Introduction to Hadoop

Amol Nater (CSE)
Chandigarh University, Mohali
Punjab, India
anatersh@gmail.com

Ms. Manpreet Kaur
Assistant Professor,
Department of CSE,
Chandigarh University
manpreet.cse@cumail.in

***Abstract*— In Computer Science any raw facts and figures is data. Data is the basic component of Computer Science. When data is in huge volume that a single processor can't process is known as Big Data.**

Keywords—Data, Big Data, Hadoop, Hdfs, Map Reduce.

I. INTRODUCTION: *BIG DATA*

Any raw facts and figures is data. Data in processed form is information. Information is used to perform several activities/operations. Without information nothing can be achieved. When data is in huge volume that is difficult to handle(storage, processing) by a single processor is **BIG DATA.** Now big data is used almost everywhere because life is depending more on technology. Example: Big data is used in YouTube, in Social media websites like facebook, twitter etc.

II. *BIG DATA : CHARACTERISTICS*

There are 5 V's in Big Data:

1. Volume
2. Velocity
3. Variety
4. Veracity
5. Value

- Volume: Amount of data.
- Velocity: Rate of collecting data.
- Variety: Type of data(videos, audios, images, text etc.).
- Veracity: Validness of data.
- Value: Conclusion that we get after analysis of data.

These are the characteristics of big data.

III. CLASSIFICATION OF BIG DATA

Big Data is classified into 3 categories:

1. Structured Big Data
2. Unstructured Big Data
3. Semi Structured Big Data

- Structured Big Data: Structured data is the data which can be stored in a table i.e, with rows and columns. Example: RDBMS. It is highly organized and is in sequential manner and is easy to process.

- Unstructured Big Data: Unstructured data is data which has it's own internal structure. Example: images, audios, videos etc.

- Semi Structured Big Data: Semi Structured data is the data which doesn't reside in relational database but has some structure/ organization which makes it easier to analyse. Example: XML, NoSQL etc.

IV. BIG DATA: ADVANTAGES

- Data Procurement
- Data Quality and Integration
- Data Governance
- Customer Feedback
- Data Segmentation
- Risk Identification
- Data Modeling
- Operational efficiency
- Business Intelligence

V. BIG DATA: CHALLENGES

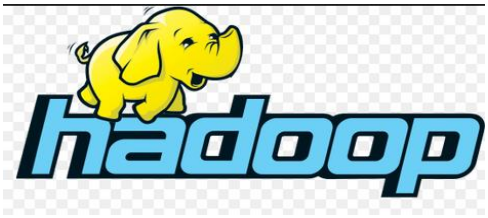
Big data generally faces 2 major challenges:

1. Storage

2. Processing

Big data that is data in huge volume faced challenges to store that data. If stored then problem arises how to process this data? To solve this problem Hadoop framework is introduced.

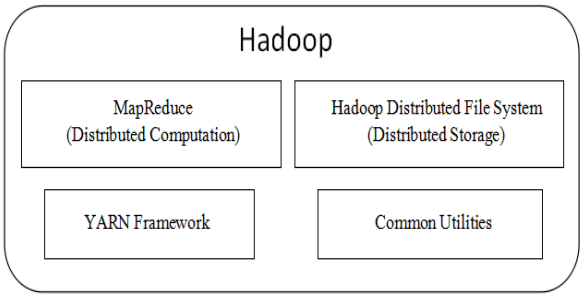
VI. Introduction to Hadoop



Hadoop

- Open source framework (Apache foundations).
- It is a Java Based Platfom.
- It is used to perform distributed computing and parallel processing.
- It follows Master- Slave Architecture.
- It is highly scalable and fault tolerant.

VII. Hadoop Components

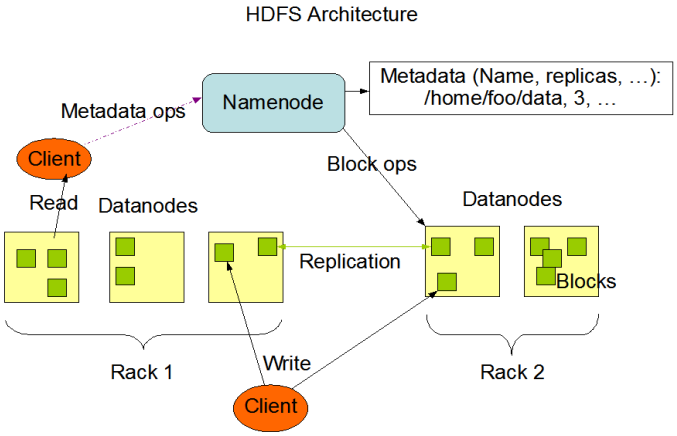


Hadoop has basically four components:

1. HDFS common
2. HDFS(Hadoop Distributed File System)
3. MapReduce
4. YARN(Yet Another Resource Negotiator)

- **HDFS Common:** It includes all the library files included in Hadoop.
- **HDFS:** It is the storage component of the Hadoop.
- **MapReduce:** It is an algorithm used for processing and analysis of Big Data.
- **YARN:** It is similar to MapReduce but has a new feature Resource Allocation. It is introduced in V.Hadoop 2.x .

VIII. HDFS: Hadoop Distributed File System



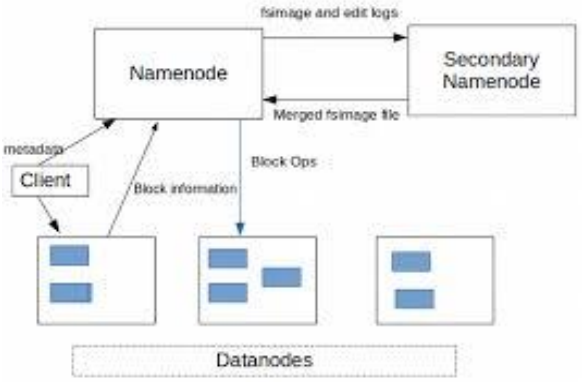
It is primary storage component of hadoop in which data is stored. It is a distributed file system that provides high-performance.

Further HDFS has 3 components:

- Name Node
- Data Node
- Secondary Name Node

- **Name Node** is a node which stores the metadata of Data Node.

  Metadata is data about the data.

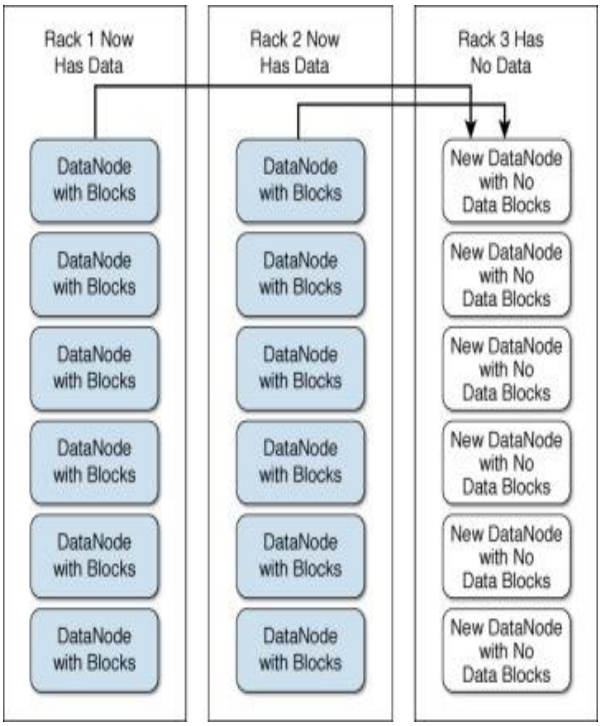


- **Data Node** is the node which stores the data. There are several data nodes in a cluster but there is only on Name Node connected to several data nodes.

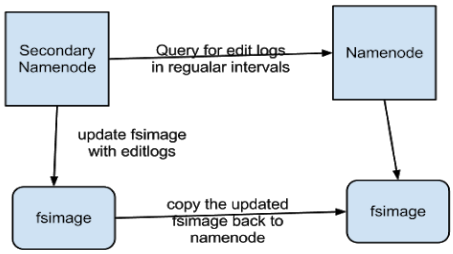
  Data Nodes are enclosed within racks.

  Data Node keeps on sending **heart beat signals** to the Name Node that it is working well without any faults and also tells whether the data node is idle or working. These are acknowledgement signals.

  Name Node waits for 10 minutes in case it is not receiving heartbeat signals. If it fails to get heart beats for 10 minutes it marks the Data node not working or there is a fault in the rack.



- **Secondary Name Node** is the node which stores the copy of FsImage and edit log. It periodically keeps on tracking these files and keep storing.



IX. Result And Discussion

Big data is used almost everywhere because life is depending more on technology. More use of technology means more data to store and process. There comes the need of Big Data. Hadoop Framework(by Apache Foundations) provided a solution to challenges faced in big data storage and processing.

X. Conclusion

Big Data provided several advantages like Data Procurement, Data Quality and Integration, Data Governance, Customer Feedback, Data Segmentation, Risk Identification, Data Modeling , Operational efficiency and Business Intelligence.

XI. References

[1] Sumit Kumar, Nishant Sharma, Gagan Sharma "**Li-Fi Technology in Wireless Communication**" Published in International Journal of Trend in Research and Development (IJTRD), ISSN: 2394-9333, Volume-4 | Issue-3 , June 2017, URL: http://www.ijtrd.com/papers/IJTRD8584.pdf

[2] Kumari Neha et al. "Using Reconfigurable Directional Antenna in MANET." Procedia Computer Science 125 (2018): 194-200.

[3] Kumai, Neha, et al. "Mobile ad hoc networks and energy efficiency using directional antennas: A Review." *Intelligent Computing and Control Systems (ICICCS), 2017 International Conference on*. IEEE, 2017.

[4] Singh, VK, Kumar, R. "Multichannel MAC Scheme to DeliverReal-Time Safety Packets in Dense VANET". Procedia computerscience ISSN 1877-0509, 2018.

[5] S. Kapil et al, et al. "Analysing the Role of Risk Mitigation andMonitoring in Software Development (2018).