Big Data and Hadoop

Manoj Valeti
B160091CS

Department of Computer Science and Engineering
NIT Calicut

16 September 2019

Outline

1. Big Data
    - Introduction
    - Characteristics of Big Data
    - Big Data Use Cases
    - Problems with Big Data Processing
2. Hadoop
    - What is Hadoop?
    - Hadoop Architecture
    - Who uses Hadoop?
    - Pros and Cons of Hadoop

Outline

1. Big Data
   - Introduction
   - Characteristics of Big Data
   - Big Data Use Cases
   - Problems with Big Data Processing
2. Hadoop
   - What is Hadoop?
   - Hadoop Architecture
   - Who uses Hadoop?
   - Pros and Cons of Hadoop

Big Data

Big Data

Big Data refers to the data set that are not possible to process using traditional methods of processing.

Outline

1. Big Data
   - Introduction
   - Characteristics of Big Data
   - Big Data Use Cases
   - Problems with Big Data Processing

2. Hadoop
   - What is Hadoop?
   - Hadoop Architecture
   - Who uses Hadoop?
   - Pros and Cons of Hadoop

The 5V's of Big Data

- Volume - .The size of data

The 5V's of Big Data

- Volume - .The size of data
- Velocity - The speed at which data is generated.

The 5V's of Big Data

- Volume - .The size of data
- Velocity - The speed at which data is generated.
- Variety - The different types of data that are generated.

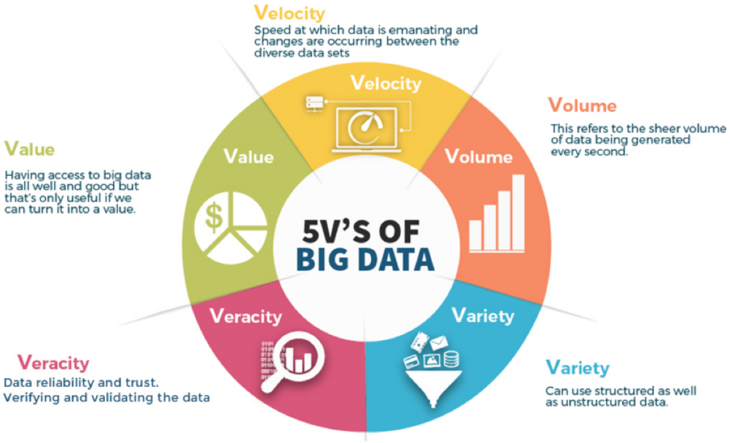
The 5V's of Big Data

- Volume - .The size of data
- Velocity - The speed at which data is generated.
- Variety - The different types of data that are generated.
- Value - The worthiness of data being abstracted.

The 5V's of Big Data

- Volume - .The size of data
- Velocity - The speed at which data is generated.
- Variety - The different types of data that are generated.
- Value - The worthiness of data being abstracted.
- Veracity - The quality or trustworthiness of data.

The 5v's of Big Data



Velocity
Speed at which data is emanating and changes are occurring between the diverse data sets

Volume
This refers to the sheer volume of data being generated every second.

Value
Having access to big data is all well and good but that's only useful if we can turn it into a value.

Veracity
Data reliability and trust.
Verifying and validating the data

Variety
Can use structured as well as unstructured data.

Outline

1. Big Data
   - Introduction
   - Characteristics of Big Data
   - Big Data Use Cases
   - Problems with Big Data Processing
2. Hadoop
   - What is Hadoop?
   - Hadoop Architecture
   - Who uses Hadoop?
   - Pros and Cons of Hadoop

Use Cases

- **Security Intelligence** - Helps many organisations to identify the internal and external information to help them prevent, detect and mitigate the attacks.
- **Recommendation engines** - Helps to provide recommendations based on the historical data.
- **Banking** - Helps to understand customers, manage sources, minimise risk and fraud.
- **Price Optimization** - Helps companies to know which price points have yielded the best overall results under the various historic market conditions.
- **Internet of Things** - The data collected by the sensors can be analyzed to achieve actionable insights.
- **Government** - Helps in managing utilities, running agencies, traffic congestion, crime prevention.

Outline

1. Big Data
   - Introduction
   - Characteristics of Big Data
   - Big Data Use Cases
   - Problems with Big Data Processing

2. Hadoop
   - What is Hadoop?
   - Hadoop Architecture
   - Who uses Hadoop?
   - Pros and Cons of Hadoop

Problems with Big Data Processing

- Heterogeneity and Incompleteness
- Scale
- Privacy
- Human Collaboration

Outline

1. Big Data
   - Introduction
   - Characteristics of Big Data
   - Big Data Use Cases
   - Problems with Big Data Processing
2. Hadoop
   - What is Hadoop?
   - Hadoop Architecture
   - Who uses Hadoop?
   - Pros and Cons of Hadoop

Hadoop

Hadoop

Hadoop is an open source distributed processing framework that manages data processing and storage for big data applications running in clustered systems

Outline

1. Big Data
   - Introduction
   - Characteristics of Big Data
   - Big Data Use Cases
   - Problems with Big Data Processing
2. Hadoop
   - What is Hadoop?
   - Hadoop Architecture
   - Who uses Hadoop?
   - Pros and Cons of Hadoop

Hadoop Architecture

The 4 most important components that play an important role in Hadoop architecture are

- Hadoop Common
- Hadoop Distributed File System (HDFS)
- Hadoop MapReduce
- Yet Another Resource Negotiator (YARN)

Hadoop Common

- It is the core/base of Hadoop framework.
- It contains all the necessary JAR files/scripts to start Hadoop
- Provides source code and documentation
- Provides essential services and basic processes such as abstraction of the underlying operating system and it's file system

HDFS

- It is a distribute file system that runs on commodity hardware.
- It follows Master Slave architecture.
- Scales up incrementally to store data
- Survive the failure of significant parts of the storage infrastructure without losing data
- Data Node and Name Node are two crusial components of HDFS.

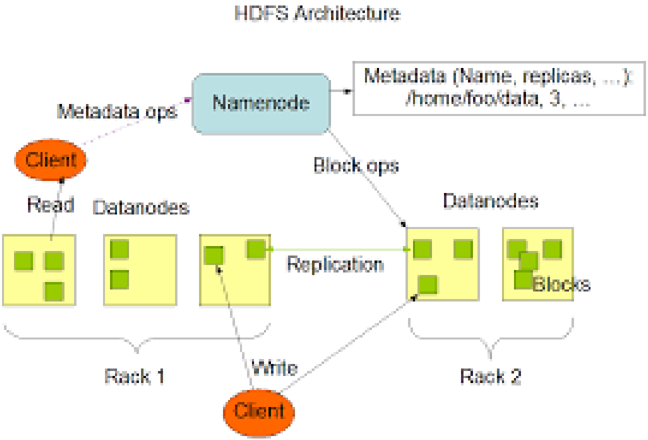
Name Node

- It stores meta-data.
- It is treated as master node.
- NameNode knows the list of the blocks and its location for any given file in HDFS. With this information NameNode knows how to construct the file from blocks.
- NameNode is so critical to HDFS and when the NameNode is down, HDFS/Hadoop cluster is inaccessible and considered down.

Data Node

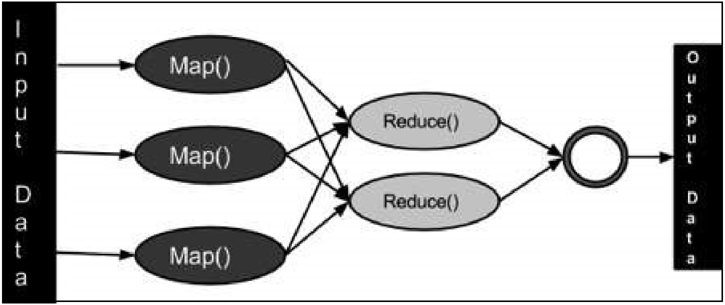
- It stores the actual data in HDFS.
- It is treated as a slave node.
- It performs operations like read, write, block creation, deletion and replication
- Name Node and Data Node are in constant communication.
- When a DataNode is down, it does not affect the availability of data or the cluster. NameNode will arrange for replication for the blocks managed by the DataNode that is not available.

HDFS Architecture



Hadoop MapReduce

- Splits the data into chunks and process them parallelly
- It has 3 stages - Map ,Shuffle and Sort, Reduce
- Map : Generates key-value pairs for the data present locally.
- Reduce : It reduces the output of mapper function and sends them to Output file.



Hadoop YARN

- YARN stands for Yet Another Resource Negotiator.
- Resource management is done by YARN
- It consists of Resource Manager, Node Manager and Application Master
- Resource Manager arbitrates resources among all competing applications.
- Node Manager monitors the resource manager and reports to Resource Manager.
- The Application-Master negotiates resources with Resource-Manager and works with Node-Manager to execute and monitor the job

Outline

1. Big Data
    - Introduction
    - Characteristics of Big Data
    - Big Data Use Cases
    - Problems with Big Data Processing

2. Hadoop
    - What is Hadoop?
    - Hadoop Architecture
    - Who uses Hadoop?
    - Pros and Cons of Hadoop

Who uses Hadoop?

- Financial companies
- Retailers to get better understanding of customers
- Large scale Pre-processing of raw data
- Data agility

Outline

1. Big Data
   - Introduction
   - Characteristics of Big Data
   - Big Data Use Cases
   - Problems with Big Data Processing
2. Hadoop
   - What is Hadoop?
   - Hadoop Architecture
   - Who uses Hadoop?
   - Pros and Cons of Hadoop

Pros of Hadoop

- It is open source
- Automatic data replacement and rebalancing when data grows
- Saves lot of time to extract valuable data from given raw data sets.
- Supports replication and machine failover.

Cons of Hadoop

- Not suitable for large number of small files
- easily vulnerable
- Not efficient for stream processing
- It has high processing overhead for large datasets

Summary

- In modern world data is increasing exponentially. In order to process such a huge data Big Data is becoming more crucial.
- Hadoop is one of the open source framework which helps to process and analyze Big Data easily.

For Further Reading I

Harshawardhan S. Bhosale , Prof. Devendra P. Gadekar
A REVIEW PAPER ON BIG DATA AND HADOOP.

Deepak Motwani, V.K. Chaubey, A. S. Saxena
Hadoop based Information Extract from Text Document

https://www.javatpoint.com/hadoop-tutorial

https://en.wikipedia.org/wiki/Big$_d ataDefinition$

Thank You