

A Review on Big Data Analytics : An Eminent Approach for Handling an Outsized Data

Sangeetha.T.G.R¹, Renuka Devi D²

PG Scholar, Department of Computer Science, Stella Maris College, Chennai, India¹

Assistant Professor, Department of Computer Science, Stella Maris College, Chennai, India²

ABSTRACT: The volatile increase of data volume and the growing demands of data mining have stimulated us into the era of big data. Many research scholars are drawn their desirability towards the research areas of big data mining, machine learning, computational intelligence and social networking. The big data technologies with conventional data mining approaches have posed many challenges in the field of big data analytics. These newest challenges are determined mainly on the issues related to the pre-processing of data, representation of data, data storage management, big social data mining and visualization. This paper reviews the concept of big data, technologies, architecture and applications. The open challenges and future directions are also stated at the end for the research community in this field of study.

KEYWORDS: Big data mining, visualization, social networking, big data analytics

I. INTRODUCTION

Big data analytics is the method of examining enormous datasets to discover the unseen patterns, unknown correlations and other useful information that can be used to make better decisions. Big data defines a knowledge used to record and execute the data set and it has the structured, semi structured and unstructured data that has to be mined for meaningful insights. On the other hand, mining through the high dimensional data from the search space is very challenging issue.

Manyika et al [1] describes Big Data as “datasets whose size is beyond the capacity of conventional database tools to capture, accumulate, control, and analyze”. Similarly, Davis and Patterson [4] states that “Big data is very large amount of data that cannot be handled and analyzed by conventional database protocols such as SQL”; and the same view is shared by [2, 5, and 8].

Big Data is identified by its characteristics velocity, volume, value and veracity. Big data streams are continuously evolving and changing, the traditional data mining algorithms infeasible for real time mining. Big Data have number of challenges that are related to its complexity which are: how to understand the unstructured data, capturing of important data, how to store, analyse and visualize the data. The characteristics of big data are shown in the figure 1.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

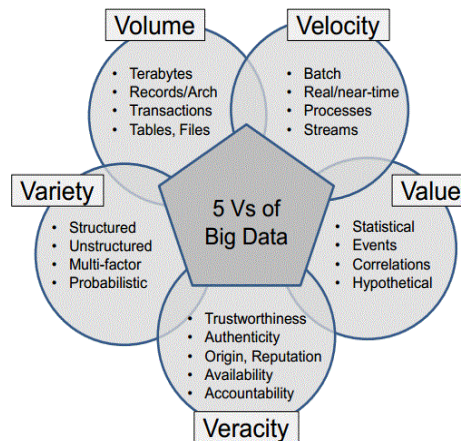


Fig- 1: 5 Vs of Big Data

Volume: Processing large amount of data is intended to gain meaningful insights from it. The sources of data are social networks data, health care data and biological data etc., the different big data sources are depicted in the figure 2.

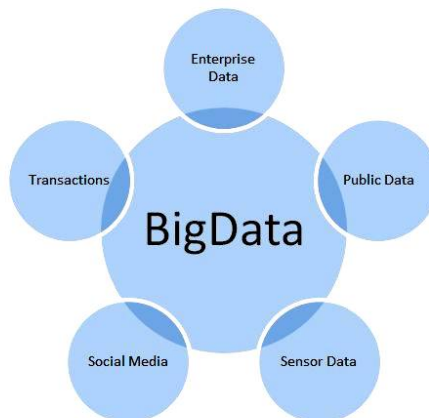


Fig-2: Sources of Big Data

Variety: These data are not in fixed structures which are enormous in size to be processed. In fact, such data can be very structured (relational databases), semi-structured (feeds, email, etc.) or unstructured (video, images, audio) [11]. This is shown the figure 3.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

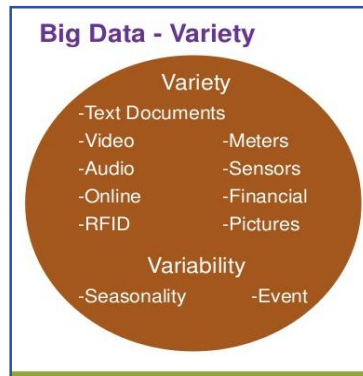


Fig -3 : Variety of Big Data

Velocity: This describes about the rate at which data streams are gathered for processing. The significance lies in the speed for taking data from input through to decision making [11].

Value: This characteristic is the objective of Big Data technology. It is divided into two categories: analytical use (decision making and discovering pattern) and business intelligence models [3,6].

Veracity: This is uncertainty that exists in data.

II. BIG DATA ARCHITECTURE

Fundamentally, data processing is defined as gathering, processing and management of data for producing novel insights [9]. The key issues addressed at this stage are storage, transportation and processing of enormous volume of data.

Big Data analysis divided into four fundamental stages: Acquisition of data, organization of data, analysis and visualization [14]. It is shown in the figure 4.

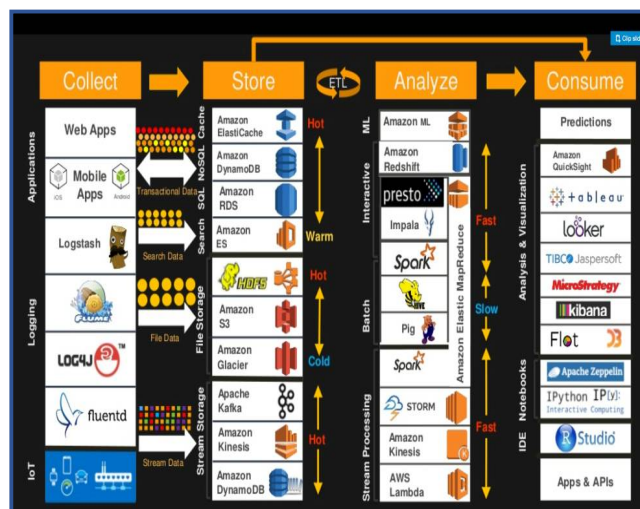


Figure 4. Big Data Architecture

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

Acquisition or collect:

The architecture processes very high streams of Big Data as it comes from variety of resources. Indeed, only the essential data with low extent of uncertainty is acquired for processing [15].

Storage:

Big data storage systems are generally a distributed architecture which addresses the issue of handling massive volume of data. This permits the augmented storage requirements by scaling out to new nodes with computational power and storage. New systems can flawlessly be added to a storage cluster and the data storage system takes the concern of distributing the data between entity nodes evidently [17].

Analyze:

Big data are insignificant until is processed and analyzed. Its impending value is considerable only when it is constructive for decision making. To facilitate such evidence-based decision making, organizations necessitate competent processes to turn the elevated volumes of rapid-moving and divergent data into evocative insights [16]. Some of the analytics techniques are text analytics, audio analytics, video analytics, social media analytics and predictive analytics [12].

Visualization:

Building an important decision is the vital objective of Big Data analysis and the realization of this purpose requires fine visualization techniques of Big Data. For this motive, there is an existent attention in the field of visualization [7,8]. Some of the visualizations techniques are Tag cloud, Clustergram, History Flow and Spatial Information flow [15].

III. BIG DATA TECHNOLOGIES

There are a variety of tools which can be used in Big Data organization from data collection to data analysis. It is depicted in the figure 5.

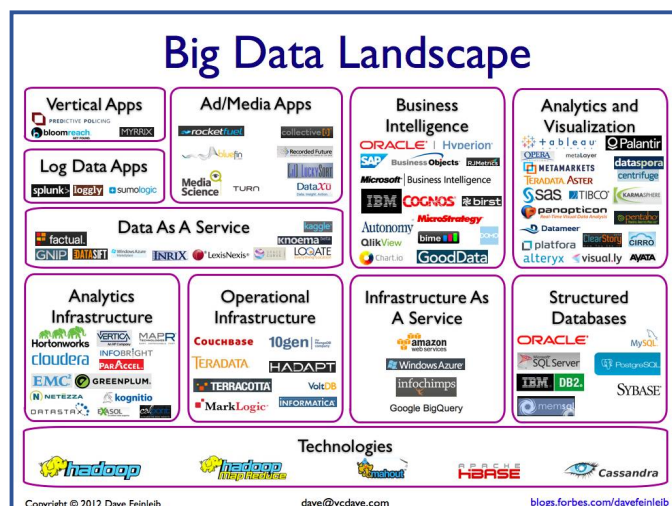


Fig -5: Big Data Technologies

1. Hadoop

It is most widely used frameworks today. It consists of tools and technologies including Pig, Hive, Flume and HDFS.

2. Spark

Spark is yet another processing domain for handling Big Data. The Hadoop ecosystem can contain the Spark processing engine in position of MapReduce, leading to accommodate all tools and technologies common for both eco systems.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

3. Flink

Apache Flink is especially designed for stream mining big data streams over distributed framework. Flink is efficiently handles both a batch and real-time processing and also provides a number of application programming interface for java and scala .The additional features, including: higher performance, continuous streaming of data models and handling fault tolerance.

4. Storm

Apache Storm is also an distributed framework for real-time computing. Storm is exclusively designed for processing big data streams, and independent of programming language. The features of this framework are fast, scalability, fault tolerance, reliability and ease of use.

5. Samza

Apache Samza is yet another processing framework for streaming data. Samza is built on Apache Kafka for messaging and YARN for cluster resource management. The feature of this framework includes simple application programming interface, durability, fault tolerance and scalability.

IV. BIG DATA APPLICATIONS

In recent times, big data and big data analysis has been proposed for describing datasets and advanced analytical methods for handling large-scale complex data sets [16]. Some of the industries driven by big data analytics are public sector services, healthcare, E- Learning, insurance, industry, transportation services, banking sectors and fraud detection[17].

Health care

The big data is in comprehensive use in the domain of medicine and healthcare. As there is ever growing technological advancement in the field of information technology, the cost of health care is also escalating further. Big data is an immense helping hand in this concern. It is an immense aid for even physicians to keep track of all the patients' electronic health care records (EHR) and intern leads to efficient decision making in treatment.

Public Sector

In the public sectors, the major confrontations are the combination and ability of the big data to address the issues of various public sector units and linked unions. Big data provides a great variety of services to the government sectors including the power, deception detection, health care, economy and ecological reinforcement.

Education

Big data has great impact in the education and learning. Along with the online learning, there are many examples of the use of big data in the education industry. Applications named as the Bubble Score allow teachers to conduct on line assessments tests through the mobile devices. Adaptive learning provides not only the grading development and data-driven approach for students learning practices. Enterprises construct digital courses that use analytics to locate the learning ability of the learner and what components of a lecture plan most exceptionally ensembles them at those situations.

Transportation

In recent times, infinite volumes of statistics from community networks and towering speediness statistics from telecoms have influenced journey policies a lot. Big data finds its application in public sector traffic management,

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

intellectual transportation and direction preparation etc., Private sector finds the big data applications in income administration, industrial improvements, and logistics and for rational profit. Personal use of the big data comprises direction forecasting, financial crisis, travel, pricing, product preferences etc.,

Fraud detection

Big data is vastly used in the fraud detection in the banking sectors. These technologies used to detect the abuse of credit cards, debit cards, archival of inspection tracks, credit risk treatment, business clarity, customer statistics revision, public analytics for business, IT action analytics, and IT strategy analytics.

V. OPEN CHALLENGES

The analysis of big data is confronted with numerous challenges. Significant research efforts are required to advance the competence of data collection, data storage, data analysis and visualization. As big data research is highly developed, novel issues on big data processing arise from the conventional data analysis, including [18]:

Reuse of data: Big data features are of huge value but with low density, with the increase of data scale, additional values may be mined from re-utilization of available data.

Data re-organization: Datasets gathered from different sources can be re-organized, with the total re-organized data values generously proportioned than the total datasets' value.

Unstructured information: Sources of unstructured data from the online activities of internet users poses an extremely great challenge in handling the data streams.

Security: In the big data era, safety and privacy are constantly two key concerns. As the data streams are growing more rapidly, there are added severe safety risks, while the conventional data protection methods are infeasible in handling such an enormous data.

The big data technology is still under rising stage. Many key technical issues, such as cloud computing, grid computing, stream computing, parallel computing, big data architecture, big data model, and software systems supporting big data, etc., should be fully investigated.

VI. CONCLUSION

In the big data era, big data technologies provide decision making through mining and analytical processing. With the emergence of IoT, expansion of mobile technology, and progress of data acquisition technology, individuals are not only the user and consumer of big data, but also its producer and participant. Many applications are directly related to big data will be ever more concerned and will definitely root the massive changes of social activities in the society. In this review paper, the concepts of big data, architecture, applications and potential challenges are presented which would give greater perceptiveness of Big Data technology.

REFERENCES

- [1] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H. Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011.
- [2] P. Zikopoulos, C. Eaton, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Education, 2011.
- [3] J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A.H. Byers, Big Data: The Next Frontier for Innovation, Competition, and Productivity, McKinsey Global Institute, 2011.
- [4] K. Davis, D. Patterson, Ethics of Big Data: Balancing Risk and Innovation, O'Reilly Media, 2012.
- [5] K. Krishnan, Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013.
- [6] A. Reeve, Managing Data in Motion: Data Integration Best Practice Techniques and Technologies, Morgan Kaufmann, 2013.

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 6, Special Issue 11, September 2017

- [7] K. Krishnan, Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013.
- [8] A. Reeve, Managing Data in Motion: Data Integration Best Practice Techniques and Technologies, Morgan Kaufmann, 2013.
- [9] K. Krishnan, Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013.
- [10] X. Dong, D. Srivastava, Big data integration, in: Data Engineering (ICDE), 2013 IEEE 29th International Conference on, 2013, pp. 1245–1248.
- [11] I. O'Reilly Media, Big Data Now: 2014 Edition, O'Reilly Media, 2014.
- [12] Chen, Min, et al. Big data: related technologies, challenges and future prospects. Heidelberg: Springer, 2014.
- [13] Chen, Min, et al. "Big data applications." Big Data. Springer International Publishing, 2014. 59-79.
- [14] Emani, Cheikh Kacfeh, Nadine Cullot, and Christophe Nicolle. "Understandable big data: A survey." Computer science review 17 (2015): 70-81.
- [15] H.V. Jagadish, D. Agrawal, P. Bernstein, E. e. a. Bertino, Challenges and Opportunities with Big Data, The Community Research Association, 2015.
- [16] Gandomi, Amir, and Murtaza Haider. "Beyond the hype: Big data concepts, methods, and analytics." International Journal of Information Management 35.2 (2015): 137-144.
- [17] Strohbach, Martin, et al. "Big data storage." New Horizons for a Data-Driven Economy. Springer International Publishing, 2016. 119-141.
- [18] <https://intellipaat.com/blog/7-big-data-examples-application-of-big-data-in-real-life>