A Comprehensive Study on Characteristics of Big Data and the Platform Used in Big Data

Anusha Medavaka
Software Programmer
Seven Hills IT Solutions LLC, NJ

Abstract— Because of extensive use of several computer tools such as smartphones, laptops, wearable computing devices; the data handling online has actually gone beyond greater than the modern-day computer systems can manage. Because of this high development price, the term Big Data is imagined. Nonetheless, the rapid development price of huge data will certainly create countless difficulties, such as data inconsistency and also incompleteness, scalability, timeliness, and also security. This paper offers a quick intro to the Big data technology as well as its significance in the modern globe. This paper addresses numerous obstacles as well as concerns that require to be stressed to provide the complete impact of big data. The devices utilized in Big data technology are additionally gone over carefully. This paper additionally goes over the qualities of Big data as well as the platform made use of in Big Data i.e. Hadoop.

Keywords: Big Data, Hadoop, MapReduce

I. INTRODUCTION

Big Data has actually obtained much interest from the last couple of years in the IT sector. As we can locate billions of individuals are attached to the internet worldwide, creating a big quantity of data at the quick price. The generation of this huge quantity of engenders numerous difficulties. In Addition To Big Data's substantial advantages to several companies, the obstacles, as well as concerns, need to additionally be brought right into the light. A projection from International Data Corporation (IDC), the Big Data technology and also solutions market stands for a fast-growing multi-billion- buck around the world possibility. Actually, the current IDC projection is revealing that the Big Data technology and also solutions market will certainly expand at a 26.4% substance yearly development price to $41.5 billion with 2019, or regarding 6 times the development price of the general infotech market. Furthermore, by 2020 IDC thinks that the line of company purchasers will certainly aid drive analytics past its historic wonderful area of relational (efficiency administration) to the double-digit development prices of genuine- time knowledge as well as exploration/discovery of the disorganized globes. [1].

Fig. 1: Growth rate of Big Data from 2011-2017[2]

II. BIG DATA OVERVIEW

Big Data is a mix of big datasets that cannot be refined utilizing standard computer methods. It is not a strategy that can be serviced it's very own or alone; instead, it includes numerous locations of organization and also technology. The residential properties of indicating Big Data are volume, Variety, Velocity, Variability as well as Complexity as displayed in number 2 [3].

Sr. No.	Properties	Description
1.	Volume	Many factors contribute towards increasing Volume streaming data , live streaming data and data collected from sensors etc.,
2.	Variety	Data comes in all types of formats-from traditional databases ,text documents, emails, video, audio, transactions etc.,
3.	Velocity	This means how fast the data is being produced and how fast the data needs to be processed to meet the demand.
4.	Variability	Along with the Velocity, the data flows can be highly inconsistent with periodic peaks.
5.	Complexity	Complexity of the data also needs to be considered when the data is coming from multiple sources. The data must be linked, matched, cleansed and transformed into required formats before actual processing.

Fig. 2: Properties of Big Data

Big data involves the data produced by different devices and applications. Some of the sources of Big Data are shown in the figure.

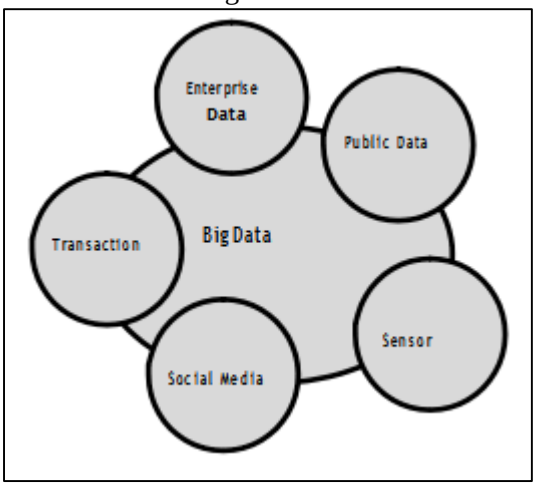

Fig. 3: Some Sources of Big Data

III. PHASES IN BIG DATA PROCESSING

Prior to handling Big data, it needs to be tape-recorded from different data creating resources. After tape-recording, it needs to be a filtering system as well as pressed. Just the pertinent data must be tape-recorded using filters that dispose

of worthless info. In order to promote this job, specialized devices are utilized such as ETL. ETL devices stand for the ways in which data really obtains filled right into the storehouse. The number 3 shows various phases while doing so.

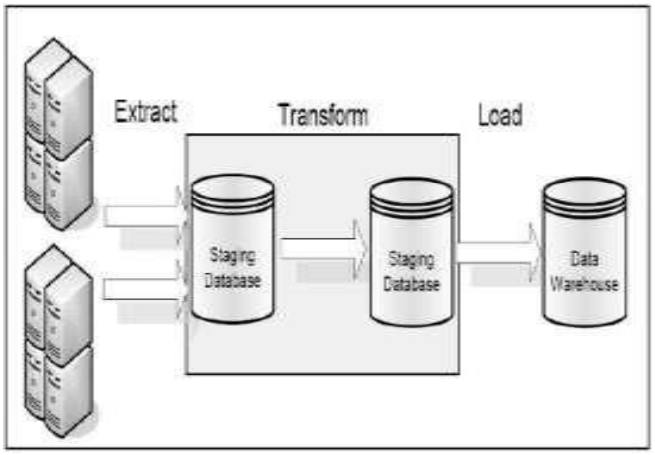

Fig. 4: ETL process [4]

Sr. No	Phase	Description of the Phase
1.	Extraction	In this phase relevant information is extracted. To make this phase efficient, only the data source that has been changed since recent last ETL process is considered.
2.	Transformation	Data is transformed through various phases [10] The phases are 1. Data analysis; 2. Definition of transformation workflow and mapping rules; 3.Verification; 4. Transformation; and 5. Backflow of cleaned data.
3.	Loading	At the last, after the data is in the required format, it is then loaded into the data warehouse.

Table 1: Various phases in ETL [5]

IV. BIG DATA CHALLENGES

Big data because of its numerous residential properties like volume, velocity, variety, variability, worth and also complexity advanced lots of difficulties. Figure 5 programs numerous difficulties in big data [12] Figure 6 checklist a few of the obstacles in Big data in addition to its influence as well as dangers entailed. [6]

Fig. 5: Challenges in Big data

Challenge	Impact	Risk
Uncertainty of the market landscape	Difficulty in choosing technology components Vendor lock-in	Committing to failing product or failing vendor
Big data talent gap	Steep learning curve Extended time for design, development, and implementation	Delayed time to value
Big data loading	Increased cycle time for analytical platform data population	Inability to actualize the program due to unmanageable data latencies
Synchronization	Data that is inconsistent or out of date	Flawed decisions based on flawed data
Big data accessibility	Increased complexity in syndicating data to end-user discovery tools	Inability to appropriately satisfy the growing community of data consumers

Fig. 6: Challenges, its impact and risk involved in Big data

V. TECHNIQUES FOR BIG DATA HANDLING

There are numerous methods readily available for data administration. That consists of Google Big Table, Straightforward DB, Not Just SQL (NoSQL), Data Stream Administration System (DSMS), Mem cache DB, and also Voldemort [7] Yet these typical techniques are just relevant to conventional data and also not Big data as it can not be kept on a solitary maker. The Big Data taking care of strategies and also devices consist of Hadoop, MapReduce, as well as Big Table. Out of these, Hadoop is among one of the most commonly utilized modern technologies.

A. Hadoop

HDFS (Storage space layer): - Hadoop has actually a dispersed Data System called HDFS, which represents Hadoop Distributed File System. It is a Data System utilized to keep huge documents with streaming data accessibility patterns, working on collections on asset equipment. [8] There are 2 sorts of nodes in the HDFS cluster, specifically name node as well as data nodes. The name node will certainly keep the documents system tree and also the metadata for all the data as well as directory sites in the tree. The data node shops as well as obtain blocks based on the directions of customers or the name node. The data gotten is reported back to the name node with listings of blocks that they are keeping. Without the name node, it is not feasible to access the data. So, it ends up being really essential to make name node resistant to failing. [11]

B. MapReduce Processing/Computation layer

It is the standard of a program which is indicated for handling applications on numerous dispersed web servers. In Map Lower divide and also dominate technique is utilized to damage the huge complicated data right into tiny devices as well as refine them. It will certainly review the data from HDFS. Nonetheless, it can review the data from various other locations consisting of installed neighborhood data systems, the internet, as well as data sources. It separates the calculations in between various web servers or nodes. It is additionally fault-tolerant. If any individual of the node stops working, Hadoop can comprehend just how to proceed with the calculation by reassigning the insufficient job to one more

node as well as tidying up after the node that can not finish its job. It additionally understands just how to integrate the outcomes of the calculation in one location. [9] The various other core parts in Hadoop design consists of Hadoop THREAD, it is a structure for work organizing as well as cluster source monitoring. The various other element is the cluster which is the collection of host machines (nodes).

Hadoop is an Apache open resource structure which is composed in java. High quantities of data, in any kind of framework, are refined by Hadoop. Hadoop enables dispersed storage space and also dispersed handling for large data collections. The major elements of Hadoop are:

1) Hadoop distributed file system (HDFS).
2) MapReduce.

The style of Hadoop has received the figure 7. Hadoop has 3 layers. Both significant layers are MapReduce as well as HDFS.

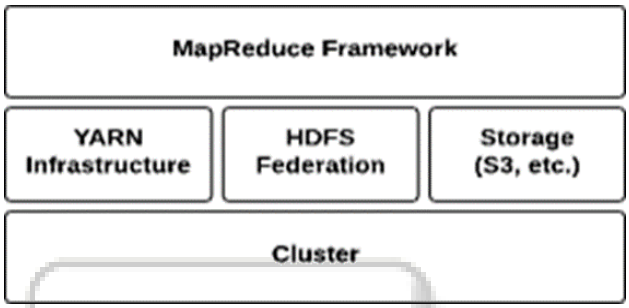

Fig. 7: Hadoop Architecture

VI. Conclusion

As there are significant quantities of data that are created daily, so such plus size of data it comes to be extremely tough to attain reliable handling making use of the existing typical methods Big data is data that surpasses the handling capability of standard data source systems. In this paper, basic ideas regarding Big Data exist. These principles consist of Big Data features, difficulties and also methods for dealing with big data.

References

[1] Golfarelli, M., & Rizzi, S. (2009). Data warehouse design: modern principles and methodologies. Columbus: McGraw-Hill
[2] Almeida, F., and Calistru, C, "The Main Challenges and Issues of Big Data Management", International Journal of Research Studies in Computing, 2(1), 2013, pp. 11-20.
[3] M. Chen, S. Mao, and Y. Liu, "Big data: a survey," Mobile Networks and Applications, vol. 19, no. 2, pp. 171–209, 2014
[4] Amrit pal, Pinki Aggrawal, "A Performance Analysis of Map Reduce Task with Large Number of Files Dataset in Big Data using Hadoop" Forth International Conference on Communication Systems and Network Technologies, 2014.
[5] Rahm, E., & Hai Do, H. (2000). Data cleaning: problems and current approaches. Bulletin of the Technical Committee on Data Engineering, 23(4), 3-13.).
[6] Apache Hadoop (2013). HDFS Architecture Guide [Online]. Available: https://hadoop. apache.org/docs/r1.2.1/hdfs_design.ht
[7] Anusha Medavaka, P. Shireesha, "Analysis and Usage of Spam Detection Method in Mail Filtering System" in "International Journal of Information Technology and Management", Vol. 12, Issue No. 1, February-2017 [ISSN : 2249-4510]
[8] Anusha Medavaka, P. Shireesha, "Review on Secure Routing Protocols in MANETs" in "International Journal of Information Technology and Management", Vol. VIII, Issue No. XII, May-2015 [ISSN : 2249-4510]
[9] Anusha Medavaka, P. Shireesha, "Classification Techniques for Improving Efficiency and Effectiveness of Hierarchical Clustering for the Given Data Set" in "International Journal of Information Technology and Management", Vol. X, Issue No. XV, May-2016 [ISSN : 2249-4510]
[10] Anusha Medavaka , P. Shireesha, "Optimal framework to Wireless Rechargeable Sensor Network based Joint Spatial of the Mobile Node" in "Journal of Advances in Science and Technology", Vol. XI, Issue No. XXII, May-2016 [ISSN : 2230-9659]
[11] Anusha Medavaka,"Enhanced Classification Framework on Social Networks" in "Journal of Advances in Science and Technology", Vol. IX, Issue No. XIX, May-2015 [ISSN : 2230-9659]
[12] Anusha Medavaka, P. Shireesha, "A Survey on TraffiCop Android Application" in "Journal of Advances in Science and Technology", Vol. 14, Issue No. 2, September-2017 [ISSN : 2230-9659]
[13] Anusha Medavaka, Dr. P. Niranjan, P. Shireesha, "USER SPECIFIC SEARCH HISTORIES AND ORGANIZING PROBLEMS" in "International Journal of Advanced Computer Technology (IJACT)", Vol. 3, Issue No. 6 [ISSN : 2319-7900]
[14] Anusha Medavaka, "Monitoring and Controlling Local Area Network Using Android APP" in "International Journal of Research", Vol. 7, Issue No. IV, April-2018 [ISSN : 2236-6124]
[15] Yeshwanth Rao Bhandayker, "AN OVERVIEW OF THEINTEGRATION OF ALL DATA MINING AT CLOUD-COMPUTING" in "Airo International Research Journal", Volume XVI, June 2018 [ISSN : 2320-3714]
[16] Siripuri Kiran, 'Decision Tree Analysis Tool with the Design Approach of Probability Density Function towards Uncertain Data Classification', International Journal of Scientific Research in Science and Technology(IJSRST), Print ISSN : 2395-6011, Online ISSN : 2395-602X,Volume 4 Issue 2, pp.829-831, January-February 2018. URL : http://ijsrst.com/IJSRST1841198
[17] Yeshwanth Rao Bhandayker , "Artificial Intelligence and Big Data for Computer Cyber Security Systems" in "Journal of Advances in Science and Technology", Vol. 12, Issue No. 24, November-2016 [ISSN : 2230-9659]
[18] Sugandhi Maheshwaram, "A Comprehensive Review on the Implementation of Big Data Solutions" in "International Journal of Information Technology and Management", Vol. XI, Issue No. XVII, November-2016 [ISSN : 2249-4510]
[19] Ajmera Rajesh, Siripuri Kiran, " Anomaly Detection Using Data Mining Techniques in Social Networking" in

"International Journal for Research in Applied Science and Engineering Technology", Volume-6, Issue-II, February 2018, 1268-1272 [ISSN : 2321-9653], www.ijraset.com

[20] Sugandhi Maheshwaram , "An Overview of Open Research Issues in Big Data Analytics" in "Journal of Advances in Science and Technology", Vol. 14, Issue No. 2, September-2017 [ISSN : 2230-9659]

[21] Siripuri Kiran, Ajmera Rajesh, "A Study on Mining Top Utility Itemsets In A Single Phase" in "International Journal for Science and Advance Research in Technology (IJSART)", Volume-4, Issue-2, February-2018, 637-642, [ISSN(ONLINE): 2395-1052]

[22] Yeshwanth Rao Bhandayker, "Security Mechanisms for Providing Security to the Network" in "International Journal of Information Technology and Management", Vol. 12, Issue No. 1, February-2017, [ISSN : 2249-4510]

[23] Sugandhi Maheshwaram, S. Shoban Babu , "An Overview towards the Techniques of Data Mining" in "RESEARCH REVIEW International Journal of Multidisciplinary", Volume-04, Issue-02, February-2019 [ISSN : 2455-3085]

[24] Yeshwanth Rao Bhandayker , "A Study on the Research Challenges and Trends of Cloud Computing" in "RESEARCH REVIEW International Journal of Multidisciplinary ", Volume-04, Issue-02, February-2019 [ISSN : 2455-3085]

[25] Sriramoju Ajay Babu, Dr. S. Shoban Babu, "Improving Quality of Content Based Image Retrieval with Graph Based Ranking" in "International Journal of Research and Applications", Volume 1, Issue 1, Jan-Mar 2014 [ISSN : 2349-0020]

[26] Dr. Shoban Babu Sriramoju, Ramesh Gadde, "A Ranking Model Framework for Multiple Vertical Search Domains" in "International Journal of Research and Applications" Vol 1, Issue 1,Jan-Mar 2014 [ISSN : 2349-0020].

[27] Mounika Reddy, Avula Deepak, Ekkati Kalyani Dharavath, Kranthi Gande, Shoban Sriramoju, "Risk-Aware Response Answer for Mitigating Painter Routing Attacks" in "International Journal of Information Technology and Management", Volume VI, Issue I, Feb 2014 [ISSN : 2249-4510]

[28] Sugandhi Maheshwaram, "A Review on Deep Convolutional Neural Network and its Applications" in "International Journal of Advanced Research in Computer and Communication Engineering", Vol. 8, Issue No. 2, February-2019 [ISSN : 2278-1021], DOI 10.17148/IJARCCE.2019.8230

[29] Yeshwanth Rao Bhandayker. "An Overview : Security Solutions for Cloud Environment." International Journal for Scientific Research and Development 7.2 (2019): 1596-1598.

[30] Yeshwanth Rao Bhandayker. "An Overiew Of Cyber Security", International Journal of Research, vol. 8, Issue. 3 (2019): 2236-6124.

[31] Sugandhi Maheshwaram, "A Study On The Challenges In Handling Big Data", International Journal of Research, vol. 8, Issue. 3 (2019): 2236-6124.

