International Journal of Intelligent Computing and Technology
ISSN: 2457-0249 (Online)
Website: https://ijict.com 2019 (Jan)−Vol.2, Iss.2 Page: 01−10

Big Data and its Applications: A Review

E Boopathi Kumar[1] & Dr V Thiagarasu[2]

[1]Research Scholar, [2]Associate Professor Department of Computer Science, Gobi Arts and Science College, Gobichettipalayam, Tamilnadu, India

edboopathikumar@gmail.com, profdravt@gmail.com

Article History: Received: August 2018; Published: January 2019

ABSTRACT

Big Data analytics is a process of analyzing and mining large amount of data. Big Data describes data sets that are too large, too unstructured or too fast changing for analysis. Due to increase in number of complicated targeted threats and rapid growth in data, the analysis of data becomes too difficult. In present world, every small device is a potential data source, adding to the huge data bank and every bit of data generated is practically valued, be it in enterprise data or personal data, historical or transactional data. At present, the leading challenge is to collect and analyze the big data fast enough to perform operations. Various frameworks, languages and techniques are available to handle these huge amounts of datasets. In this paper, an introduction to Hadoop framework and concepts of distributed file system is going to discuss in upcoming chapters.

KEYWORDS: Big data, Hadoop, MapReduce, Pig, Hive

INTRODUCTION

Big data analytics is the process of analyzing big data to find hidden patterns, unknown correlations and other useful information that can be extracted to make better decisions [1]. As the data coming into enterprises continue to reach extraordinary levels due to the volume, variety and velocity of data, analysis of big data is a big problem [2]. Due to this unexpected growth, the analyzer should understand big data in order to process the information that truly counts as well as analyze the possibilities of what one can do with the massive amount of data. Organizations of the chapters are follows. Chapter 2 deals with the characteristics of big data. Basics of Hadoop architecture is presented in Chapter 3. Chapter 4 describes key components of Hadoop Eco system. Few data sets used in big data analytics are given in Chapter 5 and conclusion is given in Chapter 6.

6V'S OF BIG DATA

Big Data is relatively a new concept and a lot of definitions have been given to it by researchers, organizations and individuals. Very first, industry analyst Doung Laney [3] expressed the mainstream of definition of Big Data as three Vs; Volume, Velocity and Variety. Oracle defined Big Data in terms of four Vs; Volume, Velocity, Variety and Value [4]. At present, it extends to six Vs; Volume, Velocity, Variety, Veracity, Variability and Value as shown in Figure 1. Further, Complexity and Visualization are added to the BigData scenes to visualize the data differently [5].

Volume: Denotes the size of the data ranges from terabytes and petabytes, to even Exabytes. In the modern era of technology, it has become very difficult to talk about data volume in any absolute sense. As technology grows, numbers get quickly outdated, so the data volume should be treated in a relative sense instead.

Velocity: Denotes data, which is generating at unpredictable speed. As the generation of data is rapid, the process of acquiring, processing and analyzing it requires fast mechanisms. The rate at which data is being received and has to be acted upon is becoming much more real-time. While it is unlikely that any real analysis will have to be completed in the same time period, delays in execution will inevitably limit the effectiveness of campaigns, limit interventions or lead to sub-optimal processes.

Variety: Variety define various data types which includes structured and unstructured data such as text, audio, video, sensor data, posts, log files and many more. Since data is either structured or unstructured, managing, merging or governing it is a challenging task for organizations.

Veracity: Refers to the requirement of correct form of data as it is relies upon accurate datasets for all further analysis such as availability, authenticity, accountability and so on.

Variability: Data can be in the same form but having different semantics. In addition to the increasing velocities and varieties of data, amount of data flow is highly variable.

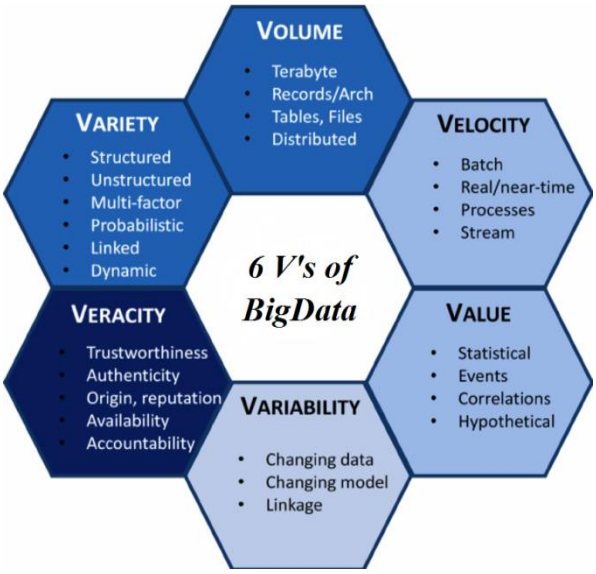

Fig1. 6 Vs of Bigdata [6]

Value: The value of data in analytics turns the input data into information followed by knowledge.

Data Visualization: As the output of data analytics should be presented to the users in a convincing form, data visualization acts an important component in big data analytics. By predicting customer purchasing behavior patterns, the organization can influence them to make purchases which used to be an incomprehensible task for most companies.

Big data analytics is a one-stop solution for business experts to question and interpret data according to their business requirements irrespective of the complexity and volume of the data. In order to achieve this requirement, data scientists need to efficiently visualize and present this data in a comprehensible manner. Giants like Google, Facebook, Twitter, EBay, Wal-Mart etc., adopted data visualization to ease complexity of handling data. Data visualization has shown immense positive outcomes in such business organizations. The resultant analytics can be presented to the users in the form of charts, slides, plots, tables, presentations and text contents [7].

Big data contains different data types and also having different datasets which is in the form of structured, unstructured and semi structured. In order to handle those kinds of data, Big data had a technology called Hadoop framework. Framework contains Hadoop, Pig, Hive and HDFS.

Single system is not capable to perform analytics with massive data sets. To rectify that, Hadoop framework is introduced with the concept of distributed file system to handle large number of data under its framework.

HADOOP ARCHITECURE

Apache Hadoop is an open source framework for distributed storage and processing of large sets of data on commodity hardware. Hadoop enables businesses to quickly gain insight from massive amounts of structured and unstructured data. It is used in maintaining, scaling and analyzing large scale of structured and unstructured data. Figure 2 shows high level architecture of Hadoop with working views.

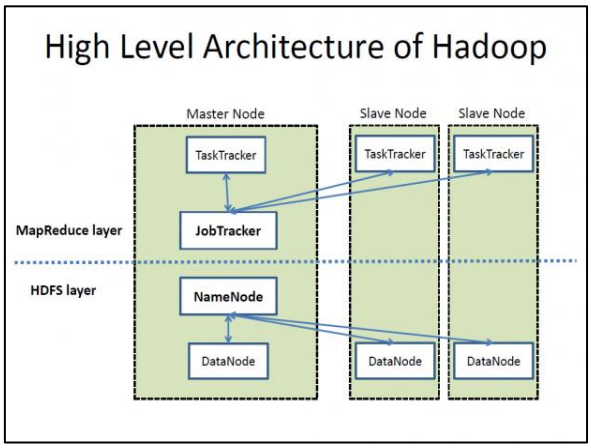

Fig2. Architecture of Hadoop [8]

Hadoop is an open-source tool build on java platform and aimed to improve the performance in terms of data processing on clusters. Hadoop comprises of multiple concepts and modules like HDFS, Map-Reduce, HBASE, PIG, HIVE, SQOOP, KAFKA, FLUME, MAHOUT, OOZIE and ZOOKEEPER to perform fast processing of huge data. It is different from Relational databases and can process high volume/velocity/variety of data to generate accuracy of data among huge datasets.

MapReduce is responsible for processing jobs in a distributed environment and it is the key algorithm that the Hadoop MapReduce engine uses to distribute work around a cluster. The architecture of the MapReduce is shown in Figure 3. The data is read from files into mappers and emitted from mappers to reducers [9]. A mapper function is comprised of map part and reduces part.

A map transform function is provided to transform an input data row of key and value to an output key/value: Map (key1, value) -> list<key2, value2>. For an input, the map function returns a list containing zero or more (key, value) pairs: The output can be a different key from the input and can have multiple entries with the same key. The Reduce function uses a reduce transform technique to take all values for a specific key, and generate a new list of the reduced output: Reduce (key2, list<value2>) -> list<value3>

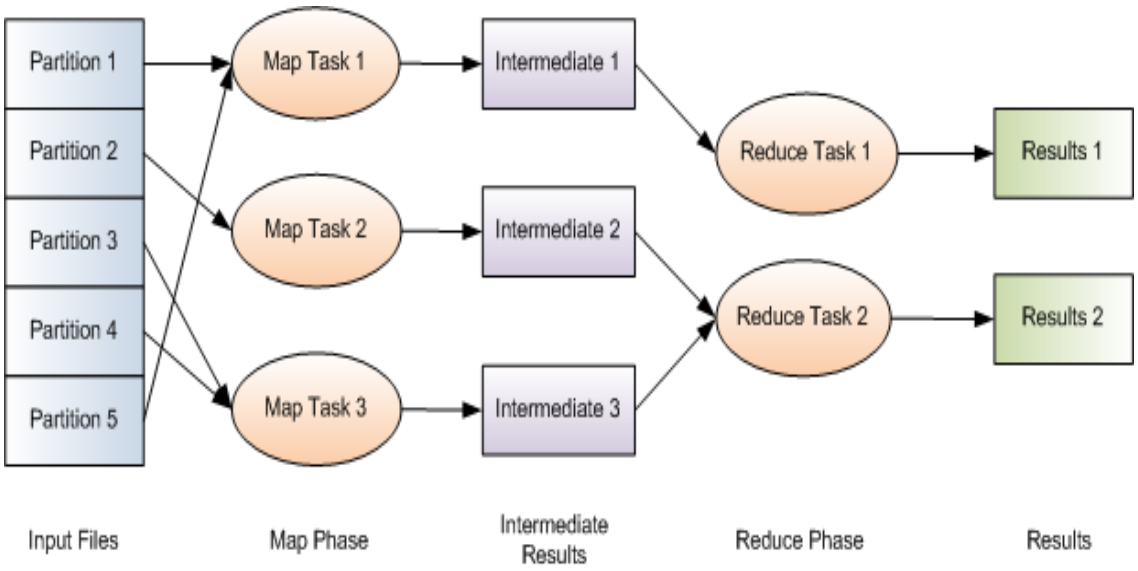

Fig3. MapReduce Architecture

KEY COMPONENTS OF HADOOP ECO - SYSTEM

PIG

Apache PIG is a platform for analyzing large data sets even in the form of unstructured data and it is a dataflow language which is also used as a scripting language. Pig works on Hadoop environment and it executes the files through Hadoop Distributed File System.

Pig runs in three different ways: Script file, Grunt command line and Java embedded form. PIG language allows for query execution over data stored on a Hadoop cluster, instead of a SQL-

like language. It operates on client side of any cluster [10]. Pig queries are similar to SQL queries and its statements are executed through MapReduce mode.

HIVE

Hive is a data warehouse tool used for querying and analyzing data in easier way and it enables traditional BI applications to run queries against a Hadoop cluster. It was developed originally by Facebook, later it was taken care by Apache software foundation. It's a higher-level abstraction of the Hadoop framework that allows anyone to make queries against data stored in a Hadoop cluster just as if they were manipulating a conventional data store. Hive supports an SQL like scripting language called HIVEQL, through which it executes MapReduce jobs effectively. Mainly, it is used for analyzing structured data and to handle large volume of data. Hive operates on server side of any cluster [11].

Like SQL databases, Hive works in terms of tables. There are two kinds of tables to be creating for processing external and internal data: managed tables whose data is managed by Hive and external tables whose data is managed outside of Hive. If the file is loaded into a managed table, Hive moves the file into its data warehouse. To drop that table, both the data and metadata are deleted. If the file is loaded into an external table, no files are moved. Dropping the table only deletes the metadata. The data is left alone [12]. Among the two kinds of tables, External tables are useful for sharing the data between Hive and other Hadoop applications.

APPLICATIONS OF BIG DATA

Big data applications are huge in nature. Every field has a set of data that needs to be processed for analytics. Some of the domains where the Big data can be apply and their challenges are listed below:

- Education & Research
  - Issues of data privacy and protection
  - Experiment sensor analysis
- Healthcare
  - Unavailability/Unusable data
  - Rising medical costs
  - Wearable devices sensor data
- Media, Entertainment and Communication

Understanding patterns of real time data

Social media monitoring

Threat analysis

- Banking

  Customer data analytics

  Card fraud detection and audit trials

  Risks & portfolio analysis

- Manufacturing and natural Resources

  Large volume of unused data from the manufacturing industry

  Clinical trials & Genomics

- Retail sale

  Unutilized data derived from customers such as loyalty cards, scanners, etc.

  Loyalty & Retention

- Transport

  Transport demand models

  Sensor analysis for optimal traffic flows

- Insurance

  Under utilization of data gathered by loss adjusters

SAMPLE DATA SET

A sample dataset in various fields used for big data analytics is given below:

No	Name	Website
1	Weather data	»cdo.ncdc.noga.gov, »earth.nullschool.net, »ncdc.noaa.gov/cdo-web, »datahub.io/datasets
2	Medical	»archive.ics.uci.edu »healthdata.gov/dataset »inf.ed.ac.uk
3	Yahoo Research Labs	»webscope.sandbox. yahoo.com »insidebigdata.com
4	War & Peace Book	»Gutenberg.org/files/2600/2600-8.txt
5	Twitter	»apps.twitter.com

6	Data Science	»kaggle.com »springboard.com »datasciencecentral.com »datascienceweekly.org
7	Indian Govt Data	»Data.gov.in »India.gov.in
8	Public datasets	»kdnuggets.com
9	Bank data	»data.worldbank.org »econ.worldbank.org »bankofengland.co.uk »globalbanking.org
10	Amazon	»aws.amazon.com/datasets »stackoverflow.com
11	Basic datasets	»dreamtolearn.com »support.minitab.com »stata.com

CONCLUSION

Big Data is an evolving field, where much of the research is yet to be done. Due to enormous growth of data in various fields, handling and storing the data become more difficult. The quantity of V's also extended like the growth of Big Data Analytics. The study presents the fundamental concepts of Big Data along with 6 Vs, Volume, Velocity, Veracity, Variety, Variability and value. Big data needs Visualization techniques to visualize the results of the analysis in pictorial form. This paper describes Hadoop framework along with its applications and sample datasets for processing of Big Data. Data sets provided in chapter 6 is useful for budding researchers in the field of analytics. Some applications and challenges of Big data are discussed in this study, and challenges can be taken for further research.

REFERENCE

1. Bijesh Dhyani and Anurag Barthwal (2014), Big Data Analytics using Hadoop, International Journal of Computer Applications, ISSN: 0975 – 8887, 108(12)
2. Soumya Shukla, Vaishnavi Kukade and Sofiya Mujawar (2015), Big Data: Concept, Handling and Challenges: An Overview, International Journal of Computer Applications, ISSN: 0975 – 8887, 114(11)
3. Ranjana Bahri (2015), Big Data: Concept, Challenges and Management Tools, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, 5(2)
4. Oracle (2014), Information Management and Big Data: A Reference Architecture, www.oracle.com/info-big-data-r
5. Samiddha Mukherjee and Ravi Shaw (2016), Big Data – Concepts, Applications, Challenges and Future Scope, International Journal of Advanced Research in Computer and Communication Engineering, ISSN: 2278-1021, 5(2)
6. Z Mo and Y F Li (2015), Research of Big Data Based on the Views of Technology and Application, American Journal of Industrial and Business Management, 5,192-197,
7. Sanjeev Dhawan and Sanjay Rathee (2013), Big Data Analytics using Hadoop Components like Pig and Hive, American International Journal of Research in Science, Technology, Engineering & Mathematics, 2(1), 88-93
8. Harshawardhan S. Bhosale and Devendra P. Gadekar (2014), A Review Paper on Big Data and Hadoop, International Journal of Scientific and Research Publications, 4(10)
9. V Bhuvaneswari and T. Devi (2016), Big Data Analytics: A Practitioner's Approach, 1-7
10. V Bhuvaneswari (2016), Data Analytics with R: Step by Step, ISBN: 978-81-929131-2-4
11. X Wu, X Zhu and G Q Wu (2014), Data mining with big data, IEEE Trans. on Knowledge and Data Engineering, 26(1),97-107
12. N Rashmi, K M Uma, K Jayalakshmi and K P Vinodkumar (2014), Big Data Security Challenges: Dealing with too many issues, International Journal of Recent Development in Engineering and Technology, 3(2)
13. Mrigank Mridul, Akashdeep Khajuria, Snehasish Dutta and N Kumar (2014), Analysis of Big data using Apache Hadoop and Map Reduce, 4(5)

14. Kiran kumara Reddi and Dnvsl Indira (2013), Different Technique to Transfer Big Data: survey, IEEE, 52(8), 2348-2355
15. M L Umasri, D Shyamalagowri and Suresh Kumar (2014), Mining Big Data: Current status and forecast to the future, International Journal of Advanced Research in Computer Science and Software Engineering, 4(1)

How to Cite:
E Boopathi Kumar & Dr V Thiagarasu, "Big Data and its Applications: A Review", International Journal of Intelligent Computing and Technology (IJICT), Vol.2, Iss.2, pp.01-10, 2019