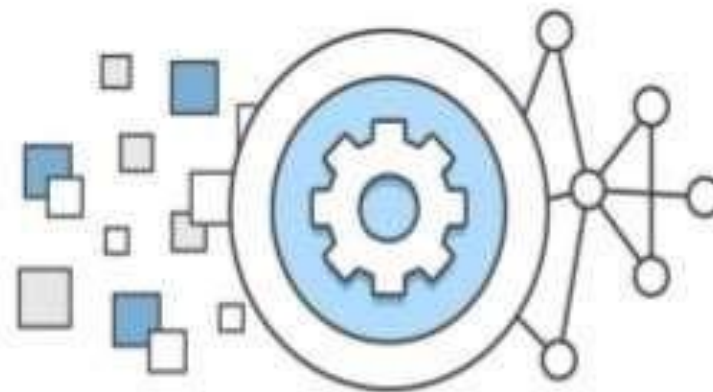# Introducing Amazon Athena

Amazon Athena is an **interactive query service** that makes it easy to analyze data directly from Amazon S3 using Standard SQL

Athena is Serverless

No Infrastructure or administration

Zero Spin up time

Transparent upgrades

Amazon Athena is Easy To Use

Log into the Console

Create a table
 Type in a Hive DDL Statement
 Use the console Add Table wizard

Start querying

Amazon Athena is Highly Available

You connect to a service endpoint or log into the console

Athena uses warm compute pools across multiple Availability Zones

Your data is in Amazon S3, which is also highly available and designed for 99.999999999% durability

Query Data Directly from Amazon S3

No loading of data

Query data in its raw format

    Text, CSV, JSON, weblogs, AWS service logs

    Convert to an optimized form like ORC or Parquet for the best performance and lowest cost

No ETL required

Stream data from directly from Amazon S3

Take advantage of Amazon S3 durability and availability

Familiar Technologies Under the Covers

Used for SQL Queries

In-memory distributed query engine

ANSI-SQL compatible with extensions

Used for DDL functionality

Complex data types

Multitude of formats

Supports data partitioning

Amazon Athena Supports Multiple Data Formats

Text files, e.g., CSV, raw logs

Apache Web Logs, TSV files

JSON (simple, nested)

Compressed files

Columnar formats such as Apache Parquet & Apache ORC

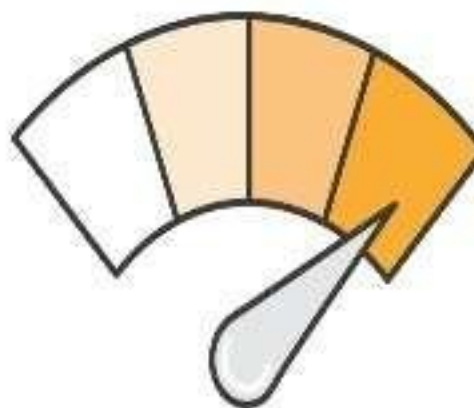AVRO support

Amazon Athena is Fast

Tuned for performance

Automatically parallelizes
queries

Results are streamed to console

Results also stored in S3

Improve Query performance
    Compress your data
    Use columnar formats

Amazon Athena is Cost Effective
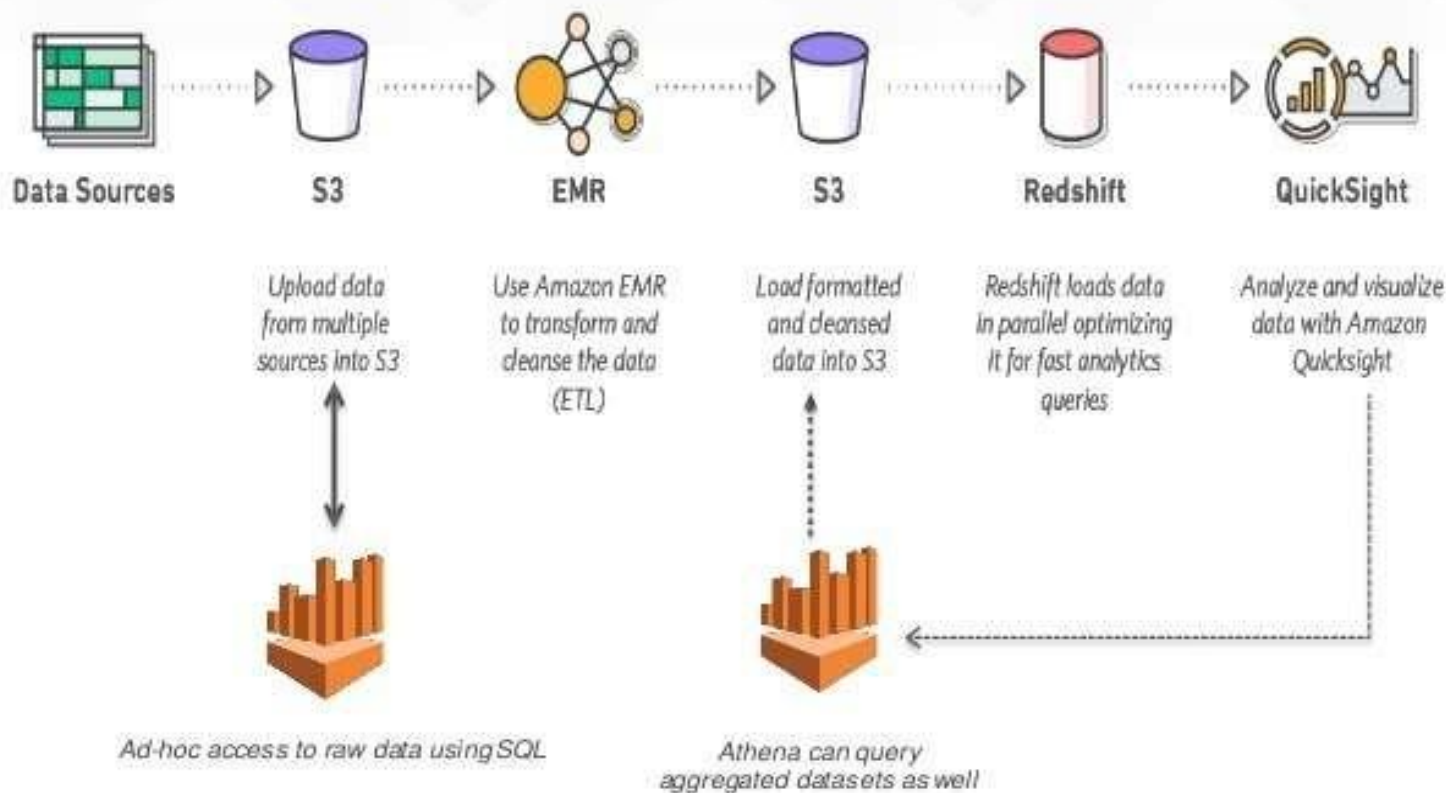
Pay per query

$5 per TB scanned from S3

DDL Queries and failed queries are free

Save by using compression, columnar formats, partitions

Building a Data Strategy on AWS

A Sample Pipeline

Data Sources	S3	EMR	S3	Redshift	QuickSight
	Upload data from multiple sources into S3	Use Amazon EMR to transform and cleanse the data (ETL)	Load formatted and cleansed data into S3	Redshift loads data in parallel optimizing it for fast analytics queries	Analyze and visualize data with Amazon Quicksight

Ad-hoc access to raw data using SQL

Athena can query aggregated datasets as well

Questions?