International Journal of Computer Science and Engineering (IJCSE)
ISSN Print: 2278–9960; ISSN Online:2278–9979;
Vol. 9, Issue 1, Dec–Jan 2020; 27–42
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

A COMPREHENSIVE REVIEW ON BIG DATA ANALYTICS

***Supriya. H. S*[1] *& Dayananda. R. B*[2]**

[1]*Assistant Professor, Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

[2]*Professor, Department of Computer Science and Engineering, Kammavari Sangha Institute of Technology, Bengaluru, Karnataka, India*

ABSTRACT

A huge amount of data regarding Terabytes are generated from digital technologies all over the world. Analysis of such massive data termed Bigdata is required to extract essential information from it. Such Bigdata analysis are required in almost every sector such as business intelligence, financial services, consumer services, healthcare services, energy management, industrial process management, IOT, etc. In this paper, we review the works related to big data analytics that is applied in the variety of applicative domains, the techniques involved and the tools and platforms involved.

KEYWORDS: ***Analytics***, *Big Data*, ***IOT***

Article History

Received: *22 Nov 2019* | ***Revised:*** *28 Nov 2019* | ***Accepted:*** *17 Dec 2019*

INTRODUCTION

Big data refers to the collection of large and complex datasets which are beyond the processing limits of the conventional database systems [1]. Such data are of different varieties (video, image, tables, etc.) and quantifies itself at a very rapid rate. Hence, they cannot be handled by traditional databases. The increased use of the Internet through social sites, blogs, e-commerce sites, continuous data generated by sensors connected to the Internet has resulted in such big data. According to the sources from IDC, the amount of global data generated every year gets doubled. [2][4]

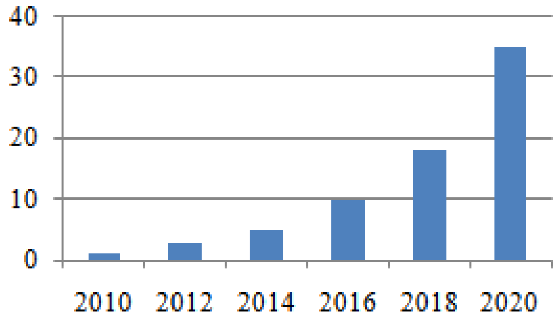

Figure 1: Prediction of Data Generation.

This big data generated from various domains are significant, as they contain information that may be used by private firms or government organizations to improve their business or try to understand people’s behavior to make appropriate business or administration decisions.

Big data differs from conventional data, which can be described, mainly regarding three Vs, namely, volume, velocity and variety. Veracity and value are recently considered as other dimensions for big data. The term *volume* implies the data size. The Big data sizes are expressed regarding terabytes, petabytes and zettabytes. The *Variety* refers to the type of datasets which may be structured, unstructured or semi-structured. The tabular data is commonly known as structured. Text, audio, video and picture, all fall under the unstructured category. Extensible markup language (XML) falls under the semi-structured category. The *velocity* refers to the data generation rate from the sources, as in the case of data generated from sensors, surveillance cameras. Veracity refers to the uncertainties, missing values, imprecision present inherently in some datasets. Value refers to the importance of the data in analyzing.

Big Data System

The operation of the big data system is segregated as four different phases – data generation, acquisition, storage and analytics. The large datasets may be generated from any of the sources, such as sensors, social sites, business applications and scientific research. The acquisition layer performs data collection, transmission and preprocessing operations. Data collection refers to a dedicated data acquisition technology that collects raw data from a specific data production environment. Data preprocessing involves filtration of irrelevant data, compressing and structuring the data. The collected data acquired are stored using big data storage layer. The storage system consists of a scalable hardware infrastructure and a software management layer. The analytics layer uses the analytic tools to model, transform, inspect and finally extract information.[4]

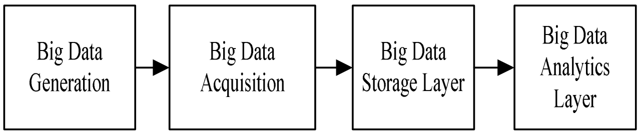

Figure 2: Big Data System Reference Architecture [4].

The big data system introduces a new set of storage, processing requirements, computation models, and visualization methodologies to handle such huge datasets for a considerable time, which cannot be achieved using traditional data management systems. This paper reviews some of the recent works that shed light on the various processing methods, the tools that are used to handle the Bigdata generated from different domains.

Big Data Analytics

Analytics involves identifying the behavioral patterns in raw data using mathematical and analytical models for knowledge discovery [5]. Big data analytics (BDA) refers to the analytics performed on large datasets consisting of different types of data (structured/unstructured) to reveal hidden patterns, hidden correlations, market trends, customer demands, etc. The main objective of BDA lies in the understanding of the huge volume of data and makes efficient and well-informed decisions. BDA tools allow data miners and scientists to analyze a large volume of data that may not be harnessed using traditional tools [6].

There are four types of Bigdata analysis that prevails in business analytics. [7]

Descriptive Analytics: Analysis performed on the past collected data is termed descriptive analysis. Data aggregation and data mining are used to analyze historic data and provide information regarding what has happened on the data. They provide analytics regarding a company's production, finance, etc.

Predictive Analytics: This analytics estimates the probability of a future outcome. Statistical models and forecast techniques are used for predicting what could happen in the future. They are used for predicting the purchase patterns, customer behaviors, end of the year sales, etc.

Prescriptive Analytics: This analytics provides suitable advice for making decisions. They use optimization and simulation algorithms and prescribe some possible actions to be carried out. They are used in companies for optimizing production and optimizing customer experience.

Diagnostic Analytics: This is an advanced analytics that portrays the causes of events and behaviors. It examines the data to find out why the events have occurred. They can be used by companies to have an insight into their employees and solve complex workforce issues.

The emerging BDA based on the technical areas are (a) Structured data analytics, (b) Text data analytics, (c) Multimedia analytics, (d) Web analytics, (e) Network analytics and (f) Mobile analytics. [4]

The types of analytics used in IOT applications are real-time analytics, offline analytics, memory level analytics, BI analytics and massive analytics according to [6].

Big data analysis methods [Olshannikova][4]

Data Visualization

Visualization methods use charts, tables, images for visualizing the analysis results. Big data visualization methods are different from the traditional visualization methods, such as bar chart, Venn diagrams, histogram, etc. Those methods should be capable of dealing with the three Vs of big data along with that there are other challenges concerning dealing with visual noise, information loss and large image perception and performance requirements. Some of the visualization methods have utilized visualizing big data based on data volume, variety and dynamics, such as Treemap, Circle Packing, Sunburst, Parallel Coordinates and Circular Network Diagram. Large-scale data are visualized feature extraction and geometric modeling for reducing the data size. Some of the visualization tools that run in Hadoop are Pentaho, Flare, Jasper Reports, Dygraphs and Tableau, etc. [8]

Statistical Analysis

The methods available for Bigdata statistical analysis can be grouped under (a) sub-sampling-based method, which includes leveraging, mean-log likelihood, (b) divide and conquer-based methods, such as aggregation, majority-voting (c) online updating.[9]

Data Mining

They analyze datasets for finding the unsuspected relationships. They generate models for extracting the implicit, hidden or unknown information from data. Data mining tasks are grouped as predictive based, descriptive based and optimization based. The most popular data mining methods available are classification, regression (predictive based), clustering and association rule techniques (descriptive based). Cluster analysis classifies objects based on similarity measures. They belong to unsupervised learning class, which requires training data for learning. Classification techniques are supervised learning methods, which use the computational model to categorize the new data point into a specific cluster by the training data. Regression techniques determine the relationship between two variables, which aid in the prediction and forecasting. Association rule mining techniques are designed for detecting interesting relationships or strong rules among variables in a database.[10]

Table 1: Data Mining Techniques Found Suitable for Handling Big Data.[10]

Data Mining	Technique	Dimension Handled
Classification	KNN	Volume Veracity
	Decision Trees	Volume, Veracity, Variety
	Neural Network	Volume
Association Mining	Apriori FP Growth	Volume, Veracity, Variety
Clustering	K-Means Clustering	Volume

Optimization Methods

They use mathematical tools for efficient analysis. It includes analytical techniques, such as genetic and evolutionary programming and PSO. Since they involve mathematical operations, using such a method is a very time-consuming task and not preferred for big data. However, convex-based optimization algorithm for big data is recently being proposed in [11].

Machine learning is a few machine learning paradigms that is suitable for big data, which was presented in [12]. *Deep learning* uses a hierarchical learning process, which posses several hidden layers. These hidden layers apply nonlinear transformation process on data, when data pass through these layers to extract the data representations. These representations enable the features to be learned. Deep learning is suited for problems in image classification and recognition. *Online learning* method uses streaming data for learning and training. This method does not require the data to be held in memory. Hence, this method can process large volumes of streaming data which is the nature of Bigdata. This method may be used for stock data prediction. *The local Learning* method is concerned with learning from a subset of the large set called local sets of interest rather than using the entire set. This methodology enables to process large datasets as in the case of big data. *Transfer learning* trains the models with datasets from multiple domains called source domain with similar problem and constraint. *Lifelong learning* poses features of online learning and transfer learning. Like online learning, the learning process is continuous, and like transfer learning, it is capable of transferring information among domains. *Ensemble learning* uses multiple learning models and a voting process for deterring the outcome from the weighted individual learning models outcome.

Big Data Tools

Some big data analytic tools are available for analyzing big data. The currently used Bigdata platforms and their supporting tools will be presented in this section.

Apache Hadoop

It is an open source platform that handles large datasets on distributed computer clusters of commodity hardware. The Hadoop architecture is composed of (i) distributed storage layer and (ii) processing layer. The storage layer uses HDFS file system which splits the data and stores it as a replication in other servers. The distributed computations on the datasets are performed using a Map Reduce programming model. The model comprises “map” and “reduce” functions. A key-value combination is formed from the large dataset using map function, and this key-value combination is reduced to form another pair using the reduce function.[1][2]

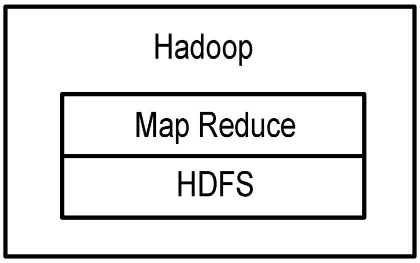

Figure 3: Hadoop Architecture.

Apache Cassandra: It is an open source distributed storage system used for storing huge amounts of structured datasets. It is used by companies such as Twitter, Facebook, eBay, Cisco, etc. The architecture is such that it handles data at multiple nodes in a cluster. Each node is interconnected to the other node and accepts read and write requests. In case of a node, failure data is served from other nodes. Cassandra uses CQL for querying.

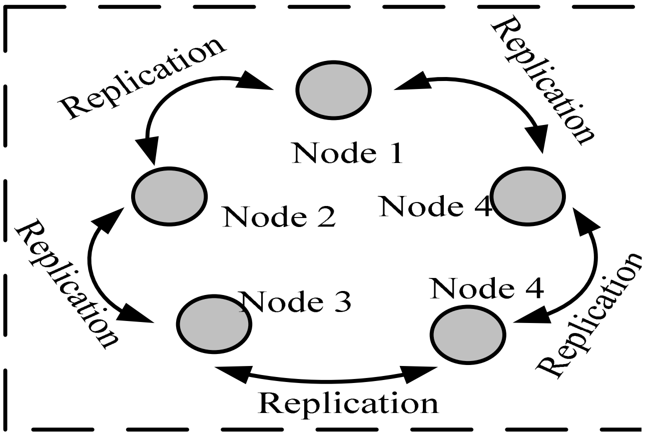

Figure 4: Shows Data Replication in Cassandra System.

Apache Spark: It is a clustering computational technology developed to improve the speed of Hadoop computation process. It increases the speed by using in-memory cluster computing in which data is maintained at each cluster node and avoids reloading. They have faster computation capability compared to MapReduce. It is capable of performing SQL Querying in real-time, analytics on streaming data, Machine-learning and graph processing using suitable libraries.

Apache Kafka: A distributed messaging system intended for passing messages from one application to another. In Kafka, a term called "topic" refers to a data stream of a specific type. When the producers publish messages and they are stored in a set of servers called brokers. Consumers subscribe to topics and pull messages from the brokers.[13][23] They are usually used for offline data processing.

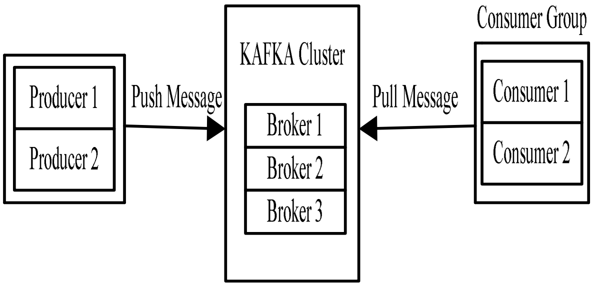

Figure 3: Messaging Model of Kafka[13].

Apache Mahout: It is a machine learning library used for creating machine learning techniques, such as recommendation, clustering and classification. The Mahout algorithms are written over Hadoop and allow analyzing of large sets of data at a fast rate. [14]

Apache Storm: It is a distributed real-time processing system capable of processing data streams flowing at high ingestion rates. Apache Storm takes in a stream of raw real-time data and feeds it through small processing units and produces the useful information. Two nodes, master and worker form the storm cluster that performs nimbus and supervisor roles, respectively.

Apache Drill: This is a distributed system used for interactive analysis of big data. It works on top of HDFS storage and uses MapReduce performing batch analysis. It supports many query language, data formats and data sources.[1]

MongoDB: It is a document-oriented database, which is capable of handling unstructured data. It uses JSON data structures for storing complex data types. It has a high-speed access to massive data, typically 10 times faster than MySql. [15]

Splunk: A real-time platform which can process data generated from business industry machines. It combines cloud technologies with big data and assists the user in monitoring the machine data via a web interface.

Existing work

The work of (Paul *et al.*, 2017) [16] used BDA for defining the human behaviors. A high-level data architecture named Smart Buddy was presented for large data processing. A SIoT-based smart city concept to collect data and pass it to Smart Buddy was used. This architecture comprised of an object domain, SIoT server domain and application domain. The system is implemented using Hadoop and Apache Spark to process data in real time and MapReduce for offline analysis.

The work of (Marjani *et al.*, 2017) [17] focused on the current research towards IoT data analytics. The author presented different methods available for performing BDA. Further IoT architecture for big data analytics was proposed here. The architecture shows the sensor devices networked wirelessly and connected to the internet through gateways. The data stored in the cloud are processed through Big data applications, which contain API management and dashboards for interacting with the processing engine.

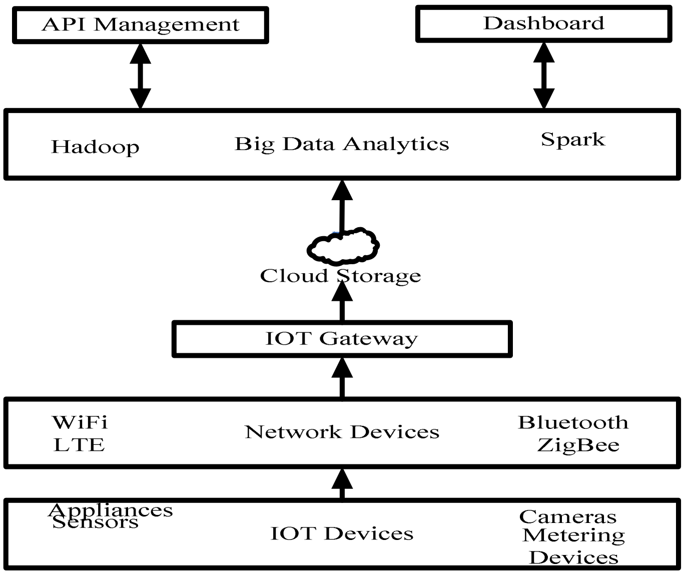

Figure 4: IOT Architecture and Big Data Analytics.

An electricity price forecasting using big data analytics was presented by (Wang *et al*., 2017) [18]. A method was proposed to handle huge price data in the grid. The proposed price forecasting model uses grey correlation analysis as a hybrid feature extractor, KPCA is used for eliminating feature redundancy and differential evolution-based SVM is used for price forecasting.

A Big data analytics for user anomaly and activity detection in the mobile wireless network was presented by Parwez *et al*., 2017 [19]. The call detail record (CDR) mobile data is used for analyzing anomaly detection and k-means clustering and hierarchical clustering for anomaly detection. The region where the anomaly is detected is identified, so that suitable actions can be taken.

A fuzzy rule-based BDA scheme for Health care services was presented by Jindal *et al*., 2017) [20]. The authors focused on the categorization of the collected healthcare data according to its context in cloud computing environment. To solve this problem, a fuzzy rule-based BDA was presented. Initially, a modified EM algorithm was used for cluster formation on the cloud and then a fuzzy rule-based classifier was applied for data classification. The evaluation of the scheme was done w.r.t. classification time, response time, accuracy and false positive rate. The result of this work is that the doctor can get related information of the diseases, the patient information by entering the symptoms of the disease.

The work of Wu and Chen, 2017 [21] focused on performing BDA on transport systems for achieving environment sustainability. A new methodology for big data analytics named *hybrid data analytics* was proposed for tackling the problems in green transport systems and generating strategies for transport companies for achieving environment sustainability. Hybrid data mining involved Topic mining and Association rule modeling. The topic mining generated the important issues in transport systems affecting environment sustainability, while the association rule model retrieved critical rules for transport systems to achieve environment sustainability.

The authors (Rathore *et al*., 2017)[22] worked on presenting a system for exploring the user data and location from the Geosocial networks. The authors proposed a system consisting of data collection, processing, application and communication and storage layers. The data processing part of the system consisted of preprocessing and analysis. Preprocessing involved filtration and classification of data based on the tie, location and content. The classified data are then analyzed using text analysis, statistical analysis, content-based analysis and other machine-learning algorithms. For processing activity, the proposed system used Hadoop ecosystem with Map Reduce programming mechanism. Apache Spark was used over Hadoop for real-time streaming data processing. From the analysis, various disaster events, such as earthquakes, fires and diseases in a particular geographical location could be determined.

The work of Schmid *et al*., 2017 [23] performed analytics on the data gathered by the software-intensive distributed systems for their self-adaptation. The analytics was used to guide self-adaptation based on real-time streaming of operational data. A smart navigation system was used for experimentation, where navigation was done using data collected, such as car sensors and traffic estimation.

The authors (Stoyanov and Kakanakov, 2017) [24] used BDA for solving the power quality problem in electricity distribution systems. An Apache Hadoop was made use for handling massive amounts of data generated by the smart meters and other grid equipments. A Map Reduce algorithm was used for processing the data.

A multi-cluster big data processing framework as designed by Wu *et al*., 2017)[25]. A Hierarchically Distributed Data Matrix (HDM) multi-cluster architecture was presented that is capable of performing analytics on large data over multiple clusters.

The work of Jagannathan, 2017[26] presented an architecture for real-time big data analysis on remote sensing data. In the architecture, the data are remotely preprocessed and further processing is performed in the earth's base station. In the base station, real-time data are passed through filtration for extracting information and load balancer server for balancing processing power. This architecture is implemented in Hadoop using map-reduce programming. The system proposed analyses the big remote data successfully.

An architecture using BDA for real-time traffic control was presented by Amini *et al*, 2017 [27]. The architecture was built using Kafka tool, which acts as a layer that separates publishers and subscribers from the data analytics engine. The data analysis results are published to subscriber topics or logged in No SQL database. The published info is used by the ITS actors.

The work of (Fiadino *et al*. 2016) [28] investigated three main applications of the cellular network for determining their content delivery dynamics, performing traffic analysis and characterizing their network infrastructure. The analysis was performed using the data collected from the cellular network of European ISP. To handle the huge data generated, DBStream BDA platform was used. The applications GEO server location, traffic flows and usage patterns were found.

The work of Li *et al*. 2016 [29] focused on reducing the inter-DC traffic that occurs in geo-distributed BDA. An optimization technique named chance-constrained optimization technique to the DC MapReduce job was presented by optimizing the input data collecting and task placement.

The work of Nicole *et al*., 2016) [30] focused on increasing the speed of data shuffling, which is required in data aggregation for BDA. A novel dynamic data shuffling strategy was presented, which consumed less memory and delivered high performance. The prototype was developed and integrated into Spark framework.

A BDA on Omic and HER data was presented by Wu *et al*., 2016)[31]. The application of BDA to enable precision medicine was demonstrated. Clustering-based methods were used for detecting the cancer types using the Omic data.

A big data architectural framework in the mobile cellular network was presented by He *et al*., 2016)[32]. The framework consisted of data collection, big-data preprocessing and analysis stage. Several case studies for big data analysis on different data of the cellular network, such as mobile signaling data, traffic data, location data and radio waveform data were performed. The authors used Hadoop platform for the analysis.

An application of data analytics to identify the trends in crime that has occurred was presented by Yetis *et al*., 2016)[33]. The crime information for a particular city over a period was collected, and a MapReduce algorithm was used for sorting the gathered data based on the year, the location of the crime and the type of crime.

A BDA for anomaly detection in the mobile cellular network was presented by Yang et al., 2016)[34]. A deep network analyzer consisting of fingerprint learning module and anomaly detection and the analyzing module was developed for detecting the anomalies and identifying the causes. DNA was implemented in Spark big data platform and an association rule was used for miming the association of quality and performance indicators. From the rule, a fingerprint database was built, from which cause was analyzed.

The author Kumar *et al*., 2016 [35] used BDA for health monitoring and diagnosis of elderly people in remote areas. In this proposed system, the patient body parameters are measured and sent to a data center, where comparison of the parameters of the collected and that of the datasets are done. The disease is identified and the respective medicines are sent to the patient. Here, the patient's dataset, disease set and medicine dataset are stored in Hadoop system and processing of datasets is done through k-means clustering.

A mobile BDA concept to deal with massive data collected from mobile devices was presented by Al-Sheikh *et al.* 2016[36]. A deep learning approach was used for understanding the raw mobile data. The deep learning method was implemented on Apache spark platform.

The work of Jeong *et al.* 2016 [37] used BDA for radiation sensor network analysis. The data collected by the sensor networks were processed using big data tools, such as Apache Pig and Hadoop. Graph indices and similarity matching techniques are exploited to compare radiation level changes.

The work of Das *et al.*, 2016)[38] focused on finding an efficient method of storing and fetching unstructured data. A big data application was built on the Hadoop platform, which takes twitter data, stores it in Hbase and performs an analysis. The data may be retrieved using REST calls.

The BDA in logistics was presented by Ayed *et al.*, 2015)[39]. A big data system for container code recognition was proposed. The container was written with unique codes and the identification was done by first capturing the codes and storing in a Hadoop system. The extraction of text regions and application of OCR is done using MapReduce programming model. The use of Hadoop framework reduced computational complexity and provided real-time performance.

The work of Yu *et al.*, 2015[40] presented the use of BDA in power distribution systems. A power distribution analytics architecture was proposed here, which collects measured data and grid data; price data from various sources are stored in the Hadoop clusters. The analytics are performed using code developed in SAS, Revolution R and Mahout will be moved to Hadoop.

The work of Kameshwari *et al.*, 2015)[41] focused on storing and analyzing the large volume of sensor data for anomaly detection in sensors from process industries. A framework consisting of the ensemble of statistical and clustering algorithms was used for anomaly detection. An Infini DB database was used here for storing the huge amount of data collected from the sensors.

The work of Pradhananga *et al.*, 2015)[42] presented a cloud-based platform for BDA. The author focused on providing a cost-effective solution for analyzing big data for small and medium enterprises. The author combined R and Hadoop over the cloud to provide the low-cost analytic platform. CBA architecture was proposed which included R statistical software, Hadoop framework, R-H Hadoop and Amazon EMR.

The work of Prabhu *et al.*, 2015[43] focused on improving the performance of MapReduce framework in Hadoop. The performance was improved by tuning the job configuration parameters, which reduces execution time and disk usage. The method proposed is an iterative process, where the job is launched first and the resource usage is analyzed. If the usage is underutilized, the parameter is adjusted, and the job is re-executed. This process is repeated until the performance criteria are satisfied.

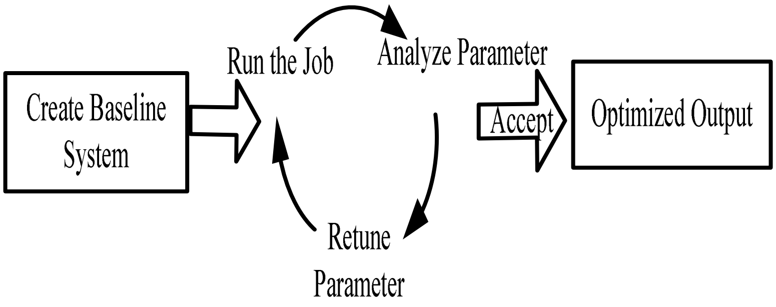

Figure 5: Hadoop Performance Tuning Process.

An efficient information retrieving in Hadoop was presented by Mathew *et al.* 2014)[44]. Two indexing methods called modified Lucene (LIndex) and Hashmap index (HIndex) in HDFS was presented for optimized text-based selection and fast retrieval. The performance of HIndex was observed to be better in the simulation results compared to both the Lucene approaches.

Predictive analytics for estimating and replacing the incorrect readings of the malfunctioned sensors was presented by Kejela *et al.*, 2014[45]. The author used H_2O big data tool, which consists of distributed and powerful machine learning methods, such as generalized linear model (GLM) and gradient boosted model for the analysis. The results showed the method could estimate the missing or incorrect readings.

Table 2: Review Summary

SI.No	Authors	Application	Problems	Tools/Technique	Performance Parameters
1	Paul *et al.*[16]	Human dynamics	Big data analytics for defining the human behaviors	Smart buddy architecture implemented with Hadoop and Apache spark	
2	Marjani *et al.*[17]	IOT	IOT data analytics	IOT architecture	
3	Wang *et al.*[18]	Electricity Grid	Electricity Price forecasting	Used GCA, KPCA and DE-SVM	Electricity price
4	Parwez *et al.*[19]	Mobile wireless network	Anomaly detection	k-means +hierarchial clustering	Mean squared error
5	Jindal *et al.*[20]	Health care services	Data classification	modified EM algorithm+ fuzzy rule-based classifier	Classification time, response time, accuracy and false positive rate
6	Wu and Chen[21]	Transport systems	Environment sustainability	Hybrid data analytics	support, confidence, and lift.
7	Rathore *et al.*[22]	Geo-social Network data	analyzing geo-social media posts	Hadoop ecosystem+ Apache Spark	Number of tweets
8	Schmid *et al.*[23]	software-intensive distributed systems	Developing model for self-adaptive software-intensive distributed systems	(RTX) tool Kafka+Spark	

Table 2 Contd.,

9	Stoyanov and Kakanakov[24]	Electricity distribution systems	Power quality problem in electricity distribution systems	Apache Hadoop	Time to run required reports and storage space
10	Wu *et al.*[25]	1)	Data analytics over multiple clusters	HDM-MC	Scheduling time, job completion time
11	Jagannathan[26]	Remote Sensing	Analyzing big remote data	Big data analytic architecture + Hadoop	
12	Amini *et al.* 2017 [27] 2)	Transport system	Real-time traffic control	Big data analytics architecture using Kafka	
13	Fiadino *et al.*[28]	Cellular network	Traffic flow, usage pattern detection using	DBStream BDA platform	Traffic flows
14	Li *et al.*[29]	Geo-distributed Big data analytics	Minimizing Inter DC traffic	MapReduce Hadoop platform	Average InterDc traffic
15	Nicole *et al.*[30]	3)	Increase in the speed of shuffling	A novel shuffling strategy implemented on Spark	Peak shuffle utilization Cpu utilization
16	Wu *et al.*[31]	Healthcare	BDA on Omic and her	Cluster-based analysis	-----
17	He *et al.*[32]	Mobile cellular network	BDA on mobile signaling, traffic data	Hadoop platform	Traffic flows
18	Yetis *et al.*[33]	Crime	Determining Type of crime and their occurrences	MapReduce algorithm	Number of crimes
17	Yang *et al.*[34]	Mobile cellular network	Anomaly detection and cause identification	Deep network analyzer using Spark	Number of anomalies detected
18	Kumar *et al.*[35]	Health care	Remote patient diagnosis	Proposed system using Hadoop infrastructure	Visual results
19	Alsheikh[36]	Mobile data	Mobile big data analytics	Deep learning on apache Spark	
20	Jeong *et al.*[37]	Sensor network	radiation level changes	Data collection in Hadoop and pg along with graph comparison	
21	Das *et al.* [38]	Social network site	Efficient storing and retrieving twitter data	Big data application on Hadoop, HBase	
22	Ayed *et al.*[39]	Logistics	Container code recognition	TEXT and OCR implemented on Hadoop	
23	Yu *et al.*[40]	Power Distribution systems	Handling big heterogeneous data	Analytics code of SAS and R on Hadoop platform	
24	Kameshwari *et al.*[41]	Process industry	Anomaly detection	Ensemble of clustering and statistical methods	Running time
25	Pradhananga *et al.*[42]	Cloud-based big data analytics	Cost effective analytics	CBA platform including Hadoop and R	

Table 2 Contd.,

26	Prabhu *et al.*[43]	4)	Performance improvement of MapReduce framework	Fine tuning job parameters	CPU execution time
27	Mathew *et al.*[44]	Text processing capabilities in Hadoop ecosystem	Improving text retrieval speed	Proposed LIndex HIndex approaches in Hadoop	CPU time
28	Kejela *et al.*[45]	Oil and gas industry	Estimating and replacing the incorrect readings	Using machine learning methods in H2O tool	Prediction error

Research Gap

From the review on the works related to big data, the following observations were made:

- The majority of the works were developed using Hadoop platform with Map Reduce programming model. The limitations with Hadoop are of slow processing nature, which does not suit real-time data processing.
- The privacy of data is an important concern in business. It was observed that there were no specific works that concern securing the data used in analytics. As data is voluminous, providing security for such vast volume data would be a challenge that needs to be addressed. The Hadoop platform does not use any encryption and hence less secured.
- The performance evaluation results for the proposed systems were not found for most of the big data related works.

CONCLUSIONS

A massive amount of data is being generated at a very fast pace. Analyzing such Bigdata for business intelligence and decision-making requires a different storage, processing and visualizing paradigms. This paper reviewed the works related to Bigdata analytics, surveyed different types of analysis performed on Bigdata, and the most widely used Bigdata tools.

REFERENCES

1. *Acharjya, Debi Prasanna and P. Kauser Ahmed. "A Survey on Big Data Analytics: Challenges, Open Research Issues, and Tools." Article in Int. J. Adv. Comp. Sci. Appl., February 2016.*

2. *Ayed, Abdelkarim Ben, Mohamed Ben Halima and Adel M. Alimi. "Bigdata analytics for logistics and transportation." Advanced Logistics and Transport (ICALT), 2015 4th Int. Conf. IEEE, 2015.*

3. *P. Bhardwaj, A. Gupta, M. Sharma, M. Gupta, S. Singhal, "A Survey on Comparative Analysis of Big Data Tools," IJCSMC, vol. 5, issue 5, pp. 789–793, 2016.*

4. *Hu, Han, et al. "Toward scalable systems for big data analytics: A technology tutorial." IEEE Access 2 (2014): 652–687.*

5. *Sruthika, S. and N. Tajunisha. "A study on the evolution of data analytics to big data analytics and its research scope." Innovations in Information Embedded and Communication Systems (ICIIECS), 2015 Int. Conf. IEEE, 2015.*

6. *Marjani, Mohsen et al. "Big IoT Data Analytics: Architecture, Opportunities and Open Research Challenges." IEEE Access 5 (2017): 5247–5261*

7. *Mujawar, Sofiya, and Aishwarya Joshi. "Data Analytics Types Tools and their Comparison."Int. J. Adv. Res. Comp. Commun. Eng. vol. 4.2, 2015. [A/Q: Please provide page range]*

8. *Wang, Lidong, Guanghui Wang and Cheryl Ann Alexander. "Big data and visualization: methods, challenges and technology progress." Digital Technol. vol. 1.1 pp. 3–38, 2015.*

9. *Wang, Chun et al. "Statistical methods and computing for big data." arXiv preprint arXiv:1502.07989 (2015).*

10. *Fawzy, Dina, Sherin Moussa and Nagwa Badr. "The Evolution of Data Mining Techniques to Big Data Analytics: An Extensive Study with Application to Renewable Energy Data Analytics." Asian J. Appl. Sci. (ISSN: 1996–3343) 4.3 (2016).*

11. *Cevher, Volkan, Stephen Becker and Mark Schmidt. "Convex optimization for big data: Scalable, randomized and parallel algorithms for big data analytics." IEEE Signal Process. Mag. vol. 31.5, pp. 32–43, 2014.*

12. *L'Heureux, Alexandra et al. "Machine Learning with Big Data: Challenges and Approaches." IEEE Access (2017).*

13. *Apache Kafka@ Copyright 2016 by Tutorials Point (I) Pvt. Ltd.*

14. *S. Gaikwad, P. Nale and R. Bachate, "Survey on big data analytics for the digital world," 2016 IEEE Int. Conf. Advances in Electronics, Communication and Computer Technology (ICAECCT), Pune, 2016, pp. 180–186.*

15. *Jing Han, Haihong E, Guan Le and Jian Du, "Survey on NoSQL database," 2011 6th Int. Conf. Pervasive Computing and Applications, Port Elizabeth, 2011, pp. 363–366.*

16. *Paul, A. Ahmad, M. M. Rathore and S. Jabbar, "Smartbuddy: defining human behaviors using big data analytics in the social internet of things," in IEEE Wireless Communications, vol. 23, no. 5, pp. 68–74, October 2016.*

17. *M. Marjani et al., "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," in IEEE Access, vol. 5, no., pp. 5247–5261, 2017.*

18. *K. Wang, C. Xu, Y. Zhang, S. Guo and A. Zomaya, "Robust Big Data Analytics for Electricity Price Forecasting in the Smart Grid," in IEEE Trans. Big Data, vol. PP, no. 99, pp. 1–1. [A/Q: Please provide year of publication.]*

19. *M. S. Parwez, D. B. Rawat and M. Garuba, "Big Data Analytics for User-Activity Analysis and User-Anomaly Detection in Mobile Wireless Network," in IEEE Trans. Industr. Inform., vol. 13, no. 4, pp. 2058–2065, Aug 2017.*

20. *Jindal, A. Dua, N. Kumar, A. V. Vasilakos and J. J. P. C. Rodrigues, "An efficient fuzzy rule-based big data analytics scheme for providing healthcare-as-a-service," 2017 IEEE Int. Conf. Communications (ICC), Paris, 2017, pp. 1–6.*

21. *P. J. Wu and Y. C. Chen, "Big data analytics for transport systems to achieve environmental sustainability," 2017 Int. Conf. Appl. Sys. Innov. (ICASI), Sapporo, 2017, pp. 264–267.*

22. *M. M. Rathore, A. Paul, A. Ahmad, M. Imran and M. Guizani, "Big data analytics of geosocial media for planning and real-time decisions," 2017 IEEE Int. Conf. Communications (ICC), Paris, 2017, pp. 1–6.*

23. *S. Schmid, I. Gerostathopoulos, C. Prehofer and T. Bures, "Self-Adaptation Based on Big Data Analytics: A Model Problem and Tool," 2017 IEEE/ACM 12th Int. Symp. Software Engineering for Adaptive and Self-Managing Systems (SEAMS), Buenos Aires, 2017, pp. 102–108.*

24. *1S. Stoyanov and N. Kakanakov, "Big data analytics in electricity distribution systems," 2017 40th Int. Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2017, pp. 205–208.*

25. *D. Wu, S. Sakr, L. Zhu and H. Wu, "Towards Big Data Analytics across Multiple Clusters," 2017 17th IEEE/ACM Int. Symp. Cluster, Cloud and Grid Computing (CCGRID), Madrid, 2017, pp. 218–227.*

26. *S. Jagannathan, "Real-time big data analytics architecture for remote sensing application," 2016 Int. Conf. Signal Processing, Communication, Power and Embedded System (SCOPES), Paralakhemundi, 2016, pp. 1912–1916.*

27. *S. Amini, I. Gerostathopoulos, and C. Prehofer, "Big data analytics architecture for real-time traffic control," 2017 5th IEEE Int. Conf. Models and Technologies for Intelligent Transportation Systems (MT-ITS), Naples, 2017, pp. 710–715.*

28. *P. Fiadino, P. Casas, A. D'Alconzo, M. Schiavone and A. Baer, "Grasping Popular Applications in Cellular Networks With Big Data Analytics Platforms," in IEEE Trans. Netw. Serv. Man ., vol. 13, no. 3, pp. 681–695, Sept. 2016.*

29. *P. Li et al., "Traffic-Aware Geo-Distributed Big Data Analytics with Predictable Job Completion Time," in IEEE Trans. Parallel Distrib. Syst., vol. 28, no. 6, pp. 1785–1796, June 1, 2017.*

30. *Nicolae, C. H. A. Costa, C. Misale, K. Katrinis, and Y. Park, "Leveraging Adaptive I/O to Optimize Collective Data Shuffling Patterns for Big Data Analytics," in IEEE Trans. Parallel Distrib. Syst., vol. 28, no. 6, pp. 1663–1674, June 1, 2017.*

31. *P. Y. Wu, C. W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman and M. D. Wang, "–Omic and Electronic Health Record Big Data Analytics for Precision Medicine," in IEEE Trans Bio-Med Eng., vol. 64, no. 2, pp. 263–273, Feb. 2017.*

32. *Y. He, F. R. Yu, N. Zhao, H. Yin, H. Yao and R. C. Qiu, "Big Data Analytics in Mobile Cellular Networks," in IEEE Access, vol. 4, no., pp. 1985–1996, 2016.*

33. *Y. Yetis, R. G. Sara, B. A. Erol, H. Kaplan, A. Akuzum and M. Jamshidi, "Application of Big Data Analytics via Cloud Computing," 2016 World Automation Congress (WAC), Rio Grande, 2016, pp. 1–5.*

34. *K. Yang, R. Liu, Y. Sun, J. Yang and X. Chen, "Deep Network Analyzer (DNA): A Big Data Analytics Platform for Cellular Networks," in IEEE Internet Things J., vol. PP, no. 99, pp. 1–1.*

35. *K. Kumar and M. R. Bagavathi, "Thriving the ills of elderly people using big data analytics," 2016 Int. Conf. Computer Communication and Informatics (ICCCI), Coimbatore, 2016, pp. 1–6.*

36. *M. A. Alsheikh, D. Niyato, S. Lin, H. p. Tan and Z. Han, "Mobile big data analytics using deep learning and apache spark," in IEEE Network, vol. 30, no. 3, pp. 22–29, May–June 2016.*

37. *M. H. Jeong, C. J. Sullivan and S. Wang, "Complex radiation sensor network analysis with big data analytics," 2015 IEEE Nuclear Science Symp. and Medical Imaging Conf. (NSS/MIC), San Diego, CA, 2015, pp. 1–4.*

38. *Das, T. K., and P. Mohan Kumar. "Big data analytics: A framework for unstructured data analysis." Int. J. Eng. Sci. Technol. 5.1 (2013): 153.*

39. *Ben Ayed, M. Ben Halima and A. M. Alimi, "Big data analytics for logistics and transportation," 2015 4th Int. Conf. Advanced Logistics and Transport (ICALT), Valenciennes, 2015, pp. 311–316.*

40. *N. Yu, S. Shah, R. Johnson, R. Sherick, M. Hong and K. Loparo, "Big data analytics in power distribution systems," 2015 IEEE Power & Energy Society Innovative Smart Grid Technologies Conf. (ISGT), Washington, DC, 2015, pp. 1–5.*

41. *U. S. Kameswari and I. R. Babu, "Sensor data analysis and anomaly detection using predictive analytics for process industries," 2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI), Kanpur, 2015, pp. 1-8.*

42. *Y. Pradhananga, S. Karande and C. Karande, "CBA: Cloud-Based Bigdata Analytics," 2015 International Conference on Computing Communication Control and Automation, Pune, 2015, pp. 47–51.*

43. *S. Prabhu, A. P. Rodrigues, Guru Prasad M S and Nagesh H. R., "Performance enhancement of Hadoop MapReduce framework for analyzing BigData," 2015 IEEE Int. Conf. Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, 2015, pp. 1–8.*

44. *B. Mathew, P. Pattnaik and S. D. Madhu Kumar, "Efficient information retrieval using Lucene, LIndex and HIndex in Hadoop," 2014 IEEE/ACS 11th Int. Conf. Computer Systems and Applications (AICCSA), Doha, 2014, pp. 333–340.*

45. *G. Kejela, R. M. Esteves and C. Rong, "Predictive Analytics of Sensor Data Using Distributed Machine Learning Techniques," 2014 IEEE 6th Int. Conf. Cloud Computing Technology and Science, Singapore, 2014, pp. 626–631.*