Big Data on AWS: The Big Picture

INTRODUCTION: BIG DATA CONCEPTS

Andrew Brust
FOUNDER & CEO, BLUE BADGE INSIGHTS

@andrewbrust www.bluebadgeinsights.com

Concepts: Big Data

Cliché:
Volume, velocity, variety

Literal:
100s of TB or higher

Credo:
Aggregations/analysis on raw data, in standalone files

Business:
Analyze data that is:
Relevant & important;
not conformed to traditional systems

Technology:
Hadoop was foundational;
Spark is successor

Concepts: Data Lakes

- Euphemism for Hadoop and Big Data?
- Storage systems and agnostic file formats together treated as virtual database
- Multiple engines against same data
- Initial importance of HDFS
- New importance of cloud object storage

Concepts: NoSQL

Big Data tie-in: semi-structured data + schema flexibility

Important producer of data for Big Data analytics

Big Data and NoSQL overlap, in HBase

Dominant indies: MongoDB, DataStax

Cloud providers taking market share

Most NoSQL platforms now support SQL!

Concepts: Internet of Things

IoT: Internet connectivity for low-powered devices

Provides telemetry, sensor data

Time series format works well for analytics

Broad use cases:

- Maintenance, remote monitoring and asset tracking

Common applications:

- Preventive/proactive maintenance
- Usage/traffic data for municipalities
- Social media sentiment analysis
- Financial market data analysis
- Consumer: thermostats, appliances

Concepts: Machine Learning/AI

Historical data, relationships can be modeled to support predictions of future outcomes

Numerous algorithms and frameworks, open source and proprietary

Picking the right algorithm and "hyperparameter" value beyond most developers

- But automation is emerging

Development, deployment, monitoring and managing are all needed.

- Some are more evolved than others

Concepts: MapReduce & Massively Parallel Processing

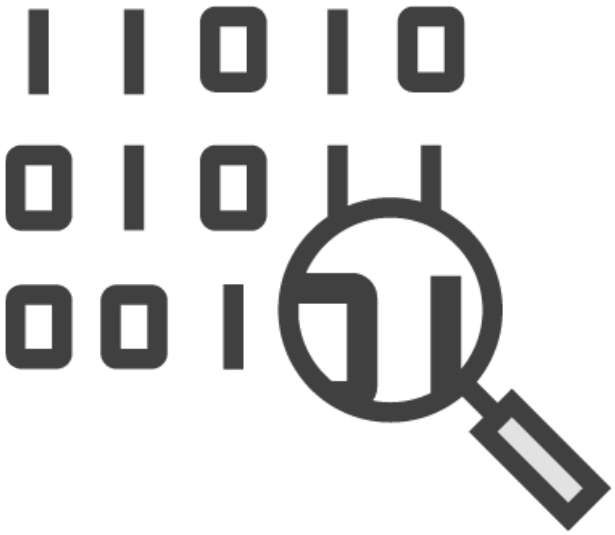
Both algorithms based on divide and conquer for large data volumes

- Create a cluster with lots of servers
- Node get subset of data to work on
- Work in parallel; output quickly
- Got more data? Add more servers
- Cloud, elasticity work well here

MR: two passes (parsing and aggregating)

MPP: partitions query across RDBMS nodes

Concepts: Streaming



Streaming data produced is high-volume/ small-payload

Bread and butter of real-time analytics

Produced in various scenarios:

- Internet of Things/sensors, financial markets, social media, Web analytics

Small number of dominant open source technologies

Major cloud providers offer proprietary streaming platforms

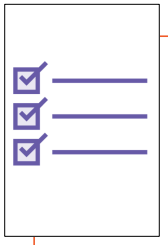
Concepts: SQL

Structured Query Language, around since the 70s

Huge pool of technologists with basic competency

Newer startups tried to abandon it; failed

Universally understood, declarative query language too valuable to abandon



Used by:

- Operational relational databases
- Data Warehouses
- NoSQL platforms
- Big data, data lake query engines

Summary

Big Data used to just mean Hadoop

Now encompasses:

- Cloud object storage
- Data warehousing
- Streaming data
- Data integration pipelines
- Data visualization
- And even machine learning/AI

AWS has services for each of these, and sometimes more than one