RESEARCH ARTICLE OPEN ACCESS

Big Data Analysis: Concepts, Tools and Applications

Poonam
Assistant Professor
A.S.College - Khanna

ABSTRACT
Nowadays, Data has been growing exponentially which is leading explosion of data. With the enhancement of technology a variety of enormous data is being generated at an extremely fast speed in various sectors. So this can simply be termed as big data. We can create 2.5 quintillions of data everyday and this data comes from various sources and has different formats. Apache Hadoop, Apache Spark, MangoDB, NOSQL are tools to handle big data. Therefore analyzing big data has become crucial and inevitable. Big data analytics is being adopted throughout the globe in order to gain numerous benefits from the data being produced. Big data analytics examines large and different types of data to uncover hidden patterns, correlations and other insights. This paper describes a brief summary of its types, tools and applications of big data analytics.
Keywords **–** Big data, Big Data Analytics, Big Data Tools, Hadoop, HDFS, MapReduce, Big Data Analytics Applications

I. INTRODUCTION

Due to use of IOTs and social media sites we are generating a large amount of data which is huge in size and not present in structured manners. Due to the data explosion caused to digital and social media, data is rapidly being produced in such large chunks. Google, Facebook, Netflix, LinkedIn, Twitter and all other social media platforms clearly qualify as big data technology centers. So it has become challenging for enterprises to store and process it using conventional methods. These are the biggest factor for evolution of big data. Big data is a term for collection of dataset so large and complex that it becomes difficult to process using traditional database system. Big data is a combination of structured, semi structured and unstructured data collected by organizations that can be mined for information. It is the massive amount of data that cannot be stored, processed and analyzed using traditional tools. Enterprises must implement modern business intelligence tools to effectively capture, store and process such large amount of data in real-time. Big data is in raw form and is no meaningful to us so we must done meaningful insight to it in order to benefit from this data. It is done by analyzing the data which is known as big data analytics.

II. BIG DATA ANALYTICS

Big data analytics is a process of collecting, organizing and analyzing large set of data. Big data analytics is the often complex process of examining big data to uncover information. The Big Data analytics lifecycle generally involves identifying, procuring, preparing and analyzing large amounts of raw, unstructured data to extract meaningful information that can serve as an input for identifying patterns, enriching existing enterprise data and performing large-scale searches. It is process to extract meaningful information from big data. Big data analytics has quickly drawn the attention of IT industry due to its application in majority of areas like healthcare, business firms, social media, education, banking [1] etc. Big data analytics helps in quicker and better decision making in organizations. Big data analytics technologies and techniques provide a means to analyze data sets and take away new information which can help organizations to make informed business decisions. Different kinds of organizations use data analytical tools and techniques in different ways. Regardless of what big data is generated from, the reality comes into challenges is to bring value to it. With the availability of advanced Big Data analyzing technologies namely, NOSQL Databases, BigQuery, Map Reduce, Hadoop, perceptions and understandings can be better achieved to enable in improving the business policies and the decision-making process [2].

A. Characteristics of Big Data

Big data is the data characterized by 10 parameters [3]. These are shown in fig 1.

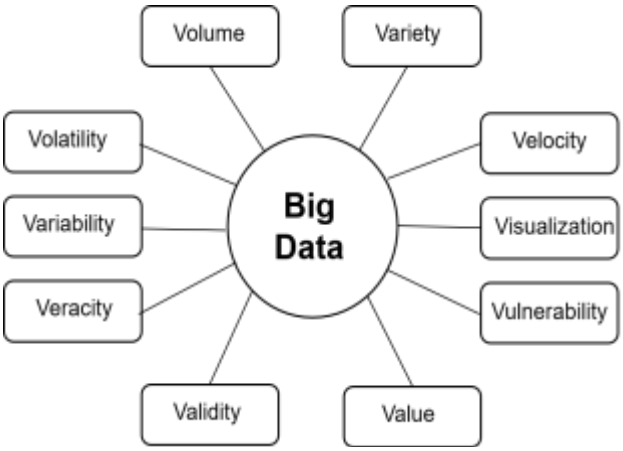


Fig 1: 10 Vs of Big Data

1. Volume: **-** Volume represents enormous amount of data that is produced. Today data is generated from various sources in different formats. Volume of data is generated exponentially. By 2020 it is expected data is rise up to 44 zettabytes. New big data tools used distributed system so that we store analyze data across databases that are dotted around anywhere in the world.

2. Velocity: - Velocity refers to the speed at which the data is generated, collected and analyzed. This is mainly due to IOTs, mobile data, social media etc. Technology allows us now to analyze the data while it is being generated, without ever putting it into databases.

3. Variety: **-** Variety refers to nature of data that is structured, semi-structured and unstructured data. In past we focus only on structured data but in fact 80% of data is unstructured which is in the form of text, video, image etc. With big data technology we can now analyze and brings together data of different types.

4. Veracity: - It refers to the assurance of quality/integrity/credibility/accuracy of the data. Since the data is collected from multiple sources, we need to check the data for accuracy before using it for business insights.

5. Value: **-** It is most important term in context of big data. We know data is huge. It has to be converted into a form from where it can be used to make analysis otherwise it will be useless.

6. Validity: **-** Validity refers to how accurate and correct the data is for its intended use. The benefit from big data analytics is only as good as its underlying data, so you need to adopt good data governance practices to ensure consistent data quality, common definitions, and metadata.

7. Variability: **-** In big data's context, Variability refers to a few different things. The number of inconsistencies in the data is one. These are required to be found by anomaly and outlier detection methods for any meaningful analytics to occur.

8. Volatility: **-** Due to the velocity and volume of big data the volatility of data needs to be considered carefully. Proper rules are established for data currency and availability so that rapid retrieval of information is done when it is required. With big data the costs and complexity of a storage and retrieval process are magnified.

9. Vulnerability: **-** Big data brings new issues regarding security. In any case, with big data, a data breach is a major violation.

10. Visualization: **-** Visualization is a new challenging characteristic of big data i.e. How to visualize the big data using current visualization tools. Because of the limitations of in memory technology and low response time, functionality, and poor scalability existing big data visualization tools visage technical challenges.

B. Types of Big Data Analytics

There are four general categories of analytics that are distinguished by the results they produce; descriptive analytics, diagnostic analytics, predictive analytics and prescriptive analytics. Different kinds of organizations use data analytics tools and techniques in different ways. The values and complexity relations between different analysis types are shown in fig 2.

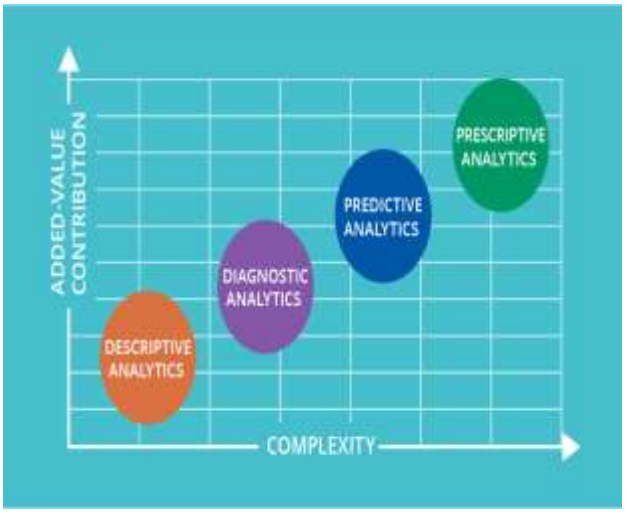


Fig 2: Value and complexity relation with analysis type [4]

1) Descriptive Analytics: **-** Descriptive analytics answers the questions what has happened. It uses data aggregation in data mining techniques to provide insight into past and then it answers what is happening now based on incoming data. It summarizes the data into a form that is understood able by humans. Google Analytical Tool is best example of descriptive analysis. This analytics helps in creating reports like company's revenue, profit, sales and so on.

2) Diagnostic Analytics: **-** Diagnostic Analysis is used to determine why something happened in the past. So it is characterized by techniques like data discovery, data mining and correlation to diagnostic analytics it takes deeper look at data. It is helpful in data mining what kinds of factors and events contributed to particular outcomes.

3) Predictive Analytics: **-** Predictive analysis uses statistical techniques and focus on techniques to understand the future. It predicts what future outcomes are. It looks into historical and present data to make predictions of the future. So predictive analysis provides the companies with actionable insight based on data. So through sensors and other machine generated data can identify when a malfunction is likely to occur.

4) Prescriptive Analytics: **-** Prescriptive Analysis uses optimization and simulation algorithm to advice on the possible outcomes and answer the questions what should we do. So basically it allows the users to prescribe a number of different possible actions and then guide them towards a solution. Thus, for prescriptive analytics, organizations optimize their business process models based on the feedback provided by predictive analytic models [5]. The result of the analytics is dependent upon the characteristics of data gathered.

III. BIG DATA ANALYTICS TOOLS

The word "Big data" can be applied to a dataset which increases at very intense rate. And it becomes difficult to store and process that data. Hence big data analytics is where the use some advance techniques which are applied on big data sets. There is a variety of tools that are used for Analytics of Big Data. The tools used for the same purpose can be categorized into different stages of lifecycle of big data as shown in Table I. That is based on their usage and implementation [6].

TYPES OF BIG DATA ANALYSIS TOOLS

TABLE I

	Data Collection tools	**Data Storage tools**	**Data Filtering and Extraction Tools**	**Data Cleaning and Validation Tools**
1	Semantria	Apache HBase (Hadoop database)	Import.io	OpenRefine
2	Opinion Crawl	Oracle NOSQL Database	OctoParse	DataCleaner
3	Trackur	MangoDB	ParseHub	MapReduce
4	OpenText	Apache Cassandra	Mozenda	Rapidminer
5	SAS Sentiment Analysis	CouchDB	Content Grabber	Talend

A) Data Collection Tools: **-** Data Collection Tools plays an important role in big data life. Some of most important tools for data collection are Semantria, Opinion Crawl, Open Text and Trackur.

1) Semantria: **-** Semantria is a cloud-based text and sentiment analysis tool offered by Lexatics. This tool is designed to help businesses collect tweets, texts, and other comments from their clients and analyze them to acquire highly valuable and actionable insights. The main benefits of Semantria are its tools and features that allow the users to gain reliable and actionable insights, its customizable features, and MS Excel Compatibility.

2) Opinion Crawl: **-** Opinion Crawl is an online web sentiments tool which is used for current events, companies, products and people. Users can enter a topic and get an adhoc sentiment assessment related to that topic. Users can get a pie chart which shows current real time sentiments. All these concepts allow the users to check what issues are derived in sentiments in a positive or negative way [7].

3) Trackur: **-**Trackur is a tool that is used to collect the information. It uses its automated sentiment analysis to look at the specific keywords that the users are supervising and after then decisions are carried out. The sentiment may be positive, negative or may be neutral with the related document. In Trackur algorithm, it could be used to observe the social sites and can outline news, to collect information through the trends and automated sentiment analysis.

4) Open text: **-** The Open Text is Sentiment Analysis module. It is a special type of engine that is used in classification to find out various subjective patterns. It is also used to evaluate the expressions of sentiment that is present in text form. First of all the analysis work is done at the topic level, sentence level, and document level. Its prime function is to acknowledge whether parts of text are realistic [8].

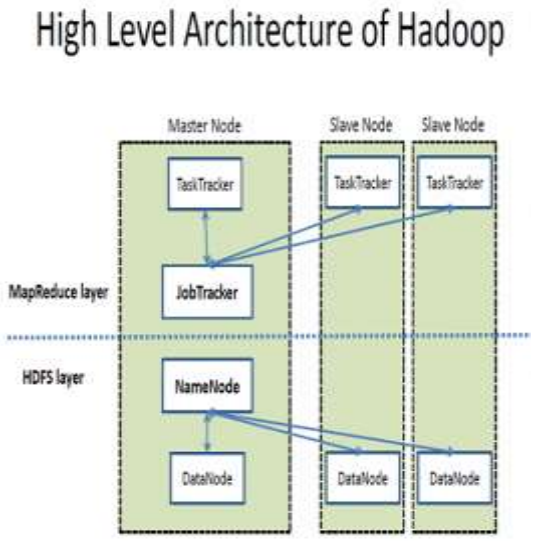
5) SAS Sentiment Analysis: **-** SAS is also sentiment analysis tool that automatically extract sentiments in real time. It performs this task with the help of various statistical modeling techniques.

B) Data Storage tools: **-** One of most important challenge of big data is how we can store it. A good storage tool provides a place to store, query and analyze big data. Some of these are as follows:

1) Apache Hadoop: **-** Hadoop is an open-source framework that is written in Java and it provides cross-platform support. Apache HADOOP is a framework used to develop data processing applications which are executed in a distributed computing environment. It works in environment that provides distributed storage and computing across clusters of computers. Hadoop runs application using the map reduces where data is processed in parallel with others. Hadoop runs the application that could perform complete statistical analysis on huge amount of data.

There are two core services which Hadoop provides:

- ***Hadoop MapReduce:*** -MapReduce is a computational model and software framework for writing applications which are run on Hadoop. These MapReduce programs are capable of processing massive data in parallel on large clusters of computation nodes [9].
- ***HDFS:*** **-** HDFS is main part of Hadoop as it provides reliable means for managing big data. It was closely related to MapReduce. When HDFS takes the data it breaks the information down into separate blocks and distributes them into different nodes in a cluster. It employs NameNode and DataNode architecture to implement distributed file system that provides high performance access to data highly scalable Hadoop clusters.



2) NOSQL: **-** NOSQL is an alternative to traditional databases which does not require any kind of fixed table schema like SQL. The original intention of NOSQL is the modern web scale databases. NOSQL database can also be referred to as structured storage which consists of relational database has a subset. Compared to relational database NOSQL are more superior and provides superior performance. It can scale out the data easily and has share nothing architecture which is capable of running on large number of nodes. It also provides nonlocking concurrency mechanism. There are four different types of NOSQL database – key value store, column based store, document based store and graph based store.

3) MangoDB: **-** MangoDB is a document database that provides high performance, high availability, and easy scalability. It is a cross-platform document-oriented database system classified as a NOSQL database. The indexing in case of MangoDB is done using document key structure. It bridges the gap between key-value and traditional RDBMS systems. It provides flexibility during initial phase of development and design. It is a database that supports online real time applications.

4) Cassandra: **-** Apache Cassandra is the leading NOSQL distributed data management systems that drive many of today's modern business applications by offering continuous availability, high scalability and strong performance security. Cassandra handles large amount of data with its distributed structure. The main goal of Cassandra is to handle big data workloads across multiple nodes without any single point of failure. Cassandra has multiple nodes in a

cluster which are identical in terms of their software architecture. All the nodes are symmetric and do not need a master node. This feature allows linear scalability.

5) CouchDB: **-** Apache CouchDB is an open source document oriented NOSQL. CouchDB is also a clustered database that allows you to run a single logical database server on any number of servers or VMs.A CouchDB cluster improves on the single-node setup with higher capacity and high-availability without changing any APIs.

C. Data Filtering and extraction Tools: **-** Data filtering and extraction tools are used to create structured output from unstructured data gathered in various stages. Some of these are as follows:

1) Import.io:- It is the one of the best and most reliable web scarping software on internet. If you want to scrape the contents from different web pages and have short of time then you can use this tool. This tool allows you to perform multiple data scraping tasks at a time. The most interactive feature of import.io are web crawling, secure login and data extraction. You can import the contents to Google sheets, Excel and plot.

2) Octoparse: **-** Octoparse is a cloud-based web crawler that helps you easily extract any web data without coding [10]. Octoparse is an ultimate tool for data extraction which allows you turn the whole internet into a structured format. It provides an easy user friendly interface which can easily deal with any type of websites. It is powerful tool to deal with dynamic websites and interact with many sites in various ways.

3) ParseHub: **-** ParseHub is the web browser extension that turns your dynamic websites into APIs. It also converts poorly structured websites into APIs without writing a code. Parsehub is supported in various systems such as Windows, Mac OS X, and Linux. It works with any interactive pages and easily searches through forms, opens dropdowns, logins to websites, clicks on maps and handles sites with infinite scroll, tabs, and pop-ups, etc.

4) Mozenda: **-** Mozenda is the most powerful and advanced data scraping and web extraction tool. It is best known for its user friendly interface. Mozenda is suitable for programmers, webmasters, journalists, scholars and enterprises. You can easily scrape, manage and store your data without compromising on quality. Mozenda has different interactive options and features to ease your work. This tool takes the hassle out of publishing data. You just have to highlight the contents and Mozenda will publish it to your site automatically.

5) Content Grabber: **-** Content Grabber is the best choice if you want to extract your data by web scraping and web automation. This tool ensures the provision of scalable and readable data. Content Grabber fixes all the minor errors in your data and is the next evolution in data scraping technology. This software can handle travels portals and new websites easily.

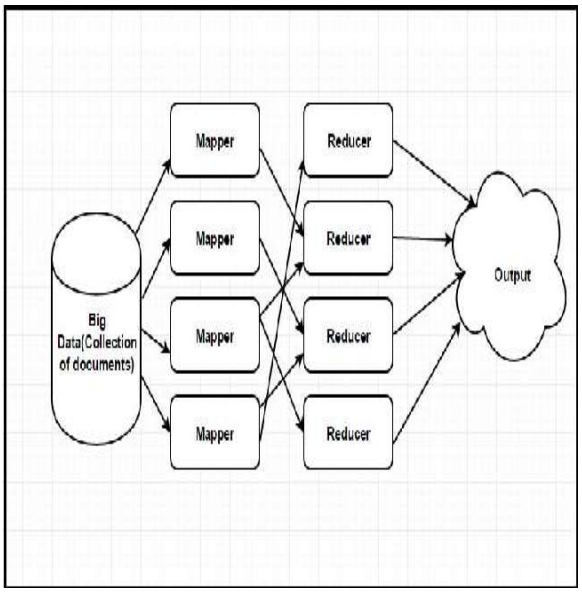
D. Data Cleaning & Validation Tool: **-** Data Cleansing is the act of detecting and correcting corrupt or inaccurate records from a record set, table or databases. Data cleaning tools are very helpful because they help in minimizing the processing time. The goal of data cleansing is not just to clean up the data in database but also brings consistency to different sets of data. They also reduce the computational speed of data analytics tools. Various validation rules are used to confirm the necessity and relevance of data extracted for analysis. Sometimes it may be difficult to apply validation constraints due to complexity of data.

1) Open Refine: **-** Open Refine which is formally known as Google Refine is a free open source tool. It is a powerful tool for working with messy data, cleaning it and transforms it from one format to another. It is extremely powerful for exploring, cleaning and linking data. It is a sophisticated tool for working on big data and performs analytics. Open Refine define explore data feature that explore large dataset with very ease. The clean and transform feature enables to clean big data and transform it from one form to another. It also provides reconcile and match data feature that extends the dataset with several web services. Open Refine always keeps your data private on your own computer until you want to share or collaborate.

2) DataCleaner: **-** DataCleaner is a data quality analysis application and a solution platform. It has strong data profiling engine. DataCleaner is a tool that is integrated with Hadoop. Data transformation, validation and reporting are its main features. It is a tool which is an application for data quality analysis. There is a profiling engine in its core to profile the data. This can be extensible by adding data cleansing, transformations, reduplication, matching merging and enrichment. It profiles and analyses the database within minutes, discovers patterns with the Pattern Finder, finds frequency of data using Value Distribution profile, filters the contact details, detects duplicates by using fuzzy logic, Merge the duplicates values etc.

3) MapReduce: - MapReduce is a powerful paradigm for parallel computation. Hadoop uses MapReduce to execute jobs on files in HDFS. Hadoop will intelligently distribute computation over clusters. During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster. The MapReduce algorithm contains two important tasks, namely Map and Reduce. MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

- **Map stage** − the map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
- **Reduce stage** − this stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.



4) RapidMiner: - With RapidMiner Studio accessing, loading and analyzing of any type of data is possible. Data can be both structured and unstructured like text, still images, and media. RapidMiner is really fantastic to perform fast ETL processes and work on your data as you want, no matter what is the source. You will really save a lot of time when you learn how to use it. You can create mining analysis with several algorithms, and thanks to add-ons, you can apply a lot of techniques.

5) Talend: - Talend is very good tool for quickly data integration and it makes the development time so short and easy. Talend provides a unified approach that combines rapid data integration, transformation, and mapping with automated quality checks. It is well suited for all kinds of data migration between various systems.

IV. APPLICATIONS OF BIG DATA ANALYTICS

The Big Data analytics is indeed a revolution in the field of Information Technology. The use of Data analytics by the companies is enhancing every year. There are various factors where big data analytics is actively used. Some of these are explained s below. Big data analytics helps organizations to work with their data efficiently and use that data identify new opportunities. Different techniques and algorithms can be applied to predict from data.

Applications of Big Data Analytics

A. Ecommerce: - The first sector of big data analytics is Ecommerce. Nearly 45% of world is online and they can create a lot of data every day. Big data can be used smartly in the field of Ecommerce by predicting customer trend, forecasting demand, adjusting price and so on. Online retailers can have opportunity better shopping experience and generate high sales if big data analytics is used properly.

B. Marketing: - Big data does not lead to have high marketing strategy. Meaningful insights need to be

derived from it in order to make right decisions. By analyzing big data we can have personalized marketing campaigns which can result in better and higher sales. Multiples business strategies can be applied for future success of the company and that leads to smarter business moves, more efficient operations and higher profits.

C. Education: **-** The next biggest where big data analytics used is Education. The usability of big data is also increased in educational sector. There are new options for research and analysis using data analytics. In the field of education depending on market requirements new courses are to be developed. The market requirements needs to be analyzed with respect to scope of course and according to scope new courses are to be developed. Hence to a analyze market requirement and to develop new courses big data analytics is used.

D. Healthcare: **-** There are number of uses of big data analytics in the field of healthcare. One of its uses is to predict patient health issues with big data analytics. With the help of patient's health history big data analytics is used to predict how likely they are to have particular health issues in future.

E. Media & Entertainment: **-** In the field of media & entertainment big data analytics is used to understand the demand of shows, movies, songs and so on to deliver a personalized recommendation lists to its users.

F. Banking: - There are various uses of big data analytics in banking sectors. One of its uses is risk Management. By using big data analytics there are many advantages. Big data analytics is used for risk management. Risk management is an important concept in any organization especially in the field of banking. Risk management analyzes a series of measures which helps the organization to prevent any sort of unauthorized activities. In addition to risk management it also used to analyze customer income and expenditures. It helps the bank to predict if a particular customer is going to choose for various bank offers like loans, credit card schemes and so on. This way the bank is able to identify the right customer who is interested in its offers.

G. Telecommunication: **-** Big data analytics is used in the field of telecommunication in order to gain profit. Big data analytics can be used to analyze network traffic and call data records. It can also improve service quality and customer experience in the field of telecommunication.

H. Government: **-** Big data analytics has been used widely in the field of government in all over the world. In the field of law enforcement big data analytics can be used to analyze all the available data in order to understand crime pattern. Intelligent services can use predictive analytics to forecast the crime that could be committed. The police department was able to reduce crime rate using big data analytics. With the help of data police could identify whom to target, where to go and how to investigate crime. Big data analytics helps them to discover pattern of crime in emerging area.

V. CONCLUSIONS

In the present scenario there are millions of sources which generate data very rapidly. These data sources present across the world. All that data together make big data. We must derive meaningful insights to it in order to make benefit from big data. It is done by analyzing big data which is known as big data analytics. This paper presents the concepts of big data analytics, its types and various applications of big data analytics. So big data analytics is being adopted by various organizations which help in quicker and better decisions making in organization.

REFERENCES

1. M. Chen, S. Mao, and Y. Liu, "Big data: a survey", Mobile Networks and Applications, vol. 19, No. 2, pp. 171–209, 2014.
2. Shweta Sinha, "Big Data Analysis: Concepts, Challenges And Opportunities", International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN: 2347-5552, Volume-8, Issue-3, May 2020.
3. https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx
4. Erl, T., Khattak, W. and Buhler, P., 2016. Big data fundamentals: concepts, drivers & techniques. Prentice Hall Press
5. Bihani, P. and Patil, S.T., 2014, "A comparative study of data analysis techniques", International journal of emerging trends & technology in computer science, 3(2), pp.95-101.
6. J.Nageswara Rao, M.Ramesh, "A Review on Data Mining & Big Data, Machine Learning Techniques" , International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-7 Issue-6S2, April 2019, pp 914-916.
7. http://www.opinioncrawl.com/aboutOpinionCrawl.htm
8. Ritu Ratra, Preeti Gulia, "Big Data Tools and Techniques: A Roadmap for Predictive

Analytics", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-9 Issue-2, December, 2019

9. Dr. Urmila R. Pol, "Big Data Analysis Using Hadoop Mapreduce", American Journal of Engineering Research, e-ISSN: 2320-0847 p-ISSN : 2320-0936 Volume-5, Issue-6, pp-146-151
10. Online source, [Available] https://www.octoparse.com/blog/yes-there-is-such-thing-as-a-free-web-scraper/, 2018.