A Study of Big Data Practices in Various Open Source Tools

Sampathrajan S
Assistant Professor, Department of Computer Science,
Shanmuga Industries Arts and Science College, Tiruvannamalai.
sampathrajan79@gmail.com

Abstract: Big data is defined as large amount of data which requires new technologies and architectures so that it becomes possible to extract value from it by capturing and analysis process. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques. Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. Since Big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues associated in bringing and adapting to this technology are brought into light.

Keywords: Big data, Big data Analysis, Hadoop Architecture, Open source tool.

I. INTRODUCTION

Big Data today influences our lives in the most unexpected ways, and organisations are using it extensively to gain that competitive edge in the market[1,2]. So let’s get acquainted with the open source tools that help us to handle Big Data.

Gone are the days when banks used to store customer information (such as names, photographs and specimen signatures) in individual postcard-like data sheets. That was an era where thick registers were used in different government offices like post offices, property tax collection centres, etc, to store customers’ details or maintain the daily attendance records of employees. If an employee had to update any of the registered customer’s details, the task could take up the whole day. Hours were wasted searching for that particular customer’s details and then creating a new record to replace the old one[3]. The customers, too, had to wait for hours for such minor tasks to be completed. Apart from the tediousness of searching for data from piles of ledgers, such paper files could be lost at any time due to disasters like floods or fire, apart from the degradation of the very paper on which the data was recorded[4].

Today, on our journey towards a digital India, all government offices are switching to digitisation instead of manual record keeping. As we proceed along this path, we see a tremendous increase in the size of data. There are around 230 billion tweets posted on Twitter per day, 2.7 billion Likes and comments added to Facebook daily, and around 60 hours of video uploaded to YouTube every minute[5]. All this leads to about 2.5 exabytes of data being generated on a daily basis by different online applications, transactional data sources, IoT devices, etc. The term that encapsulates such immense volumes of information is Big Data. Existing hardware and software systems are unable to handle such volumes of different types of data being created at such enormous speed. The data has also become too complex and dynamic to be stored, processed, analysed and managed with traditional data tools[6]. So Big Data is now analysed by computer systems to reveal specific trends, patterns and associations, especially those relating to human behaviour and interactions. We are making our machines smart enough to 'think and decide' on what action needs to be performed and when by using Big Data techniques like predictive analytics, user-behaviour analytics, etc.

II. BIG DATA IN COMPANIES

1. The techniques listed later in this article help to extract useful insights from large data sets, which are leveraged further for different surveys, statistics and case studies[7].

2. Flipkart, Amazon and other such online e-commerce sites make use of these techniques to study the behaviour of their customers and help them get what they want.

3. Facebook supposedly knows more about each one of us than our own therapists do and it's possible only because of the different Big Data techniques it implements. It continuously keeps track of different user actions, photo uploads, comments, shares, etc. using these techniques.

4. MNCs like Walmart make use of Big Data to improve their 'employee intelligence quotient' and 'customer emotional intelligence quotient'.

5. Family restaurants like Dominos, McDonald's and KFC use predictive and user-behaviour analytics to increase the efficiency of their marketing and continuously improve the customer experience.

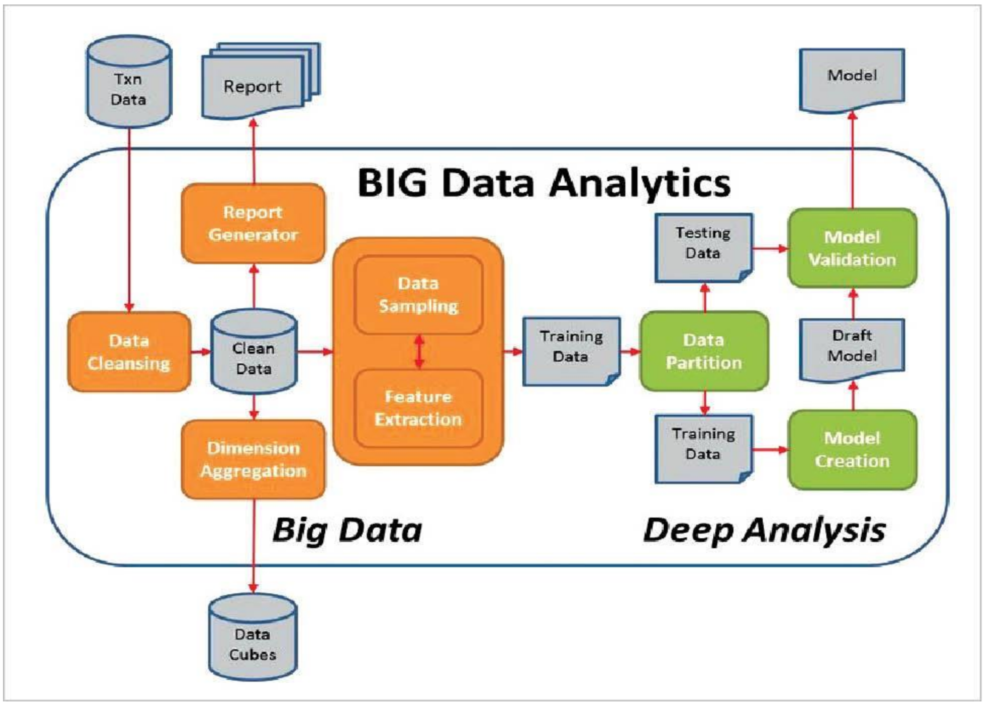

Figure 1. Architecture for Big Data analysis

III. FACTORS INFLUENCE WITH BIG DATA

The different factors that an analyst must consider while working with Big Data.

A. Ingestion:

This process is about moving data (especially unstructured data) from where it originated, into a system where it can be stored and analysed. Data ingestion can be continuous or asynchronous, in real-time or batched, or even both[8].

B. Harmonisation:

This deals with the improvement of data quality and its use with the help of different machine learning capabilities. It also interprets the characteristics of the data and the different actions taken on it, subsequently using that analysis to improve data quality.

C. Analysis:

This deals with the analysis of the data sets in order to understand the behaviour of data and identify specific patterns or trends, so that different actions can be performed on that data set[9].

D. Visualisation:

Data visualisation is the process of presenting the data in a pictorial or graphical format. This helps decision makers to grasp difficult concepts or identify new patterns in the data set.

E. Democratisation:

This is the ability for specific information in a digital format to be accessible to the end user. This is used to enable non-specialists to gather and analyse larger data sets without requiring outside help.

IV. TECHNIQUES FOR BIG DATA ANALYSIS

The popular techniques that are being used for analyzing large data sets. All of them generate useful insights that can be used further for diverse applications.

i. **A/B testing:** This is a method of comparing the two versions of an application to determine which one performs better. It is also called split testing or bucket testing. It refers to a specific type of randomised experiment, under which a group of users is presented with two variations of some product (an email, advertisement or Web page)—Variation A and Variation B. The users exposed to Variation A are referred to as the control group, because their performance is considered as the baseline against which any improvement in the performance observed from presenting the Variation B is measured. Variation A sometimes acts as the original version of the product being tested against what existed before the test. The users exposed to Variation B are called the treatment group[10].

ii. **Association rule learning:** Association rule learning is rule-based machine learning used to discover interesting relationships between variables in large databases. It uses a set of techniques for discovering the interesting relationships, also called 'association rules', among different variables present in large databases [11]. All these techniques use a variety of algorithms to generate and test different possible rules. One of its common applications is market basket analysis. This enables a retailer to determine the products frequently bought together and hence use that information for marketing (for example, the discovery that many supermarket shoppers who buy diapers also tend to buy beer). Association rules are being used today in Web usage mining, continuous production, intrusion detection and bioinformatics. These rules do not consider the order of different items either within the same transaction or across different transactions [12, 15, 28].

iii.**Natural language processing:** This is a field of computational linguistics and artificial intelligence concerned with the interactions between computers and human languages. It is used to program computers to process large natural language corpora. The major challenges involved in natural language processing (NLP) are natural language generation (frequently from machine-readable logical forms), natural language understanding, connecting the language and machine perception, or some combination thereof [13, 26]. NLP research has relied mostly on machine learning. Initially, many language-processing tasks involved direct hand coding of the rules, which is not suited to natural language variation. Machine-learning pattern calls are now being used instead of statistical inferences to automatically learn different rules through the analysis of large different real-life examples. Many different classes of machine learning algorithms have been applied to NLP tasks. These algorithms take large sets of 'features' as input. These features are generated from the input data [14, 27].

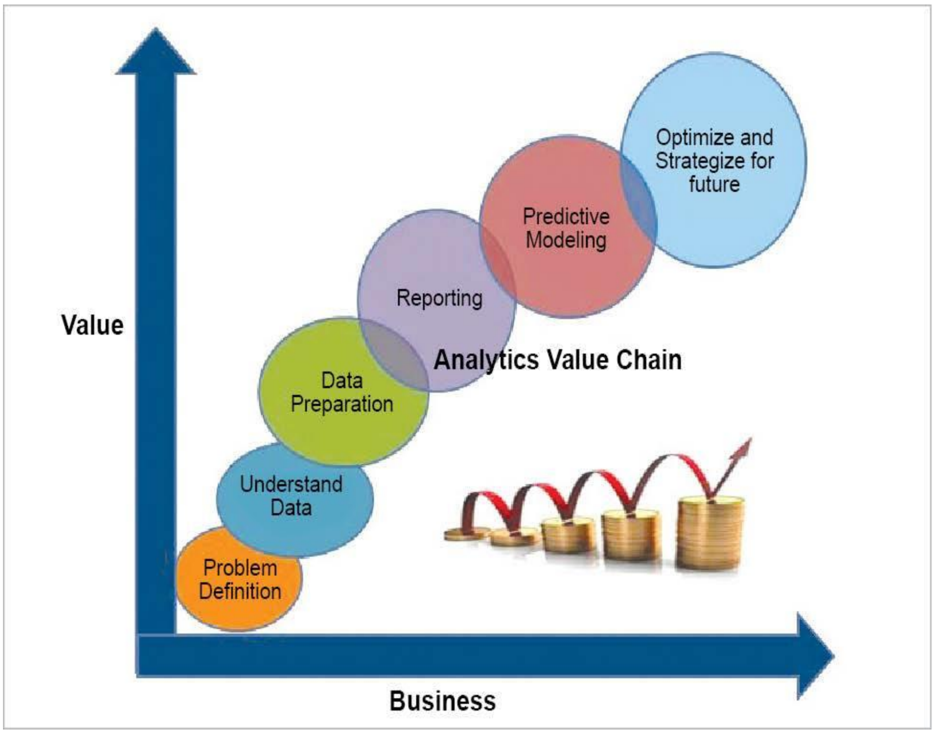

Figure2. Business Value increases using Big data Analysis

V. BIG DATA OPENSOURCE TOOLS

Big data is processed and analyzed with different techniques. The open source tools that can be used to handle Big Data in order to get some significant value from it.

A. Apache Hadoop

Apache Hadoop is an open source software framework used for the distributed storage and processing of large data sets using the MapReduce programming model. It consists of computer clusters built using commodity hardware [15, 25]. All the different modules in Hadoop are actually designed with the assumption that different hardware failures are commonly observed occurrences and they should be automatically handled by the framework.

Features:

- The Hadoop framework is mostly written in Java, with some of its native code in C. Its command line utilities are written as shell scripts.
- Apache Hadoop consists of a large storage part, known as the Hadoop Distributed File System.
- It uses the MapReduce programming model to process large data sets.
- Hadoop splits different files into large blocks and then distributes them across various nodes in a cluster.
- It transfers packaged code into different nodes to process the data in parallel.
- Apache Hadoop makes use of the data locality approach, where nodes manipulate all the data they have access to. This allows the large dataset to be processed faster and even more efficiently.
- The base Apache Hadoop framework is composed of different modules: Hadoop Common, HDFS, Hadoop YARN and Hadoop MapReduce.

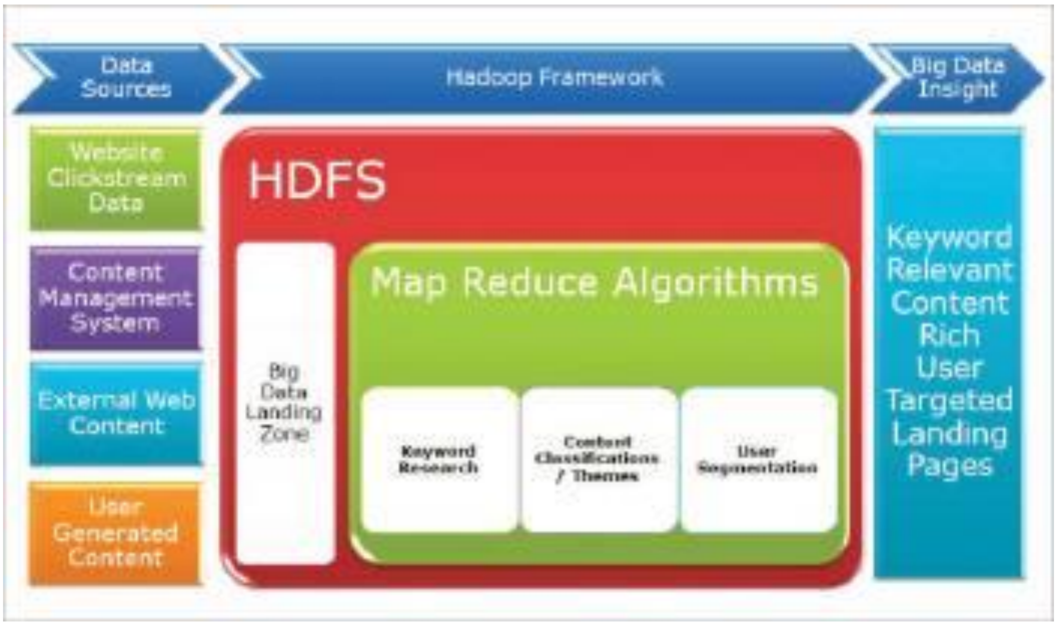

Figure 3. Big Data Hadoop architecture

B. Cassandra

This is an open source distributed NoSQL database management system. It's designed to handle large amounts of data across many different commodity servers, hence providing high availability with no single point of failure [16, 24]. It offers strong support for clusters that span various data centres, with its asynchronous master less replication allowing low latency operations for all clients.

Features:

- It supports replication and multiple data centre replication.
- It has immense scalability.
- It is fault-tolerant.
- It is decentralised.
- It has tunable consistency.
- It provides MapReduce support.
- It supports Cassandra Query Language (CQL) as an alternative to the Structured Query Language (SQL).

C. KNIME

Also called Konstanz Information Miner, this is an open source data analytics, integration and reporting platform. It integrates different components for data mining and machine learning through its modular data pipelining concept [17, 21]. A graphical user interface allows the assembly of nodes for data pre-processing (which includes extraction, transformation and loading), data modelling, visualisation and data analysis. Since 2006, it has been widely used in pharmaceutical research, but now it is also used in areas like customer data analysis in CRM, financial data analysis and business intelligence [22, 23].

Features:

- KNIME is written using Java and is based on Eclipse. It makes use of its extension capability to add plugins, hence providing additional functionality.
- The core version of KNIME includes modules for data integration, data transformation as well as the commonly used methods for data visualisation and analysis.
- It allows users to create data flows and selectively execute some or all of them.
- It allows us to inspect the models, results and interactive views of the flow.
- KNIME workflows can also be used as data sets to create report templates, which can be exported to different document formats like doc, PPT, etc.
- KNIME's core architecture allows the processing of large data volumes which are only limited by the available hard disk space.
- Additional plugins allow the integration of different methods for image mining, text mining as well as time series analysis.

D. Rapid Miner

This is basically a data science software platform. It is used for business and commercial applications as well as for education, research, rapid prototyping, training and application development. It supports all the steps of the machine learning process including data preparation, model validation, results visualisation and optimization [18]. It has been developed on an open core model. It provides a graphical user interface to design and execute different analytical workflows. All such workflows are called 'processes' and they consist of multiple 'operators'. Each of these operators perform a single task within their respective processes, and the output of each operator forms the input for the next one. Also, the engine can be called from other programs or can be used as an API[19,20].

Features:

- Uses a client or server model with the server offered either on premise, or in private or public cloud infrastructures.
- Ninety-nine per cent of this advanced analytical solution is provided through different template-based frameworks that accelerate delivery and reduce errors by almost eliminating the need to write code.
- Provides various machine learning and data mining procedures including data loading and transformation, predictive analytics and statistical modelling, data pre-processing and visualisation, evaluation and deployment, etc.
- Is written using the Java programming language.
- Provides different learning schemes, algorithms and models, which can be extended using Python and R scripts.
- Its functionality can be easily extended with additional plugins, which are made available via its 'Marketplace', which provides a platform for developers to create data analysis algorithms and then publish them for the community.

VI. Conclusion

In this paper, we try to give the basic concept of big data by first providing the definition and hadoop architecture and then the definition of some related terms. We give some examples to elaborate the concept. Then we give different factors that influence big data and techniques to implement that approaches. The various open source tools of big data are also discussed with its features to understand the big data analytics and techniques to use in various fields.

Acknowledgment

I thank our colleagues from Shanmuga Industries Arts and Science College, Department of Information systems and Management and Department of Computer science who provided insight and expertise that greatly assisted the research for this paper. I thank Dr. K. Anandaraj, Principal of our college for assistance with technique, methodology and Dr. Praveen Sundar, Director of MCA Department of our college for valuable comments that greatly improved the manuscript.

REFERENCES

[1] John Walker, S. (2014). Big data: A revolution that will transform how we live, work, and think.
[2] Zikopoulos, P., & Eaton, C. (2011). Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media.
[3] Dumbill, E. (2013). Making sense of big data..
[4] Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. IEEE transactions on knowledge and data engineering, 26(1), 97-107.
[5] Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010, May). The hadoop distributed file system. In Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on (pp. 1-10). Ieee.
[6] Sundar, P. P., & Kumar, A. S. (2016). A systematic approach to identify unmotivated learners in online learning. Indian journal of science and technology, 9(14).

[7] Taylor, R. C. (2010, December). An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics. In BMC bioinformatics (Vol. 11, No. 12, p. S1). BioMed Central.

[8] Liu, X., Han, J., Zhong, Y., Han, C., & He, X. (2009, August). Implementing WebGIS on Hadoop: A case study of improving small file I/O performance on HDFS. In Cluster Computing and Workshops, 2009. CLUSTER'09. IEEE International Conference on (pp. 1-8). IEEE.

[9] Borthakur, D. (2008). HDFS architecture guide. Hadoop Apache Project, 53, 1-13.

[10] Bhandarkar, M. (2010, April). MapReduce programming with apache Hadoop. In Parallel & Distributed Processing (IPDPS), 2010 IEEE International Symposium on (pp. 1-1). IEEE.

[11] Nandimath, J., Banerjee, E., Patil, A., Kakade, P., Vaidya, S., & Chaturvedi, D. (2013, August). Big data analysis using Apache Hadoop. In Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on (pp. 700-703). IEEE.

[12] Sundar, P. P., & Kumar, A. S. (2013). Evaluation of Regional Benchmark Impact in EDM. International Journal of Computer Science Issues (IJCSI), 10(2 Part 2), 531.

[13] Russom, P. (2011). Big data analytics. TDWI best practices report, fourth quarter, 19(4), 1-34.

[14] Srinivasa, S., & Bhatnagar, V. (2012). Big data analytics. In Proceedings of the First International Conference on Big Data Analytics BDA (pp. 24-26).

[15] Zakir, J., Seymour, T., & Berg, K. (2015). BIG DATA ANALYTICS. Issues in Information Systems, 16(2).

[16] Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-347.

[17] Kumar, V., & Chadha, A. (2011). An empirical study of the applications of data mining techniques in higher education. International Journal of Advanced Computer Science and Applications, 2(3).

[18] Xiaofeng, M., & Xiang, C. (2013). Big data management: concepts, techniques and challenges [J]. Journal of computer research and development, 1(98), 146-169.

[19] Katal, A., Wazid, M., & Goudar, R. H. (2013, August). Big data: issues, challenges, tools and good practices. In Contemporary Computing (IC3), 2013 Sixth International Conference on (pp. 404-409). IEEE.

[20] Sundar, Praveen. "Quasi Framework: A new student disengagement detection in online learning." International Journal of Engineering Research & Technology (IJERT) 1, no. 10 (2012).

[21] Jeyakumar, Balajee, MA Saleem Durai, and Daphne Lopez. "Case Studies in Amalgamation of Deep Learning and Big Data." In HCI Challenges and Privacy Preservation in Big Data Security, pp. 159-174. IGI Global, 2018.

[22] Sethumadahavi R., Balajee J. (2017). “Big Data Deep Learning in Healthcare for Electronic Health Records.” International Scientific Research Organization Journal, 2(2), pp.31-35.

[23] Hinneburg, A., & Keim, D. A. (1998, August). An efficient approach to clustering in large multimedia databases with noise. In KDD (Vol. 98, pp. 58-65).

[24] Ranjith, D., Balajee, J. M., & Kumar, C. Trust computation methods in mobile ADHOC network using glomosim: A Review. International Journal of Scientific Research and Modern Education, 1, 777-780.

[25] Saravanan N., Sathish G., Balajee J M. (2018). "Data Wrangling and Data Leakage in Machine Learning for Healthcare", International Journal of Emerging Technologies and Innovative Research, Vol.5, Issue 8, page no. pp553-557.

[26] Rajeshwari, A., Prathna, T. C., Balajee, J., Chandrasekaran, N., Mandal, A. B., & Mukherjee, A. (2013). Computational approach for particle size measurement of silver nanoparticle from electron microscopic image. Int. J. Pharm. Pharm. Sci, 5(2), 619.

[27] Ushapreethi, P., Jeyakumar, B., & BalaKrishnan, P. (2017). Action Recongnition in Video Survillance Using Hipi and Map Reducing Model. International Journal of Mechanical Engineering and Technology 8 (11), 368-375.