International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 5 | ISSN : 2456-3307

A Survey on Big Data and Hadoop Ecosystem Components

P. Bastin Thiyagaraj[1], B. Latha[2], Dr. A. Aloysius [3]

[1]Department of Information Technology, St. Joseph's College (Autonomous), Trichy, TamilNadu, India

[2]Department of Information Technology, St. Joseph's College (Autonomous), Trichy, TamilNadu, India

[3]Department of Computer Science, St. Joseph's College (Autonomous), Trichy, TamilNadu, India

ABSTRACT

Big data describes techniques and technologies to store, distribute, manage, analyse large size of data set with high velocity. Here this paper is effort to present the basic understanding of the big data and its usefulness to an organization from the performance, also evaluate the challenges faced by a small organization. Hadoop is popular tool for big data implementation. Hadoop is the core platform for structuring Big Data, and solves the problem of making it useful for analytics purposes. Hadoop and HDFS by Apache are widely used for strong and managing data. Hadoop is the platform for structuring big data and solves the problem of making it useful for analytic purpose. This paper is dealt with big data challenges, hadoop components and problem with hadoop were described in detail.

Keywords: Hadoop, Big Data, HDFS, Apache.

I. INTRODUCTION

Big data is a blanket term for the non-traditional strategies and technologies needed to gather, organize, process, and gather insights from large dataset. An exact definition of "big data" is difficult to nail down because projects, vendors, practitioners, and business professionals use it quite differently. With that in mind, generally speaking, big data is large datasets termed as "Big Data" due to its sheer Volume, Variety, Velocity and Veracity. The data are in the form of unstructured, quasi structured or semi structured and it is heterogeneous in nature. The need of big data generated from the large companies like face book, yahoo, Google, YouTube etc for the purpose of analysis of enormous amount of data. Hadoop Amis an open source java based programming framework. It supports the processing and storage of extremely large data sets in a distributed computing environment. It is a part of the apache project sponsored by the apache software foundation. The components of hadoop is hadoop common, hadoop yarn, hadoop map reduce and HDFS.

II. BIG DATA CHARACTERISTICS

Big data can be described by following Ten characteristics
Of big data.

Volume: Big data volume means "scale of data or large amount of data generated in every second. Now a day's data volume is increasing from gigabytes to pet bytes.

Variety: Variety refers to the many sources and types of data both software and instructed. It is used to store the data from sources like spreadsheets and databases.

Veracity: Big data veracity refers to the noise and abnormality in big data. The data that is being stored and meaningful to the analysed [6].Veracity means uncertainty or accuracy of big data.

Velocity: Velocity is speed at which data is generated. Sources like business process, machine, networks and human interaction

Value: The value of big data is huge. Value is main source for big data it is important for business. It is used to store large amount of values in database.

Variability: There are several potential meanings for variability. It contains many extreme values it presents statistical problem to determine the outlier values and a new important signal or just noisy data.

Viscosity: It is used to describe the lag time in the data relative to the event being described. It easily understood as an element of velocity.

Virality: Defined by some users as the rate at which the data spreads. It is picked up and repeated by other user or events.

Visualization: Big data visualization tools face technical challenges due to limitations of in memory technology and poor scalability, functionality and a response time. The multitude of variables resulting from big data is variety and velocity and the relationship between developing a meaningful visualization is not easy.

Validity: Validity refers to accurate and correct the data is for its intended use. The benefit from big data analytics is only as good as its underlying data, good data governance to ensure consistent data quality, common definitions and Meta data.

III. CHALLENGES OF BIG DATA

It is not an easy task to find ways to make this data useful in every large company. The amount of data produced makes it very difficult to store, manage, analyse and utilize it. The development of various big data analysis tools have helped with data handling to a great extent.

Data storage and quality: Companies and Organizations are growing at a very fast pace. The growth of the companies rapidly increases the amount of data produced. The storage of this data is becoming a challenge for everyone. Options like data lakes warehouses are used to collect and store massive quantities of unstructured data in its native format. Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges.

People understand Big Data Analysis: Data Analysis is very important to make the huge amount of data being produced, useful. Therefore, there is a huge need for Big Data analysts and Data Scientists. The storage of quality data scientists has made it a job in great demand. This is another challenge faced by companies. The number of data scientists available is very less in comparison to the amount of data being produced.

Quality analysis: The companies and organizations use big data produced to make the best decisions possible. The data they are using should be accurate. If the data used to make decisions is not accurate, it will result in ill-advised decisions that would ultimately be detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue. This requires many resources to ensure the accuracy of the information provided. The process of creating accurate data is very time consuming and requires the use of tools that can be expensive.

Security and privacy of the data: It also involves big risks when it comes to the security and the privacy of the data. The tools used for analysis, stores, manages, analyses, and utilizes the data from a different variety of sources. This ultimately leads to a risk of exposure of the data, making it highly vulnerable. Thus making it essential for analysts and data scientists to consider these issues and deal with the data in a manner that will not lead to the disruption of privacy.

IV. HADOOP ECOSYSTEM COMPONENTS

Hadoop ecosystem is a programming language. It is a platform or framework it solve the big data problems. Hadoop big data components like,

- ✓ Hadoop common
- ✓ HDFS
- ✓ Map reduce

- ✓ Yarn
- ✓ Hive
- ✓ Apache pig
- ✓ Mahout
- ✓ Sqoop
- ✓ Zookeeper

Hadoop common: These are java libraries and utilities required by other hadoop modules Libraries provides file systems and Os level abstraction and contains the necessary java files are scripts. For example ,If Hbase and Hive to access HDFS they need to make of jar files that are stored in Hadoop common.

HDFS: Hadoop File System was developed using distributed file system design. It is run on commodity hardware. Unlike other distributed systems, HDFS is highly fault tolerant and designed using low-cost hardware. HDFS holds very large amount of data and provides easier access. Hadoop provides a command interface to interact with HDFS[6].The built-in servers of name node and data node help users to easily check the status of cluster.

Hadoop map reduce: Map reduce is frame work to run distributed computations write two functions. The two function are map() and reduce() function.map and reduce is a programming language. Hadoop is the open source Map reduces implementation. The input and output of map reduce programs are HDFS files[1]. The map() function it takes an input key value pairs. it produce a list Of intermediate key value pair. The map reduce runtime system groups all intermediate pair base on the keys and passes to reduce () function for producing the final results. Map Reduce framework automatically runs many map per and reducer jobs on the cluster, on splits on the input files.

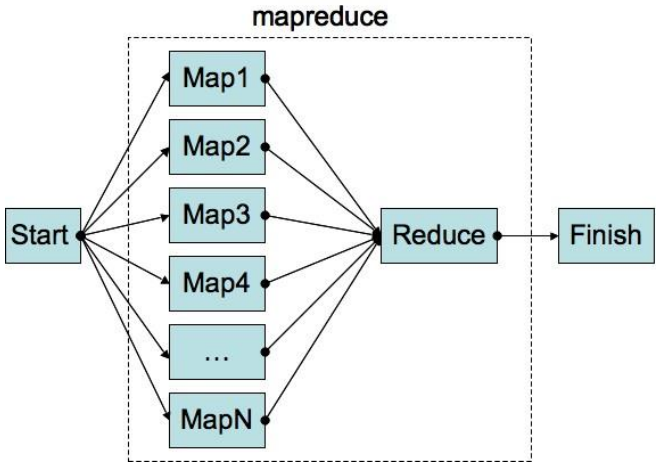

Figure 1

Yarn: It is a framework for job scheduling and cluster resource management. Yarn is also one the most important component of Hadoop Ecosystem. YARN determines which jobs run when on which machines and why. It keeps track of which machines have CPU and memory free. It looks for each job at where the data that it needs resides, to make this decision, trying to run jobs close to the data.

Hive: Apache hive is an open source data warehouse system for querying and analysing large datasets stored in Hadoop files. Hive do three main functions: data summarization, query, and analysis. Hive allows writing sql like queries to process and analysing the big data stored in HDFS Hive Query language is similar to standard sql statement. It is designed for OLAP.

Hbase: Hbase is a hadoop application which runs on top of HDFS .Hbase system represents set of table but hbase is column oriented database management system such that different from the row oriented database system[10]. Hbase is not relational database at all and also it doesn't support sql. The features of HBase are to real time read or write to large data sets. for example read and write operations involve all rows but only a small subset of columns.

Pig: Pig is a tools developed by yahoo for analysing huge data sets efficient and easily. It provides a high level data flow language pig Latin that is optimized

extensible and easy to use.pig layer consists of a compiler that produces of map reduce programs for which the large scale parallel implementation pig language[11]. It consists of a textual language.

Mahout: Mahout is open source framework for creating scalable machine learning algorithm and data mining library. Once data is stored in Hadoop HDFS, mahout provides the data science tools to automatically find meaningful patterns in those big data sets.

Sqoop: Sqoop imports data from external sources into related Hadoop ecosystem components like HDFS, Hbase or Hive. It also exports data from Hadoop to other external sources. Sqoop works with relational databases such as Teradata, Netezza, oracle, MySQL.

Zookeeper: Apache Zookeeper is a centralized service and a Hadoop Ecosystem component for maintaining configuration information, naming, providing distributed synchronization, and providing group services. Zookeeper manages and coordinates a large cluster of machines.

V. PROBLEM WITH HADOOP

Hadoop is a big suite of Big Data tools and as such is non-trivial to setup, maintain and write programs. Programming Map-Reduce jobs using Java is non-trivial and cumbersome and this motivated a new gen of Hadoop frameworks like Hive and Pig to simplify those tasks[3].

A team with good knowledge of Garbage Collection (GC) tuning should support a production Hadoop. The Java Virtual Machine itself is a big burden to support on scale. Hadoop is oriented towards non-interactive data processing. This is changing with new Hadoop technologies like Impala, but Hadoop is still the batch processing of the XXI century. Spark is more appropriate for interactive big data processing[2]. Hadoop tools like HDFS and Map-Reduce are not optimized for performing tasks quickly, but on crunch Terabytes of data in a reasonable time and with sustained throughput. Again, Spark addresses this performance penalty incurred by the use of Hadoop for big data.

VI. CONCLUSION

This paper described the concept of big data, hadoop ecosystem and the concept of big data along with ten characteristic of big data. It is the challenging one that the big data must be addressed for efficient and fast processing of big data. The hadoop ecosystem used to process the big data to make it clear and to take decision.

VII. REFERENCES

[1]. Hadoop, http://hadoop.apache.org/mapreduce/.

[2]. Edward Riberahttps://www.quora.com/What-are-some-problems-with-Hadoop-and-where-does-it-fail-to-deliver.

[3]. https://www.bloorresearch.com/2011/11/problems-hadoop/

[4]. Apache hadoop://hadoop.apache.org

[5]. S.vikramn phaneendra, E.Madhusughan Reddy"Big data-solutions for RDBMS problems-A survey paper

[6]. Harshawarthan .S Bhosala.Prof Devendra P.Gadekar "Big data and Hadoop" in international journals of scientific and Research publications, volume 4.issue10,octomber 2014.

[7]. Big data in big companies ,Thomas Davenport and Jill Dyche, SAS Institute Inc. May 2013.

[8]. Bijesh Dhyani and Anurag Barathwal, "Big Data Analytics using Hadoop". International Journal of computer applications, volume 108,December 2014.

[9]. Rahul Beakta " A Review paper Big data and Hadoop" CSE Dept., Baddi university of Emerging sciences and technology India, volume 2,issue 2(2015).

[10]. Varsha B.Bobade "survey paper on Big data and hadoop", International Research journal of Engineering and Technology volume 3, Jan 2016 .

[11]. Dr E.Lakxmi Lydia, Dr M.Ben swarup "Analysis of Big data and hadoop Ecosystem components like flume, map reduce, pig and hive". Depart of computer science and engineering volume 5, Jan 2016.

AUTHOR'S PROFILE

P. BASTIN THIYAGARAJ is working as an Assistant Professor in the Department of Information Technology, St.Joseph'scollege(Autonomous), Tiruchirappalli, TamilNadu, India. I am having 7 years of experience in teaching and 2 years in research.

B.LATHA is studying II MSc Computer science in the Department of Information Technology St.Joseph'scollege (Autonomous), Tiruchirappalli, TamilNadu, India.

Dr. A. ALOYSIUS is working as an Assistant Professor in the Department of Computer Science, St. Joseph's College (Autonomous), Tiruchirappalli, Tamil Nadu, India. He has 16 years of experience in teaching and research. He has published many research articles in the National / International conferences and journals. He has acted as a chairperson for many national and international conferences. Currently, eight candidates are pursuing Doctor of Philosophy Programmed under his guidance.