International Journal of Advances in Computer and Electronics Engineering
Volume: 01 Issue: 02, June, 2016, pp. 1 – 4

A Survey on Big Data, Challenges and Related Technologies

[1]S. Sridevi, [2]K. R. Kundhavai

[1]Assistant professor, Department of Computer Science & Engineering, New Horizon College of Engineering, Bangalore, India
[2]Assistant professor, Department of Computer Science & Engineering, New Horizon College of Engineering, Bangalore, India
[1]ssdevikumar@gmail.com, [2]kundhavai@gmail.com

Abstract: *According to recent surveys, the digital data that is available electronically are roughly in zettabytes which will grow in future. This data is arrived from various sources. One best example is Facebook which hosts approximately 10 billion photos, and occupies around one petabyte of storage. The issues here are how to collect, store, manage and analyze this large datasets. big data refers to the large datasets whose size is more which is beyond the ability of traditional database software to collect, store, manage and analyze. It is one which works well beyond the traditional limits of data based on three constraints Volume, Variety and Velocity. The process of examining these large datasets which contains variety of data types and to find the hidden patterns according to customer preferences with useful information is called as Big data Analytics. The Tool that is used for big data Analytics includes Hadoop with the related technologies YARN, MapReduce, Hive, Pig and NoSQL databases. This paper gives thorough information about Hadoop and the related technologies.*

Keyword: *Big data, Hadoop, YARN, MapReduce, Hive, Pig, NoSQL ;*

1. INTRODUCTION

Big data is same like small data but larger in size. Being larger, it increases the complexity to gather, store, manipulate and analyze it. To work out with this complexity we have come out with a technology called big data which characterizes the data by using 3Vs Volume, Variety and Velocity.

Volume is the quantity of data that is generated, Velocity is the speed of data collection and processing irrespective of the challenges in big data and variety is whether the data is structured, semi-structured or unstructured data. The main reasons for the growth of big data are data availability, to increase the storage capacity and to increase the processing power.

Apart from these 3vs the complexity increases when you try to do complex operations on it. Big data doesn't specify the quantity of data and always it is referred in petabytes and zettabytes. The large volume of data cannot be handled like the traditional data. Hence it uses a tool called Hadoop which is an Apache project started in 2006.

It can process large volume of data of different type. The concept used behind big data and hadoop is data distribution in more number of systems and working it out using parallel processing [1].

2. BIG DATA ANALYTICS

The main goal of big data analytics is to help companies in decision making, predictive analysis and other analytics professionals to analyze large volumes of transaction data. Some people consider that big data analytics is exclusively meant for semi-structured and unstructured data. Few examples are social media contents, social network activity reports, mobile-phone call detail records etc. The Semi-Structured and unstructured data cannot be handled using the

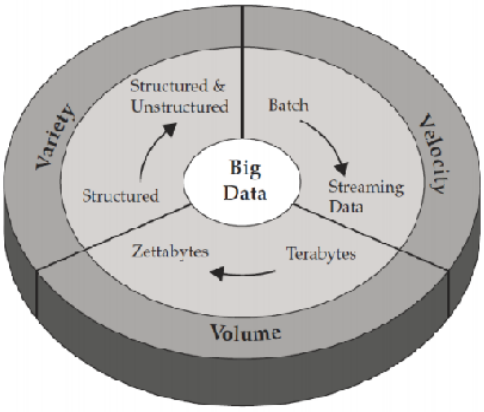

Figure 13 V's that influence big data

traditional data warehouses based on relational databases and at the same time the processing demands also cannot be posed by these relational databases. So to analyze this big data we are moving towards hadoop with the related tools YARN, Mapreduce, Hive, Pig and NoSQL databases [2].

3. HADOOP

Hadoop is an open source java framework which allows distributed processing on large datasets across cluster of computers using simple programming models. It needs an environment that provides distributed storage and computation across clusters of computers. It provides scalability which allows from single server to thousands of machines, with separate local computation and storage.

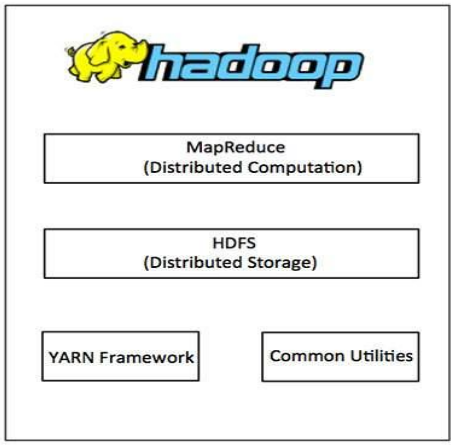

Figure 2 Hadoop

Hadoop includes four modules:

Hadoop Common: It includes libraries and utilities used by other modules. It contains the necessary java files and scripts which are needed to start Hadoop. These libraries allow file system and OS level abstractions.

Hadoop YARN: This is a framework for job scheduling and cluster resource management.

HDFS: This is a distributed file system that provides high-throughput access to application data.

MapReduce: This is YARN-based system for parallel processing of large data sets.

Apart from these base modules, Hadoop provides some additional software packages which can be installed on top of Hadoop, such as Pig, Hive, Hbase, Spark etc.

4. MAPREDUCE

It is a software framework used to write applications which does parallel processing on large amount of data stored in thousands of nodes. It performs two different tasks as follows:

Map Task: This task collects the input and converts it into data blocks, where the individual elements are broken down into tuples called key/value pairs.

Reduce Task: It is after the Map task. It takes the output of Map task as input and combines those output tuples into smaller set of tuples[3].

The Map Reduce framework contains two types of nodes called the master node Job tracker and the slave node Task tracker for each cluster node. The master node manages the resources, tracks resource availability, allocates tasks to the slave nodes, monitors them and re-executes the failed tasks. The task tracker executes the task assigned to it and periodically sends the information to its master node. The failure of the job tracker is a single point failure, which halts all the running jobs if the Job tracker goes down.

5. PIG

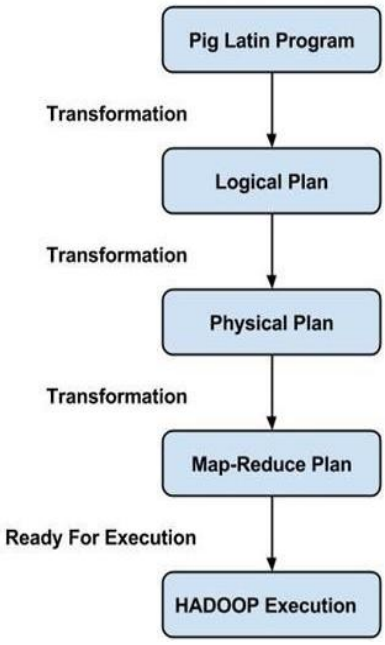

Figure 3 Series of MapReduce jobs

Pig is a high-level programming language which does the work of analyzing large datasets. It helps in translating the Map Reduce framework programs into a series of map and Reduce stages. It helps the programmers to spend time in analyzing the large datasets and not in writing the Map Reduce programs. The Pig programming language is designed to work on any kind of data.

Pig consists of two components: Pig Latin, which is the language that includes set of operations to process the input and produce the output. These set of operations describe a data flow which translates the program into executable representation. The result of these operations is the series of Map Reduce jobs.

Runtime Environment to run the Pig Latin programs [4]. It has two execution modes:

Local Mode: This mode of pig analyzes only small datasets. Here Pig runs in a single JVM and uses the local file system.

Map Reduce mode: This mode of pig analyzes large datasets. The queries written in Pig Latin are translated to Map Reduce jobs and are allowed to run on hadoop and generate the output.

6. HIVE

It is a tool that is used by Hadoop in order to process structured data. It does the summarization of big data which makes querying and analyzing easy. Hive is not

-A relational database

-A design for OnLine Transaction Processing (OLTP)

-A language for real-time queries and row-level updates

Features of Hive

-It stores schema in a database and processed data into HDFS.

-It is designed for OLAP.

-It provides SQL type language for querying called HiveQL or HQL.

It is familiar, fast, scalable, and extensible.

Architecture of Hive [7].

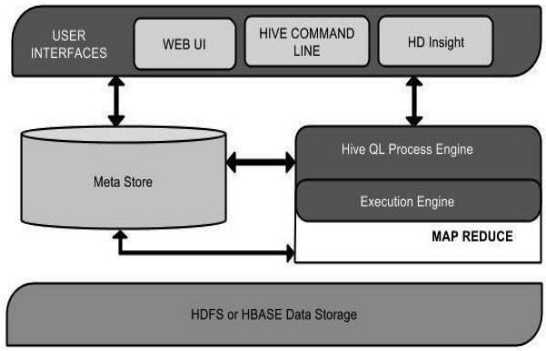

Figure 4 Hive Architecture

The following component diagram depicts the architecture of Hive:

The different units of the component diagram are as follows:

-User Interface: It allows the interaction between User and the HDFS.

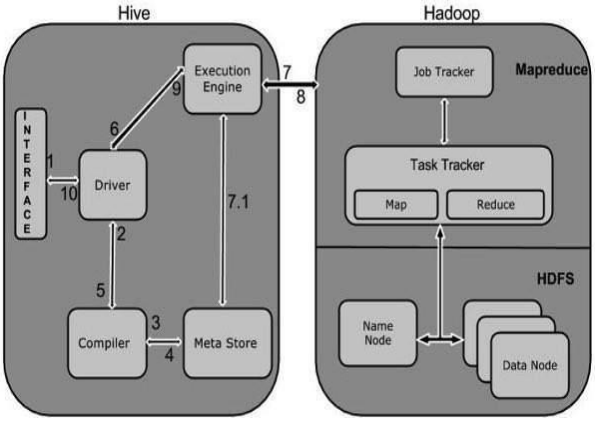

Figure 5 Workflow between Hive and hadoop

-Metastore: The storage of schema or metadata of tables, databases, HDFS Mapping etc

-Hivonal neQL process engine: It is a replacement of traditional approach for Map Reduce program. It is same like SQL which allows querying on schema info on metastore.

-Execution Engine: It is the conjunction part of HiveQL process engine and Map Reduce. It is responsible for processing the query and generates the results.

-HDFS or HBase: Data storage techniques to store the data in the file system.

Working of Hive

The following diagram depicts the workflow between Hive and Hadoop [5].

The steps that define how Hive interacts with Hadoop framework are Execute query, get plan, Get Metadata, Send metadata, Send Plan, Execute Plan, Execute Jobs, Metadata operations, Fetch Results and Send results [6].

7. CONCLUSION

The importance of big data and the information about the related topics are clearly depicted using diagrammatic representation in this paper. It gives the basic knowledge about the emerging technology. Since large amount of data is getting accumulated everywhere, it becomes very essential to study about big data.

REFERENCES

[1] http://www.planet/data.eu/sites/default/files/presentations/Big_Data_Tutorial_part4.pdf-Diagram-

[2] https://en.wikipedia.org/wiki/Big_data

[3] http://www.hindawi.com/journals/tswj/2014/712826/

[4] http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

[5] http://www.tutorialspoint.com/hadoop/hadoop_introduction.htm

[6] http://www.tutorialspoint.com/hive/hive_introduction.htm
Papers

[7] Lyman P, Varian H. How much information 2003? Tech. Rep, 2004. [Online]. Available:http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/printable_report.pdf Xu R, Wunsch D. Clustering. Hoboken: Wiley-IEEE Press; 2009.Google Scholar

[8] Ding C, He X. K-means clustering via principal component analysis. In: Proceedings of the Twenty-first International Conference on Machine Learning, 2004, pp 1–9.

[9] Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets.

[10] Fisher D, DeLine R, Czerwinski M, Drucker S. Interactions with big data analytics. Interactions. 2012;19(3):50–9.View ArticleGoogle Scholar

[11] Laney D. 3D data management: controlling data volume, velocity, and variety, META Group, Tech. Rep. 2001. [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[12] van Rijmenam M. Why the 3v's are not sufficient to describe big data, BigData Startups, Tech. Rep. 2013. [Online]. Available: http://www.bigdata-startups.com/3vs-sufficient-describe-big-data/

[13] Borne K. Top 10 big data challenges a serious look at 10 big data v's, Tech. Rep. 2014. [Online]. Available:https://www.mapr.com/blog/top-10-big-data-challenges-look-10-big-data-v.

[14] Press G. $16.1 billion big data market: 2014 predictions from IDC and IIA, Forbes, Tech. Rep. 2013. [Online]. Available: http://www.forbes.com/sites/gilpress/2013/12/12/16-1-billion-big-data-market-2014-predictions-from-idc-and-iia/.

[15] Big data and analytics—an IDC four pillar research area, IDC, Tech. Rep. 2013. [Online]. Available:http://www.idc.com/prodserv/FourPillars/bigData/index.jsp.

[16] Taft DK. Big data market to reach $46.34 billion by 2018, EWEEK, Tech. Rep. 2013. [Online]. Available:http://www.eweek.com/database/big-data-market-to-reach-46.34-billion-by-2018.html.