A Study on Big Data Tools and Applications

Gurwinder Singh[1], Dr. Anil Sharma[2]

[1]Research Scholar, [2]Associate Professor,

School of Computer Applications, Lovely Professional University,Phagwara, Punjab, India

Abstract- Big Data being a very rousing area of research has appealed a lot of responsiveness from industry, academia and government too. Big Data is group of huge and multifaceted data sets that are hard to process using traditional tools of data processing. Trillions of data is being generated on planet every day and this data is coming from various sources like Web, Social sites, technical experimentations, mobile talks, sensors etc… The organizations, companies, enterprises and business are using Big Data and that is why it has lot of applications areas such as agriculture, banking, science, data mining, cloud computing, marketing, health management. To handle such a huge amount of data and vast applications lot of tools like MapReduce, Hadoop etc. are available. This paper gives an explanation of some tools and applications of Big Data.

Keywords- Big Data, Big Data Tools, Big data Applications, Hadoop.

I. INTRODUCTION

Big Data means large amount of data that may be structured, semi-structured and unstructured having prospective to be mined for generating information. There are no clear limits on big data in terms of quantity, but the term is used frequently when talking about the petabytes and Exabyte of data. Big Data Analytics is the process of observing huge data sets that consisting of varied data types that can be named as Big Data Analytics. These investigative outcomes can result in better marketing, improved customer service, fresh revenue opportunities, better operational efficiency, viable advantages over competing organizations and various other business profits. Primary aim of big data analytics is to assist companies to make profitable business decisions with help of data scientists, predictive modeler's and other analytics professionals who can analyse huge amount of data including other different forms of data that may be untouched by more traditional Business Intelligence programs[1].The term Big Data coined in the first decade of the 21st century, and the first organizations to clinch it were start-up and online firms. Big Data consists of three V's: Volume, Velocity and Variety. Volume refers to the large amount of data as company's repositories have significant growth from megabytes to petabytes. Velocity refers to the speed at which data is generating and processing. Variety refers to different types of data i.e. structured, un-structured and semi-structured [2].

II. TOOLS FOR BIG DATA

Numbers of tools are available in market to process big data. This section introduces some popular tools used for analysing Big Data.

i) Hadoop: Hadoop is an open-source distributed file system which is capable of storing and processing large volumes of data in parallel using commodity hardware. Google and Yahoo introduced Hadoop because they need a cost effective mode to build search indexes. At present, many companies are implementing Hadoop software from Apache as well as third-party providers such as Cloud era, Horton works, EMC, and IBM. Two main components of Hadoop are HDFS and MapReduce. HDFS is used to store huge amount of data while MapReduce is used to process this data [1].

ii) Windows Azure HDInsight: It is used to deploy and provision clusters of Apache Hadoop cloud. It provides a software framework which intends to report, analyse and manage big data. It also provides the required software framework consisting HDFS, Pig, Hive, MapReduce and Sqoop in a scalable and cost-efficient environment. Azure Blob Storage is used as default file system by Windows Azure [3].

iii) NoSQL: NoSQL (Not Only SQL) is used to handle unstructured data. Unstructured data is stored in NoSQL databases with no schema in particular. A variety of NoSQL databases are available which falls in categories: Key-value data stores, Document stores, Wide-column NoSQL databases and Graph databases. Scalability, Global and High availability, and Flexible data modelling are key features of NoSQL [4].

iv) Hive: It is a data warehousing software which primarily aims at how data is structured and queried in distributed Hadoop clusters. It also provides ETL operation tools and SQL-like abilities to the environment. Hive can be used to develop applications for Hadoop environment, but it is not able to handle real-time queries, row-level updates and OLTP workloads. Various components of Hive includes: HCatalog, WebHCat and HiveQL. Queries in Hive can run from the Hive shell, JDBC, or ODBC [5].

v) Sqoop: This tool is used to connect Hadoop or Hive with different relational databases for purpose of structure data transfer. It is a command-line interface used to transfer bulk of data to relational databases and additional structured data stores from Hadoop. It also substitutes the necessity to develop scripts to import and export data [6].

vi) PolyBase: It is used to access the data stored in Parallel Data Warehouse (PDW) and it works on top of SQL Server 2012 Parallel Data Warehouse (PDW) and. This PDW is used for data warehousing to process large amount of relational data and to access non-relational data it provides integration with Hadoop [7].

vii) Big data in EXCEL: The data stored in Hadoop can also be accesses with help of Microsoft EXCEL 2013 using Hortonworks Enterprise Apache Hadoop. To perform this operation the Power View feature of EXCEL 2013 can be used to summarise data easily [8].

viii) Presto: It is an open source distributed SQL query engine used to run the interactive analytic queries for varying size data sources. Facebook is using this Presto to query different internal data stores and approximately more than one thousand employees of Facebook uses Presto daily to run over thirty thousand queries per day. It quickly retrieves data and doesn't depend on MapReduce technique [9].

III. APPLICATIONS OF BIG DATA

a) Enterprise: The in-database analytics are used and is of great benefit to many industries around the globe. Data can provide efficient and quicker perceptions when it need not go back and forth to work which results in better decisions instantaneously in a smaller amount of expenditure than customary data analysis tools for business people [10].

b) Big Data Analytics: Applications of Big Data Analytics are used to analyse big data using huge framework for parallel processing. Before deploying the applications in extensive cloud environment developers these applications usually prepare them using a lesser data in a pseudo-cloud environment [11].It is a novel group of software applications that effect extensive data which is characteristically enormous to appropriately fit in one hard drive or the memory, to expose information for parallel-processing infrastructures [11][12].

c) Clustering: Grounded on precise dimensions of data users can spontaneously catch sets within data through clustering algorithm using simple plug and click dialog. Clustering makes it simple to recognize and report masses by text documents, products, client type, user click path, customer behavior, purchasing patterns, etc [13].

d) Data Mining: Data mining in big data means the ability to extract valuable data from huge datasets using suitable technique. Using Datameer's decision trees users can straightforwardly know that what groups of data elements produce the required. The hidden structure of data can be reflected using decision tree Decision trees clarify the degree of associations and dependencies within data in addition to define what common characteristics effect the consequences such as risk of disease, fraud, purchases, log data etc. [14].

e) Banking: Using client data invariably raises the issue of privacy by exposing the hidden links between apparently isolated sections of data. Due to this privacy issue about 62% bankers are vigilant in using big data due to privacy issues. Further, data analysis outsourcing or distributing customer data across departments for the generation of wealthier perceptions also amplifies security risks [15].

f) SAP: SAP consists of a various technologies that talk about use-cases and necessities of Big Data. Technologies of SAP span the specified scale of Big Data defined by International Data Corporation such as 100TB+ data sets, small but rapidly developing data sets, real-time data sets with streaming access, utilize complex event processing, and varied format data sets. The technologies of SAP are able to deal with data at rest as well as data in motion and can also be categorized on basis of memory and disk [16].

g) Credit Cards: To identify fake transactions, Credit card companies hang on two attributes speed and accuracy of in-database analytics. Before permitting apprehensive activities to cardholders they flag uncommon amounts, locations, and retailers from the data stored worth years [16].

h) In Telecom: Telecom industry is having adequate amount of data, means that they are sitting on gold mine. Telecom industry is using big data to adopt the methodology for introducing new product, Improve customer experiences, Predicting and planning more accurately capacity of network to fulfil fast demands and moderate customer churn. To get clear and complete image of their operations and their customers, big data provides telecom operators a real chance and to promote their business objectives. Real-time prognostic analytics can support leveraging the data that exist in their multitude systems, make it immediately accessible and comparing data to generate vision that can help them drive their industry forward [17][18].

i) Consumer Goods: Data mined from surveys, web logs, reviews about products, purchases, telephonic conversations with customer, raw text selected from internet helps manufacturers of products to predict preference and purchasing behavior of consumer. By collecting data about all being assumed and talked in public about the products and extracting meaning from it makes company to understand failure and success of products. This understanding can be used to set the trends that can assist in making right decisions about marketing the product [18].

j) In Health care: Due to unreliable returns several health care investors have fewer awards in IT, because of limited ability to homogenize and associate data. As there are many players in health care industry, there is no mode to straightforwardly share data among diverse providers or facilities due to privacy concerns. Now Stakeholders of health care have access to new threads of knowledge and this info is a form of "big data," with complexity, diversity, and timelines. Researchers can apply data mining on the data to check about the type of more effective treatment for specific disorders, side effects of drugs, and to know additional vital evidence that can help patients and moderate costs. Modern technology in the industry have improved its ability to work with such data, even though the files are huge and have dissimilar database structures and technical characteristics [19][20][21].

k) Finance & Economy: Presence of big data is growing in financial industry in various ways like: Observing and surveillance of employee, in predictive models for loan decisions, financial markets prognosis and illiquid assets pricing. For evaluation of fresh credit requests the major financial companies use third party credit scoring. Banking sector, for existing customers, now days implements their personal credit score analysis with an extensive collection of data, including checking, credit cards, mortgages, savings, and investment data [22].

IV. CONCLUSION

The focus of paper is to put some light on the tools and applications of Big Data. Big data is used in number of applications like banking, agriculture, data mining, finance, marketing, health care and many more. Brief introduction of few tools such as Hadoop, Sqoop, SAP, NoSQL, Hive, Azure HDInsight and PolyBase used for storage and processing of big data is presented.

V. REFERENCES

[1]. H. S. Bhosale and D. P. Gadekar, "A Review Paper on Big Data and Hadoop," *Int. J. Sci. Res. Publ.*, vol. 4, no. 10, pp. 1–7, 2014.

[2]. C. L. Philip, Q. Chen and C. Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on big data, Information Sciences, 275 (2014), pp.314-347.

[3]. http://www.jamesserra.com/archive/2014/02/what-is-h dinsight/

[4]. http://basho.com/resources/nosql-databases

[5]. Y. Huai *et al.*, "Major technical advancements in apache hive," in *Proceedings of the 2014 ACM SIGMOD international conference on Management of data - SIGMOD '14*, 2014, pp. 1235–1246.

[6]. http://www.tutorialspoint.com/sqoop/sqoop_pdf_versio n.htm

[7]. D. J. DeWitt *et al.*, "Split query processing in polybase," in *Proceedings of the 2013 international conference on Management of data - SIGMOD '13*, 2013, pp. 1255–1266.

[8]. A. Ara and A. Ara, "Cloud for Big Data Analytics Trends," *IOSR J. Comput. Eng.*, vol. 18, no. 5, pp. 01–06, 2016.

[9]. D. Plase, "A Systematic Review of SQL-on-Hadoop by Using Compact Data Formats," *Balt. J. Mod. Comput.*, vol. 5, no. 2, pp. 233–250, 2017.

[10]. Thomas H. Davenport, Jill Dyche, "Big Data in Big Companies," in International Institute for Analytics May 2013.

[11]. N. Wingfield, "Virtual product, real profits: Players spend on zynga's games, but quality turns some off," Wall Street Journal.

[12]. Brandon Bunker, Senior Director of Customer Analytics and Intelligence, Vivint, "Big Data Insights Platform for Rapid Data Discovery," 2013.

[13]. D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker, "Interactions with big data analytics," interactions, vol. 19, no. 3, pp. 50–59, May 2012

[14]. Weiyi Shangy, Zhen Ming Jiangy, Hadi Hemmatiy, Bram Adamsz, Ahmed E. Hassany, Patrick Martinx, "Assisting Developers of BigData Analytics Applications When Deploying on Hadoop Clouds" Database Systems Laboratory, School of Computing, Queen's University, Kingston, Canada.

[15]. By Steve Lucas, Executive Vice President and General Manager, Database and Technology, SAP, "Big Data Analytics Guide 2012.

[16]. "Big Data: Trends, Strategies, and SAP Technology" by Carl W.Olofson, Dan Vesset August 2012.

[17]. Formerly Booz & company, "Benefiting from big data: A new approach for the telecom industry," in 2013.

[18]. Ari Banerjee senior analyst, heavy reading, "Big data and advanced analytics in Telecom: A Multi-Billion-Dollar Revenue Opportunity," December 2013.

[19]. "Big data is the future of Healthcare" by Cognizant 20-20 insights September 2012.

[20]. "Data-driven healthcare organizations use big data analytics for big gains" by IBM software.

[21]. Peter Groves, Basel Kayyali, David Knott, Steve Van Kuiken, "The big data revolution in health care," enter for US Health System Reform Business Technology Office, published in January 2013.

[22]. "Deep learning applications and challenges in big data analytics" by Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar,