Big Data : An Overview

Mansi [1], Tanishka Garg [1], Dr. Deepak Chahal[2]

[1]MCA Student, [2]Professor,
[1,2] Department of IT, Jagan Institute of Management Studies, Sector-05, Rohini, New Delhi, India

Abstract

This paper aims to give complete information about Big Data in easy to comprehensible language. It begins with giving the brief introduction about what Big Data is and why is it needed in today's era of technological advancement. Popular tools that are commonly used to work in the world of big data are also explained briefly. Main focus remains on the major players in the market now a days- Apache Hadoop and Apache Spark. Detailed information is given so that reader gets a clear picture about the case where which tool can be used on the basis of reliability as well as performance in real time scenario.

Keywords—*Big Data, RDD, Hadoop, Spark, Storm, MongoDB, Spark Vs Hadoop*

I. INTRODUCTION

Big data is a huge collection of data and yet growing exponentially by time. Basically, the data is so large and complex that it cannot be handled by the traditional approaches to process them efficiently.

Exciting part is what organizations do with such type of data, they use them to analyze and strategize the future business moves.

Some of the examples on big data are below-

A. Everyday *Facebook* database get ingested by *500+ terabytes* of new data by the statistics. The data is majorly in terms of messages exchange, photos, videos uploads etc.
B. In *30 minutes* of a flight a single jet can create *10+ terabytes* of data. Now, imagine thousand flights per day creates Petabytes data.

II. V'S OF BIG DATA

A. *Volume* – Volume is one of the characteristics of big data which helps to recognize that is that data has to be considered as Big Data, because it depends on volume of the data. The data can be unknown any data like Twitter data feeds, clicks on a webpage.
B. *Variety* – Variety means heterogeneous nature of data both structured and unstructured. Earlier spreadsheets and databases were only the sources of data, But now data is in variety of forms such as emails, videos, audios, photos etc. which require additional processing to derive the meaning of the data.
C. *Velocity* – Velocity refers to the speed at which data is received and acted on. The potential of data is determined by how fast the data is processed to meet the demands.
D. *Variability* – The inconsistency shown by data is known as Variability. This can be found by the detection techniques for meaningful analytics.

III. WHY BIG DATA? NEED OF BIG DATA?

Businesses use Big Data to win the competitive market by improving their operations, providing better customer service, creating personalized marketing tactics and ultimately increasing profitability. As Big data can provide immense insights of their customers that can be beneficial for marketing campaigns to increase customer engagement.

Moreover, by the big data companies can access the preferences of customers and work on the updates and strategize marketing campaigns accordingly. Not only business sector is using Big Data but in medical, researchers are using to identify the disease and diagnose the illness. This provides data from electronic health records, healthcare organizations and government agencies with up-to-date information on all types of threats and infections.

In the energy industry, big data helps oil and gas companies identify potential drilling locations and monitor pipeline operations; likewise, utilities use it to track electrical grids. Financial services firms use big data systems for risk management and real-time analysis of market data. As well as Manufacturers and transportation companies rely on big data to manage their supply chains and optimize delivery routes. Other government agencies uses as emergency response, crime prevention and smart city initiatives. Today, networks are complex arrangements that cannot be left to chance. Each design must be researched, assembled, proven, and then implemented [1].

IV. POPULAR TOOLS USED BY BIG DATA

A. ***Hadoop***: Apache Hadoop is the most popular tool in the industry because of its capability of large-scale processing data.

B. ***Apache Spark***: Apache Spark is also one of the popular Big Data tool. It fills the gap of Apache Hadoop concerning data processing. Apache spark can handle batch data as well as real-time data. It is flexible in working with HDFS as well as other data stores.

C. ***Apache Storm***: It is a distributed real-time framework for reliably processing the unbounded data stream and this framework supports multiple programming languages. Unique features of Apache Storm:

   i. Fault-tolerance
   ii. “fail fast, auto restart” approach
   iii. Written in Clojure
   iv. Runs on JVM
   v. Huge scalability
   vi. Supports JSON protocol

D. ***MongoDB***: MongoDB is fast and real-time based appropriate for business needs. It have many built-in features. It runs on MEAN stack. Some features of MongoDB are:

   i. It stores any type of data- integer, string, array, object, boolean, date, etc.
   ii. It provides flexibility in cloud-based infrastructure.

E. ***R Programming Tool***: The most interesting part of this Big Data tool is user don’t have to be expert in statistics. R has its own library CRAN (Comprehensive R Archive Network) which has inbuilt modules and algorithms for statistical analysis of data.

V. APACHE HADOOP

Apache Hadoop is an open-source framework designed for distributed storage and processing of huge data sets across clusters of computers. It is basically designed to scale up from single servers to thousands of machines. It consists of components given below:

A. Hadoop Distributed File System (HDFS), the bottom layer component for storage. HDFS breaks up files into chunks and distributes them across the nodes of the cluster.

B. Yarn for job scheduling and cluster resource management.

C. MapReduce for parallel processing.

D. Common libraries needed by the other Hadoop subsystems.

VI. FEATURES OF HADOOP

A. ***Hadoop is Open Source:*** Hadoop is open source hence its code is freely available for modification and analysis. Many companies are using this feature to modify the code according to their company’s requirement.

B. ***Hadoop cluster is Highly Scalable:*** Any number of nodes can be added, also the hardware capacity of nodes can be increased in order to achieve high computational power hence making the Hadoop cluster highly scalable.

C. ***Fault Tolerance:***One of the important feature required in todays world is being able to handle fault and overcome the aftereffects of that particular fault. Apache Hadoop provides this feature of fault tolerance using a replication mechanism. Replica of each block is maintained and stored in multiple devices so if any of the machine leads to fault , still data can be accessed from the backup carrying or replica carrying machine without creating any hinderance in the processing of data.

D. ***Hadoop is very Cost-Effective***: Hadoop is an open source product hence there is fuss of any licensing issue. The Hadoop cluster node hardware is inexpensive hence making it a cost-effective solution for storing as well as operating and processing big data.

E. ***Hadoop is Faster in Data Processing:*** Hadoop provides fast data processing since it stores data in a distributed fashion which in turn allows data to be processed on a cluster of nodes in a distributed fashion only.

F. ***Hadoop is based on Data Locality concept:*** Hadoop reduces the bandwidth utilization in a system and that is possible to presence of data locality feature in Hadoop framework. This data locality means that the computation logic is going to be moved to data rather than moving data to computational logic.

G. ***Hadoop provides Feasibility:*** Unlike traditional systems there is no restriction on the data format. which can be processed by Apache Hadoop.

H. ***Hadoop is Easy to use:*** Framework is handling the processing hence making Hadoop easy to use for the client since they won't have to look into the issues related to distributed computing.

I. ***Hadoop ensures Data Reliability:*** Since Data replication mechanism is used in the cluster, this makes Hadoop highly reliable because even if one machine fails another one carrying replica will be ready to save the cause.

VII. APACHE SPARK

Apache Spark is open source frame that is used to process huge volume of data quickly. Until now Apache Hadoop was majorly used to handle and operate on big data, but now-a-days companies in-memory database and computation is in high demand. Apache Spark is a solution to these demands as it provides in-memory capabilities giving 100 times faster output as compared to Hadoop.

It can be used with a Hadoop environment as well as a standalone platform also in cloud. Cloud computing has become a popular model for reducing cost of business, improvise quality of services, and provide good & secure computing [2].

Developed by University of California, Spark, later was given to Apache Software Foundation. Spark has become a prominent bigdata distributed processing framework all over the world. Spark provides native bindings with multiple programming languages like Java, Scala, Python and R. It supports SQL, machine learning and graph processing because of which is has become a great competition to Hadoop.

VIII.FEATURES OF SPARK.

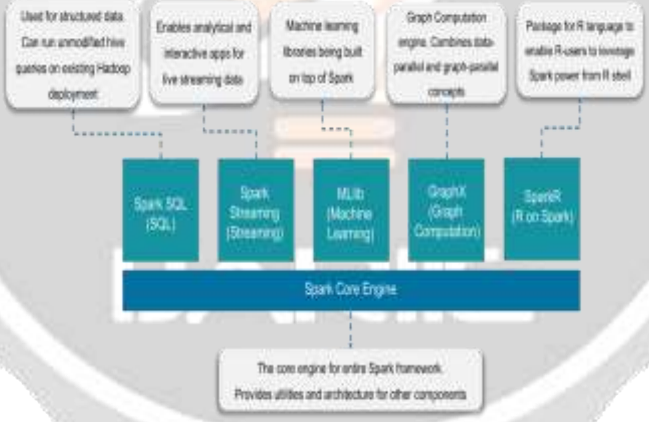

Used for structured data. Can run unmodified hive queries on existing Hadoop deployment
Enables analytical and interactive apps for live streaming data
Machine learning libraries being built on top of Spark
Graph Computation engine. Combines data-parallel and graph-parallel concepts
Package for R language to enable R-users to leverage Spark power from R shell
Spark SQL (SQL)
Spark Streaming (Streaming)
MLlib (Machine Learning)
GraphX (Graph Computation)
SparkR (R on Spark)
Spark Core Engine
The core engine for entire Spark framework. Provides utilities and architecture for other components

Figure 1: Features of Spark

A. ***In-Memory Technology:*** Spark offers In-Memory processing due to which huge amount of data can be processed in very less time. It loads the data into computer memory and this processed data can be unloaded into disk space later. This allows it to be extremely fast.

B. ***Spark's Core:*** Core manages the important function like tasks and the interaction along with managing the I/O operations. Spark's core manages several important functions like putting and setting tasks and interactions as well as producing input/output operations. It can be said to be an RDD, or resilient distributed dataset.

C. ***Spark's SQL:***Spark uses SchemaRDD, this allows data to be arranged in many levels and allows data query using a particular language.

D. ***Graphx Service:***Apache Spark can process graphs or information that is graphical in nature. This allows easy data analysis.

E. ***Streaming:***This feature of Spark allows data chunks of data to be streamed. It is done by breaking down the larger data into smaller size packets, then the transformation is done on these small packets in turn increasing the speed of creation of RDD.

F. ***MLib (Machine Learning Library):*** MLib is a framework created for structured machine learning. It is faster in implementation as compared to Hadoop.

IX. KEY FACTORS OF APACHE SPARK

A. ***Swift Processing:*** Spark reduces the number of read-write operations on disk due to which the high data processing speed is achieved which is much more faster than Hadoop.

B. ***Dynamic in nature***: Spark provides 80-level operators and due to this parallel applications can be developed.

C. ***Reusability :*** Spark code can be reused for batch processing, to run ad-hoc queries on stream state , and many more tasks could be performed using this features.

D. ***Fault Tolerance in Spark:*** Fault tolerance is done by using Spark abstraction-RDD, which are designed to handle any node failure present in the cluster.

E. ***Real-Time Stream Processing :*** In-memory computation makes Spark very fast hence allowing it to handle streams of data coming at a very fast rate.

F. ***Support Multiple Languages:***In Spark, there is Support for multiple languages like Java, R, Scala, Python.

G. ***Support for Sophisticated Analysis:*** Spark contains dedicated tools which makes it a better performing platform. Tools are available for streaming data, for querying data, machine learning ,etc..

H. ***Cost Efficient:*** Apache Spark unlike Apache Hadoop doesn’t need large storage and large data center and this makes it cost efficient.

X. SPARK VS HADOOP

Apache Hadoop and Apache Spark are both open-source frameworks for big data processing. Spark has the advantage of speedily processing the data. Spark uses in-memory processing that can perform tasks hundred time faster than the MapReduce paradigm used by Hadoop.

Spark API is developer friendly which makes it preferable. Hadoop has a distributed file system (HDFS), whereas Spark does not provide a distributed file storage system, so it is mainly used for computation, on top of Hadoop. Spark does not need Hadoop to run, but can be used with Hadoop since it can create distributed datasets from files stored in the HDFS.

XI. CAN SPARK REPLACE HADOOP?

Technological innovations such as robotics, machine learning, cloud technology etc have established themselves very fast over the last few years and have now become a key element of the commercial and social economy [3].

It has industry recognized practices, and Apache Spark is comparatively fresher hence it’ll need time to get as recognized as Hadoop.

MapReduce follows certain industry standards which are used when operating on large amount of data and performing complete operations. Although Spark’s fast speed, it is considered less reliable since its not so long presence in the market.

MapReduce is easier to configure, whereas Apache Spark is offering a new platform because of that hasn’t tested any rough patches.

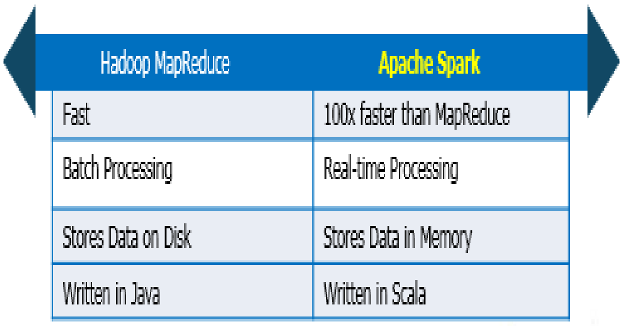
XII. WHEN TO USE HADOOP

A. ***Analysing Archive Data:***Apache Hadoop YARN is a resource management and job scheduling technology present in Hadoop framework. It allows parallel processing of huge amounts of data. To achieve this parallel processing of data, smaller parts are made and separately maintained on different DataNodes. The result is gathered from each NodeManager.

B. ***When instant results is not the goal:*** Hadoop MapReduce is a good and economical solution for batch processing.

XIII. WHEN TO USE SPARK

A. *Real-Time Big Data Analysis:* **Apache Spark offers very fast data processing speed because of which it seems perfect to handle real time data stream coming at the rate of millions per second. It supports distributed processing while it supports data streaming.** Real-time data can still be processed on MapReduce but its speed is nowhere close to that of Spark. Spark claims to process data 100x faster than MapReduce, while 10x faster with the disks.



Hadoop MapReduce	Apache Spark
Fast	100x faster than MapReduce
Batch Processing	Real-time Processing
Stores Data on Disk	Stores Data in Memory
Written in Java	Written in Scala

Figure 2: A performance comparison of Spark and Hadoop

B. ***Graph Processing:*** Spark has a brilliant graph computation library namely- GraphX. In memory computation provided by Apache Spark along with graph support improves the performance of algorithm that is being used by a measure of one or two degrees when compared to MapReduce programs.

C. Iterative Machine Learning Algorithms:

Spark uses Mesos which is a distributed system kernel to store the intermediate dataset generated in each iteration. Spark runs multiple iterations on this stored data hence reducing the I/O and eventually helping the algorithm to run fast and efficiently along with great fault tolerance. Spark uses MLib library to perform machine learning tasks . It contains top class

REFERENCES

1. Bhatnagar A. et al. Juniper Networks: An Overview of the Concept, IJSRD - International Journal for Scientific Research & Development| Vol. 6, Issue 10, 2018

2. Kharb L. et al. A Comprehensive Study of Security in Cloud Computing, International Journal of Engineering & Technology, 7 (4) (2018) 38973901.

3. Chahal D. et al. The Developing Role Of Block Chain(R)Evolution, EPRA International Journal of Multidisciplinary Research (IJMR), Volume: 4 | Issue: 11November2018