Before the year 2000, the data processing was done through the computing power of local processor. Later as data grew, larger size computation replaced them. However, post 2000 the data growth outpaced the computation growth and it became not feasible. As the years went there was been increasing growth in the amount of data being generated.

Hence there needs to be a solution to manage this ever growing data. We can use multiple machines called distributed systems to process the data. As the number of machine are added by a factor of X, the total time needed decreases by a factor of X. Thus distributed systems needs less time to process Big Data.

There are however challenges of Big Data. There are high chances of system failure and also limit on bandwidth programming. The programming complexity is also high since it is difficult to synchronize data and process.

Hadoop is a framework that allows for distributed processing of large data sets across clusters of commodity computers using programming models. Hadoop resolves the three challenges of distributed systems through following characteristics:

1. Economical

Its system are highly economical since ordinary computers can be used for data processing.

1. Reliable

It is reliable since it stores copies of the data on different machines and is resilient to hardware failure.

1. Scalable

It is easily scalable since it allows both vertical and horizontal scaling.

1. Flexible

It is flexible since you can store as much structure and unstructured data as you need to and decide to use it later.

In Hadoop, the program goes to the data and not vice versa. It initially distributes the data to multiple systems and runs the computation locally wherever the data is present.

Hadoop Ecosystems comprises of 12 components.

1. Hadoop Distributed file system

HDFS is a storage layer for Hadoop suitable for distributed storage and processing. HDFS provides streaming access to file system data and finally provides file permissions and authentication. Hadoop uses a command line interface to interact with HDFS.

1. HBase

HBase stores data in HDFS. It is a NoSQL database. It is mainly used when you need random, real time, read/write access to Big Data. It provide support to high volume of data and high throughput.

1. Sqoop

Data is ingested or transferred to HDFS via Sqoop. It is used to import data from Relational Databse to HDFS and export Data from HDFS to relational database.

1. Flume

If you want to ingest event data, then you use Flume. It is distributed service for ingesting streaming data.

1. Spark

After the data is transferred in HDFS, it is processed. One of the framework that processes the data is Spark. It provides up to 100 times faster performance for few applications within memory primitives. It can run in the Hadoop cluster and processes data in HDFS. It supports wide range of workload. It has following components:

* Spark Core and Resilient Distributed Datasets
* Spark SQL
* Spark Streaming
* Machine Learning Library
* GraphX

1. Hadoop Map Reduce

Hadoop Map Reduce is the other framework that process data. It was original processing engine and is Java based. It is based on Map and Reduce Programming Model. It has extensive and and fault tolerant framework. It is still commonly used.

1. Pig

After the data is processes, it is analyzed by an open source high level data flow system called pig. Pig reduces its code to Map-Reduce code. It is best for ad-hoc queries such as join and filter.

1. Impala

One can also use Impala to analyze the data. It is a high performance SQL engine which runs on Hadoop cluster. It is ideal for interactive analysis and has very low latencies. It supports a dialect of SQL

1. Hive

Hive can also be used to perform data analysis. It is abstraction layer on top of Hadoop. It is similar to Impala. However it is preferred to Impala for data processing and data extract transform load.

1. Cloudera search

After the data is analyzed it is ready for the users to access. What supports the search of data, it can be done using cloudera search. Search is one of cloudera near real-time access products. It enables non-technical users to search and explore data stored in or ingest it into Hadoop and HBase. Users do not need SQL or programming skills to use cloudera search because it provides a simple full-text interface for searching. Another benefit of cloudera search compared to standalone search solutions is a fully integrated data processing platform. Cloudera search uses the flexible scalable and robust storage system included with CD 8 or cloudera distribution including Hadoop. This eliminates the need to move large data sets across infrastructures to address business tasks.

1. Hue

Another component in Hadoop ecosystem is Hue. Hue is an acronym for Hadoop User Experience. It is an open source web interface for Hadoop. You can perform the following operations using hue: upload and browse data, query a table in hive and Impala, run spark and pig jobs and workflows, search data. Hue makes Hadoop easier to use. It also provides SQL editor for hive, Impala, MySQL, Oracle, PostGRE SQL, spark SQL and solar SQL.

1. Oozie

Hadoop jobs such as MapReduce, Pig, Hive and sqoop have workflows. Oozie is a workflow or coordination system that you can use to manage the Hadoop jobs. Multiple actions occur between the start and end of the workflow.

There are four stages to process Big Data:

1. Ingest

The first stage of big data processing is ingest. The data is ingested or transferred to Hadoop from various sources such as relational databases systems or local files.

1. Processing

The second stage is processing. In this stage the data is stored and processed. As discussed earlier that the data is stored in the Hadoop distributed file system - HDFS and the NoSQL distributed data – Hbase. Spark and MapReduce perform the data processing

1. Analyze

The third stage is analyze. Here the data is analyzed by processing frameworks such as pig, hive and Impala. Pig converts the data using map and reduce and then analyzes it. Hive is also based on map and reduced programming and is most suitable for structured data.

1. Access

The fourth stage is access which is performed by tools such as Hue and cloudera search. In this stage the analyzed data can be accessed by users. Hue is the web interface whereas cloudera search provides a text interface for exploring data.

Total Time : 30 minutes