ISSN 2347 - 3983
Volume 7, No. 11 November 2019
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter257112019.pdf
https://doi.org/10.30534/ijeter/2019/257112019

A Comprehensive Review of Tools & Techniques for Big Data Analytics

Amita Dhankhar [1], Kamna Solanki[2]
[1,2]University Institute of Engineering and Technology, Maharshi Dayanand University, Rohtak-124001, India
[1]Amita.infotech@gmail.com, [2]Kamna.mdurohtak@gmail.com

ABSTRACT
The changing technology prospect is described by the phrase Big Data that resulted in a large amount of data, a greater variety of data sources, a continuous flow of data and multiple data formats. As such data is growing rapidly, there is a need for advanced analytic techniques that operates on such data and extracts effective information, unknown patterns, and relationships that help in making decisions. Big Data Analytics provides such valuable insight. In this paper, we start with a definition of Big Data. Then we provide the systematic structure that divides the Big Data system into five sections namely data generation, acquisition and storage, data processing, data querying, and data analytics. This paper is also providing a brief overview of different techniques and technology used in Big Data Analytics.

Keywords: Big Data, Analytics, Machine learning, Comprehensive Review.

1. INTRODUCTION

The term “Big Data” refers to data that are not only huge but high in diversity and speed. There are different definitions of Big Data from different points of view. Some of them are: [1] defines big data as: “Big data is a collection of data from traditional and digital sources inside and outside your company that represents a source of ongoing discovery and analysis.” Mill et al defines “Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety ; requires new technologies and techniques to capture, store, and analyse it; and is used to enhance decision making, provide insight and discovery, and support and optimize processes”[2]. [3] defines as “Big data is a term describing the storage and analysis of large and/or complex data sets using a series of techniques including, but not limited to NoSQL, MapReduce, and machine learning." [4] defines Big data as “Big data is where the data volume, acquisition velocity, or data representation limits the ability to perform effective analysis using traditional relational approaches or requires the use of significant scaling (more nodes) for efficient processing”. [7] Defines Big data as “datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze”. In 2011 [27] defines Big data as “Big data technologies, describe a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data, by enabling high –velocity capture, discovery, and/or analysis". Big data is further classified into a big data framework and big data science. Big data frameworks are "software libraries along with their associated algorithms that enable distributed processing and analysis of big data problems across clusters of computer units" whereas big data science is "the study of techniques covering the acquisition, conditioning, and evaluation of big data". All these definitions emphasize size, diversity, and speed. And the requirement of analytical tools and technologies to get an understanding of what data implies. The 5 V's of Big Data are volume, variety, and velocity [5]. Two new dimensions are added to the existing namely veracity and value. The sizes of the Big Data are constantly increasing as a result of which capturing, storing, searching, sharing, analyzing and visualizing data becomes difficult. Advanced analytic methods are applied to such data sets using big data analytics. The following sections will discuss the systematic structure for Big Data system and the techniques associated with the Big data Technologies.

2. BIG DATA SYSTEM

The Big Data System is divided into five consecutive layers namely data generation, acquisition and storage, data processing, data querying, data access, and data analytics. Different Big Data Analytic technologies are given for various big data layers. The following section discusses technologies used in each layer of the Big Data System.

2.1 *Data Generation*

‘How data is generated' is the main concern of the data generation phase. The huge, high in diversity and complex datasets are generated from different sources. These datasets are related to various "domain-specific values". Due to technological growth, the rate of data generation is increased. The sources of data generation are web systems like social media and networking sites, blogs, online groups, forums, e-commerce businesses, web search engines, etc, mobile devices like smartphones, tablets, etc, Internet of things, scientific data, etc.

2.2 *Acquisition and Storage*

Big data analytics deals with vast amounts of unstructured and heterogeneous data. The traditional data techniques and infrastructures fail to collect, integrate and store these data. These obstacles have stimulated a lot of research on Big data collection and storage. “Data acquisition is a process to aggregate information in a well-organized digital form for further storage and analysis “[14].

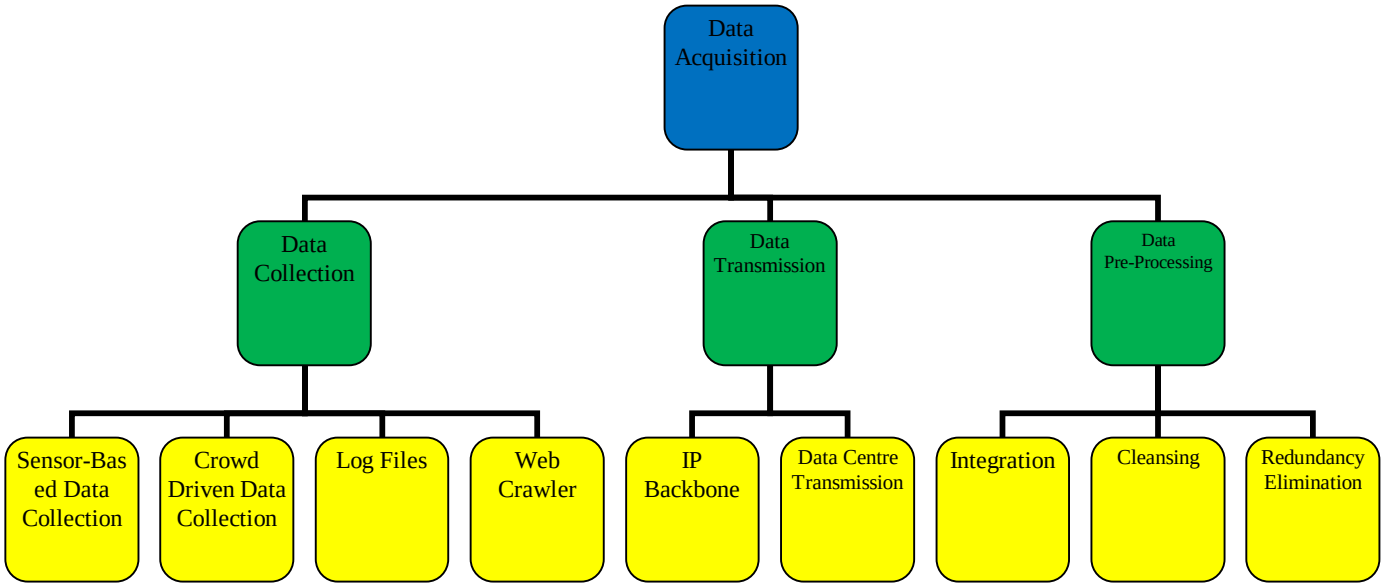

Figure 1: The data acquisition phase consists of three sub-phases: Collection, Transmission, Pre-Processing.

X. Chen et al [15] define data acquisition as data collection, communication and pre-processing as shown in figure 1. Different kinds of data collection methods consist of sensor-based data collection such as inventory management, video surveillance, etc, Log Files such as clickstream, weblog, etc and Web Crawler such as SNS analysis, search, etc. Sensor-based data collection approaches are used to collect and transfer information in many data studies and applications [16][17]. The application of sensor-based data gathering is hampered by the high initial installation and maintenance. To deal with these challenges, some researchers and firms advised crowd focused data gathering as an alternative to sensor-based data gathering [18]. After data collection comes to the Big data transmission methods. It is decomposed into two phases, IP backbone transmission and data center transmission which include data center network architecture and transportation protocols. After data transmission comes data pre-processing which includes integration, cleansing and redundancy elimination of data. Table I gives the advantages and disadvantages of various data acquisition tools.

Table 1: DATA ACQUISITION TOOLS

Name	Description	Advantages	Disadvantages
Flume	It is a distributed system. It collects log data from different sources, aggregate, and transfers to HDFS	. Fault-Tolerance . Failure recovery services . It runs various processes on a single machine . Tuneable reliability mechanism	. It does not assure that message reaching is unique . If the backing store is not selected wisely than scalability and reliability under question . Weak ordering
Sqoop	Import and export data between structured data stores and Hadoop	. Fault-Tolerant . It provides fast performances . Optimal system utilization to reduce processing loads to the external system	. Architecture is based on connectors . Agent-based architecture
Chukwa	It is a data collection system for monitoring large distributed systems.	. To make the best use of the collected data it includes an interface for displaying, monitoring and analyzing results	. There is currently no HDFS file compaction and retention support.

To store and manage huge datasets with availability and reliability is the main focus of Big data storage. The challenges to conventional relational database systems are the growing number of data sources and large –scale features of Big Data. To deal with such data, several technologies, ranging from Massive Parallel Processing databases and distributed systems for providing platform scalability and high query performance to non-relational databases, have been used for Big Data. One of the well-known Big Data technology is Apache Hadoop. It stores and analyzes Big Data. The key component of Hadoop to store data in the Hadoop Distributed File System (HDFS). The data spread across the different servers are managed by HDFS. It manages so many servers in parallel. The main advantage of HDFS is its portability across various hardware and software platforms. By moving computations near to data storage, HDFS helps to reduce network congestion and increase system performance. For Fault tolerance, it also ensures data replication. It based on master-slave architecture. There is a unique master that manages file system operation (also known as Name Node) and many slaves that coordinate and manage data storage on individual compute nodes (also known as Data Node) [19]. Another distributed non-relational database used to store data in HBase. It provides many features such as linear and modular scalability, natural language search, real-time queries and consistent access to Big Data sources. Many data-driven websites and Big Data solutions include HBase. Non-Relational (NoSQL) Database used to store data of any structure. The aim of NoSQL databases is data model flexibility, massive scaling, and simplified application development and operation. Three main types of NoSQL databases are identified namely document-based stores, column-oriented databases and key-value stores [20]. Popular NoSQL databases like Cassandra, Simple DB, Big Table, Mongo DB, and Couch DB belong to one of these types. Table II compares various NoSQL storage systems.

Table 2: COMPARISON OF NOSQL STORAGE SYSTEMS

Name	Developer	Data Model	Advantages	Disadvantages
Dynamo	Amazon	Key-Value	. Automatic data replication . Built-in fault tolerance . Cost-Effective . Easy Administration . Flexible . Distributed . Scalable	. Poor query comparison operators . No triggers . Latency in reading/write . 64KB limit on row size . 1MB limit on querying
BigT able	Google	Column	. No limit for row length . Offers high availability . No special query language needed thus query optimization is not necessary . Operations are performed only at the level of the line, so join operations are not required	. The secondary index is not supported . Possibility of multiple copies of the same data . Lack of advanced features for data security . Data loss can occur . No ACID properties
Voldemort	LinkedIn	Key-Value	. Easy to configure . Only Efficient queries are possible . Easy to distribute across a cluster . Clean separation of storage and logic . Can be used in parallel with a SQL DB	. No built-in support for “multiple data center” aware routing. . No Triggers . No foreign key constraint . No complex query filters . No ACID properties
Simple DB	Amazon	Document	. Low set up cost . No DB administrators needed . Automated indexing . It reduced maintenance	. No support for Text searching . Data size limits . No dumps and backups are not built-in
Redis	Salvatore Sanfilippo	Key-Value	. Exceptionally fast . Supports rich data types and set of operations to work with these types . Very easy to set up and has no dependencies	. No joins or query language . Data set has to fit comfortably in memory

Cassandra	Facebook	Column	. Peer to Peer architecture . Data replication feature makes it highly available and faults tolerant . Elastic scalability . Supports Eventual and strong consistency . For large sets of data, it gives high performance	. Does not support ACID property . No support for aggregation . Does not support relational data property . Does not support Range Based row-scans
MangoDB	10gen	Document	. Good for Realtime analytics . Full index support . Replication and failover functions . High availability . Rich document-based queries for easy readability . Support auto-sharing for easy scalability . For web applications, it is a good replacement of RDBMS	. It consumes more memory . No joins results in less flexibility with querying . Not good for highly transactional systems . Does not support applications with traditional DB requirements such as foreign key constraint
Hbase	Apache	Column	. Good for range-based scan . Takes a small amount of time to read/write with scalability . Strong consistency . Used extensively for online analytical operation . Random read/write operations	. Does not support cross data and joining operations . Not good for applications that need a classic transaction, relational analytics or full table scan . Difficult to store large size of binary files . No exceptional handling mechanism
Couch DB	Couchbase	Document	. No, read locks . Flexible indexes supported . Replication is easy and bidirectional	. Support only eventual consistency . In-place updates require server-side logic . Temporary views on large data sets are very slow
Neo4j	Neo Technology	Graph store	. Connected data representation is very easy . Semi-structured data is easily represented . The data model used is simple and powerful . No need for Joins for retrieval of connected data . Retrieval, traversal, navigation is easy and fast	. Shading is not supported . limitation of supporting numbers of nodes, relationships, and properties

2.3 *Data Processing*

Big data processing stage comes after the Big data storage. According to [21] there are four main requirements of Big data Processing: “Fast data loading, fast query processing, highly efficient utilization of storage space and adaptive to dynamic workload patterns". To deal with such data leads to the development of many new frameworks like GraphLab, prequel, Apache Strom, Simple Scalable Streaming system and Yarn. MapReduce is a parallel programming model that includes “map function” and “reduce function”. The large data processing tasks are divided into smaller tasks by the map function. Then these tasks are assigned with suitable key-value pairs. The output of the map function is the input of the reduce function. The reduce function then performs the aggregation of the output of those having similar key-value. Then it gives a set of combined output values [14]. The main concept t of the MapReduce is to separate the large computational task into different steps and then paralleling execute these steps. This will reduce the time required to complete the task. MapReduce used to solve problems in job scheduling, online aggregation, text processing, distributed computation, high-performance computing, log processing, workload balancing, graph analysis and database system optimization. YARN is another framework that is used for data processing on Hadoop. It is more general than

MapReduce. As compared to MapReduce it provides enhanced parallelism, better scalability, and advanced resource management [22]. Table III compare various programming models.

Table 3:. COMPARISON OF VARIOUSPROGRAMMING MODELS

Name	Advantages	Disadvantages
Map Reduce	. Parallel in nature . Work very fast with both structured and unstructured data . Require minimal amount of memory . Used for problems like a computational, graph	. Jobs run in isolation . Complex algorithm . Does not support iterations . Huge code . High latency which makes it not good for real-time data processing
Pregel	. Efficiently supports iterative computations . Easy to build . Deterministic . Supports automatic fault recovery by checkpointing	. Entire computation state resides in main memory . The performance lacks in case of dense graph
GraphLab	. Supports various data sources . For ML problems it provides various toolkits to give easy and fast solutions . It handles large data sets that result in scalable machine learning . Faster and more efficient runtime performance . Allow dynamic asynchronous scheduling	. Non-deterministic execution . Complicated to implement
Storm	. Allows real-time data processing . Easy to use, highly scalable and fault-tolerant system . Has operational intelligence . Good for continuous computation, online machine learning, real-time analytics	. Not able to run scheduled jobs

2.4 *Data Querying*

Pig Latin is a high-level scripting language generated by an open-source framework called Apache Pig [23]. Pig Latin was developed by Yahoo. HiveQL is a declarative language given by Apache Hive. It provides users to access and manipulate data stored in HDFS and HBase. HiveQL is suited for structured data [24]. There is another declarative language called JAQL. It supports massive data processing. JAQL was designed to query semi-structured data based on Java-script object notion (JSONs) formats [25].

2.5 *Data Analysis*

The function of Big Data analysis is to obtain knowledge from huge data for achieving a better prediction of future and decision making. The huge, heterogeneous and multiple sources of data lead to actual challenges that hinder the application of data mining and information discovery. The researchers proposed a lot of approaches to coping with such challenges like dynamic data-mining methods etc. Blackett defines data analytics into descriptive analytics, predictive analytics and prescriptive analytics [26]. Analytical tools like Apache Mahout and R are used for efficient implementation of machine learning applications and algorithms over large data sets.

3. DATA ANALYSIS METHOD

In this section, we present some common methods that are used for the data analysis. These methods are applied according to the purpose and application domains of the problem.

3.1 *Machine Learning*

Machine learning aims to discover information and make intelligent decisions. It is used for various real-world applications for example recognition system, differentiates between spam and non-spam e-mail messages, recommender system. Machine learning is divided into three main categories: supervised learning, unsupervised learning and reinforcement learning [6]. Different research fields in machine learning are deep learning, incremental and ensemble learning granular computing and genetic algorithm.

3.2 *Data Mining*

According to [7] data mining is defined as "combining methods from statistics and machine learning with database management." Another definition of "searching or 'digging into' a data file for information to understand a particular phenomenon" given by [8]. Some of the data mining methods are clustering, association rule learning, classification algorithm, etc.

3.3 *Social Network Analysis*

Recently social media has played a vital role in the sphere of social networking. Social Network Analysis focuses on the social associations and samples between networks of people. SNA is a tool to know the formal and casual connections and to understand the factors responsible for the transmission of information among the related persons viz who knows whom and utilizing what [9].

3.4 *Text Analysis*

Nowadays Text analytics has become important since a large portion of data generated is in text form such as internet searches, blogs, e-mails, corporate documents, and web page content, etc. To understand the meaning of the information contained by a document or set of documents, Text analytics is used. [10].

3.5 *Advanced Data Visualization*

ADV is a technique to discover knowledge from data. It is a systematic approach that focused on data. It is useful in a situation where analysts have little information about the data. In ADV, the data analysis process is combined with interactive visualization to facilitate a comprehensive data examination [11]. ADV provides comparative statistical graphics and a point-and-click to facilitate faster analysis, effective presentation and better determination and understanding of results [12].

3.6 *Sentimental Analysis*

With the tremendous growth of online opinion data like forums, blogs, movie reviews, product reviews and data from social media sites such as Facebook and Twitter, sentiment analysis is becoming important. It helps researchers to analyze and understand emotions from subjective text patterns. It uses text analytics and Natural language Processing (NLP) to recognize and take out subjective information from texts that are suggestive of a sentiment. It also discovers the connections between words to identify the sentiments accurately [13].

4. CONCLUSION

Today in every sector and industry, Big data has made a strong impact. In this paper, we have given Big data concepts and highlighted the systematic structure for Big data systems. The Big data system consists of five phases namely data generation, acquisition, and storage, data processing, data querying, data analysis. Then we studied different data analysis methods. We have discussed the advantages and disadvantages of various acquisition tools, NoSQL storage systems, and programming models. The selection of these tools, storage systems, and programming models depends on the requirements of the application that need to be developed. Big data deals with data with large volumes, diversified data types, and high velocity. Due to these features of Big data, traditional methods fail to process it. So, there is a need for advanced analytical tools and technologies which can harness the complexity of Big data. Analytical tools and technologies help in extracting useful information that can be used for better prediction of future events and decision making. If Big Data Analytics is properly applied and utilized than it has the power to give a basis for advancement on the technological and scientific level.

REFERENCES

1. http://www.forbes.com/sites/lisaarthur/ 2013/08/15/what -is -big-data/.
2. H. J. Watson, **"Tutorial: Big data analytics: Concepts, technologies, and applications"**, *Communication of the Association for Information System s*, Vol.34, Article 65, pp.124–168, 2014. https://doi.org/10.17705/1CAIS.03465
3. J. S. Wardand, A Baker, **"A Survey of Big Data Definitions"**, arxiv.org /abs/ 1309.5821 VI.
4. **NIST definition of big data and data science,** www.101.datascience.community /2015/nist-defines-big -data-and-data- science.
5. Rick. Smolan, Jennifer. Erwitt, **"The Human face of Big Data, Ed. Against all odds production",** Sausalito, CA 2012.
6. Qiu J., Wu Q., Ding G., Xu Y., Feng S**., "A survey of machine learning for big data processing"**. *EURASIP J. Adv. Signal Process*, 1–16, 2016.
7. Manyika. J, Chui .M, Brown .B, Bughin .J, Dobbs. R, Roxburgh .C, & Byers .A.H**, "Big data : The next frontier for innovation, competition and productivity"** ,pp.1-143, 2011.
8. Picciano .A.G, **"The Evolution of Big Data and Learning Analytics in American Higher Education"**. *Journal of Asynchronous Learning Networks,* 16(3), 9-20, 2012. https://doi.org/10.24059/olj.v16i3.267
9. Serrat .O**: "Social Network Analysis Knowledge Network Solutions"** 28, 1– 4, 2009.

10. Sanchez.D, Martin Bautista M.J, Blanco I, Torre C: **"Text Knowledge Mining: An Alternative to Text Data Mining"** in *IEEE International Conference on Data Mining Workshops*, pp. 664–672, 2008.
11. Shen. Z, Wei J, Sundaresan. N, Ma. K.L: "**Visual Analysis of Massive Web Session Data"** *in Large Data Analysis and Visualization (LDAV),* pp. 65–72, 2012. https://doi.org/10.1109/LDAV.2012.6378977
12. Cebr Data equity, **"Unlocking the value of big data",** In SAS Reports, pp. 1–44, 2012.
13. Mouthami. K, Devi. K.N, Bhaskaran. V.M*:* **"Sentiment Analysis and Classification Based on Textual Reviews"** in *International Conference on Information Communication and Embedded Systems (ICICES),* pp. 271– 276, 2013.
14. Cuzzocrea A, Song I.Y, Davis K. C. **"Analytics over large-scale multidimensional data: The big data revolution",** in Song I. Y. (ed.), *in Proceedings of the ACM 14th international workshop on data warehousing and OLAP* (pp. 101 –104), New York, USA, ACM, 2011.
15. Chen X, Lin X. **"Big data deep learning: Challenges and perspectives",** *IEEE Access*, 2, 514–525, 2014.
16. Barrenetxea G, Ingelrest F, Schaefer G, Vetterli M, Couach O, & Parlange M: **Sensorscope Out-of-the-box environmental monitoring** *in Bill K. (Ed.), in Information processing in sensor networks*, IPSN '08, pp. 332 –343, Washington, DC, USA, IEEE, 2008. https://doi.org/10.1109/IPSN.2008.28
17. Kim S, Pakzad S, Culler D, Demmel J, Fenves G, Glaser S*:* **"Health monitoring of civil infrastructures using wireless sensor networks"** *in information processing in sensor networks*, pp. 254 –263, Washington, DC, USA, IEEE, 2007.
18.] Mondal A., Vasudha B., Srinath S. (Eds.)*:* **"The role of incentive-based crowd-driven data collection in big data analytics"***: A perspective in Big Data Analytic*s, pp. 86 –96, 2013.
19. Bakshi, K., **"Considerations for Big Data: Architecture and Approaches"**, *in Proceedings of the IEEE Aerospace Conference*, pp. 1–7, 2012.
20. Leavitt, N., *Will NoSQL databases live up to their promise Computer*, 43,12 –14, 2010. https://doi.org/10.1109/MC.2010.58
21. He, Y., Lee, R., Huai, Y., Shao, Z., Jain, N., Zhang, X., Xu, Z.: *RCFile:* **"A Fast and Space-efficient Data Placement Structure in MapReduce- based Warehouse Systems",** *in IEEE International Conference on Data Engineering (ICDE),* pp. 1199–1208, 2011. https://doi.org/10.1109/ICDE.2011.5767933
22. White, T. **"Hadoop: The Definitive Guide"**. O'Reilly Media Inc,2012.
23. Sakr, S., **"General-purpose big data processing systems",** *In Big Data 2.0 Processing System s*. Springer, pp. 15–39, 2016 b. https://doi.org/10.1007/978-3-319-38776-5_2
24. Mazumder, S., "**Big data tools and platform s in Big Data Concepts, Theories, and Applications",** *Springer*, pp. 29–128, 2016. https://doi.org/10.1007/978-3-319-27763-9_2
25. Beyer, K.S., Ercegovac, V., Gemulla, R., Balm in, A., Eltabakh, M., Kanne, C.C., Ozcan, F., Shekita, E.J., *Jaql:* **a scripting language for large scale semi-structured data analysis** *in Proceedings of VLDB Conference*, 2011.
26. Blackett, G., Analytics network O.R. & analytics. Retrieved from https://www.theorsociety.com/Pages/SpecialInterest/AnalyticsNetwork_analytics.aspx, 2013.
27. J.Gantz and D. Reinsel, **"Extracting value from chaos",** in *Proc. IDC view*, 2011, PP.1-12.