Hadoop_Udemy_Notes

February 25, 2019

1 Hadoop Notes

1) How to say if the problem is BigData problem? - We need to consider 2 parameters - Volume - Velocity - **Volume:** We need to understand how big is our data at present and in near future - __ Velocity:__ How speed our data is growing.

- Based on these 2 parameters we can decide if the problem is BigData problem or not i.e for example if we got data of 1TB in 1month and expecting same growth in near future then for one year we will have 12TB of data, if growth rate is exponential then we will have very huge data, which we cant handle with our HDD or SSD, then it is considered as Big Data Problem.

2) Why Hadoop came into picture? - Lets, consider a Scenario where we need to get Max closing price of all Stocks which are Listed in NSE, lets suppose we have **1TB** of data i.e history data of all stocks, if we are handling with normal Laptop or Desktop and if Data is present on the network, we need to get that to local drive it takes hours of time lets say 3+ hours approx and computation time may be around 1+ hours that totals to 4+ hours, its very huge in Business point of view, if we want to get this done in minutes, we need to Consider Hadoop, because Hadoop converts the data into batches/chunks and distributes this fixed size blocks to different nodes/instances and replicates these blocks in multiple nodes for fail safe and maintains info on which node has which blocks, then it parlallely computes Max Stock price of stocks in the instances/nodes present in the network, which gives the final result in very less time.

3) Why we need different file system HDFS(Hadoop Distributed File System)? - The traditional File Systems like **NTFS**(used in windows, which can handle **16Exa Bytes(1 EB= 1000 Peta Bytes)** of data), **EXT4**(used in Linux, which can handle **1 Exa Byte** of data), but in these systems we **doesn't have Image of global distribution of Blocks in the Nodes/instances** i.e we will not have info on which node has which block info and which block is replicated in which other node, so **this is handled by Hadoop Distributed File System(HDFS)**

- HDFS takes care of distributing Data to multiple nodes in fixed size **Block** of **128MB** each
- HDFS takes care of **replicating** the data in **Multiple Nodes(by default 3 nodes)** for **Fail Safe**
- HDFS keeps track of info on which node has which block

Benefits of HDFS - Supports Distributed Processing - Handle Failures(by replicating data in multiple nodes) - Scalability i.e we can add nodes depending on requirements - Cost Effective i.e we dont need any super special computers

What is MapReduce? - Distributed programming model for processing large data sets - Created by Google - Can be implemented in any programming language - MapReduce is not a programming language - Hadoop implements MapReduce - It manage communication, data transfers, parallel execution across distributed servers.