AWS Big Data Services

Andrew Brust
FOUNDER & CEO, BLUE BADGE INSIGHTS

@andrewbrust www.bluebadgeinsights.com

The Major Big Data Services

Simple Storage Service (S3)	Athena	Elastic MapReduce (EMR)	Redshift
Glue	Data Pipeline	DynamoDB/ NoSQL	Relational Database Service (RDS)

Services: S3

Amazon's cloud object store...

...And its data lake, too

"Special relationship" with EMR

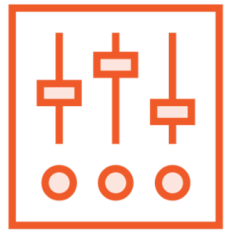
Virtually all AWS data services connect to S3

Buckets, folders, files

Services: Athena

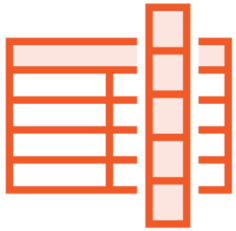
SQL query layer over S3

Based on Presto, Apache Hive



EMR = lots of options
S3 + Athena = all you need

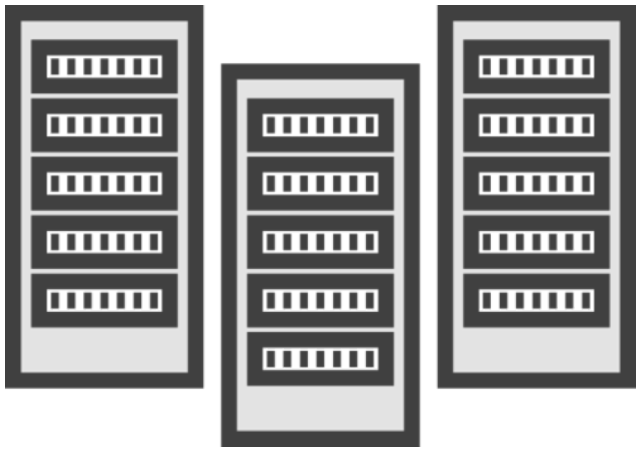
Consumption-based pricing



Works great with columnar file formats, e.g. Parquet, ORC

Leverages Glue data catalog

Services: Elastic MapReduce (EMR)



The big daddy of AWS Big Data

Apache Hadoop, Spark, and lots more

Tightly integrated with S3, via EMRFS

Primarily about big data/data lake analytics but also handles other workloads:

- Streaming data
- Data integration
- Machine learning/AI

And then there's the ecosystem

Services: Redshift

AWS' cloud data warehouse; pioneer in category

Was AWS' fastest-growing data service for years

Elastically scalable but does not use S3 for storage

Clusters run 24/7, with associated costs

Huge ecosystem support

Big competitor is Snowflake, which can run on AWS *and* use S3

Services: Glue

Data catalog and integration/prep platform

Crawlers, Tables and Jobs

Tight integration with S3, DynamoDB, Redshift, RDS and external databases (last three via JDBC)

Jobs' visual interface generates code that runs on (serverless) Spark

- Code uses high-level Glue API
- Editable, to a point

Glue data catalog is strategic

Services: Lake Formation

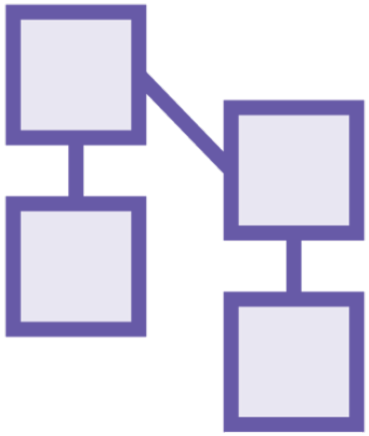
Service for data lake creation

Automation of:

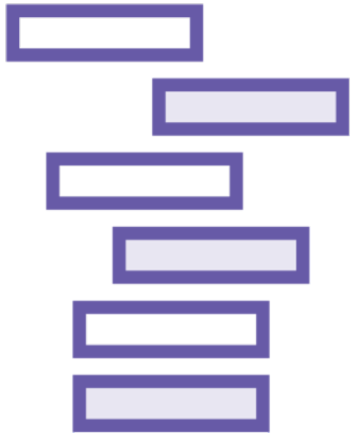
- Access controls
- Partitioning
- Deduplication (ML-based)
- Cleansing
- Classification

Builds on/heavily leverages Glue

Services: Data Pipeline



Visual “boxes-and-lines” service for data integration



Tight integration with S3, Redshift, DynamoDB

Jobs can be scheduled or run on-demand

Code-free but not simple

NoSQL: The Full AWS Story

DynamoDB

Key-value store that integrates with numerous other AWS data services

DocumentDB

Document store

Neptune

Graph database

Apache HBase on Elastic MapReduce

Column family store

SimpleDB

Deprecated key-value store, superseded by DynamoDB

What About Operational Databases?

Relational Database Service (RDS)

- Oracle, SQL Server
- MySQL, MariaDB
- PostgreSQL (aka “Postgres”)
- Aurora
  - Cloud-native/serverless
  - MySQL- and Postgres-compatible

Mapping the Services

Big Data Technology	Amazon Service
Data Warehouse	Redshift
Data Lake	S3, Athena
Batch Analytics	EMR
Relational	RDS
NoSQL	DynamoDB
Streaming Data	Kinesis, MSK
Data Integration	Glue, Data Pipeline
Artificial Intelligence	SageMaker

Demo

Provisioning an EMR cluster

Choosing OSS analytics components