Hadoop Intro

Data Explosion

IBM reported that 2.5 billion gigabytes of data was generated every day in 2012

40,000 search queries on Google every second.

300 hours of video uploaded to YouTube every minute.

31.25 milion messages sent and 2.77 milion videos views by Facebook users every minute.

By 2017, nearly 80% of photos will be taken on smartphones.

By 2020, at least a third of all data will pass through Cloud.

By 2020, about 1.7 megabytes of new information for every human.

Method to handle data

Distributed Systems

Challenges of Distributed Systems

High chances of system failure

Limit on bandwidth

High programming complexity

Hadoop

Hadoop is a framework that allows for distributed processing of large data sets across clusters of commodity computers using simple programming models.

Inspired by technical document published by Google.

1. Hadoop solve distributed system high failures.
2. Limit on band width.
3. Programming complexity

Hadoop Key Characteristics

1. Reliable: Stores copies of the data on different machines and is resistant to hardware failure.
2. Economical: Ordinary computers can be used for data processing
3. Scalable: Can follow both horizontal and vertical scaling.
4. Flexible: Can store as much of the data and decide to use it later.

Taditional Database System vs. Hadoop

Traditional Database System: Data sent to program

Hadoop: Program sent to Data

Hadoop Components

Hadoop HDFS

1. A storage layer for Hadoop
2. Suitable for the distributed storage and processing

3. Hadoop provides a command line interface to interact with HDFS
4. Streaming access to file system data
5. Provides file permissions and authentication

What store in HDFS ---- HBase

1. HBase stores data in HDFS
2. A NoSQL database or non-relational database
3. Mainly used when you need random, real-time, read/write access to your Big Data.
4. Provides support to high volume of data and high throughput
5. The table can have thousands of columns.

How data transfer in HDFS ---- Sqoop & Flume

1. Sqoop is a tool designed to transfer data between Hadoop and relational database servers.

2. It is used to import data from relational databases such as, Oracle and MySQL to HDFS and export data from HDFS to relational databases.

3. Flume is a distributed service for ingesting streaming data.

4. Ideally suited for event data from multiple systems.

How to process data ---- Spark & MapReduce

1. An open-source cluster computing framework
2. Provides 100 times faster performance as compared to MapReduce

3. Supports Machine learning, Business intelligence, Streaming, and Batch processing

4. The original Hadoop processing engine which is primarily Java based

5. Based on the map and reduce programming model

6. An extensive and mature fault tolerance framework

7. Commonly used, but less than Spark

How to analyze data ---- Pig & Impala & Hive

1. An open-source dataflow system
2. Converts pig script to Map-Reduce code
3. An alternate to writing Map-Reduce code

4. Best for ad-hoc queries like join and filter

5. High performance SQL engine which runs on Hadoop cluster

6. Ideal for interactive analysis

7. Very low latency - measured in milliseconds

8. Supports a dialet of SQL (Impala SQL)

9. Similar to Impala

10. Best for data processing and ETL

11. Executes queries using MapReduce

How to access data -- Cloudera Search

1. One of Cloudera's near-real-time access products

2. Users do not need SQL or programming skills to use Cloudera Search
3. Enables non-technical users to search and explore data stored in or ingested into Hadoop and HBase
4. A fully intergrated data processing platform.

How to manage Hadoop jobs ---- Oozie

1. Oozie is a workflow or coordination system used to manage the Hadoop jobs

Make Hadoop easy to use ---- Hue

1. Hue is an acronym for Hadoop User Experience
2. It provides SQL editors for Hive, Impala, MySQL, Oracle, PostgreSQL, Spark SQL, and Solr SQL

How these components work together.

4 Steps:

1. Ingest: sqoop, Flume
2. Processing: HDFS, HBase, MapReduce, Spark
3. Analyze: Impala, Hive, Pig
4. Access: Hue, cloudera search