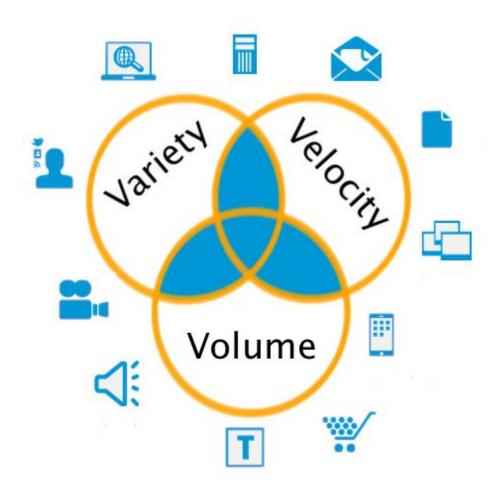
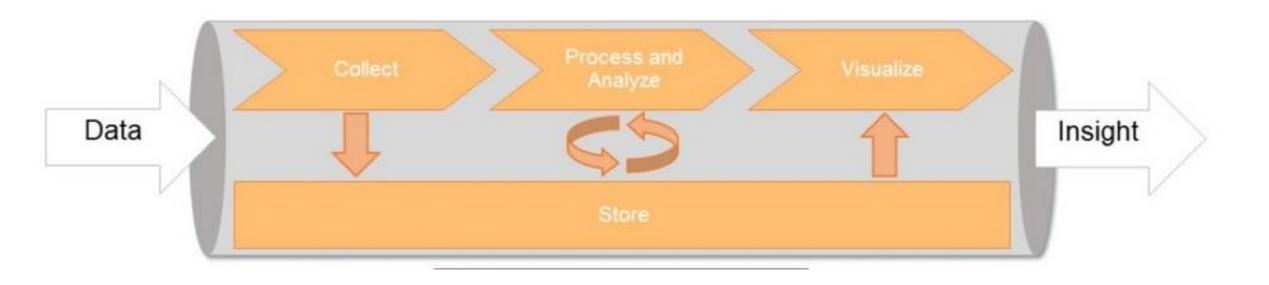
# Big Data Processing

Anusuriya Devaraju 01.03.2017

# What is Big Data?



# The Big Data Pipeline



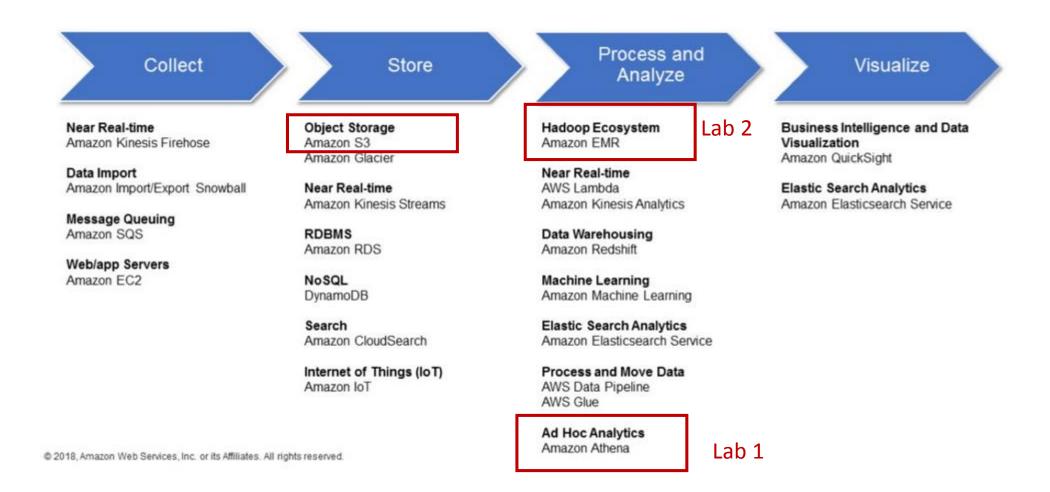
Source: 2018, Amazon Web Services, Inc.

#### Outline

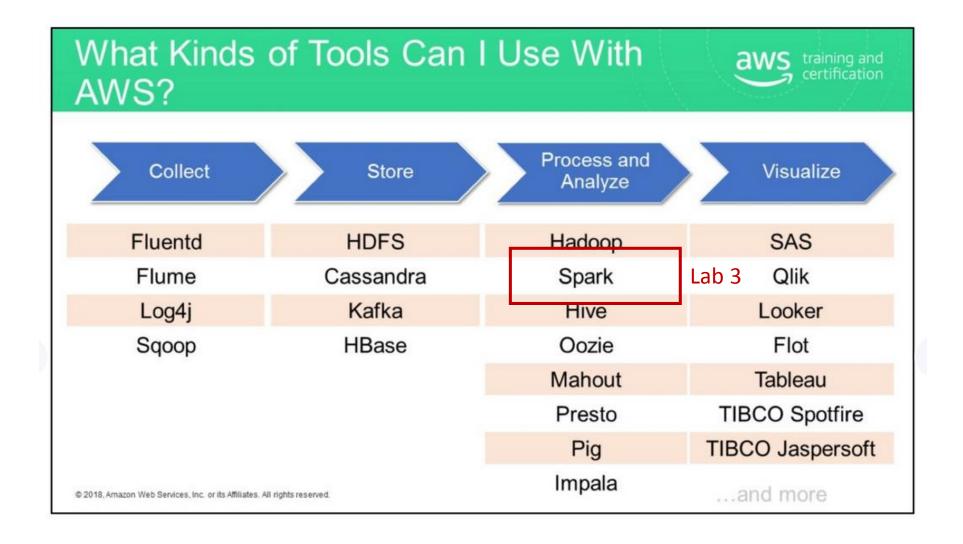
- Basic Concepts
- Hands-on & Review Questions
  - Lab 1 Using Amazon Athena to Analyze Log Data
  - Lab 2 Processing Server Logs with Hive on Amazon EMR
  - Lab 3 Apache Spark on Amazon Web Services EMR

Reference: Some of the following slides were modified from AWS. Big Data on AWS 3.2 (EN): Student Guide. AWS/Gilmore. VitalBook file.

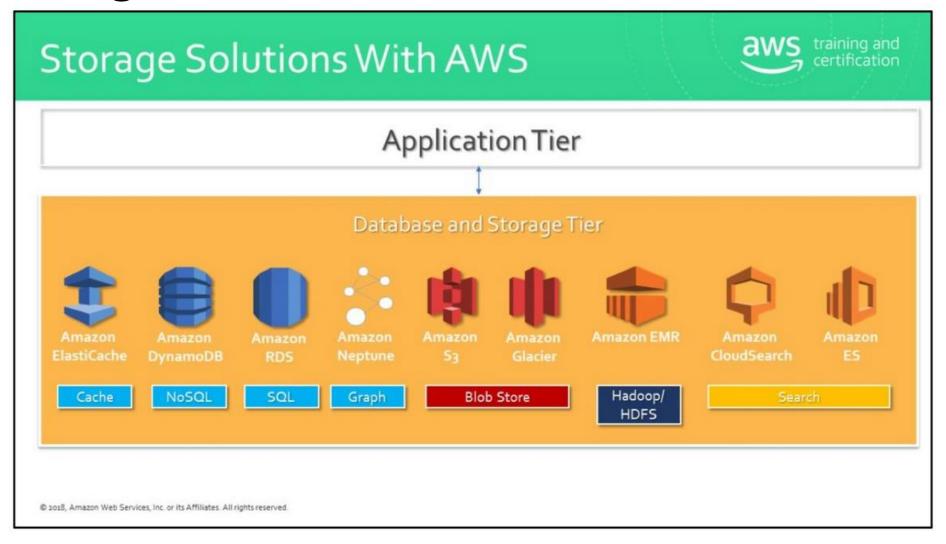
## From AWS Solutions to the Big Data Pipeline



#### Other Tools That Can be Used With AWS



## Storage Solutions With AWS



#### Outline

- Basic Concepts
- Hands-on & Review Questions
  - Lab 1 Using Amazon Athena to Analyze Log Data
  - Lab 2 Processing Server Logs with Hive on Amazon EMR
  - Lab 3 Apache Spark on Amazon Web Services EMR

# Lab 1 - Using Amazon Athena to Analyze Log Data

 The following log data is provided in an Amazon S3 bucket stored in the US East (North Virginia) Region.

```
2015-01-01T00:00:00.022719Z elb_demo_005 244.218.91.244:2255 172.36.231.239:443 0.000878 0.000803 0.000891 200 200 0 1886 "GET https://www.example.com/jobs/376 HTTP/1.1" "Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.111 Safari/537.36" DHE-RSA-AES128-SHA TLSv1.2
```

- Raw data stored in plain text
  - a) Raw data stored in plain text
  - b) Compressed data in gzip format
  - c) Compressed and Partitioned data split into sub-directories
  - d) Columnar data, stored in Parquet format (compressed using Snappy compression)

# Lab 1 - Results

TASKS	RUNTIME & DATA SCANNED
Task 1: Raw data stored in plain text	(Run time: 19.57 seconds, Data scanned: 90.72GB)
Task 2: Compressed data in gzip format	(Run time: 24.5 seconds, Data scanned: 13.08GB)
Task 3: Compressed and Partitioned data	(Run time: 19.54 seconds, Data scanned: 432.46MB)
split into sub-directories	
Task 4: Columnar data (Parquet with	(Run time: 2.57 seconds, Data scanned: 18.44MB)
Snappy compression)	

### Lab 1 – Conclusions and Q&A

Q1. Name few benefits of Amazon Athena

Q2. In your Lab 1, which data format improve the performance and save cost in Amazon Athena, and why?

Q3. Where does Amazon Athena store information and schemas about database?

#### Outline

- Basic Concepts
- Hands-on & Review Questions
  - Lab 1 Using Amazon Athena to Analyze Log Data
  - Lab 2 Processing Server Logs with Hive on Amazon EMR
  - Lab 3 Apache Spark on Amazon Web Services EMR

### Apache Hadoop

#### Apache Hadoop: Primary Components



#### MapReduce

- Distributed, key-value based programming framework
- Can be written in Java natively or Python, Bash, or other languages with add-ons
- Used to analyze huge, distributed data sets using parallel, distributed processing

### HDFS (Hadoop Distributed File System)

- Fault tolerant, distributed storage system
- Primary storage system used by the Apache Hadoop framework
- Creates replicas of data blocks and distributes them through clusters

#### YARN

- Hadoop 2.x only
- Schedules jobs and manages cluster resources

#### Hadoop Common

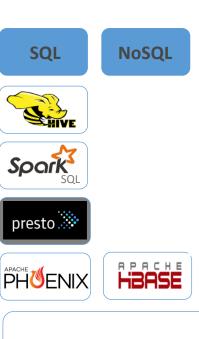
- Libraries and utilities used by other modules
- Package includes HDFS, Hadoop engine, and necessary JAR files to run Hadoop

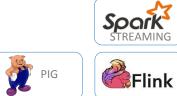
© 2018, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

#### Amazon EMR

- Managed cluster platform
- Run Hadoop, Spark and other applications
- Easy launch cluster in minutes!
- Use HDFS and S3 file systems
- Resize a running cluster
- Low cost –only pay for the resources you actually use!









**Streaming** 



Spark

Machine

Learning



HUE

IDE



**Utilities** 

Ganglia





**Transform** 



AmazonS3



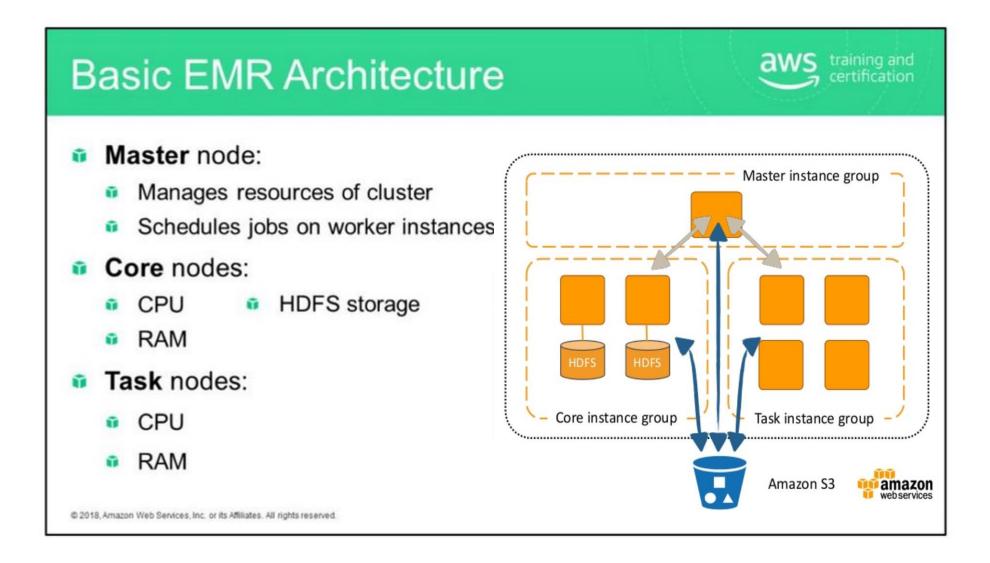
**Applications** 

Framework

**Process Layer** 

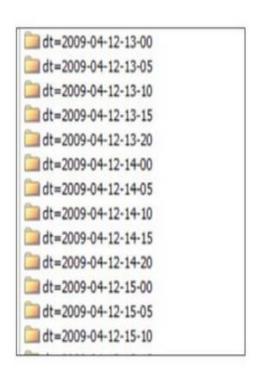
Data Layer

#### Amazon EMR Architecture



# Lab 2 - Processing Server Logs with Hive on Amazon EMR

- Launch an Amazon EMR cluster with Apache Hive installed
- Use Hive commands to create external relation database tables from log data stored in Amazon S3
- Use Hive to query the tables you create and persist in Amazon S3.

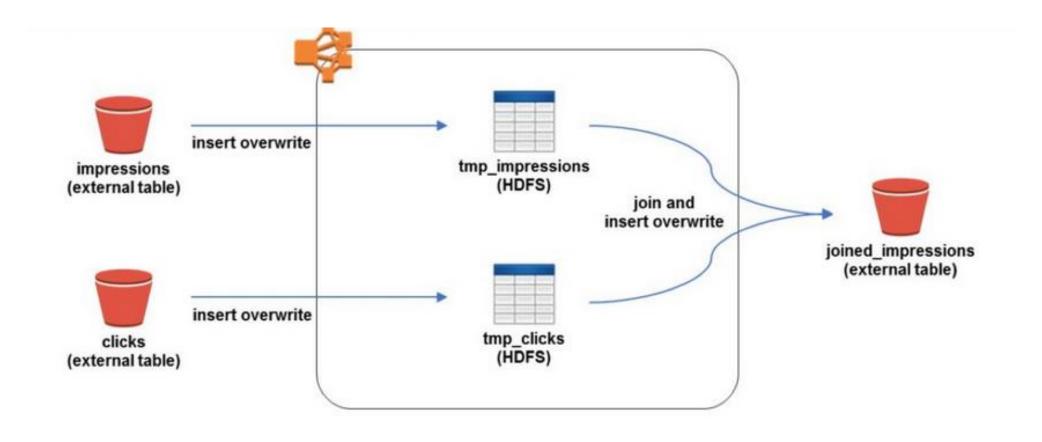


Stored log structure in Amazon S3

"number": "95164", "referrer": "palpal.com", "processid": "1762", "adid": 96blt4f0Mx5u38Qk9x8doGXgkactee\*, "browserCookie": "rmbetcpihd", "userCookie": "xBm25DcX9FlbHm1FJ3CWtJTBN45oam", "requestEndTime": "1239541871000", 'impressionId": "2xF7V9TgDdDaStmFAqFCwHBL9cV9Ln", "userAgent": "Mozilla/4.0 compatible: MSIE 8.0: Windows NT 5.1: Trident/4.0)\*, "timers": "modellookup": "0.277", "requestTime": "0.8606"), "threadId": "13", "ip": "20.143.145.26", "modelId": "hxxiuxduab", "hostname"; "ec2-0-51-75-39.amazon.com", "sessionId": "UphjCmFISBtBELaxF2MSnlp9lbbOXPS", requestBeginTime\*: \*1239541870000\*) "number": "97742", "referrer": "paipal.com", "processId": "1601", "adId": "nUSSOND2:NE; @M4prfu9GEX8v7dk12", "browserCookie": "hivwtkeghd", "userCookie": "dhQFvtSeAFdQdsFN93219bPv68N38X", "requestEndTime": "1239541797000", "impressionId": "DBLF6TfBpgDm7gFRxsGCjB9dtcEOnV", "userAgent": "Mozilla/4.0 compatible: MSIE 6.0: Windows NT 5.1: SV1: .NET CLR 1.1.4322: .NET CLR 2.0.50727: .MET CLR 3.0.04506.50: InfoFat\*, "timers": ("modellookup": \*0.2844\*, "requestTime": "0.6756"), "threadId": "51", "ig": "65.130.207.41", "modelId": "bxxiuxduab", "hostname": "ec2-0-51-75-39.amazon.com", "sessionId": \*0Xr0@oFBpgLUMJg197W09589X6Jr3s\*, \*requestSeginTime\*: \*123954179@000\*) "number": "102384", "referrer": "pancramic.com", "processid": "1646", "adid": \*C6V3aHPmnKBgpdc3R72H32wDF8tJr2\*, \*browserCookie\*: \*hicibsnika\*, \*userCookie\*: "wFioSbsVQk13nEJ1kWLILjoqoCSosx", "requestEndTime": "1239541516000", "impressionId": "omdbgsgHfljlknwdQmankCCK4nifve", "userAgent": "Mozilla/5.0 Windows: U: Windows NT 5.1: en-U5) AppleWebKit/550.5 (XHTML, like Gecko) Chrome/2.0.172.37 Safari/530.5", "timers": ("modelLockup": "0.2726", "requestTime": "0.784"), "threadId": "70", "ip": "51.140.241.78", "modelId": "bxxiuxduab", "hoginame": "ec2-0-51-75-39.amazon.com", "sessionId": "HTCcOjSVeN6daXqRMAxjVG22ROcsbt", "requestSeginTime": "1239541516000")

Log data (sample)

# Lab 2 - Processing Server Logs with Hive on Amazon EMR



#### Lab 2 - Review

- What are the primary Apache Hadoop components?
- List applications for which Apache Hadoop is not suitable.

• What is the difference between a transient and a long-running cluster?

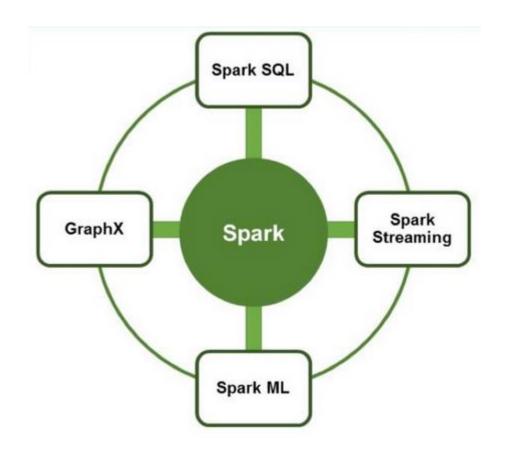
List advantages of running Hive and Amazon EMR

#### Outline

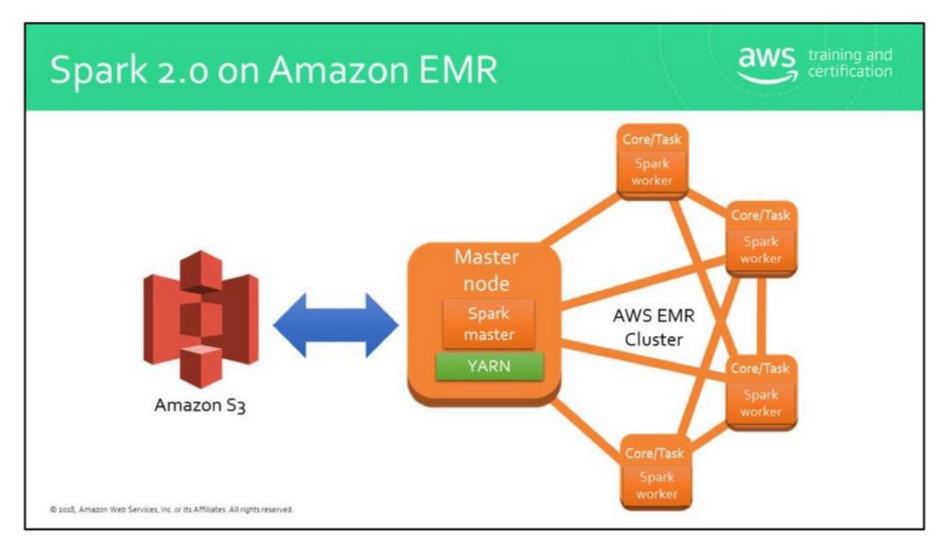
- Basic Concepts
- Hands-on & Review Questions
  - Lab 1 Using Amazon Athena to Analyze Log Data
  - Lab 2 Processing Server Logs with Hive on Amazon EMR
  - Lab 3 Apache Spark on Amazon Web Services EMR

## Apache Sparks

- Allows in-memory data mining and querying big datasets at fast speeds
- Resilient distributed datasets (RDD)
- Faster than MapReduce!
- Supports batch, interactive and streaming applications
- Can read and store data in HDFS, Amazon S3, and other databases.
- JDBC/ODBC connector compatible



## Spark 2.0 on Amazon EMR



# Lab 3 - Apache Spark on Amazon Web Services EMR

- 1. Create an AWS S3 storage bucket to hold data inputs, data outputs and logs
- 2. Create, Configure and Launch an Amazon EMR Cluster
- 3. Log in to the cluster master node
- 4. Run Spark-SQL Command Line Interface and Issue commands to create tables and run queries
- 5. Shut down cluster and remove any temporary resources

#### Lab 3 - Review

- SparkQL Exercise: Find the total trading volume (trade\_size) for each stock before 12:00 noon
- What are the alternatives to load data to S3 from your local PC?

# Data61 Research Big Data Cluster

