Volume 6, Issue 6, June 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Big Data: An Overview

[1]Anusha B. Dhakite, [2]Prof. Nitin V. Wankhade
[1] Research Scholar, [2] Assistant Professor,
[1, 2] P. G. Department of Computer Science and Technology, DCPE, (Autonomous College), HVPM, Amravati, Maharashtra, India

Abstract-Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data duration, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. Accuracy in big data may lead to more confident decision making. And better decisions can mean greater operational efficiency, cost reductions and reduced risk. Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, and combat crime and so on." Scientists, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, connectomics, complex physics simulations, and biological and environmental research.

Keywords: Big Data, Map Reduce, Hadoop, Architecture, Big Data Analytics.

I. INTRODUCTION

Data sets grow in size in part because they are increasingly being gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing),software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte (2.5×1018) of data were created; The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.Work with big data is necessarily uncommon; most analysis is of "PC size" data, on a desktop PC or notebook that can handle the available data set. Relational databasemanagement systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make Big Data a moving target. Thus, what is considered to be "Big" in one year will become ordinary in later years. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

A.What is Big Data?

Big data is a buzzword, catch-phrase, used to describe a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques.Big data is typically described by the first three characteristics. The term big data is believed to have originated with Web search companies who had to query very large distributed aggregations of loosely-structured[23] data. Big data analytics requires capturing and processing data where it resides. This paper explores the value of data at the edge of networks, where some of ―biggest‖ big data is generated. As the use of sensors and devices as well as intelligent systems [4] [5] [6] continues to expand, the potential to gain insight from the flood of data from these sources becomes a new and compelling opportunity.

Fig: Big Data

Businesses that can harness the power of big data at the edge and unlock its value to the organization will outperform their competitors with greater capabilities to innovate creatively and solve complex problems whose solutions have been out of reach in the past. Below-sometimes referred to as the three Vs. However, organizations [6] [7] [12] need a fourth—value—to make big data work.

II. CHARACTERISTICS

Big data can be described by the following characteristics:

A. Volume – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.

B.Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analysing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.

C.Velocity - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.

D.Variability - This is a factor which can be a problem for those who analyse the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

E.Veracity - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

F.Complexity - Data management can become a very complex process, especially when large volumes of data come from multiple sources. Thesedata need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data.

III. ARCHITECTURE

In 2004, Google published a paper on a process called MapReduce that used such architecture. The MapReduce framework provides a parallel processing model and associated implementation to process huge amounts of data. With MapReduce, queries are split and distributed across parallel nodes and processed in parallel (the Map step). The results are then gathered and delivered (the Reduce step). The framework was very successful, so others wanted to replicate the algorithm. Therefore, an implementation of the MapReduce framework was adopted by an Apache open source project named Hadoop.MIKE2.0 is an open approach to information management that acknowledges the need for revisions due to big data implications in an article titled "Big Data Solution Offering". The methodology addresses handling big data in terms of useful permutations of data sources, complexity in interrelationships, and difficulty in deleting (or modifying) individual records.Recent studies show that the use of multiple layer architecture is an option for dealing with big data. The Distributed Parallel architecture distributes data across multiple processing units and parallel processing units provide data much faster, by improving processing speeds. This type of architecture inserts data into a parallel DBMS, which implements the use of MapReduce and Hadoop frameworks. This type of framework looks to make the processing power transparent to the end user by using a front end application server.Big Data Analytics for Manufacturing Applications can be based on 5C architecture (connection, conversion, cyber, cognition, and configuration). Big Data Lake - With the changing face of business and IT sector, capturing and storage of data has emerged into a sophisticated system. The big data lake allows an organization to shift its focus from centralized control to a shared model to respond to the changing dynamics of information management. This includes the storage, servers, and network as the base, inexpensive commodities of the big data stack. This stack can be bare metal or virtual (cloud). The distributed file systems are part of this layer. Platform as a Service (PaaS): The NoSQL data stores and distributed caches that logically queried using query languages form the platform layer of big data. This layer provides the logical model for the raw, unstructured data stored in the files. Data as a Service (DaaS): Theentire array of tools available for integrating with the PaaS layer using search engines, integration adapters, batch programs, and so on in this layer. Big Data Business Functions as a Service (BFaaS): Specific industries—like health, retail, ecommerce, energy, and banking—can build packaged applications that serve a specific business need and leverage the DaaS layer for cross-cutting data functions.

IV. TECHNOLOGIES

A 2011 McKinsey Global Institute reportecosystem of big data as follows:

Techniques for analysing data, such as A/B testing, machine learning and natural language processingBig Data technologies, like business intelligence, cloud computing and databasesVisualization, such as charts, graphs and other displays of the dataMultidimensional big data can also be represented as tensors, which can be more efficiently handled by tensor-based computation, such as multilinear subspace learning. Additional technologies being applied to big data include massively parallel-processing (MPP) databases, search-based applications, data mining,[44] distributed file systems, distributed databases, cloud-based infrastructure (applications, storage and computing resources) and the Internet.[citation needed]Some but not all MPP relational databases have the ability to store and manage petabytes of data. Implicit is the ability to load, monitor, back up, and optimize the use of the large data tables in the RDBMS.DARPA's Topological Data Analysis program seeks the fundamental structure of massive data sets and in 2008 the technology went public with the launch of a company called Ayasdi.The practitioners of big data analytics processes

are generally hostile to slower shared storage,[47] preferring direct-attached storage (DAS) in its various forms from solid state drive (Ssd) to high capacity SATA disk buried inside parallel processing nodes. The perception of shared storage architectures—Storage area network (SAN) and Network-attached storage (NAS) —is that they are relatively slow, complex, and expensive. These qualities are not consistent with big data analytics systems that thrive on system performance, commodity infrastructure, and low cost.Real or near-real time information delivery is one of the defining characteristics of big data analytics. Latency is therefore avoided whenever and wherever possible. Data in memory is good—data on spinning disk at the other end of a FC SAN connection is *not*. The cost of a SAN at the scale needed for analytics applications is very much higher than other storage techniques.

V. MANUFACTURING

Based on TCS 2013 Global Trend Study, improvements in supply planning and product quality provide the greatest benefit of big data for manufacturing. Big data provides an infrastructure for transparency in manufacturing industry, which is the ability to unravel uncertainties such as inconsistent component performance and availability. Predictive manufacturing as an applicable approach toward near-zero downtime and transparency requires vast amount of data and advanced prediction tools for a systematic process of data into useful information. A conceptual framework of predictive manufacturing begins with data acquisition where different type of sensory data is available to acquire such as acoustics, vibration, pressure, current, voltage and controller data. Vast amount of sensory data in addition to historical data construct the big data in manufacturing. The generated big data acts as the input into predictive tools and preventive strategies such as Prognostics and Health Management (PHM).

VI. TYPES AND SOURCES OF BIG DATA

Executives need to be cognizant of the types of data they need to deal with. There are three main types of data, regardless of whether or not a company is using big data – unstructured data, structured data, and semi structured data. Unstructured data are data in the format in which they were collected; no formatting is used (Coronel, Morris, & Rob, 2013). Some examples of unstructured data are PDF's, e-mails, and documents (Baltzan, 2012). Structured data are formatted to allow storage, use, and generation of information (Coronel, Morris, & Rob, 2013). Traditional transactional databases store structured data (Manyika et al., 2011). Semistructured data have been processed to some extent (Coronel, Morris, & Rob, 2013). XML or HTML-tagged texts are examples of semistructured data (Manyika et al., 2011). Business executives with traditional database management systems need to broaden their data horizons to include collection, storage, and processing of unstructed and semistructured dataData collection of unstructured and semistructured data is done through several internet-based technologies. Chui, Löffler, and Roberts (2010) describe sensors providing big data as being part of the Internet of Things. The Internet of Things is described as sensors and actuators that are embedded in physical objects that provide data through wired and wireless networks (Chui, Löffler, & Roberts, 2010). Some industries that are creating and using big data are those that have recently begun digitization of their data content; these industries include entertainment, healthcare, life sciences, video surveillance, transportation, logistics, retail, utilities, and telecommunications (Chui, Löffler, & Roberts, 2010). Devices generating data in these industries include IPTV cameras, GPS transceiver, RFID tag readers, smart meters, and cell phones (Chui, Löffler, & Roberts, 2010).

VII. BIG DATA ANALYTICS

Big data analytics [5] refers to the process of collecting, organizing and analysing large sets of data ("big data") to discover patterns and other useful information. Not only will big data analytics help you to understand the information contained within the data, but it will also help identify the data that is most important to the business and future business decisions. Big data analysts basically want the knowledge that comes from analysing the data.

A. An Example of Big Data?

(The Apache Hadoop Framework and MapReduce) New technologies are emerging to make big data analytics possible and cost-effective [31]. The Apache Hadoop* framework is evolving as the best new approach. The Hadoop framework redefines the way data is managed and analysed by leveraging the power of a distributed grid of computing resources.The Hadoop open-source framework [5] [6] [7] [21] uses a simple programming model to enable distributed processing of large data sets on clusters of computers. The complete technology stack includes common utilities, a distributed file system, analytics and data storage platforms, and an application layer that manages distributed processing, parallel computation, workflow, and configuration management. In addition to offering high availability, the Hadoop framework is more cost- effective for handling large, complex, or unstructured data sets than conventional approaches, and it offers massive scalability and speed.

B.Hadoop, The Open Source heart of Big Data Analytics:

According to Forrester, Hadoop is the nucleus of the next generation enterprise data warehousing by delivering cloud-facing architectures. Created by Doug Cutting, the creator of Apache Lucene, Hadoop provides a comprehensive toolset for building distributed systems, including data storage, data analysis and coordination. Hadoop originates from Apache Nutch, an open source web search engine. After realizing that existing architectures would not scale to the billions of pages on the web, the initiators wrote an open source implementation based on Google's distributed file system, called Nutch Distributed Filesystem (NDFS). In 2004 Google released a paper that introduced MapReduce, a parallel

programming model and an associated implementation for processing, analyzing and generating large data sets across a cluster of commodity machines (Dean &Ghemawat,2008), to the public. Nearly a year later all Nutch algorithms ported to use MapReduce and NDFS. In 2006, Nutch became a separate subproject under the name Hadoop and two years later it became a top-level project at Apache, confirming its success. In that year, Hadoop used by many international organizations such asLast.fm and Facebook. For many, Hadoop is a synonym for big data because of its powers to store and handle huge amounts of (unstructured) data within a smaller time frame in an economically responsible way. Soothe Hadoop ecosystems play a major role in big data analytics. Figure 2illustrates the "mountain of data" commonly finds within organizations. With traditional data analytics, only the peak analyzed and utilized to create value or support value creation. This peak often consists of highly structured data stored in traditional data warehouses. Since the amount of unstructured data is growing rapidly as described earlier, this peak is becoming relatively smaller. With Hadoop, it is possible to store and analyze unstructured data in a much smaller time frame using the power of distributed and parallel computing on commodity hardware.

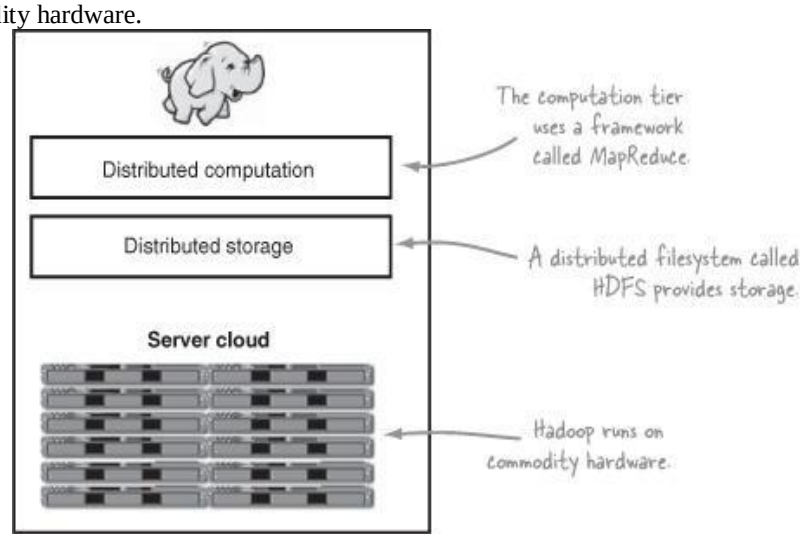

Fig:Hadoop Overview

C. MapReduce:

MapReduce is a framework for efficiently processing the analysis of big data on a large number of servers. It was developed for the back end of Google's search engine to enable a large number of commodity servers to efficiently process the analysis of huge numbers of webpages collected from all over the world. Apache [8] [13] developed a project to implement MapReduce, which was published as open source software (OSS), this enabled many organizations, such as businesses and universities, to tackle big data analysis.erms of efficiency, productivity, revenue, and profitability. The Age of Big Data is here, and these are truly revolutionary times if both business and technology professionals continue to work.

VIII. GOALS AND CHALLENGES OF ANALYSING BIG DATA

Two main goals of high-dimensional data analysis are to develop effective methods that can accurately predict the future observations and at the same time to gain insight into the relationship between the features and response for scientific purposes. Furthermore, due to large sample size, Big Data give rise to two additional goals: to understand heterogeneity and commonality across different subpopulations.In other words, Big Data give promises for: (i) exploring the hidden structures of each subpopulation of the data, which is traditionally not feasible and might even be treated as 'outliers' when the samplesize is small; (ii) extracting important common features across many subpopulations even when there are large individual variations. Together and deliver on the promise.

IX. APPLICATIONS

Big data has increased the demand of information management specialists in that Software AG, Oracle Corporation, IBM, Microsoft, SAP, EMC, HP and Dell have spent more than $15 billion on software firms specializing in data management and analytics. In 2010, this industry was worth more than $100 billion and was growing at almost 10 parent a year: about twice as fast as the software business as a whole.Developed economies make increasing use of data-intensive technologies. There are 4.6 billion mobile-phone subscriptions worldwide and between 1 billion and 2 billion people accessing the internet Between 1990 and 2005, more than 1 billion people worldwide entered the middle class which means more and more people who gain money will become more literate which in turn leads to information growth. The world's effective capacity to exchange information through telecommunication networks was 281 petabytes in 1986, 471 petabytes in 1993, 2.2 exabytes in 2000, and 65 exabytes in 2007and it is predicted that the amount of traffic flowing over the internet will reach 667 exabytes annually by 2014. It is estimated that one third of the globally stored information is in the form of alphanumeric text and still image data, which is the format most useful for most big data applications. This also shows the potential of yet unused data (i.e. in the form of video and audio content).While many vendors offer off-the-shelf solutions for Big Data, experts recommend the development of in-house solutions custom-tailored to solve the company's problem at hand if the company has sufficient technical capabilities.

X. CONCLUSION

Throughout this paper we have discussed Big Data. Accuracy is one of the best feature in big data that may lead to more confident decision making. The technologies like Big data and cloud based analytics are the cost effective, faster and better solutions for the large organizations. Various new service models are discovered with the use of Big data. One of its platforms Hadoop is a highly scalable storage platform, which can store and distribute very large data sets across hundreds of low cost servers that operate in parallel. Hadoop enables businesses to run applications onthousands of nodes involving thousands of terabytes of data. The data is replicated to other node when the data is sent which means that in the event of failure, there is another copy available for use. We found that the Big data is best suited for big businesses as it deals with very large sets of information from disparate sources.

REFERENCES

[1] Dr. Siddaraju1, Sowmya C L2, Rashmi K3, Rahul M4,*Efficient Analysis of Big Data Using Map Reduce Framework*,International Journal of Recent Development in Engineering and Technology,Volume 2, Issue 6, June 2014,(ISSN 2347-6435(Online).

[2] Prashant Kumar b,KhushbooPandeya, *Big Data and Distributed Data Mining: An Example of Future Networks*, International Journal of Advance Research and Innovation, Volume 1, Issue 2 (2013) 36-39, ISSN 2347 – 3258.

[3] Munesh Kataria1, Ms.Pooja Mittal2, *BIG DATA: A Review*, International Journal of Computer Science and Mobile Computing, Vol.3 Issue.7, July- 2014, pg. 106-110, ISSN 2320–088X.

[4] Bernice Purcell, *The emergence of "big data" technology and analytics*, Journal of Technology Research, Holy Family University.

[5] G. Aloisioa,b, S. Fiorea,b, Ian Fosterc, D. Williamsd,Scientific big data analytics challenges at large scale.

[6] https://en.wikipedia.org/wiki/Big_data accessed on dated 13/03/2016.

[7] Baltzan, P. (2012). Business driven information systems, (3rd ed.). New York: McGraw-Hill.

[8] Business & Finance Week Editors. (2012, 12 May). "Data analytics: 34 percent of mid-market businesses using business intelligence are planning to adopt big data analytics; Lack of expertise among SMBs is main barrier." Business & Finance Week.

[9] Chui, M., Löffler, M., & Roberts, R. (2010, March). "The Internet of things."McKinsey Quarterly. Retrieved from https://www.mckinseyquarterly.com /The_Internet_of_Things_2538

[10] Coronel, C., Morris, S., & Rob, P. (2013). Database Systems: Design, Implementation, and Management, (10th Ed.). Boston: Cengage Learning.

