INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZIG YOUR INNOVATIVE WORK

A REVIEW ON TOOLS FOR BIG DATA PROCESSING

SUNIL SAINI

Assistant Professor, Deptt. of Computer Sc. & Engg, Kurukshetra Institute of Technology & Management, Kurukshetra, Haryana.

Accepted Date: 15/07/2017; Published Date: 01/08/2017

Abstract: - Big data is very sensational topic from past few decades, it attracts many researchers, analysts to dive deep into the sea of data and bring back the pearls of knowledge for themselves. As every individual is connected with internet and from different sources of data, data is increasing at exponential rate. This data is structure, semi-structured and unstructured in nature due to which new tools and techniques with high processing requirements is needed. Now a days a number of tools are used for big data processing. This paper presents a overview of tools which can be used for big data processing

Keywords: Big data, Volume, Velocity, Variety

Corresponding Author: MR. SUNIL SAINI

Access Online On:

www.ijpret.com

How to Cite This Article:

Sunil Saini, IJPRET, 2017; Volume 5 (12): 13-19

PAPER-QR CODE

INTRODUCTION

Any data which is satisfying the five V's i.e. volume, velocity, variety, veracity & value is termed as big data. Internet is ubiquitous in our society due to which every individual who is using internet is generating data. From the survey it is concluded that 2.5 quintillion bytes of data is created on daily basis. There are so many sources of data like sensors, social networks like Facebook, twitter, weather forecasting data, YouTube, web forms, emails etc. One real life example of big data is servers of Walmart tackle more than one million customer's transaction every hour which is about 2.5 PB (Peta- Bytes) of data [1]. Another example is from CERN i. e. Large Hadron Collider data centers which process everyday around one petabytes of data which is equivalent to 210,000 DVDs [2].

Conclusion is colossal data i.e. big data needs new architecture, analytics and techniques to extract valuable information from the data's sea. This is exactly similar to finding gold nuggets from goldmine [3]. Therefore it is difficult to process big data with traditional desktop applications and database management system and necessitate parallel computing softwares. In section 2 big data characteristics will be discussed. Different tools which are used to mine big data will be discussed in section 3.

II. Big Data's V

Big data is distinguished from data in respect of high dimension, dynamic data & variety of data (audio, video, images etc) and having multiple optimization objectives. As the scale of data increases exponentially, traditional processing tools are unable to process it [7].

Challenges include in processing of big data is storage, how to capture data, how to analyze, how to share and visualize [8]. Volume, velocity and variety are the 3Vs of big data. But later this is extended to fourth & fifth V, i.e. Veracity and value [4].

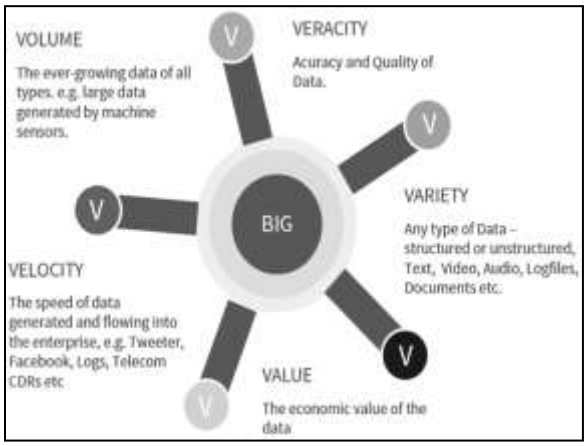

Figure1: 5 V's of Big Data.

These five V's that is volume, velocity, variety, veracity and value is summarized in figure 1.

A Volume

Volume is the amount data having very large scale ranging from petabytes to yottabytes. Source of this data is endless like social site, data generated by machines, transactional data and so on [5] [6].

B Variety

Variety is defined as opulence of data representation. Variety includes structured, semi-structured and unstructured data. For example relational databases, email, web pages, audio and videos represents variety in data [5] [6].

C Velocity

Velocity is too fast data which is not tackled with traditional algorithms and techniques. This real time, high speed data is need to be processed for detecting fraud and for suggesting items on the basis of customer purchase history, without storing it into databases. [5] [6].

D Veracity

Veracity represents the unreliability inherent in some sources of data. For example, customer sentiments in social media are uncertain in nature, since it entails human judgments [5-6].

E Value

Value is the worthness of data extracted from large scale data. Ultimate goal of big data mining is finding insights [5-6].

III. Big Data Tools

Big data tools are used for preprocessing of data after that dimensionality is reduced. By using reduced data a model is build which will analyze the data from different perspectives to mine associations, pattern etc. Then evaluation of model is done to find true positive and true negative values. Different mining tools are discussed next.

A MATLAB

MATLAB stands for MATrix LABoratory and it is language for technical computing. Basic data element of MATLAB is an array. This array does not require dimensioning. Initially MATLAB was written for matrix software developed by LINPACK & EISPACK projects. MATLAB is used for development of algorithm, computation in math, analysis of data, its exploration and its visualization and application development & so on. MATLAB is collection of toolboxes which consists of different functions for solving various classes of problems. Various toolboxes available are neural networks, control systems, fuzzy logic, signal processing, and image processing and so on. MATLAB language, environment, graphics system, mathematical function library and application program interface are five main part of MATLAB system. MATLAB workspace is shown in figure 2 below [11].

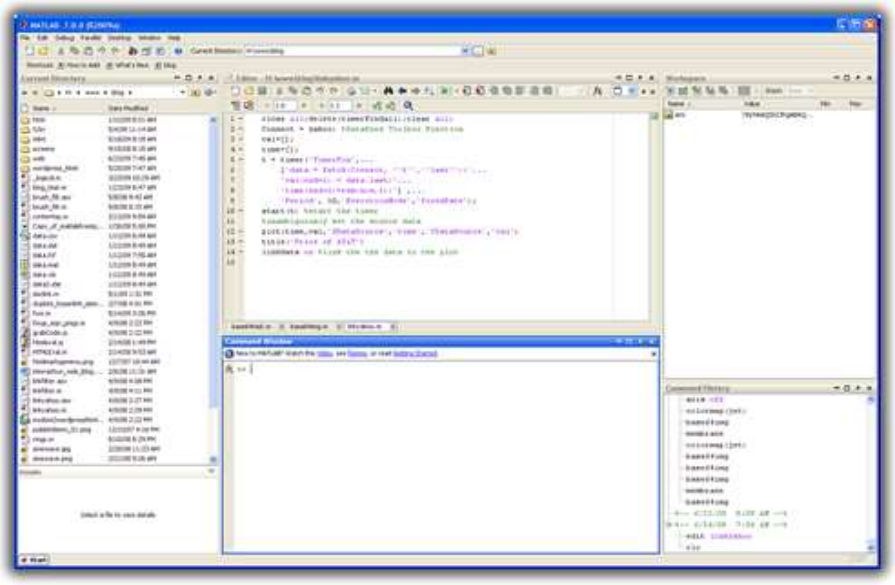

Figure 2: MATLAB Workspace

B Hadoop

Hadoop is majorly used to solve the issues related to big data. Currently everyone is generating loads of data and it is very difficult to manage such a colossal data. For this Google invented new data processing methodology known as map reduce and a year later Doug Cutting & Mike Cafarella invented Hadoop. Hadoop is open source and java based framework. Ecosystem of Apache Hadoop is shown below in figure 3.

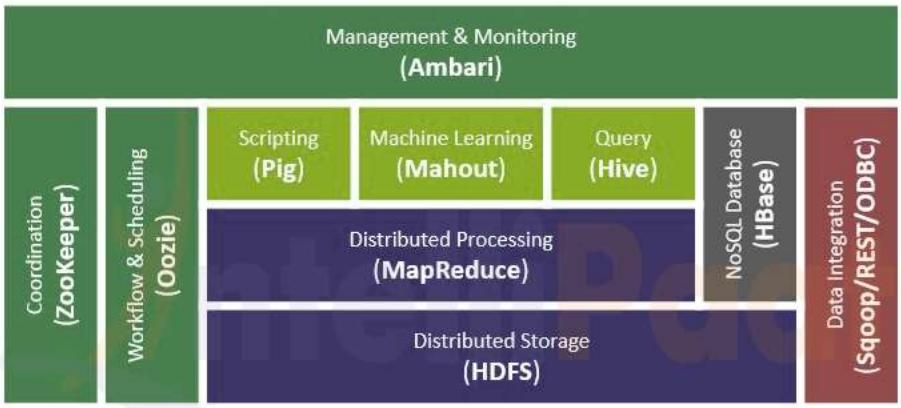

Figure 3: Apache Hadoop Ecosystem

Hadoop is able to tackle different data ranging from text data, transactional data, statistical data, audios, videos, social media, and news update and so on. Heart of Hadoop is map reduce, where the data is processed parallel. Various modules of Hadoop are Hadoop Distributed File System, Map Reduce, Hadoop YARN, Ambari, HBAse and so on [10].

C Weka

Weka is a data mining tool which is open source. It consists of various machine learning algorithms. Weka contains different tools for data preprocessing, clustering, classification, association rule mining and visualization. It provides facility of supervised and unsupervised learning with inbuilt algorithms. Weka explorer is shown in figure 4.

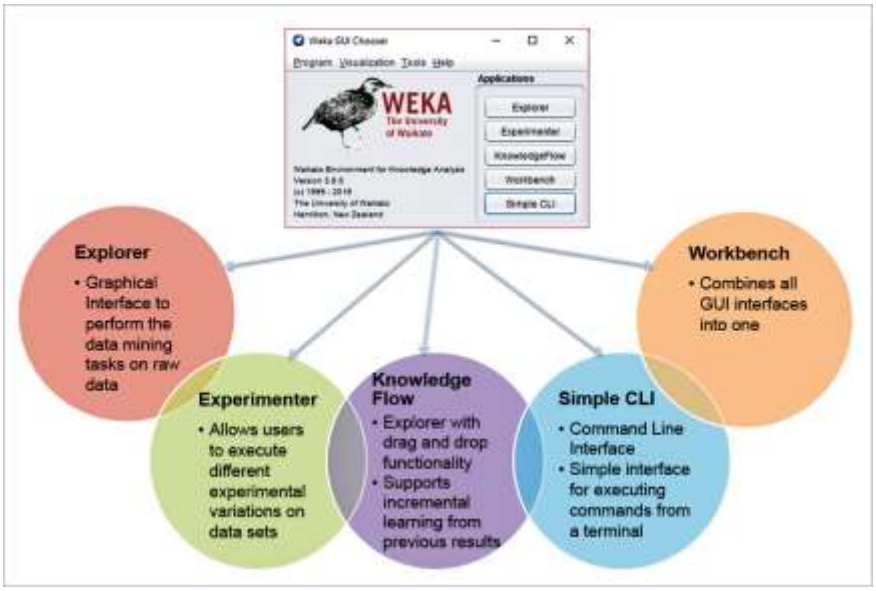

Figure 4: Weka Explorer

Weka supports different format for datasets i.e. CSV, ARFF etc. Different inbuilt classification algorithms are decision tree, support vector machines, instance based classifiers, regression etc. Clustering schemes includes k-means, X-means, farthest first etc. For rule mining it includes Apriori algorithm. In visualization it supports 1-D (one-dimensional), 2-D (two-dimensional) and 3-D (three-dimensional) views according to the need of user [9].

IV. CONCLUSION

Big data is the burning topic in the market due to its distinguishing properties that is volume, velocity and variety. For processing this data new techniques and tools are required. MATLAB, HADOOP & WEKA are used for distributed processing, computation and to analyze data for finding the hidden patterns. With the help of this review paper researchers will be able to understand the different tools for big data processing.

REFERENCES

1. https://www.dezyre.com/article/how-big-data-analysis-helped-increasewalmarts-sales-turnover/109,accessed on 31 March 2017.

2. https://home.cern/about/computing

3. https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/#1f6aaa441157, accessed on 31 March 2017

4. Book: “Metaheuristics for big data”, by Clarisse Dhaenens, Laetitia Jourdan. Ed. 1, 2016.

5. Satyanarayana Lenka Venkata, “A Survey on Challenges & Advantages in Big Data”, International Journal of Computer Science & Technology, Vol 6, Issue 2, June 2015.

6. https://www.xsnet.com/blog/bid/205405/the-v-s-of-big-data-velocity-volume-value-variety-and-veracity, accessed on 31 March 2017

7. Chang Shi etal. ,”Swarm intelligence in Big Data Analytics”, Springer, 2013.

8. Xindong Wu etal.,” Data Mining with Big Data”, IEEE Transaction on Knowledge & Data Engineering, 2013.

9. Weka: https://opensourceforu.com/2017/01/an-introduction-to-weka/accessed on 31 March 2017

10. Apache Hadoop: https://intellipaat.com/tutorial/hadoop-tutorial/introduction-hadoop/, accessed on 31 March 2017

11. MATLAB: http://cimss.ssec.wisc.edu/wxwise/class/aos340/spr00/whatismatlab.htm, accessed on 31 March 2017

12. Houcque David,” Introduction to MATLAB for Engineering Students”, PP No.-1-2, version 1.2, August, 2005