Research Journal of Computer and Information Technology Sciences E-ISSN 2320 – 6527
Vol. **4(2),** 1-4, February **(2016)** *Res. J. Computer and IT Sci.*

Big Data Analytics

Hemlata and Preeti Gulia
M.C.A. Department, M.D. University, Rohtak, Haryana, India
hemlatachahal@gmail.com

Available online at: www.isca.in, www.isca.me
Received 20th June 2015, revised 7th October 2015, accepted 29th January 2016

Abstract

Big data analytics refers to the method of analyzing huge volumes of data, or big data. The big data is collected from a large assortment of sources, such as social networks, videos, digital images, and sensors. The major aim of Big Data Analytics is to discover new patterns and relationships which might be invisible, and it can provide new insights about the users who created it. There are a number of tools available for mining of Big Data and Analysis of Big Data, both professional and non-professional. In this paper, we have summarised different big data analytic methods and tools.

Keywords: Big Data, Big Data Mining, R, Rapid-I Rapid Miner, KNIME.

Introduction

Big data means the datasets which cannot be recognized, obtained, managed, analyzed, and processed by present tools. Different definitions of big data have been given by different users of Big Data and different analysts of Big Data like research scholars, data analysts, and technical practitioners.

According to Apache Hadoop "Big data is a dataset which could not be captured, managed, and processed by general computers within an acceptable scope"[1].

Actually big data was defined in 2001 for the first time. Doug Laney, defined the 3Vs model, i.e., Volume, Variety and Velocity [2]. In spite of the fact that the 3Vs model was not used to define big data, Gartner and many other organizations, like IBM[3] and Microsoft[4] still uses the "3Vs" model to define big data[5]. In the "3Vs" model, Volume means, the dataset is so big and large that it is very difficult to analyze; Velocity means the data collected and gathered so rapidly to utilize it to the maximum; Variety shows different types of data like structured, semi-structured and unstructured data i.e. audio, video, webpage, and text. IDC (International data Corporation), one of the most dominant leaders in the research fields of Big Data, is of different view about Big Data. According to an IDC report of 2011 "Big Data technologies describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling the high-velocity capture, discovery, and/or analysis"[6]. According to this definition, big data characteristics can be: Volume (huge volume), Variety (various types and structure of data), Velocity (quick creation), and Value (great value but very low similarity).

This 4Vs definition draws light on the meaning of Big Data, i.e., examining the concealed values. The definition specifies the most crucial point of big data, i.e. to explore new values from datasets[7,8].

Big Data Processing Framework[2]

Big Data processing framework: META Group Research gave a three tier structure of "Big Data mining platform" (Tier I). Tier I emphasizes on low-level data accessing and computing. Tier II emphasizes on information sharing and privacy, and the domains and knowledge of Big Data application. Tier III emphasizes on mining algorithms.

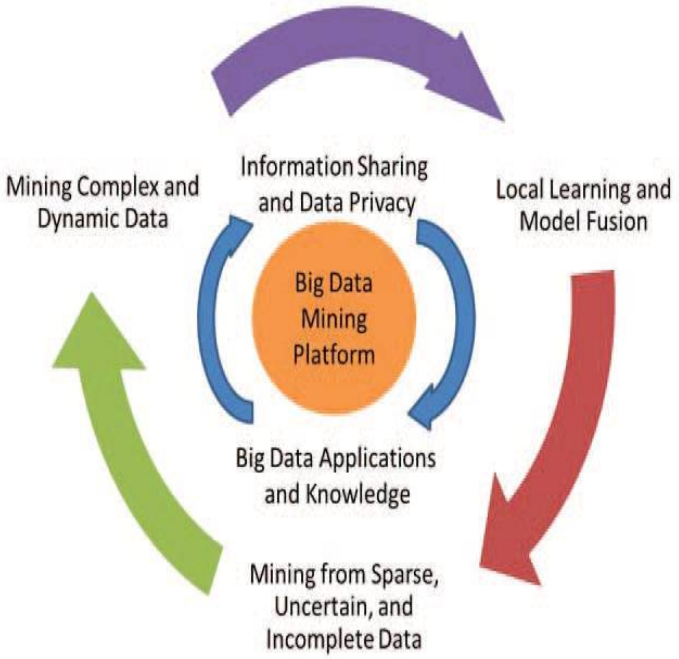

Figure-1
Big Data Framework Processing[2]

Big Data Analysis[1]

Big Data Analysis mainly involves analytical methods of big data, systematic architecture of big data, and big data mining and software for analysis. Data investigation is the most important step in big data, for exploring meaningful values, giving suggestions and decisions. Possible values can be explored by data analysis[7]. However, analysis of data is a wide area, which is dynamic and is very complex.

Traditional Data Analysis: Traditional data analysis means the proper use of statistical methods for huge data analysis, to explore and elaborate the hidden data of the complex dataset, so that value of data can be maximized. Data analysis guides different plans of development for a country, predicting demands of customers, and forecasting the trends of market for organisations. Big data analysis may be stated as a technique of analysis of a special data. So, most of the traditional methods are still used for big data analysis. Many traditional data analysis methods are represented here from statistics and computer science. Factor Analysis, Cluster Analysis, Correlation Analysis, Regression Analysis, A/B Testing, Statistical Analysis, Data Mining Algorithms.

Big Data Analytic Methods[1]

In the Big Data era, everybody wants to concentrate on extracting key value and information from the huge dataset to achieve objectives of their organisation. Now a days, the main methods of big data analysis used are:

Bloom Filter: Bloom Filter method is collection of Hash functions. Main concept of this method is that bit arrays are used to store data Hash values. Bit arrays are actually the bitmap index for the storage of lossy compression of Hash functions. Its advantages can be high space efficiency and high query speed. Its disadvantage is misidentifying values.

Hashing: Hashing method mutates data into smaller index and numeric values. Hashing has advantages like fast reading, writing, and querying speed, but it is very difficult to calculate a correct Hash function.

Index: Index is an efficacious method for cutting the disk reading cost and disk writing cost, and increasing the speed of query insertion, deletion, and modification. Disadvantage of this method is the extra cost of storage of index files.

Triel: A derived form of Hash Tree, is also called trie tree. This method is mostly used for fast retrieval. In this method, to improve efficiency of query, the common prefixes of strings of character are used to reduce comparison.

Parallel Computing: In contrast to the serial computing, parallel computing refers to utilisation of resources simultaneously to complete a task. The main idea behind this method is to fragment a problem and allocate them to different processes for achieving co processing. Some parallel computing models and low level tools are MPI (Message Passing Interface), Map Reduce, and Dryad. These low level tools are very difficult to use and learn. Some high level parallel computing tools are developed like Map Reduce uses Sawzall, Pig, and Hive, and Dryad uses Scope and Dryad LINQ.

Tools for Big Data Mining and Analysis[1]

Different commercial and open source software are available for Big Data Mining and Analysis. Five most frequently used software are:

R[1]: R is an open source environment. It is proposed for visualization, analysis and data mining. R is a collection of software facilities for[9], i. Reading and manipulating data, ii. Computation, iii. Conducting statistical analyses and iv. Displaying the results.

R is the next version of S language which was developed by AT&T Bell Labs for data extraction and statistical analysis. When complex tasks are processed, the module in C,C++ and Fortran can be called in R environment. We can also directly call objects of R in C. According to KDNuggets survey of 2012, R is more popular as compared to S. In a survey of "Design languages you have used for data mining/analysis in the past year" of 2012, it was on the top rank, above Java and SQL. After the success of R, Teradata and Oracle also launched the products which supported *R*.

Excel[1]: Excel of Microsoft Office, has robust data computing and statistical analysis capabilities. Some plug-ins like Analysis ToolPak and Solver Add-in are installed with Excel which have many capabilities of data analysis. Excel is a commercial software.

Rapid-I RapidMiner[1]: According to KDnuggets in 2011, Rapidminer is ranked at number 1 and also more frequently used as compared to R. R is open source software which is used for machine learning, data mining, and predictive analysis. It was developed in the University of Dortmund in 2001 and has been further maintained by Rapid-I GmbH. Data mining programs developed in RapidMiner follow the process of Extract, Transform and Load (ETL). Written in Java Rapid-Miner combines the WEkA's methods and implements them in R. The flow of process may be represented as a series of production of a factory in which data is considered as input and model as output. RapidMiner is a flexible analysis tool which bestow upon a large variety of methods like statistical analysis, correlation analysis, regression analysis, cluster analysis etc.[10]

KNIME[1]: KNIME (Konstanz Information Miner) is a open-source platform for data consolidation, data processing, analysis, and data mining[11]. KNIME creates data flows visually,

to execute the procedures, provides results and creating models and views.

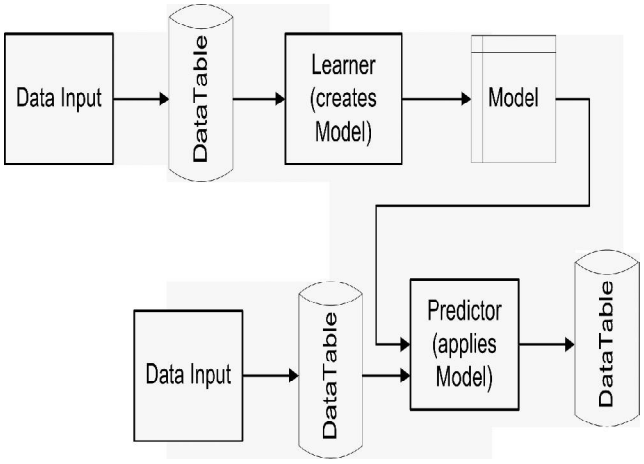

Figure-2
Flow of data in a Knime [12]

The three main principles of KNIME are[12]: i. *Visual and interactive framework*: Drag and drop option can be used for combining various data flows of a variety of processing units. A variety of application models can be achieved by data pipelines. ii *Modularity*: In order to enable easy distribution of computation and allow for independent development of different algorithms modularity should be followed.

Written in Java, KNIME provides many functionalities which can be used as plug-ins. Users can process different files, pictures by using plug-ins, and can apply into different open source environments, like R and Weka.

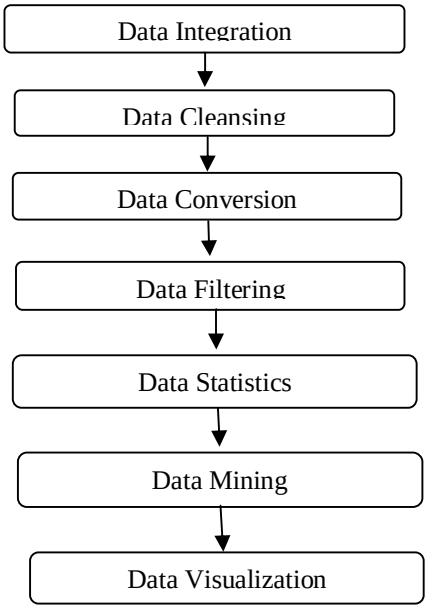

Figure-3
KNIME follows the following steps

This process is implemented in a envisioned environment. KNIME is a module-based architecture which can be expanded. Its processing units are not dependent on data containers. KNIME nodes and views can be expanded.

Weka/Pentaho[1]: Waikato Environment for Knowledge Analysis abbreviated as WEKA, is an open-source data mining software which written in Java. Weka allows capabilities like data processing, classification, regression, clustering, and visualization, etc. Pentaho is a popular open-source software for Business Intelligence. It has several tools for analysis, data integration, and data mining, etc.

Conclusion

Big Data Analytics is a hot research topic among the database researchers as well as the business community. However, currently we have different methods to analyse big data which we have mentioned in our paper but there is a lot of scope to create or invent new method of analytics. There are different tools and open source software available. Some of which we have mentioned briefly in the paper. There is a scope for the future research to compare the tools and find out the best in a particular situation by applying it. Also new can always be searched and invented. There are many more issues which can be further investigated like: Big data privacy and security, completeness, Data Quality etc.

References

1. Min Chen, Shiwen Mao and Yunhao Liu (2014). Big Data: A Survey, © Springer Science+Business Media New York 2014, published online: 22 january.

2. Laney D 3-d data management: controlling data Volume,velocity and variety. META Group Research Note, 6 February (2001)

3. Olaiya Folorunsho (2013). Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database. *International Journal of Advanced Research in Computer Science and Software Engineering* 3(3), March 2013 ISSN: 2277 128X.

4. Zikopoulos P and Eaton C et al (2011). Understanding big data: analyticsfor enterprise class hadoop and streaming data. McGraw-Hill Osborne Media(2011)

5. Beyer M, Gartner says solving big data challenge involves more than just managing volumes of data. Gartner. http://www.gartner.com/it/page.jsp.

6. O. R. Team Big data now: current perspectives from OReilly Radar. OReilly Media Gantz J, Reinsel D (2011) Extracting value from chaos. IDC iView, 1–12 (2011)

7. Mayer-Sch¨onberger V and Cukier K (2013). Big data:a revolution that will transform how we live, work, and think. Eamon Dolan/Houghton Mifflin Harcourt.

8. Duren Che, Mejdl Safran and Zhiyong Peng (2013). From Big Data to Big Data Mining: Challenges, Issues and Opportunities, © Springer-Verlag Berlin Heidelberg.

9. Petra Kuhnert and Bill Venables," An Introduction to R: Software for StatisticalModelling & Computing", CSIRO Mathematical and information Sciences Cleveland, Australia (2011)

10. Sebastian Land and Simon Fischer (2012). RapidMiner 5 RapidMiner in academic use 27th August.

11. Berthold MR, Cebron N, Dill F, Gabriel TR, K¨otter T, Meinl T, Ohl P, Sieb C, Thiel K and Wiswedel B (2008). KNIME: the Konstanz information miner". Springer.

12. Michael R. Berthold etal (2010). Knime: The Konstanz Information Miner Technical Report, Altana Chair for Bioinformatics and Information Mining.

13. Raymond Gardiner Goss and Kousikan Veeramuthu (2010). Heading Towards Big Data- Building A Better Data Warehouse For More Data, More Speed, And More Users.

14. Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding (2014). Data Mining with Big Data, *IEEE Transactions On Knowledge And Data Engineering*, 26(1).

15. Bharti Thakur and Manish Mann(2014). Data Mining for Big Data: A Review, *IJARCSSE*, 4(5).

16. Avita Katal et al (2013) Big Data: Issue, Challenge, Tools and Good Practices, IEEE.

17. Seref Sagiroglu and Duygu Sinanc, Big Data: A review, IEEE January (2013)

18. Zaiying Liu, Ping Yang and Lixiao Zhang (2013). A Sketch of Big Data Technologies IEEE Seventh International conference on Internet Computing for Engineering and Science.

19. Wei Fan, Albert Bifet (2012). Mining Big Data: Current Status, and Forecast to the Future, *SIGKDD Explorations*, 14(2).

20. Zikopoulos P, Eaton C et al. (2011). Understanding big data: analytics for enterprise class hadoop and streaming data" McGraw- Hill Osborne Media.

21. Mayer-Sch¨onberger V and Cukier K (2013). Big data: a revolution that will transform how we live, work, and think" Eamon Dolan/Houghton Mifflin Harcourt.

22. Albert Bifet "Mining Big Data in Real Time" (2010)

23. Meijer E (2011). The world according to linq. Communications of the ACM 54(10), 45–51.

24. Manyika J, McKinsey Global Institute, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C (2011). Byers AH Big data: the next frontier for innovation, competition and productivity. McKinsey Global Institute.