A Survey on Big Data Analytics Techniques

Ratnavali. V [1], Ramalakshmi.G [2]

[1], [2] Assistant Professor/ Department of Computer Science, Kamaraj College, Thoothukudi, Tamilnadu, India

Abstract- **Recently, big data plays a major role in Information Technology. It is described as VVVVV which is Volume, Variety, Velocity, Variability, and Veracity. Big data analytics system is to collect data first and then retrieve useful patterns for gain some profit in organization. Big data gives a lot of opportunities to great progress in many fields. Social networks, search engines are also tracking user information and analyze user web behavior. Ecommerce gets improved day by day through big data analytics. Researcher uses various techniques to store and analyze the data. The aim of this survey paper is to overview the techniques for analyzing big data**

Keywords: Big Data, Big Data Analytics, Algorithm

I. INTRODUCTION

In Traditional days it is impossible to analyze petabytes of data using RDBMS.But now a days we can analyses large set of structured and unstructured data using various big data tools such as hadoop, Cassandra, rapid miner etc. Large amount of data easily get flow in social media like Facebook, twitter, instagram, whatsapp etc. Every data are stored in cloud using hadoop framework and get analyses using big data tools to analyses user online behavior. Big data used in various fields like business, healthcare, social media etc.Data analyst and researcher analyzing the data using various techniques and tools. Hence this survey paper analyses some of the big data techniques used in various applications.

II. BIG DATA AN OVERVIEW

Big data the word coined first time by US Chief Scientist (SGI-Silicon Graphics) John R. Mashey in 1998 in his presentation slide "Big Data and the Next Wave of Infrastress ".The Data mining book named "Big Data" written by Weiss and Indrukya also get introduced in the year 1988. The first academic paper having the term Big Data get appeared in the year 2000 by Diebold.

Big Data which is a collection of largest volume of structured and unstructured data. Promptly almost all the fields require big data usage. Data will be found everywhere. Data is increasing day by day in internet which is collected, stored and then it get analyzed. Some of the industries impelled by big data analytics are –Public Sector Services, Healthcare contributions., Learning Services, Insurance Services, Industrialized ,Natural Resources, Transportation Services, Banking Sectors and Fraud Detection. It is extended use in the field of medicine and healthcare.

The Data which come under the big data are
Black Box Data: It records voices of the cabin crew, microphones and earphones, and the functional information of the aircraft.
Social Media Data: Social networking sites store information and the views posted by millions of users across the globe.
Stock Exchange Data : The stock exchange data have the information about the ecommerce view of the customers data
Transport Data : Transport data includes hold information about model of the vehicle, capacity ,distance and overall performance of that vehicle
Search Engine Data : Search engines fetch data from the database according to user relevant search query

III. CHARACTERISTICS OF BIG DATA

In 2001, Gartner analyst Doug Laney define big data with 3 vs. they are volume, velocity, variety.

Volume:
Which is the large amount of data get stored in database everyday in internet world.

Velocity:
Which is how rapidly data gets flow everyday in internet is known as velocity

Variety:
Data gets stored in different formats such as text, audio, video, spreadsheet and unstructured database

Additional vs. are describing now a days they are Variablity, Veracity and value [7]

Variability: Variability which refers to the consistency of data

Veracity: Veracity which refers the quality of data.Reliablity of data gets collected and analyzed

Value: Worth of the data which gets extracted. The data gets extracted can be monetized.

IV. STRUCTURED AND UNSTRUCTURED DATA

Structured data which contains particular format that is easy for machine to get understand. It is easily searchable by using some basic algorithms. It is easily to get stored and analysed.Generally, these types of data gets generated in two types they are Machine generated and Human generated. Machine generated data are arriving from medical devices, sensors and web server logs. Human generated data ,the data get manually to store in database such as age,gender,zipcode .These information are generally get stored in SQL database. Examples for structured data are spreadsheet, SQL.

Unstructured data plays major role in big data. It have own internal structure .It don't get stored in Particular format. It is complex to handle with traditional database like SQL.So various big data tools are used to handle this type of data. Unstructured data such as web logs, multimedia content, email, customer service interactions, sales automation, and social media data, mobile applications, location services, and IOT. Most Business data in internet also in unstructured data format

V. BIG DATA ANALYTICS

Important Big data Analytics need to be considered are

1) Predictive analytics
2) Descriptive analytics
3) Prescriptive analytics
4) Diagnostic analytics

1. Predictive Analytics

It uses both new and historical data and predict what may happen in future. It is an advanced analytic technique to analyze the unknown future events, behaviors and trends. It uses statistical algorithm and machine learning techniques to analyze the historical events. It is used in Business fields to identify risk and opportunities. It is used in various fields like marketing, finance, travel and health care. Regression analysis is an important technique to analyses the data in predictive analysis. Regression technique are used predict the data such as age, weight, distance and temperature. [21] Regression analysis are used in application like Trend analysis, bio medical and financial forecasting .Various algorithms are used in regression analysis generally it considered Linear model and Support vector machines.

2. Descriptive Analytic

[22] It summarizes raw data and makes it something that is interpretable by humans. Which use data aggregation and data mining to provide insight into the past and answer:"What happened?" Or what is happening? Descriptive analysis analyses the real time and historical data and it gets analyses and approach the future. Descriptive analysis are social networks to analyses the number of likes, comments, followers and posts .Statistics of pages per view, average response time, post response are get counted by simple arithmetic operation in this analysis. Descriptive analysis determines past success and failure data. Descriptive analysis is used in sales, marketing and finance fields to analyses their business strategy. It identifies the relationship between customer and products.

3. Prescriptive Analytics

It relates both predictive and descriptive analysis. In predictive analysis it describes what may happen .In descriptive analysis aims to provide insight what happened .In prescriptive it helps to determine the best solution by using known parameters. In prescriptive analysis process new data and improve the accuracy of forecasting and provide better decisions. Techniques which are implemented in prescriptive analysis are optimization, simulation, and game theory and decision-analysis methods.

4. Diagnostic Analytics

Diagnostic analytics measured historical data against other data to attempt to understand the causes of events and behaviors. This is also called root cause analysis. In diagnostic analysis it answers the question why did it happen? It examines the data deeper and analyses the cause of events and behaviors. Diagnostic analysis used the past data to determine what happened and why. Attribute importance, principle components analysis, sensitivity analysis and conjoint analysis are some examples of diagnostic analysis.

VI. BIG DATA ANALYTICS TECHNIQUES

According to IDC Canada, a Toronto-based IT research firm, Big Data have used various analytical techniques
Some of the techniques are
1) Association rule learning
2) Sentimental analysis
3) Social network analysis

1. ASSOCIATION RULE LEARNING

Association rule learning is a method for discovering relationship of unknown hidden pattern in big data. It is used in supermarket to discover user choice of their frequent items in dataset. It is an unsupervised machine learning method. Using this algorithm we can analyses customers purchasing behavior. Different types of algorithms are used to perform association rule mining .Frequently used algorithm are apriori, Fp growth and maximum frequent item set algorithms.

Apache Spark- Apriori algorithm are not suitable for large datasets so it is implemented using spark technology[1].Spark is an open source cluster computing framework to run on large set of data. It handles real time and batch processing application. It run on Hadoop, Mesos, Kubernetes, standalone, or in the cloud. It is a tool for running spark applications. It is faster in accessing and handling data.

2. SENTIMENTAL ANALYSIS

Sentimental Analysis is the process of text analyst to mine the opinion. It is also known as opinion mining. Politicians perform sentimental analysis to know what the opinion about their politics and their policy is. Through this analyses we can know positive and negative opinion on social networking sites like face book, twitter. It identifies the new internet slangs, emoji, sarcasm words, characters that are used in social media .Algorithm used in sentimental analysis are Support Vector Machine (SVM), Naïve Bayes and Maximum Entropy ML algorithms.ML algorithms are used in classify the data which is positive or negative.

The polarity measurement is handling with the help of ML techniques. Polarized opinion get stored in new dictionary format [3][4].That results then get implement in hadoop environment. Pig tool is a famous tool for the sentimal analyst to get a result.

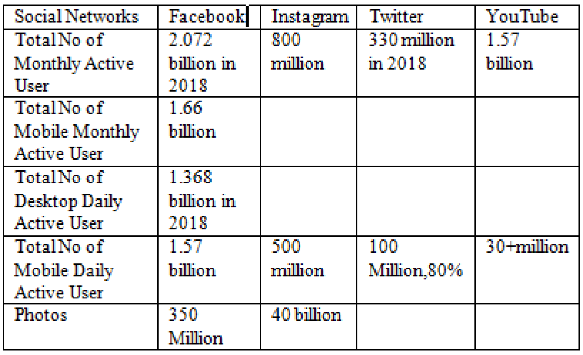
Apache Pig-Pig is a tool which is used in hadoop. Pig is used to analyses large set of data It provide high level language. Pig Latin which is used to read, write and process the data. Programmers need to write script in Pig Latin to analyses the data in pig .pig has a component known as pig engine which is then convert the pig Latin script in to map reduce forms. Using pig Latin programmers perform map reduce tasks. Apache pig uses Multi query approach to reduce the length of the code. Pig tool is used in applications like to process data in web logs, search platforms and to process time sensitive loads. pig tool is used by sentimental analyst to analyses the positive, negative, neutral data in social media.

3. Social Network Analysis

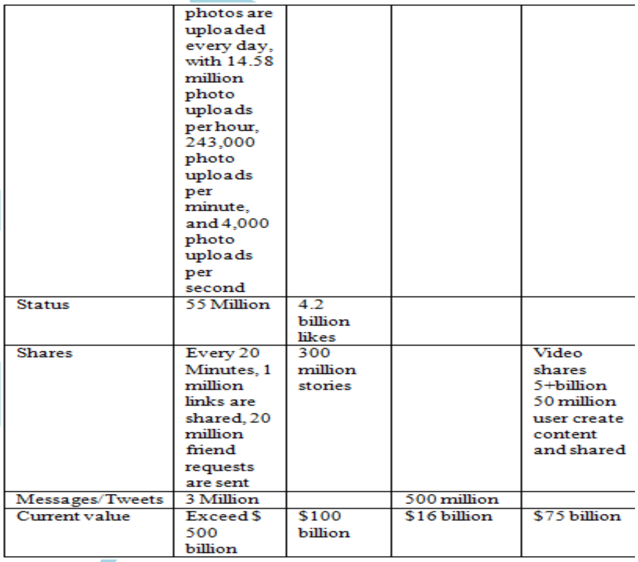
Social network can be analyzed by computer technologies [12][13][14][15].Tools such as Map reduces, No SQL, Hadoop gets supports. Big data analysis is growing day by day in this modern world. According to survey from some source [6][7][8]. Number of search everyday on Google - 3.5 billion, which equates to 1.2 trillion searches per year worldwide. In 2017 46.8% of the global population accessed the internet and by 2021.

It is projected to grow to 53.7%. For each and every 60 seconds on Facebook 510,000 comments are posted. It states that 293,000 statuses are updated, and 136,000 photos are uploaded. Every second 4.75 billion pieces of content shared daily as of May 2013 which is a 94 % increase from August 2012. At 2.07 billion, Facebook has more 2.07 million user monthly active users than WhatsApp (500 million), Twitter (284 million) and Instagram (200 million).

According to some source Whatsapp used by 1.5 billion users.300 Million daily active user using Whatsapp Status. Average number of Voice call made on whatsapp is 100 million. Average number of video call made on whatsapp is 55 million call per day.60 Billion of message getting send everyday by active users.1 billion videos are sending everyday and 4.5 billion photos are send everyday through whatsapp.Linkedin used by 500 million users. In 2018, Facebook have 2.072 billion users. Total number of mobile monthly active users 1.66billion.



Social Networks	Facebook	Instagram	Twitter	YouTube
Total No of Monthly Active User	2.072 billion in 2018	800 million	330 million in 2018	1.57 billion
Total No of Mobile Monthly Active User	1.66 billion			
Total No of Desktop Daily Active User	1.368 billion in 2018			
Total No of Mobile Daily Active User	1.57 billion	500 million	100 Million,80%	30+million
Photos	350 Million	40 billion		



	photos are uploaded every day, with 14.58 million photo uploads per hour, 243,000 photo uploads per minute, and 4,000 photo uploads per second			
Status	55 Million	4.2 billion likes		
Shares	Every 20 Minutes, 1 million links are shared, 20 million friend requests are sent	300 million stories		Video shares 5+billion 50 million user create content and shared
Messages/Tweets	3 Million		500 million	
Current value	Exceed $ 500 billion	$100 billion	$16 billion	$75 billion

VII. BIG DATA TOOLS

Serious challenge of big data is cleaning, analyses, collect and store the large volume of data. It can overcome by various big data tools. This section describe big data tools which were used in various analytical techniques.

a) Hadoop:

Doug Cutting, Mike Cafarella and team refer the solution produced by Google and started an Open Source Project called HADOOP in 2005. Doug's one of the sons named his toy Elephant as Hadoop. He placed that name as Hadoop for his open Source project because it is easy to pronounce and Google. Hadoop is an open-source java based framework that helps to store and analyses big data in a distributed environment over clusters of computers using simple programming models. It uses a master or slave structure. It runs applications using the Map Reduce algorithm, where the data get processed in parallel on different CPU nodes. Large data sets can be processed by using hadoop .Even in case of node failures it continue its normal operation .It help in its rapid Transfer rate. This process helps to reduce the risk of entire system failure. Hadoop tool is Flexible, fault Tolerant and Cost effective. Hadoop is used by top most companies like Facebook, Google, Amazon, IBM etc., to maintain their large volume of data

b) Map reduces:
It is coupled with HDFS to organize large amount of data. Map reduce is the programming model to process large set of data using distributed process algorithm. Structured and unstructured data get in basic unit before it get inject into a map reduce model. It has two functions namely map function and reduce function. In this mapper, it takes single pair as input and produces any number of pair as output. In the reduce function, it takes all of the values associated with a single key and outputs any number of (key, value) pairs. Some of the application using map reduce function are Google, Amazon, yahoo and facebook.

c) Hive
It's a tool used to develop Sql like scripts to do in map reduce operations. Hive is an data warehouse infrastructure tool to process data. It is used to process structured data. It makes the analyzing easy. Hive was developed by facebook later it get by apache software foundation and get named as Apache hive. It Support query similar like SQL which is called HIVEQL or HQL.It stores data in schema and process its data in hdfs.It is fast, scalable and flexible in processing the data.

CONCLUSIONS

This survey paper described some of the important techniques and tool which are used in the current applications of big data. Techniques and tool which we examined in this paper are used by data analyst and data researcher to analyses big data. Day by day data is increasing so future of big data gets replaceable by the term fast data and actionable data.

REFERENCES:

1) m. Dolores Ruiz&Maria j. Martin-bautista Extraction of association rules using big data technologies, Carlos Fernandez-basso, Universidad de Granada, citic-ugr. Vol. 11, No. 3 (2016) 178–185

2) M. Edison, A. Aloysius -Concepts and Methods of Sentiment Analysis on Big Data. International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 5, Issue 9, September 2016

3) Sayali Zirpe and Bela Joglekar Polarity Shift Detection Approaches in Sentiment Analysis: A survey International Conference on Inventive Systems and Control (ICISC-2017)

4) L. Jeba Sheela "A Review of Sentiment Analysis in Twitter Data Using Hadoop", International Journal of Database Theory and Application, 2016, pp: 77-86.

5) https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

6) https://zephoria.com/top-15-valuable-facebook-statistics/

7)https://expandedramblings.com/index.php/whatsapp-statistics/

8)https://www.omnicoreagency.com/instagram-statistics/

9) Shuhui Jiang, Xueming Qian, Tao Mei, Yun Fu, Personalized Travel Sequence recommendation on Multisource Big Social Media, 2016, IEEE Transactions on Big Data, Vol.2, Issue: 1

10) Vallabh Dhoot, Shubham Gawande, Pooja Kanawade and Akanksha Lekhwani, Efficient Dimensionality Reduction for Big Data Using Clustering Technique, Imperial Journal of Interdisciplinary Research (IJIR), Vol-2, Issue-5, 2016,ISSN: 2454-1362

11) Manjit kaur, Urvashi Grag. ECLAT Algorithm for Frequent Item sets Generation. International Journal of Computer Systems (ISSN: 2394-1065), Volume 01– Issue 03, December, 2014.

12) [S. Fan, R. Y. K.Lau, and J. L. Zhao, "Demystifying Big Data Analytics For Business Intelligence through the Lens of Marketing Mix." Big Data Research, vol. 2, 2015, pp. 28-32.

13) S. Mithas, M. R., Lee, S., Earley, Murugesan, S. and R. Djavanshir, "Leveraging Big Data and Business Analytics [Guest editors' Introduction]." IT Professional, vol. 15, 2013, pp. 18-20.

14) Y. Zhao, D. Li, and L. Pan, "Cooperation or Competition: An Evolutionary Game Study between Commercial Banks and Big Data-Based E-Commerce Financial Institutions in China." Discrete Dynamics in Nature and Society, vol. 8, 2015.

15) C. K. Velu, S. E. Madnick, and M. W. Van Alstyne, "Centralizing Data Management with Considerations of Uncertainty and Information-Based Flexibility." Journal o Management Information Systems, vol. 30, 2013. Pp.179-212

16) https://www.elderresearch.com/company/blog/42-v-of-big-data

17) Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A Survey, Computer Science Review, 2015, Vol: 17, pp: 71-80

18) Katrina Sin and Loganathan Muthu, Applications of big data in education data mining and learning analytics – A literature Review, ICTACT Journal on soft computing special issue on Soft computing models for big data, July 2015, Vol: 05, Iss: 04, pp: 1035-1049

19) K. Krishnan, Data warehousing in the age of big data, in: The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013

20) Mike Barlow, Real-Time Big Data Analytics: Emerging Architecture, ISBN: 978-1-449-36421-2, 2013

21) Yun Wang and Sudha Ram," Predicting Location- Based Sequential Purchasing Events by Using Spatial, Temporal, and Social Patterns", IEEE Intelligent Systems, May/June 2015.

22) https://halobi.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/

23) Doreswamy and Channabassayya M Vastrad. Performance Analysis of Regularised Linear Regression Models for Oxazolines and Oxazoles Derivatives Descriptor Dataset. Int. J of Computational Science and Informational Technology (IJCSITY). Vol. No. 4. November 2013. Pg. 111-123.

24) A1bert Bifet "Mining Big Data In Real Time" Informatica 37 (2013) IS- 20 DEC 2012.