Data lakes and analytics on AWS

Customers want more value from their data

Growing exponentially

From new sources

Increasingly diverse

Used by many people

Analyzed by many applications

aws

Companies want more value from their data

Complications

Siloed approaches don’t work anymore

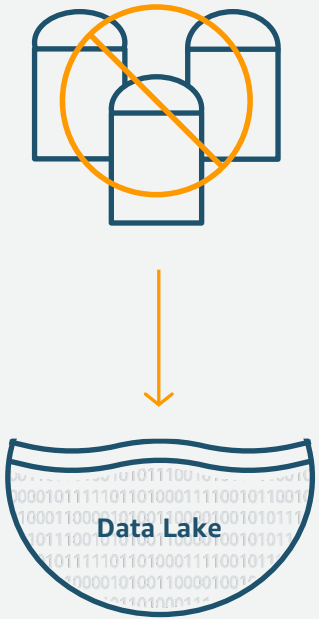
It’s too expensive and limiting
to store data on-premises

Implication

A new approach is needed to
extract insights and value

aws

Cloud data lakes are the future



Customers want:

To move to a single store; i.e., a data lake in the cloud

To store data securely in standard formats

To grow to any scale, with low costs

To analyze their data in a variety of ways

To democratize data access and analysis

aws

Most comprehensive

Broadest and deepest portfolio, purpose-built for builders

Visualization & Machine Learning

- Dashboards
- Predictive Analytics

Analytics

- Data Warehousing
- Big Data Processing
- Serverless Data processing
- Interactive Query
- Operational Analytics
- Real time Analytics

Data Lake Infrastructure & Management

- Infrastructure
- Security & Management
- Data Catalog & ETL

Data Movement

Migration & Streaming Services

aws

Most comprehensive

Broadest and deepest portfolio, purpose-built for builders

Visualization & Machine Learning

 QuickSight

 SageMaker

 Comprehend

 Lex

 Polly

 Rekognition

 Translate

 Transcribe

 Deep Learning AMIs

+ 10 more

Analytics

 Redshift

 EMR (Spark & Hadoop)

 Glue (Spark & Python)

 Athena

 Elasticsearch Service

 Kinesis Data Analytics

Data Lake Infrastructure & Management

 S3/Glacier

 Lake Formation NEW

 Glue

Data Movement

Database Migration Service | Snowball | Snowmobile | Kinesis Data Firehose | Kinesis Data Streams | Managed Streaming for Kafka

aws

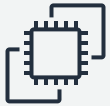
Most cost effective

Decouple compute and storage, choice of PAYG analytics services

Storage

S3 tiers &
intelligent tiering

From $0.023 per
GB/mo to as low as
$0.004 per GB/mo



Compute

Spot & reserved
instances

Save up to 90% off
on-demand prices

EMR

Autoscaling

57% less than
on-premises
per IDC report

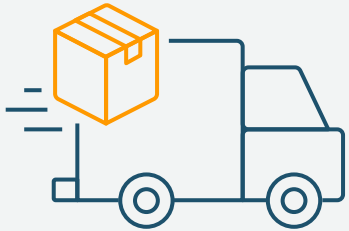
Redshift

less than a tenth
of the cost of
traditional solutions.

Athena &
QuickSight

Serverless pay
only for what is used

aws



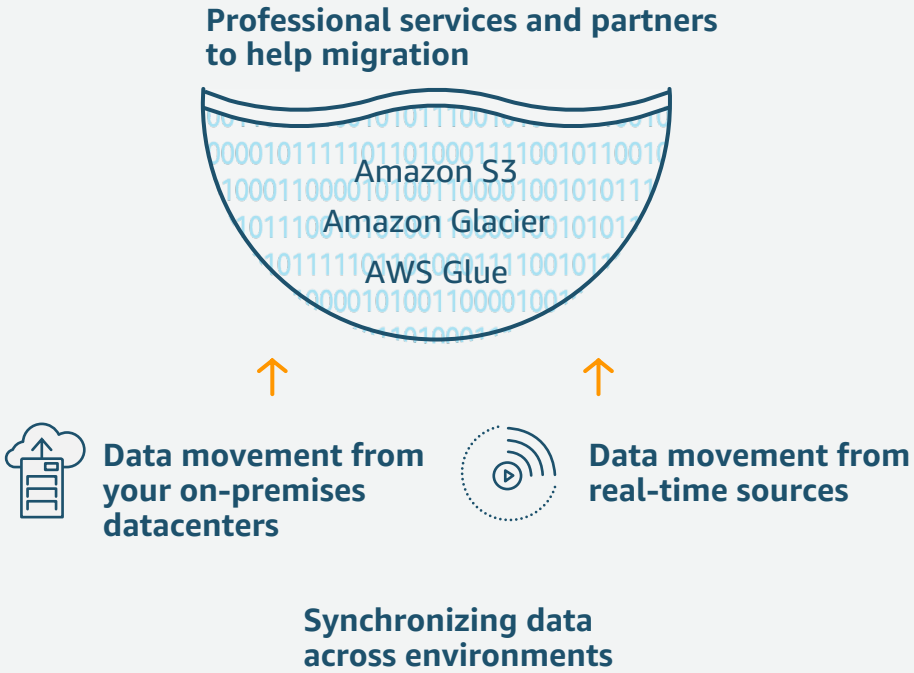
Data movement solutions

Data Movement

Migration & Streaming Services

aws

Most ways to move data to the data lake



Data movement from on-premises datacenters

Dedicated network connection

Secure appliances

Ruggedized shipping containers

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

Robust data lake infrastructure

Data lake infrastructure & management

- Durable and available; exabyte scale
- Secure, compliant, auditable
- Object-level controls for fine-grain access
- Fast performance by retrieving subsets of data
- Decoupling of compute and storage
- On-demand resources, tiering, cost choices

aws

Set up a catalog, ETL, and data prep
with AWS Glue

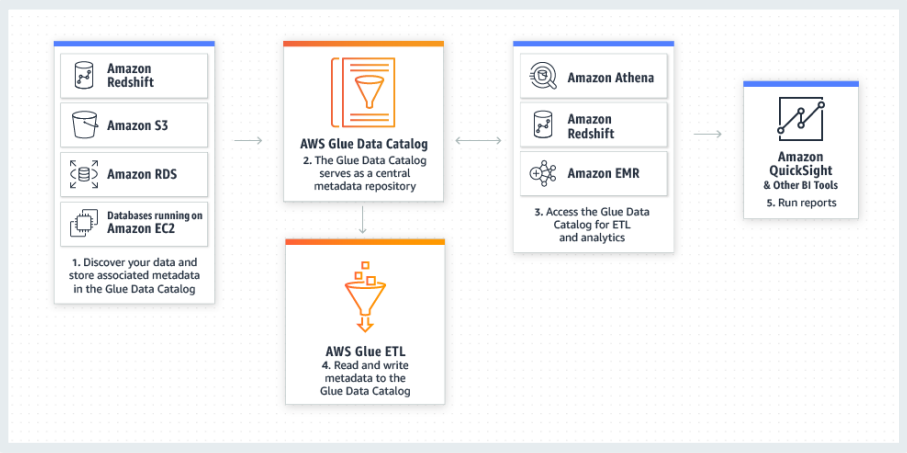
Data lake infrastructure & management

Serverless provisioning, configuration, and scaling to run your ETL jobs on Apache Spark

Pay only for the resources used for jobs

Crawl your data sources, identify data formats and suggest schemas and transformations

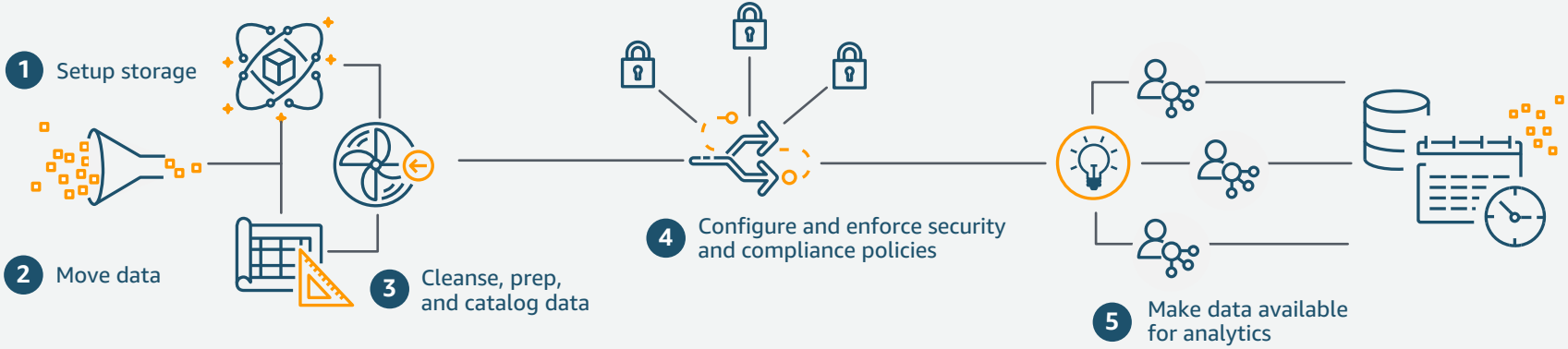
Automates the effort in building, maintaining and running ETL jobs



Challenges to making a secure data lake

Data lake infrastructure & management

Typical steps of building a data lake



aws

Build a secure data lake in days
with AWS Lake Formation

Data lake infrastructure & management

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

aws

Analytics solutions

- Data Warehousing
- Big Data Processing
- Serverless Data processing
- Interactive Query
- Operational Analytics
- Real time Analytics

Big data processing with Apache Spark & Hadoop
with Amazon EMR

Analytics

Easy to use notebooks

Low cost vs on-premises

Elastic autoscaling

Reliable 99.9% SLA

Secure with encryption and keys

Flexible, open source choice

Enterprise-grade

Easy

Lowest cost

aws

Data warehouse for business reporting
with Amazon RedShift

Analytics

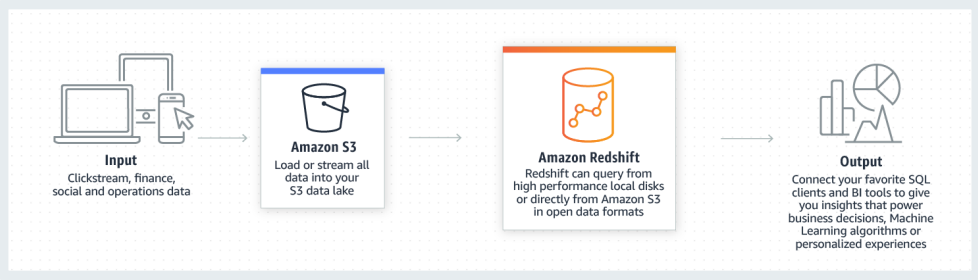
Fast—up to 10x faster than traditional data warehouses

Easy to setup, deploy and manage

Cost-effective

Scale on-demand for large data volume and high query concurrency

Query data in open formats directly from the data lake



aws

Amazon Redshift architecture

Massively parallel, shared-nothing architecture

Leader node

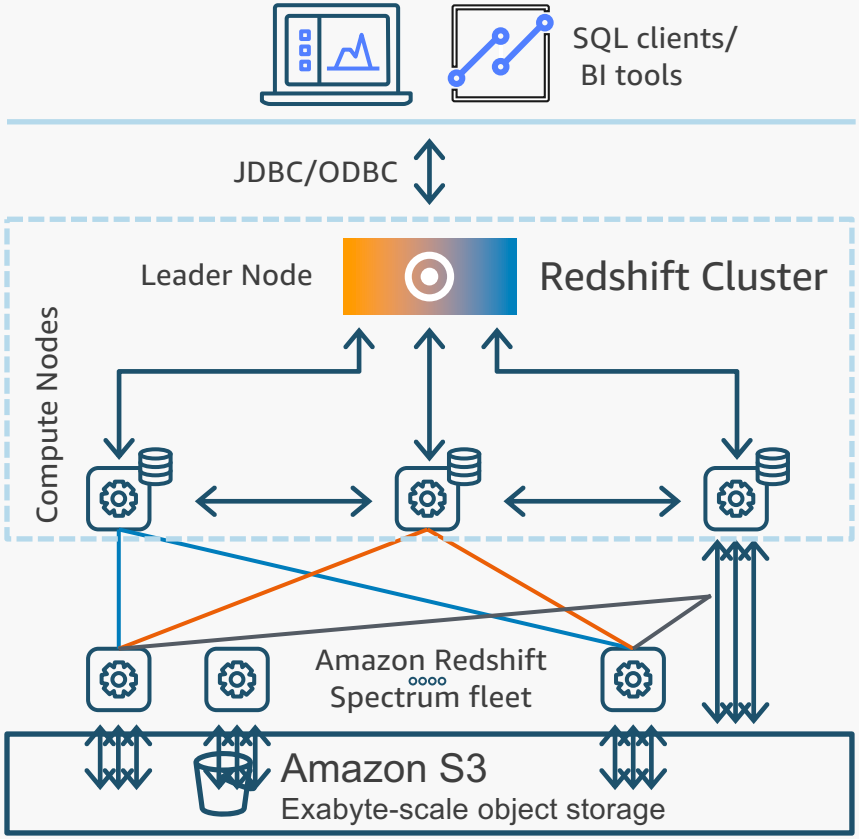
- SQL endpoint, stores metadata
- Coordinates parallel SQL processing
- Free for any cluster with two or more nodes

Compute nodes

- Local, columnar storage
- Executes queries in parallel
- Load, backup, restore

Amazon Redshift Spectrum nodes

- Serverless, not managed by customer, bring power proportional to cluster slices
- Execute queries directly against data lake



aws

Typical Redshift use cases

Modern data warehousing

Mid-Market, enterprise customers, large established customers

Deliver the same compatibility at a vastly lower price with operational ease of use and migration

Operational analytics

New entrants

Variety and volume of data coming at a high velocity—streaming data

Requirement to store and analyze for internal and external analytics

Analytics on data lake

Prefer data lake approach to managing data

Need a DWH that can-do high-performance BI/Reporting but also incorporate data from data lake

Real-time analytics for timely insights
with Amazon Kinesis

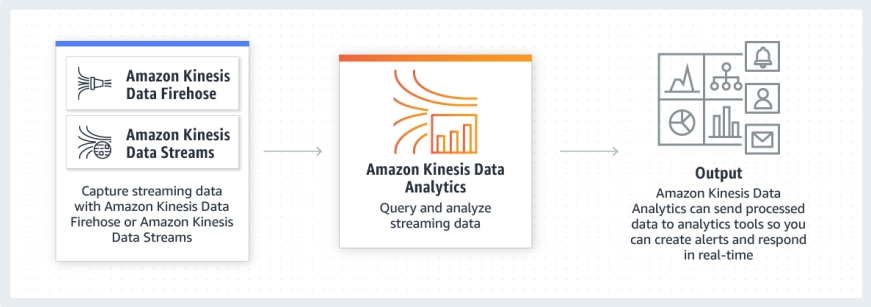
Analytics

Make streaming data available to multiple real-time analytics applications

Run streaming applications without managing any infrastructure

Durable to reduce the probability of data loss

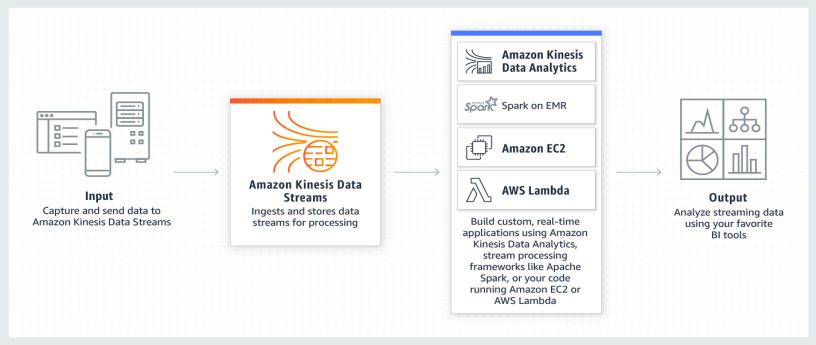
Scalable to process data from hundreds of thousands of sources with low latencies



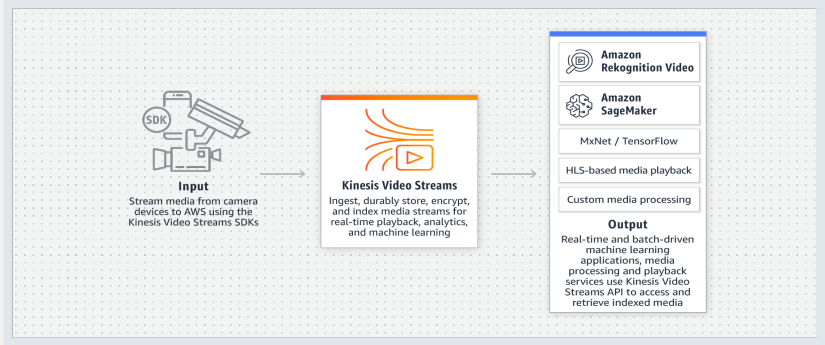
aws

Amazon Kinesis and its 4 flavors

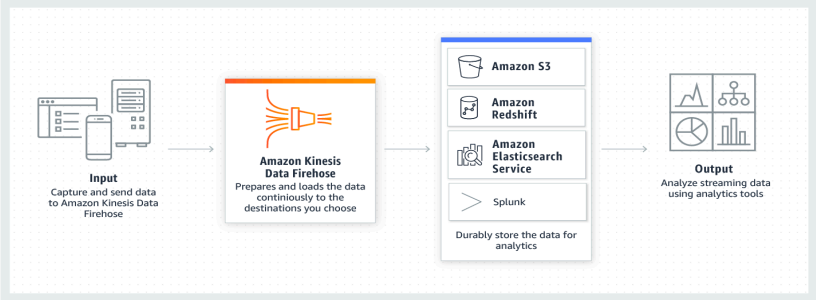
Analytics



Massively scalable real-time data streaming service



Securely stream video from connected devices



Stream data into data lakes, data stores and analytics tools

Analyze streaming data, gain actionable insights and respond in real-time.

aws

Operational analytics for logs and search
with Amazon Elasticsearch

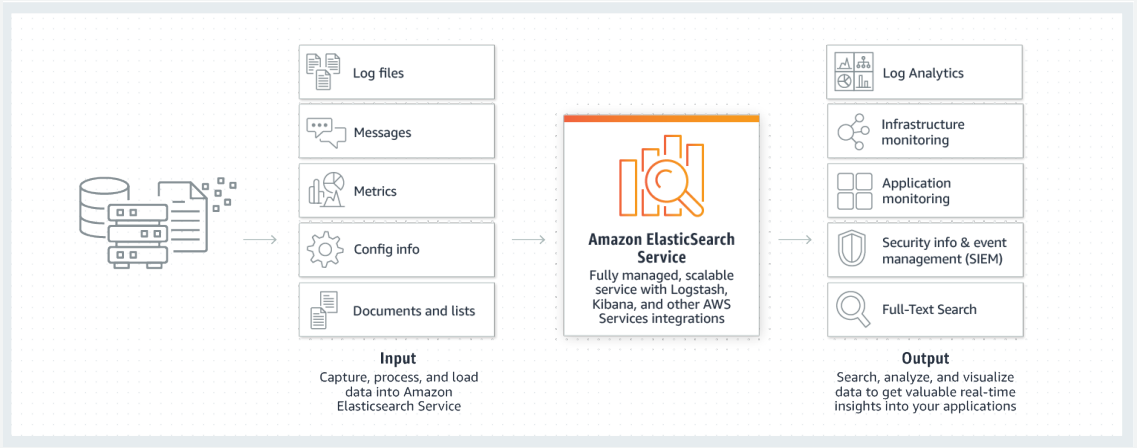
Analytics

Fully managed; deploy production-ready cluster in minutes

Direct access to Elasticsearch open-source APIs, Logstash and Kibana

VPC support; at-rest and in-transit encryption

Scale up and down easily



Interactive analysis
with Amazon Athena

Interactive query service to analyze data in Amazon S3 using standard SQL

No infrastructure to set up or manage and no data to load

Ability to run SQL queries on data archived in Amazon Glacier (coming soon)

aws

Serverless analytics

Deliver on-demand analytics on the data lake

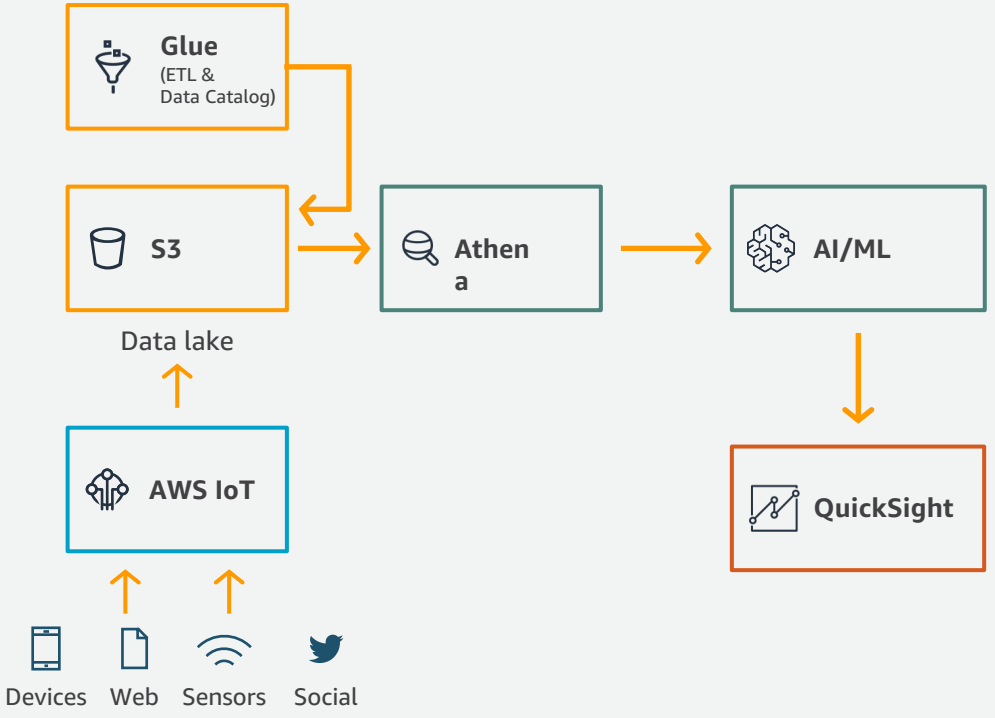
Analytics

Serverless. Zero infrastructure. Zero administration

Never pay for idle resources

Automatically scales resources with usage

Availability and fault tolerance built in



aws

When to use Athena vs EMR

Athena

- Ad-hoc querying
- Serverless
    - No setup or management of cluster
- Run queries using standard SQL
- Scales automatically based on complexity of queries

EMR

- Data processing/ETL
- Build and manage your own cluster
- Run custom applications and code
- Use big data processing frameworks
    - Spark, Hadoop, Presto, or HBase

aws

When to use Athena vs Redshift

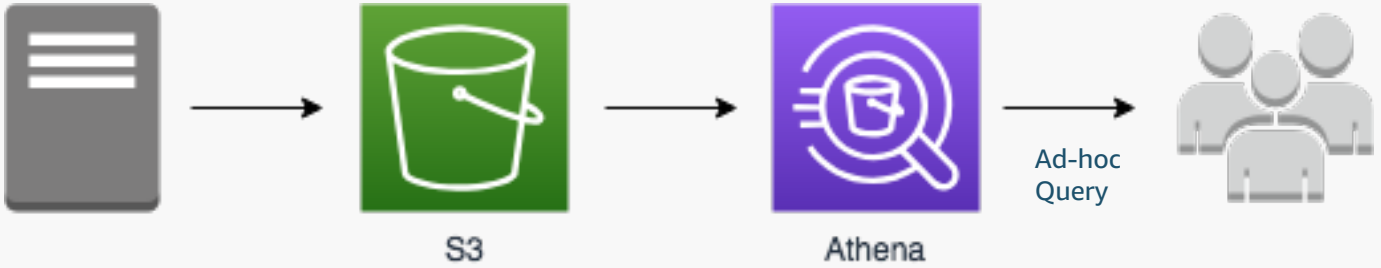
Athena

- Ad-hoc querying
- Serverless
    - No setup or management of cluster
- Run queries using standard SQL
- Scales automatically based on complexity of queries

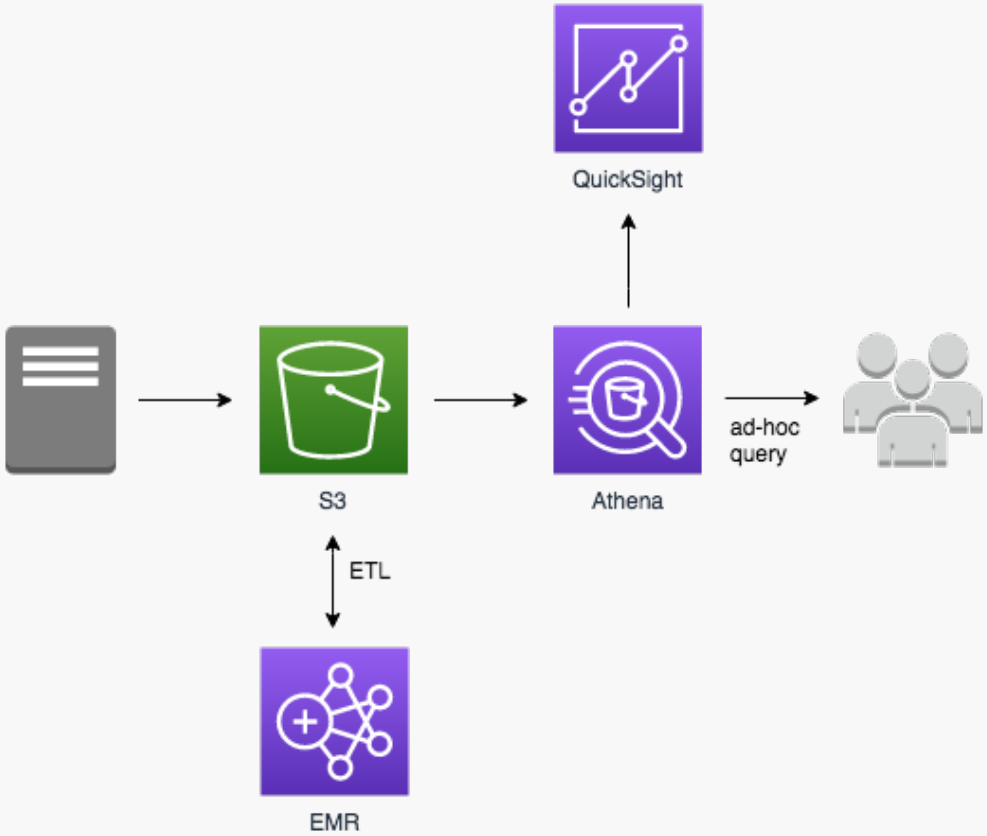
Redshift

- Data warehouse
    - Historical analysis and reporting
- Need to setup a cluster
- Run queries against highly structured data with many joins
- Can use same S3 data source as Athena

Ad-hoc Querying

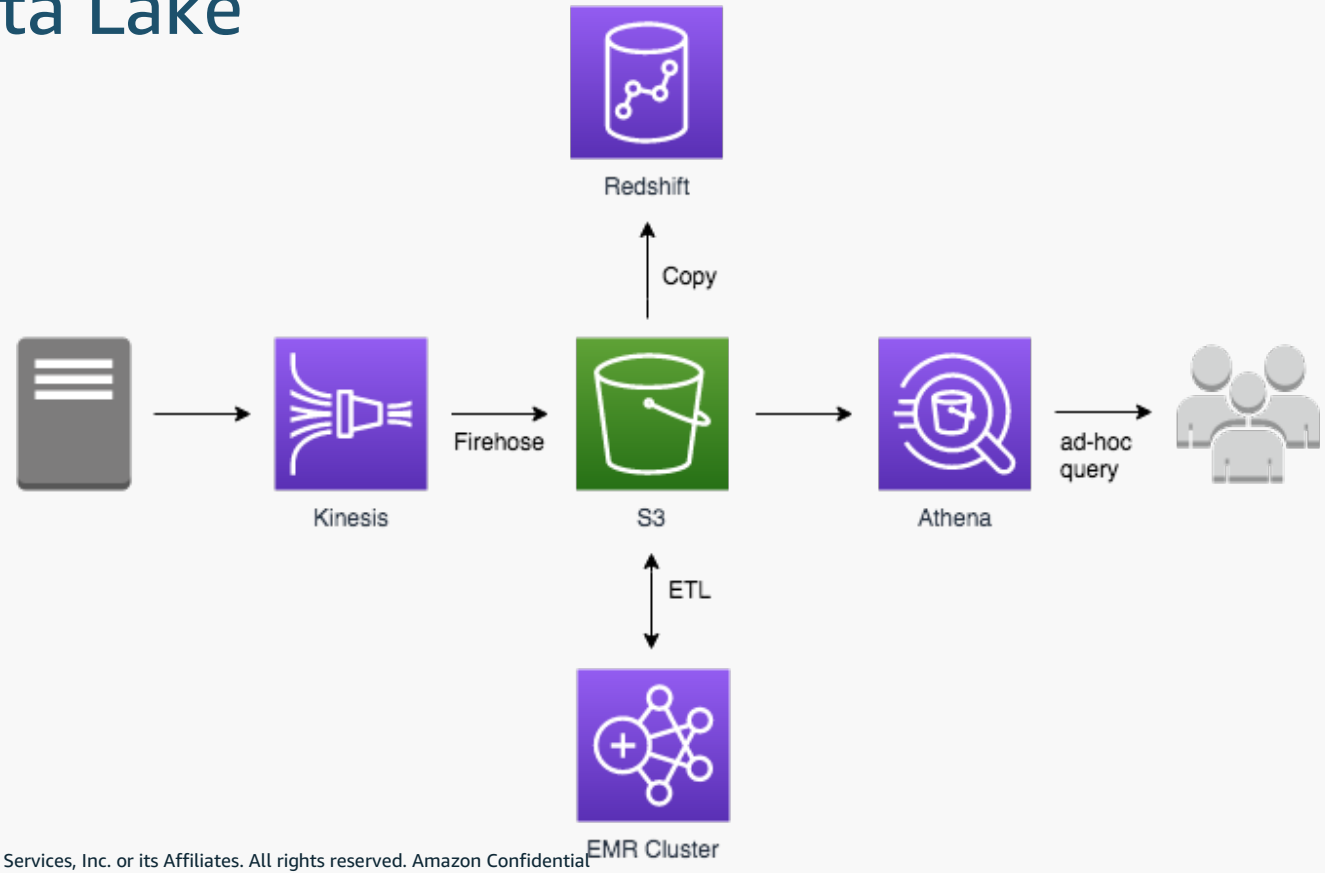


Ad-hoc Visualization and ETL

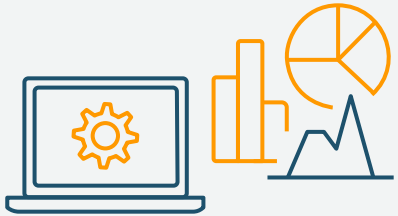


aws

ETL, Historical Analysis, and Ad-hoc Querying from S3 Data Lake



aws



Visualization & machine learning solutions

Visualization & Machine Learning

- **Dashboards**
- **Predictive Analytics**

aws

Visual insights for everyone
with Amazon QuickSight

Visualization &
Machine Learning

Pay only for what you use

Scale to tens of thousands of users

Embedded analytics

Build end-to-end BI solutions

aws

Start Visualizing Instantly

- Fully managed browser-based BI tool
  - No need to install software
  - No need to wait for updates
  - Provides full featured experience for a phone browser as well
- Serverless
  - No servers to provision
  - Scale seamlessly to tens of thousands of users
- Pay for what you use
- Fully Integrated with AWS services

aws

Connect to Existing Data Sources

On-premises

Securely connect to on-premise databases and flat files like Excel and CSV

- Excel
- JSON
- Teradata
- MySQL
- SQL Server
- PostgreSQL
- Delimited Files (CSV, TSV)
- Web Logs (CLF, ELF)

In the cloud

Connect to hosted database, big data formats, and secure VPCs

- Redshift
- RDS
- S3
- Athena
- Aurora
- Teradata
- MySQL
- Presto
- Spark
- SQL Server
- PostgreSQL
- MariaDB
- Snowflake
- IoT Analytics

Applications

Connect directly to third party business applications

- Salesforce
- Square
- Adobe Analytics
- Jira
- ServiceNow
- Twitter
- Github

Visual insights for everyone

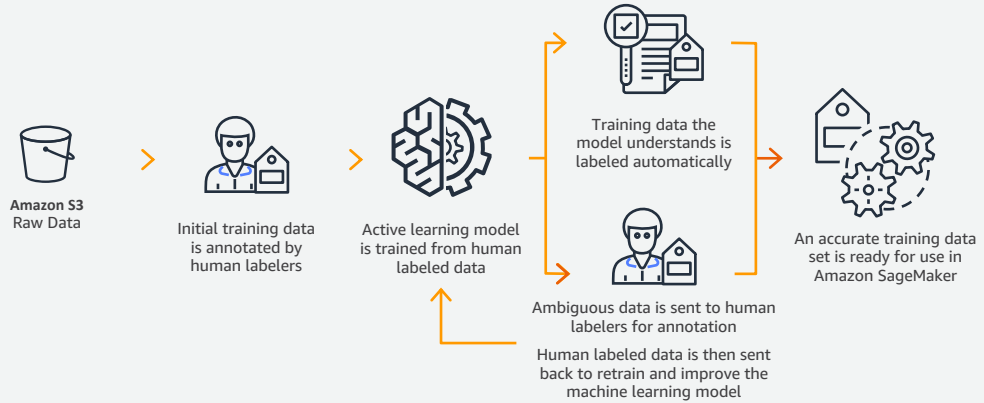
With AWS ML & AI services

Visualization & Machine Learning

Frameworks and interfaces for machine learning practitioners

Platform services that make it easy for any developer to get started and get deep with ML

Application services that enable developers to plug-in pre-built AI functionality into their apps



The Amazon ML stack

Broadest & deepest set of capabilities

AI Services

Vision

- Rekognition Image
- Rekognition Video
- Textract

Speech

- Polly
- Transcribe

Language

- Translate
- Comprehend
- & Comprehend Medical

Chatbots

- Lex

Forecasting

- Forecast

Recommendations

- Personalize

ML Services

Amazon SageMaker

Build

Pre-built algorithms & notebooks
Data labeling (Ground Truth)
Algorithms & models (AWS Marketplace for ML)

Train

One-click model training & tuning
Optimization (NEO)
Reinforcement learning

Deploy

One-click deployment & hosting

ML Frameworks & Infrastructure

Frameworks

TensorFlow
mxnet
PYTORCH

Interfaces

GLUON
Keras

Infrastructure

- EC2 P3 & P3dn
- EC2 C5
- FPGASs
- GreenGrass
- Elastic Inference

aws

Thank you!