MODULE III PART 1

At the end of the session you will be able

- Define the Hadoop Framework.
- Explain the features of Hadoop.
- Explain the components Hadoop Ecosystem.

What is Hadoop?

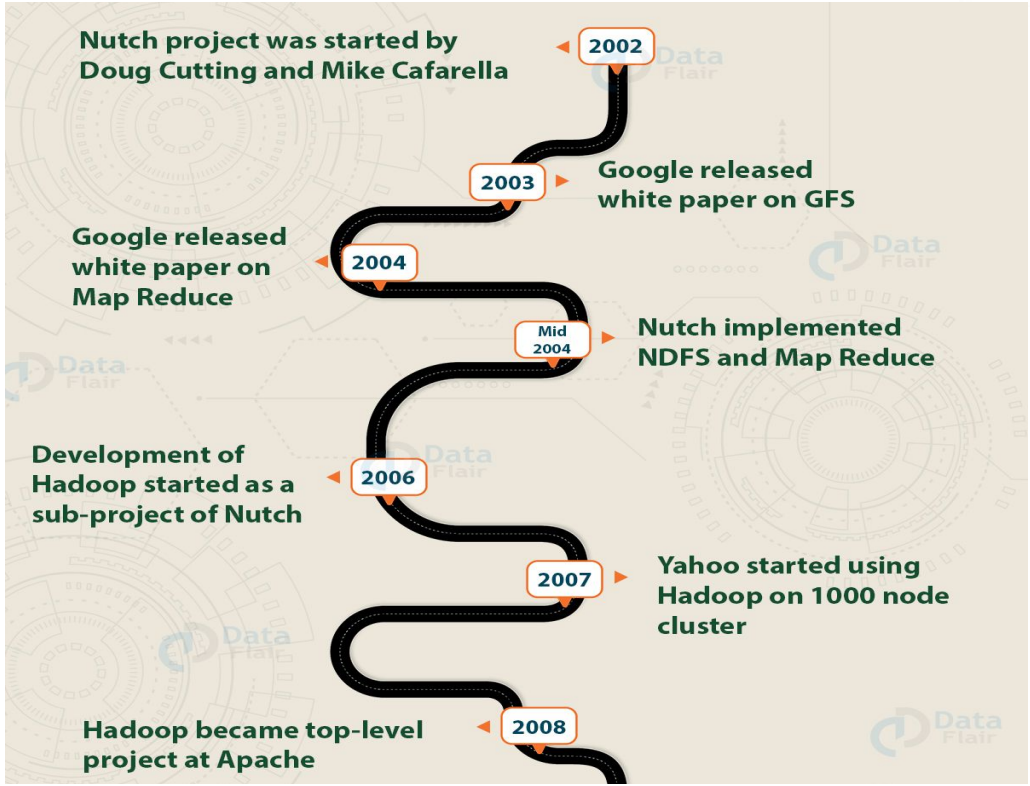
- An open-source framework
- Allows to store and process big data in a distributed environment across clusters of computers using simple programming models.
- Developed at the Apache Software Foundation.

“Hadoop is a technology to store massive datasets on a cluster of cheap machines in a distributed manner”. -Doug Cutting and Mike Cafarella.

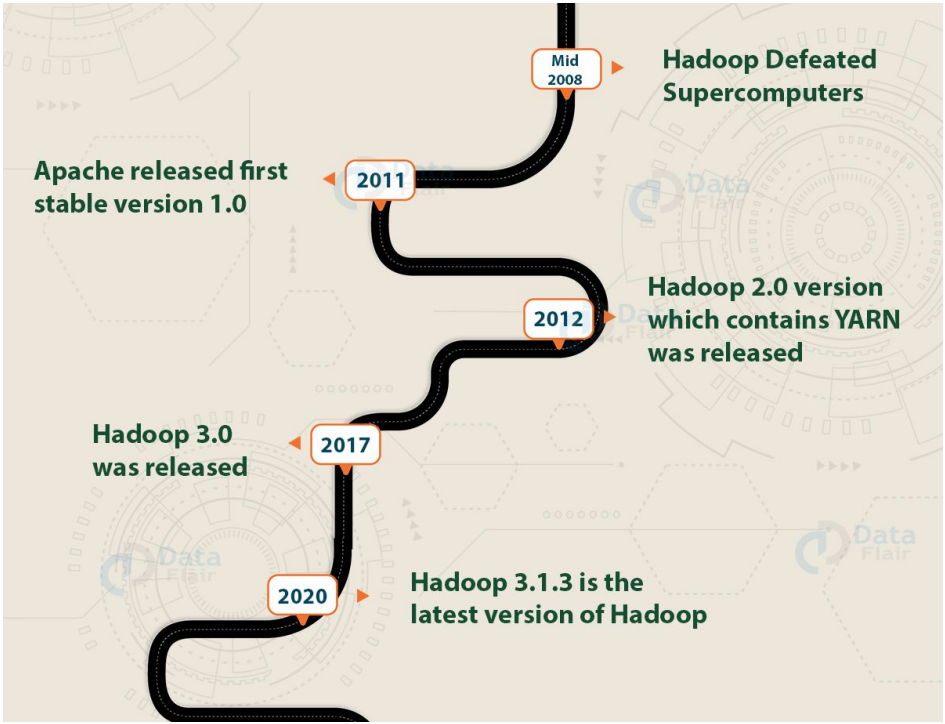
Why Hadoop?

- Shortcomings of the traditional approach
  - Storage for Large Datasets
  - Handling data in different formats
  - Data getting generated with high speed

History Hadoop



History Hadoop



Features of Hadoop

- Apache Hadoop is the most popular and powerful big data tool
  - Open Source
  - Highly Scalable
  - Provides fault tolerance
  - High availability
  - Cost-effective
  - Faster in data processing
  - Provides feasibility
  - Ensures data reliability

Hadoop Ecosystem

- Each of the Hadoop Ecosystem Components is developed to deliver explicit function.
- Each has its own developer community and individual release cycle.
- Hadoop has two major layers:
    - **Hadoop Distributed File System (HDFS)-Storage Layer**
    - **MapReduce-Processing or Computation Layer**

Hadoop Ecosystem

Data Flair

oozie

oozie

(Work flow)

Table & schema Management

Pig (Scripting)

HIve (Sql Query)

(Machine Learning)

Drill (Interactive Analysis)

AVRO (JSON)

Thrift (Cross Language Service)

HBASE

HBASE (Columnar Store)

Sqoop

(Data Collection)

Zookeeper (Coordination)

Ambari

Apache Ambari (Management & Monitoring)

Mapreduce (Data Processing)

Yarn (Cluster Resource Management)

HDFS (Hadoop Distributed File system)

FLUME

Flume

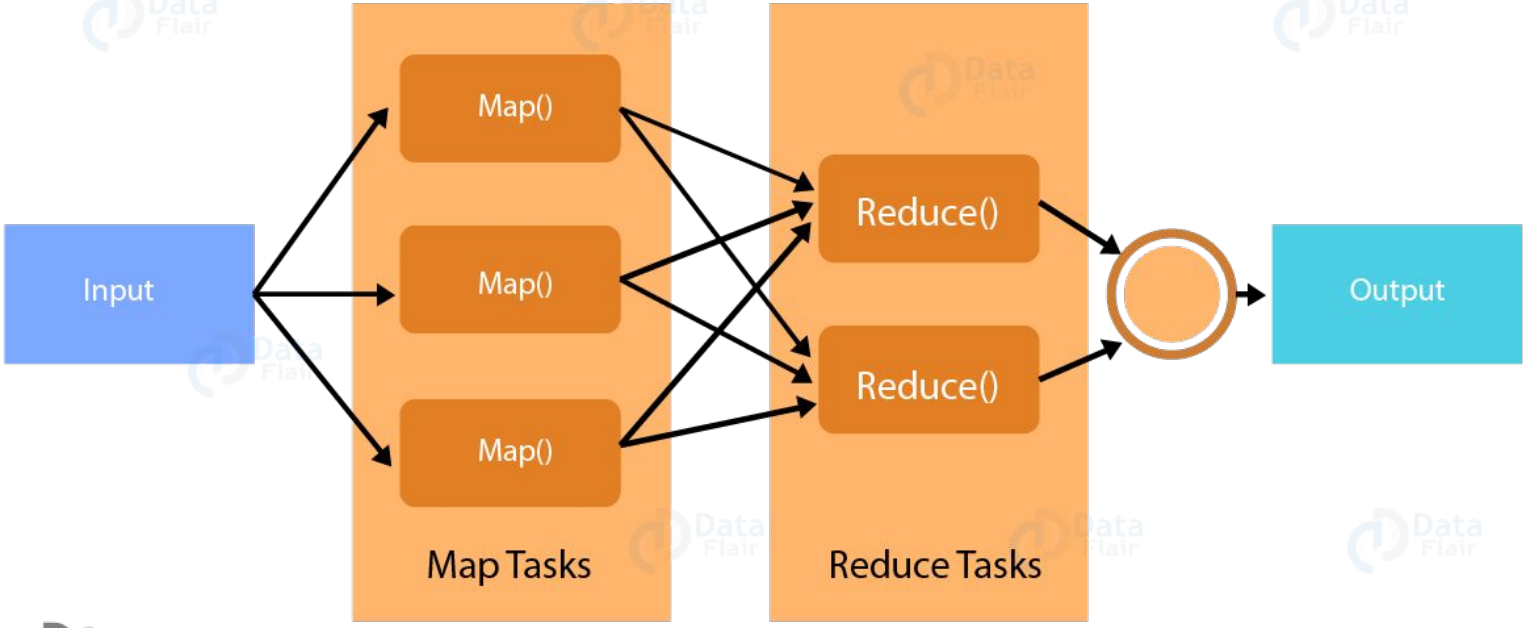
(Data Collection)

Hadoop Ecosystem

- **Hadoop Distributed File System (HDFS)-Storage Layer**
  - HDFS is the foundation of Hadoop.
  - It is Java software
  - Provides many features like scalability, high availability, fault tolerance, cost effectiveness etc.
  - It also provides robust distributed data storage.
  - Many other software frameworks are deployed over HDFS.

Hadoop Ecosystem

- **MapReduce-Processing or Computation Layer**
  - **Data processing component.**
  - **Apply computations on data.**
  - **Parallel programming model.**
  - **Works in two phases**
    - **Map phase**
      - takes input as key-value pairs and produces output as key-value pairs.
    - **Reduce phase**
      - applies the summary type of calculations to the key-value pairs.


Input
Map()
Map()
Map()
Map Tasks
Reduce()
Reduce()
Reduce Tasks
Output

Hadoop Ecosystem

- **Hadoop Yarn**
  - Yet Another Resource Manager
  - the operating system of Hadoop.
  - Manages and monitors resources.
  - Framework for job scheduling.
  - Has two components
    - Node Manager
      - Ttakes care of the individual compute nodes in a Hadoop cluster.
    - Resource Manager
      - Tracks the resources in the cluster and schedule tasks like map-reduce jobs.