A Survey On Big Data Analytics

JYOTI HAWELIYA[1], SOURABH CHOUDHARY[2], MANISH HATE[3]
[1]Assistant Professor, [2]MBA Student, [3]Trainee
[1]IET, DAVV, Indore (M.P.), India
[2]SGSITS, Indore (M.P), India
[3]Bulls Eyes, Nagpur (M. H), India

Abstract: Now a day the term big data has become universal. In various arrears like an industry, academia, media etc. there is no single unified definition of Big Data, and various stakeholders provide different and often contradictory definition of Big Data. It is the term for any collection of data sets so large and very complex that it becomes hard to process using traditional data processing applications. There are so many challenges in front of this era for big data which includes analysis, curation, capture, sharing, search, transfer, storage, visualization, and the most important privacy violations. In this paper, we review various hardware platforms, technologies and their suitability for the Big Data environment. Basically, big data deals with huge parallel software running on not only tens, hundreds even thousands of servers that's why the traditional relational database management system is not appropriate for it. Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time.

Index Terms: **Big data, Hadoop, IBM Infosphere, and Apache.**

I. Introduction

In today's scenario, the big data is a very ambiguous term. Basically, big data is some set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and on a massive scale. In an example of big data, size is being considered as petabytes (1,024 terabytes) or exabytes (1,024 petabytes) which consists of billions to trillions of records of millions of people—all from different sources (like Sales, Web, Customer contact center, Social media, Mobile data and so on). The data is typically loosely structured data that is often incomplete and inaccessible [1]. We can acquire, organize and analyze a variety of data by using a big data environment. This huge amount of the data is known as "Big data" [2]. Generally, a big data having seven V's [3], volume, velocity, variety, variability, veracity, visualization, and value.

- Volume: As the name implies it is big in terms of size and this size is measured as volume. There are lots of factors which contribute to an increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected [4]. Due to the volume of data storage issues comes in-front of us. So, we decrease the storage costs, but other issues further emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.
- Velocity: It is related to two concepts first is the speed at which new data is being created and other is the corresponding need for that data to be digested and analyzed in real near future. RFID tags, sensors, and smart metering are driving the need to deal with torrents of data in near-real time. Responding rapidly enough to deal with data velocity is a very big challenge for the most organizations.
- Variety: Variety means various formats or forms of data. It may be structured data, unstructured text documents, email, video, audio, stock ticker, numeric data etc.
- Variability: Variability means data can only be meaningfully interpreted. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks.
- Veracity: Veracity means accuracy or truthfulness. Here it means that there is a doubt of the presence of uncertain or imprecise data. Due to data inconsistency and incompleteness, ambiguities, latency, deception, [and/or] model approximations the veracity of big data is doubtful. There's widespread agreement about the potential value of Big Data, the data is virtually worthless if it's not accurate.
- Visualization: Big Data visualization involves the presentation of data of almost any type in a graphical format that makes it easy to understand and interpret as big data is a huge amount of data of many users of technology that a common user cannot understand or analyze. Visualization-based data discovery methods allow business users to mash up disparate data sources to create custom analytical views.
- Value: This is the last V but not least, big data must have value. That is, if we are going to invest in the infrastructure required to collect and interpret data on a system-wide scale, it's important to ensure that the insights that are generated are base on accurate data and lead to measurable improvements at the end of the day.

In this paper, we review the background and comparative study on widely used tools of big data analytics.

II. Background

[4] Big data is a buzzword utilized to describe a massive volume of both structured and unstructured data that is so huge that it's complicated to process using traditional database and software techniques. Apache Hadoop is open source software that enables reliable, scalable and distributed computing on clusters of low-cost servers. It uses Map-framework that was introduced by Google by leveraging the concept of the map and reduces function, so as to well known used in functional programming.
[5] Big Data is as a collection of the large dataset that cannot be processed using traditional computing techniques. There can be three types of data- Structured, Semi-structured and Unstructured data. Volume, variety, velocity, value, and veracity are the five Vs that are the limitations of big data. Apache Hadoop is software designed to scale up from single servers to thousands of machines, each offering local computation and storage Hadoop consists of two component Hadoop Distributed File System (HDFS) and MapReduce Framework.
[6] Hadoop distributed file system (HDFS) is a file system designed to store large amounts of data across multiple nodes of commodity hardware. The HDFS is not a database, but a file storage system designed for a specific purpose, and it doesn't include all of the functionality that some of the other data storage solutions have. HDFS is well known for its scalability, fault-tolerance and is a good option for historical data that does not need to be edited or accessed frequently, but there are several limitations that may impact Hadoop users, in particular, one for whom rapid random reads or writes are a priority. Hadoop itself has a wide range of tools to study and analyze the big data in a very detailed and organized manner.
[7] The IBM Streams Processing Language (SPL) is the programming language for IBM InfoSphere Streams, a platform for analyzing Big Data in motion. By providing flexible support for primitive and composite operators, SPL allows users to write a broad range of expressive analytics. By supporting efficient code with code generation and optimizations, SPL is suitable for handling big data at massive scales. SPL is the primary means for programming InfoSphere streams applications.
[8] Cloudera Impala is an Apache-licensed, real-time query engine for data stored in HDFS. Impala is well suited to use cases where real-time queries and speed is essential. Cloudera Impala is Cloudera's open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop. Using Impala, Data processing workload acceleration, with data pipelines will last seconds instead of minutes or hours, to meet tighter service-level agreement (SLA) specifications. Impala is capable of handling the vast amount of data and is more efficient than Hive. Impala is intended to handle real-time ad-hoc queries to handle data exploration and is well-suited to executing SQL queries for interactive exploratory analytics on large data sets.
[9] Hadoop can run on large clusters of commodity hardware to support big data processing. SQL-on-Hadoop systems are more cost-efficient than MPP options such as Teradata, Vertica, and Netezza, which need to run on expensive high-end servers and don't scale out to thousands of nodes. When the volume of data is really big, only some portion of data can be loaded into main memory, the remaining data has to be stored on disks. Spreading input-outputs to a large cluster is one of the merits of the MapReduce framework, which also justifies SQL-on-Hadoop systems. SQL-on-Hadoop systems not only provide SQL query capability but also provide machine learning and data mining functionalities.
[10] Apache Giraph, an open source implementation of Google Pregel which is based on Bulk Synchronous Parallel Model (BSP) is used for graph analytics in social networks like Facebook. Graph mining has become an active and important subarea in data mining with the increasing demand on the analysis of large amounts of graph formatted structured data.
[11] Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. Data volume, Velocity, Variety, Veracity, and Value are the inseparable terms of big data. Big data which are flying over the world is the master key for any new digital decorations, and parallel algorithm is the suitable solution for the big data mining techniques as they serve a large number of notable uses areas like, (Financial, Telecommunication, Company, Science and Engineering and Industrial).

III. Analysis of Tools used in Big Data

For the analysis of Big Data we have considered these four commonly used tools:

- Apache Hadoop
- IBM Infosphere
- Cloudera Impala
- Apache Giraph

Basis	Apache Hadoop	IBM Infosphere	Cloudera Impala	Apache Giraph
Mode of software	Open source and free source	Commercial	Open source	Open source
Type of Data	Unstructured data, time series, textual data	Unstructured data, semi structured data and structured data	supports structured as well as unstructured data	Semi-structured or unstructured
Data Sources	Files, the network scripted Output	IBM Warehouse	SQL Clusters	Vertex and edge data
Database Support	HBASE, Sybase, SAP	Mongo DB, DB2, Oracle	HBASE or a HDFS	Graph databases such as Infinite Graph or Neo4j or with Hadoop
Operating System	Windows, Linux	Windows	RHEL-compatible, SLES, Ubuntu, and Debian	Cross-platform
Query Speed	Slow	High	High	High
Fault Tolerance	High	Yes	Yes	Yes - by check points

Table 1.1 Comparison of Various Tools

The tools discussed above are open source software that is used to access and work with big data. Apache Hadoop is high fault-tolerant software working on Windows and Linux operating systems taking unstructured data in the form of time series, textual data etc. On the other hand, other tools i.e. IBM Infosphere, Cloudera Impala and Apache Giraph work upon both structured as well as unstructured data giving the benefit of high query speed and fault tolerance to the users. Unlike Apache Hadoop Cloudera Impala is supported by RHEL-compatible, SLES, Ubuntu, and Debian operating systems and Apache Giraph can work on cross platforms.

IV. Conclusion

As the Big Data is the next big thing in an era and will continue to be for next few decades. This paper briefly describes the most commonly used seven V's of Big Data. Hadoop is ideal for storing large amounts of data, like terabytes and petabytes. It uses HDFS (Hadoop Distributed File System) as its storage system. HDFS connect nodes (commodity personal computers) contained within clusters over which data files are distributed. In this paper, we present a comparative study of various tools used in Big Data analytics. We have taken only four tools and seven parameters for the comparison. So, there is a scope for future enhancement to take more tools and more parameter to give the comparative study.

REFRENCES:

[1] Apache HBase. Available at http://hbase.apache.org

[2] Apache Hive. Available at http://hive.apache.org.

[3] https://www.ibm.com/blogs/watson-health/the-5-vs-of-big-data/.

[4] Vibhavari Chavan et al, (2014) "Survey Paper on Big Data" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (6), pp. 7932-7939.

[5] Varsha B.Bobade (2016) "Survey Paper on Big Data and Hadoop" International Research Journal of Engineering and Technology (IRJET), Vol. 3, Issue 1, pp. 861-863.

[6] Sara Landset, Taghi M. Khoshgoftaar, Aaron N. Richter and Tawfiq Hasanin (2015) "A survey of open source tools for machine learning with big data in the Hadoop ecosystem", Journal of Big Data, a Springer open journal, 2:24 DOI 10.1186/s40537-015-0032-1.

[7] M. Hirzel H. Andrade B. Gedik G. Jacques-Silva R. Khandekar V. Kumar M. Mendell H. Nasgaard S. Schneider R. Soule´ K.-L. Wu (2013)" IBM Streams Processing Language: Analyzing BigData in motion" IBM J. RES. & DEV. VOL. 57 NO. 3/4.

[8] SahithiTummalapalli, Venkata raoMachavarapu, (2016) "Managing Mysql Cluster Data Using Cloudera Impala", Procedia Computer Science 85, pp. 463 – 474.

[9] Yueguo Chen, Xiongpai Qin, Haoqiong Bian, Jun Chen, Zhaoan Dong, Xiaoyong Du, Yanjie Gao, Dehai Liu, Jiaheng Lu and Huijie Zhang, (2014) "A Study of SQL-on-Hadoop Systems" , Springer International Publishing Switzerland, pp. 154–166.

[10] Anuraj Mohan , Remya G ,"A Review on Large Scale Graph Processing Using Big Data Based Parallel Programming Models", I.J. Intelligent Systems and Applications, 2, pp. 49-57, 2017.

[11] Nour E. Oweis, Suhail S. Owais , Waseem George, Mona G. Suliman , Václav Snášel, "A Survey on Big Data, Mining: (Tools, Techniques, Applications and Notable Uses)", Conference Paper , DOI 10.1007/978-3-319-21206-7_10, 2015.