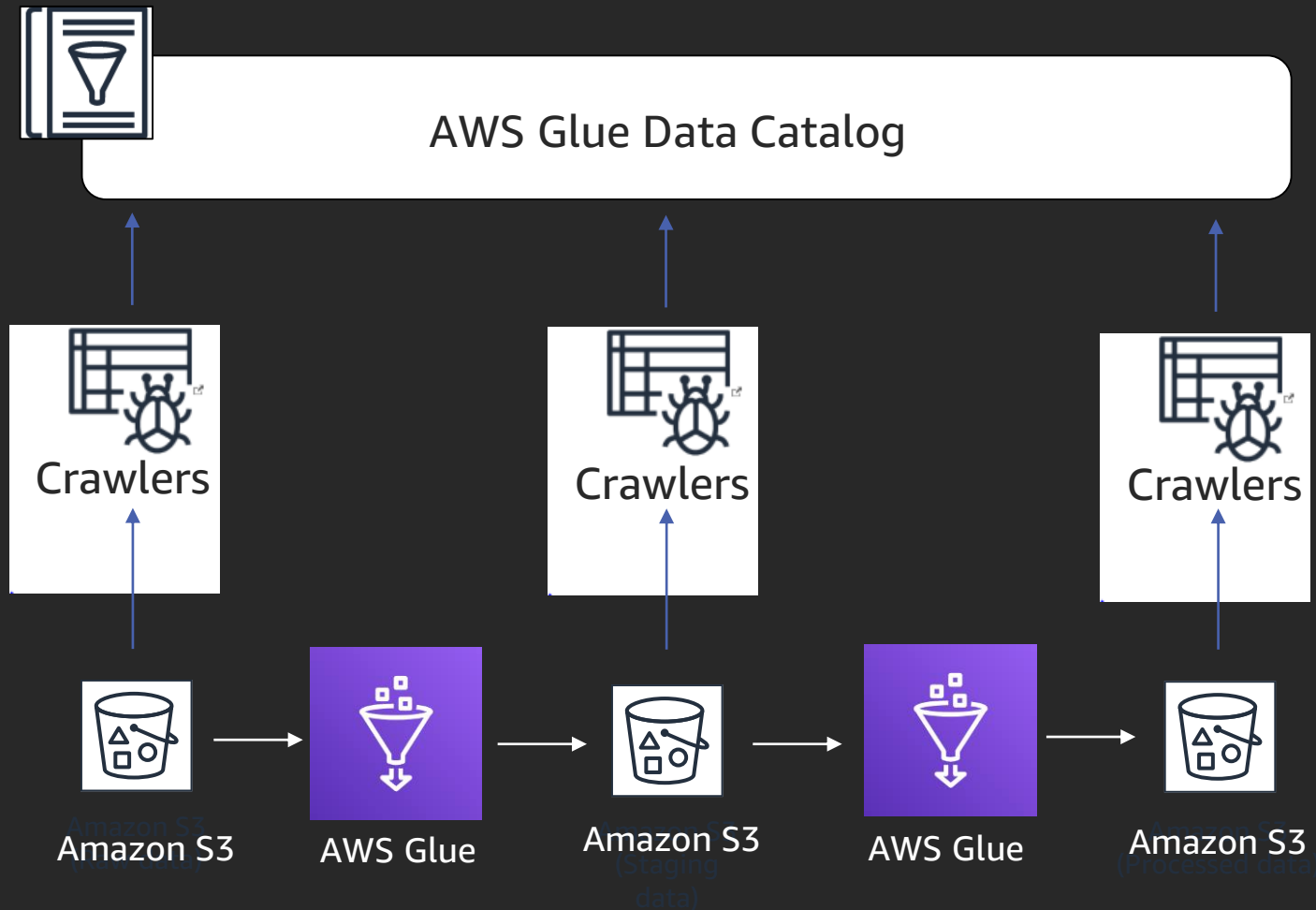
# File formats

- **Text – xSV, JSON**
  - May or may not be compressed
  - Human readable when uncompressed
  - Not optimized for analytics
- **Columnar – Parquet & ORC**
  - Compressed in a binary format
  - Integrated indexes and stats
  - Optimized read performance when selecting only a subset of columns
- **Row – Avro**
  - Compressed in a binary format
  - Optimized read performance when selecting all columns of a subset of rows

# Partitioning guidance

- Chose columns that have low cardinality (uniqueness)
  - Partitioning on day/month/year has 365 unique values per year
  - Partitioning on seconds has millions of values per year
- You can partition on any column, not just date
  - For example, `s3://abc-corp-sales-data/country=xx/state=xx/bu=xx)`
- Look at your query patterns – what data do you want to query, and what do you want to filter out?

# Use AWS Glue to cleanse, prep, and catalog



## AWS Glue Data Catalog – a single view across your data lake

- Automatically **discovers** data and stores schema
- Makes data **searchable** and available for ETL
- Contains table definitions and custom metadata

## Use AWS Glue ETL jobs to cleanse, transform, and store processed data

- Serverless Apache Spark** environment

- Use Glue ETL libraries or bring your own code

- Write code in **Python or Scala**

- Call any AWS API** using the AWS boto3 SDK



# Amazon S3

Tier 1 data lake:  
Ingestion

- Single source of truth for raw data
- Organized by Ingestion Time
- Use least transformations
- Use lifecycle policies to Amazon S3 IA or Amazon Simple Storage Service Glacier

Tier 2 data lake:  
Analytics

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

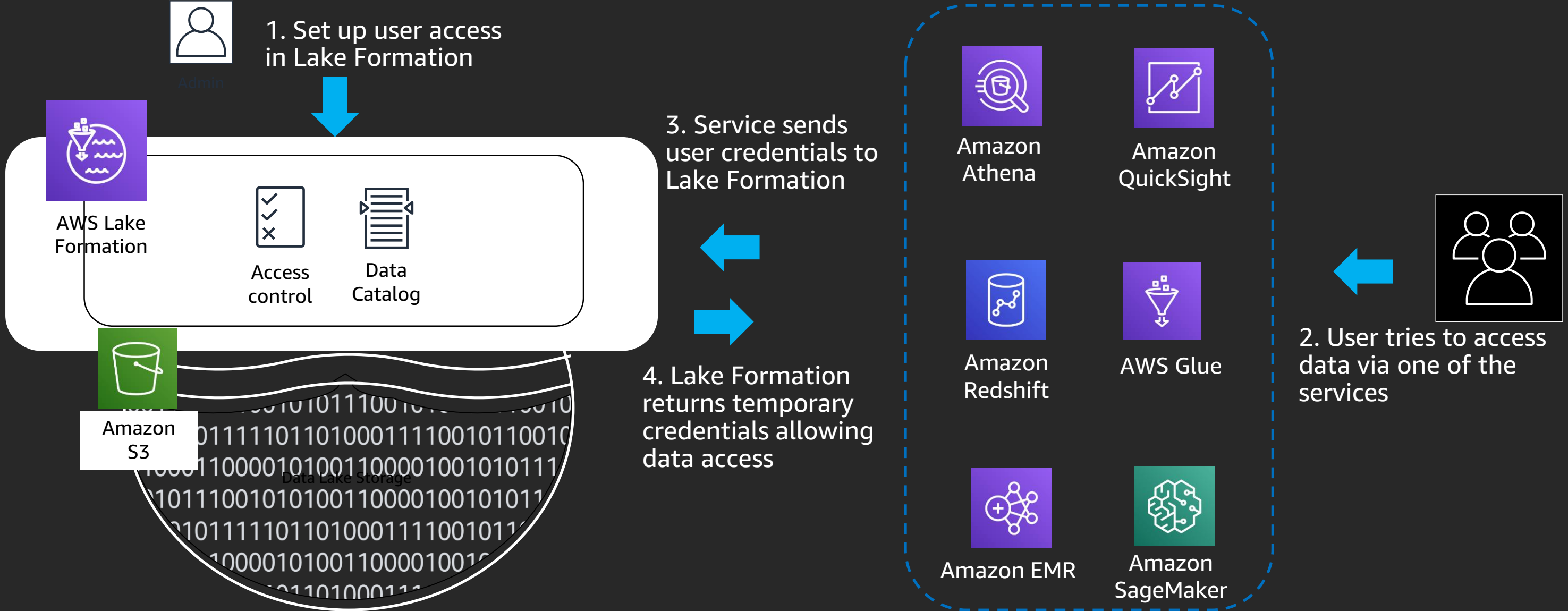
Tier 3 data lake:  
Analytics

- Domain-level data mart
- Organized by use cases
- Optimized for specialized analysis



# Secure

# Secure once, access in multiple ways



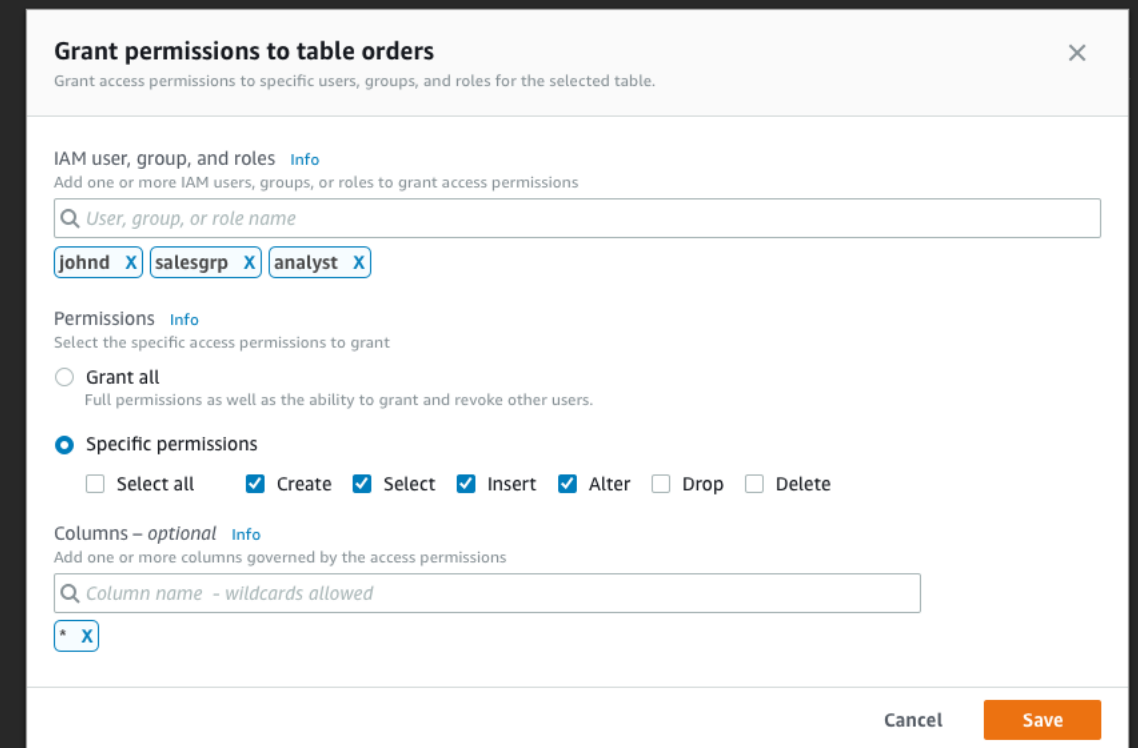
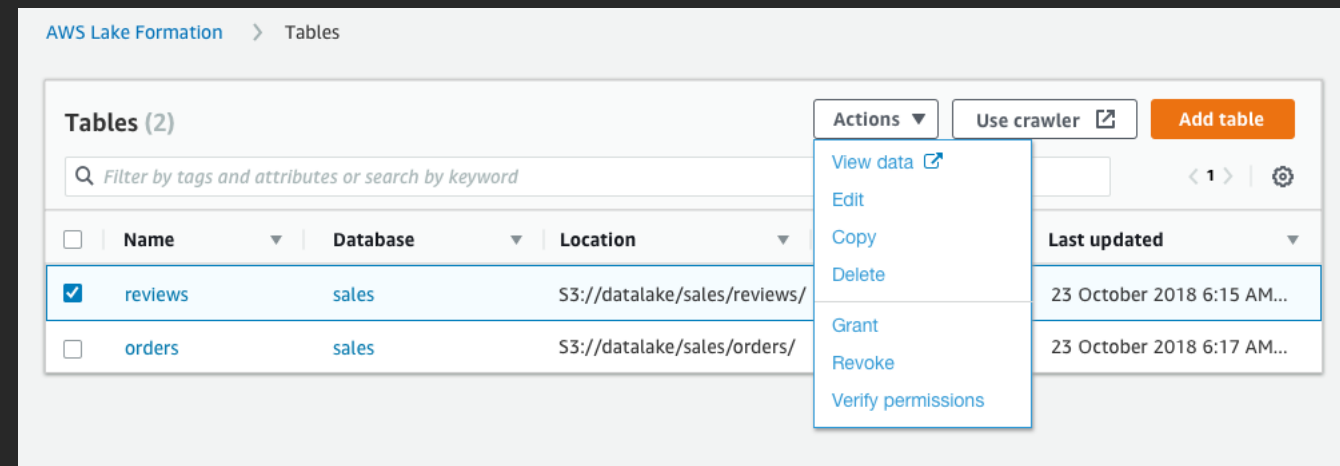
# Security permissions in AWS Lake Formation

Control data access with simple grant and revoke permissions

Specify permissions on tables and columns rather than on buckets and objects

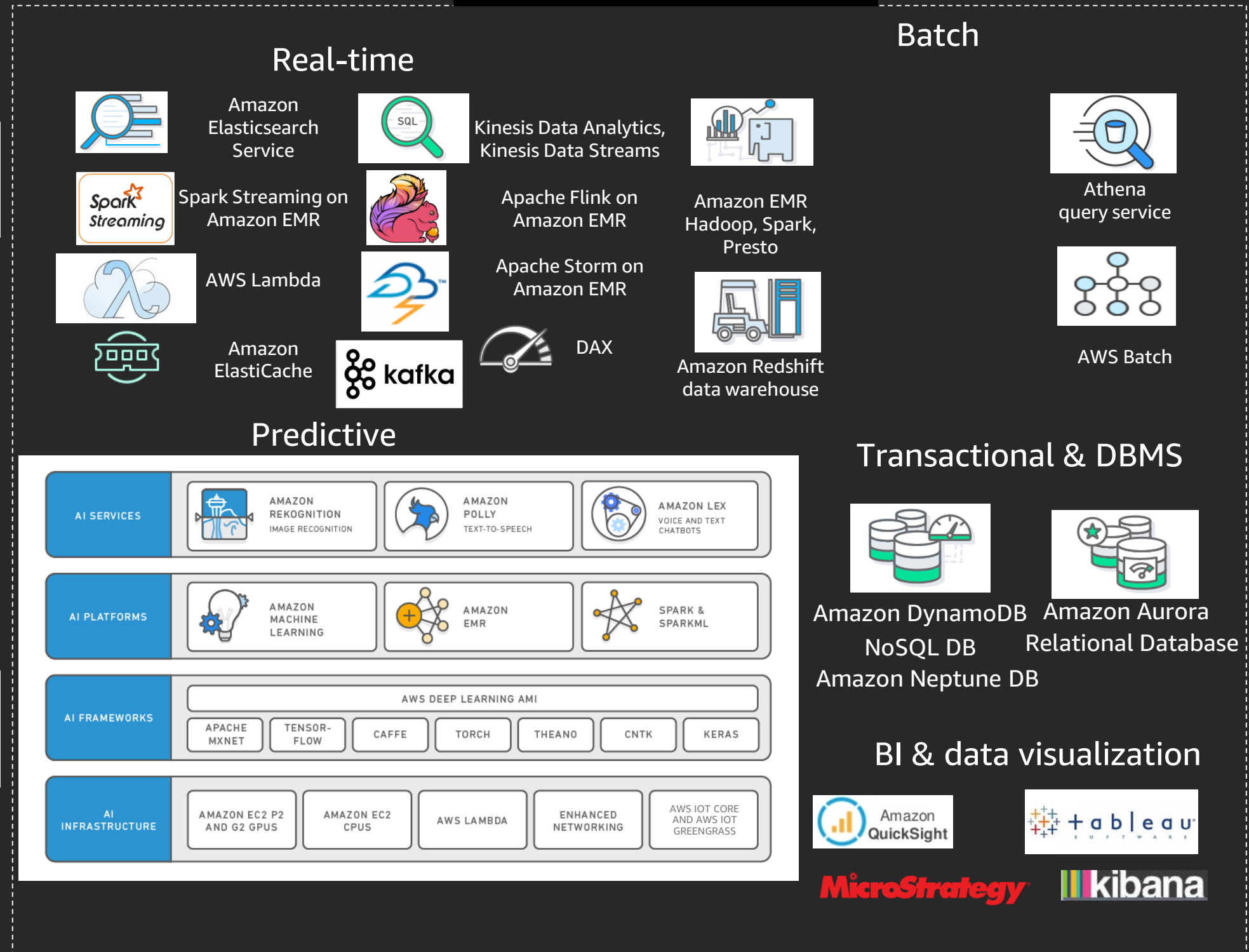
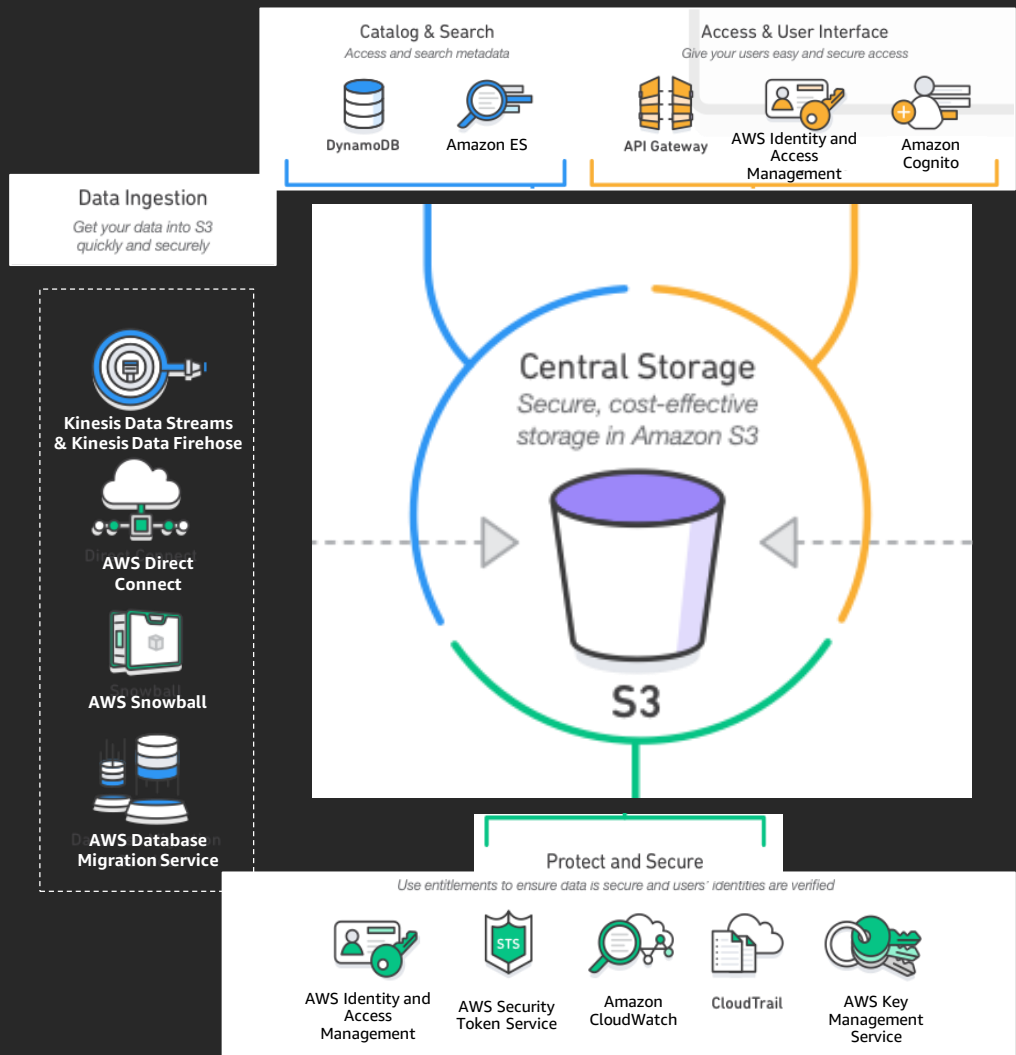
Easily view policies granted to a particular user

Audit all data access at one place



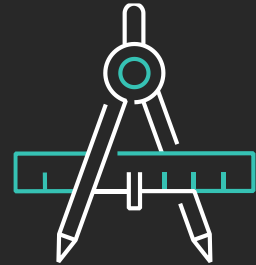
# Analyze

# Processing & Analytics

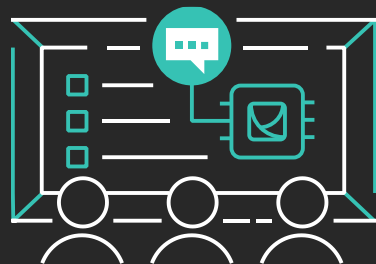


# Learn to architect with AWS Training and Certification

Resources created by the experts at AWS to propel your organization and career forward



Free foundational to advanced digital courses cover AWS services and teach architecting best practices



Classroom offerings, including Architecting on AWS, feature AWS expert instructors and hands-on labs



Validate expertise with the **AWS Certified Solutions Architect - Associate** or **AWS Certification Solutions Architect - Professional** exams

Visit [aws.amazon.com/training/path-architecting/](https://aws.amazon.com/training/path-architecting/)

# Thank you!



Please complete the session  
survey in the mobile app.