An Overview: Classification of Big Data Tools

R. Bhuvanya[1],
Assistant Professor,
Vel Tech Rangarajan

Dr. Sagunthala
R&D Institute of Science and Technology

P. Matheswaran[2],
Assistant Professor,
K.Ramakrishnan College of Technology

***Abstract* - As we know data is an essential term in various fields and without data it is not possible for the industry to operate and in order to show the continuous improvement the company will store the past data. Data analysis is a collective term of gathering, organizing and analyzing data for present and future improvements. And also manipulating and analyzing the large volume of data i.e big data is a complex process. And collecting data, analyzing, searching, storing, sharing of big data is a challenging tasks using modern big data analytics tools. This paper will suggest a solution for all the above mentioned challenges and the review of modern big data analytics tools are also described here.**

I. INTRODUCTION

Big data is a data sets which holds huge volume of data both structured and unstructured exceeding the range of Exabyte where the traditional methods of data processing software will find inadequate to deal with. In Business point of view, importance of big data can be well understood by how a organization utilizes the collected data instead of discussing how much data a company stores. And the data can be taken from any source like social interaction where the people can collect the reviews of product and process. Business interaction which will indicates the data produced as a result of business activities, Electronic files which deliberately mean the internet pages, video, audio, PDFs finally the sensor will also takes place while collecting temperature, bio metric, location etc.

By collecting all these data the company will be profited through various means such as,

1. Implementing new or innovation in product- By getting the recent trends and customer needs, satisfaction of the customer can be enhanced.

2. Dealing of online reputation: Though some Feedback the business people will get to know about their product. And it can be done based on the sentiment analysis. If the business people wants to improve their business in further online, then the big data tools will help in all this aspects.

3. Prediction of personality- With the help of sentimental analysis it is extremely useful for the data mining tool to predict about the personality based on positive and negative reviews. And even the prediction can be further extended to poetics by analyzing whether the person will win in the next election or not.

4. Reduction of time: Modern data analytical tools helps to identify new source of data which helps in business and analyzing can be done quickly which helps to make the decision based on prior learning.

5. Early identification of risk

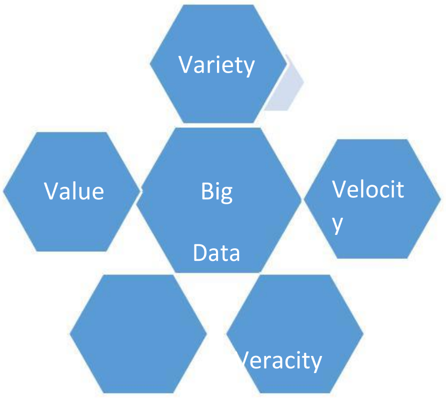
Barriers Involved:

1. If the data representation is Unstructured
– Data which is in the form of photographs, video file such as MP4 and audio file such as MP3 has not followed any meaningful order in the way of storing. And it is extremely difficult for the people to search for the particular data and to analyze it.

2. To deal with Authentication-When we try to break down the data barriers then certain unnecessary people may get access and they can do some modification over the particular data. Thus the authentication, security are a major drawback of Big Data Storage.

3. Storage Issue: Since we are dealing with extremely huge voluminous data, and also the data is getting overloaded in internet there is no recent or largest storage medium to process all at a time.

II. FIVE V'S OF BIG DATA:



Five V's such as Variety, Velocity, Volume, Veracity, and Value are generally termed to be the characteristics of Big Data.

Variety

The term variety indicates the different types of data that we collect for performing analysis. And this variety can be even structured, semi structured and at sometimes unstructured. This variety of data are collected from the social networking sites, customer feedback surveys, web logs and it may be in the form of text, image, audio and at times video. If the data format is incompatible, or if it is incomplete then it will lead to significant challenge while doing the analysis.

Velocity

As a term denotes ‘velocity’ in network it refers to the rate of data flow around the system. It means the amount of data entered as input and the extraction of data as output. Especially it becomes tedious process while we are dealing with banking transaction, social networking sites such as twitter, Facebook etc. Traditional method of data analytics tool will not perform this analysis process efficiently. Hence the modern analytics tools are introduced here to perform gathering and analyzing the data.

Volume:

It is most important feature of big data which refers to the quantity of data gathered by the particular organization. And the data acquisition can be any of the form like sensor, Reviews in Social network sites, Internet of Things, Web Pages etc.

Veracity

It deals with the degree of integrity. Since we are collecting data from various source we may not sure that all the collected data is accurate. Sometimes the data may be less accurate, low quality, less reliable and it may not be consistent all the time. But the introduction of modern analytics tools helps to achieve all these.

Value

Growth of the organization will be predicted based on the good delivered with high quality. And it also refers the ‘usefulness of data’ while making decisions. And it will be carried out by analyzing the organization’s data which in turn increases the profit of the particular organization.

III TOOLS AVAILABLE IN BIG DATA: DATA STORAGE TOOLS

1. HDInsight

Software Product of Microsoft which provides solution for the storage of big data where the Azure Blob Storage will be used as the default file. This also increases high availability of data with low cost.

2. Hadoop

Java Based free source software which store large amount of data. It has an ability to process large amount of data. It works by splitting the big data and the data will be distributed among many nodes.

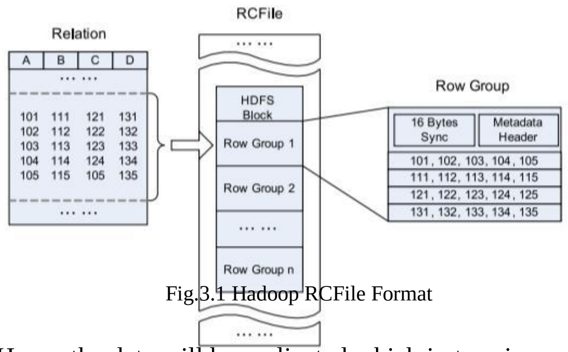


Fig.3.1 Hadoop RCFile Format

Hence the data will be replicated which in turn increases the availability of data.

3. Not Only SQL (NOSQL)

Traditional SQL deals with the large amount of structured data where the NoSQL helps for the large amount of unstructured data. It gives better performance in storing massive amount of data.

4. Hive

Tool for managing distributed data in Hadoop. It can be implemented by SQL like Query option (HSQL) to access large data. This tool is primarily used for data mining which runs on top of Hadoop. It depends on the map reducing technique for the fast retrieval of data.

5. Sqoop

In order to deal with the relational database and to transfer data but it helps to transfer large amount of structured data to Hadoop or Hive.

6. Presto

Social networking site of Facebook developed the open source query engine called Presto which helps to handle petabytes of data. It does not depend on the map reducing technique but it deals with the quick retrieval of data.

Data Visualization Tools:

1. Data Wrapper

One of the data visualization tool to prepare charts. Once the file is uploaded in the form of CSV/PDF this tool will generate a bar or line or map as the people want. This tool is extremely helpful for the reporters since it can embed the live charts in their article. And it is extremely helpful to produce well defined graphics.

2. Solver

To deal with financial reporting, budgeting and to analyses the past and present data in order to make the organization more profitable, solver will efficiently deal with all the above.

3. Tableau

One of the business intelligence software helps for the data visualization, data analytics and reporting. It can gather data from various data source such as spreadsheets, even big data which helps to perform the dynamic data analysis.

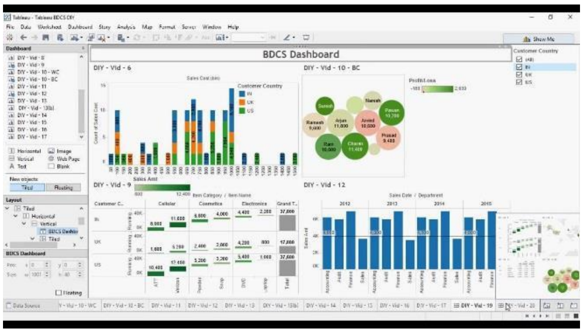


Fig. 3.2 Tableau Dashboard

Whatever the data may be, whether it is structured or unstructured organization has to perform the data analysis in order to come with best decision making. But it is not easier with traditional tools. Hence the new business intelligence software is introduced as ‘Tableau’ which makes the youngsters to perform data analytics much easier.

4. Fusion Tables from Google

To deal with the data analysis and to perform data set visualization, mapping Google fusion table is an incredible tool .It can merge two or three tables to generate a single visualization that incorporates the two sets of data.

5. Infogram

In order to represent interactive item of maps and to visualize the data more pleasantly infogram offers service for all the mentioned above. It can create various types of charts includes column, bar, pie chart. To grab the attention of audience we can even add a map to our report analysis.

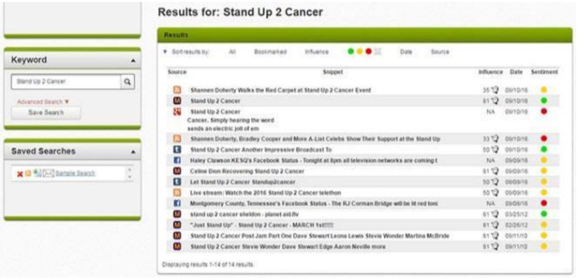
Sentiment Analysis Tools

1. Open text

In order to deal with the expressions written in the textual content and to identify the patterns of subject this tool will add its remark in analyzing the sentiment. It is performed at the topic level, sentence level and further it can be extended to the document level. Finally it refines all the above mentioned levels and the resultant can be expressed in terms of positive, negative or neutral.

2. Trackur

Especially used to perform the analysis on social media and checks whether the sentimental keyword obtained is positive, negative or neutral.



Results for: Stand Up 2 Cancer

Results

Sort results by: All Bookmarked Influence ● ● ● Date Source

Source	Snippet	Influence	Date	Sentiment
	Shannen Doherty Walks the Red Carpet at Stand Up 2 Cancer Event	35	09/10/16	●
	Stand Up 2 Cancer	61	09/10/16	●
	Stand Up 2 Cancer Cancer. Simply hearing the word sends an electric jolt of em	NA	09/10/16	●
	Shannen Doherty, Bradley Cooper and More A-List Celebs Show Their Support at the Stand Up	33	09/10/16	●
	Stand Up 2 Cancer Another Impressive Broadcast To	50	09/10/16	●
	Haley Clawson KESQ's Facebook Status - Tonight at 8pm all television networks are coming t	NA	09/09/16	●
	Celine Dion Recovering Stand Up 2 Cancer	61	09/09/16	●
	Let Stand Up 2 Cancer Standup2cancer	50	09/09/16	●
	Live stream: Watch the 2016 Stand Up 2 Cancer telethon	50	09/09/16	●
	Montgomery County, Tennessee's Facebook Status - The RJ Corman Bridge will be lit red toni	NA	09/08/16	●
	stand up 2 cancer sheldon - planet ait.flv	61	03/25/12	●
	"Just Stand Up" - Stand Up 2 Cancer - MARCH 1st!!!!!	61	02/26/12	●
	Stand Up 2 Cancer Post Jam Part One Dave Stewart Leona Lewis Stevie Wonder Martina McBride	61	09/11/10	●
	Stand Up 2 Cancer Stevie Wonder Dave Stewart Edge Aaron Neville more	61	09/11/10	●

Displaying results 1-14 of 14 results.

Fig. 3.3 Trackur Social Media monitoring

It can be done by monitoring all the social media news through the current trends, discovery of keywords given and performing automated sentiment analysis.

3. SAS Sentiment Analysis

It uses the combination of statistical modelling and natural language processing techniques associated with rule based. Extracting various types of data to judge about famous personality, about a product is termed to be sentiment analysis. Now the current trend is all about predicting the sentiment through facebook posts and tweets. If the people want to know about the personality eg: Actor Rajinikanth we can do a sentiment analysis on him to assess people’s negative and positive attitude. And we can do a prediction which helps for the election surveys to analyse whether he will win in the next election or not. Finally the classification model in sentiment analysis can be adjusted to the current trends.

4. Opinion mining through Crawl

One of the sentiment analysis tool for the ongoing trends of products and people. If the topic is entered by the user then the resultant of adhoc sentiment assessment will be retrieved. The resultant image consists of comparison chart of pie, recent news, some images. In addition it will display the tag concepts related to the particular search. So the people those who are involved in doing prediction they can extract it through recent ongoing topic and can estimate the sentiment which in turn results in the ratio of positive to negative feedback.

List of Open Source Software Tools

In addition to the above specified tools here we have some of the open source tools where the people can be benefited under various applications.

1. Apache Hadoop

To deal with the distributed data storage and processing Apache Hadoop will make tremendous performance. Since it has the large storage part which is termed to be Hadoop Distributed File System which can process dataset efficiently and quickly.

2. Cassandra

Well efficient in handling large amount of distributed database especially when the database is distributed among many server. Since it deals with the distributed data there is no possibility of failure. It use CQL(Cassandra Query language) instead of our traditional SQL.

3. KNIME

It is also known as Konstanz Information miner which includes various components of data mining and machine learning. It is especially used in health care for the purpose of pharmaceutical research, to perform data analysis in financial transaction.

4. Rapid Miner

It helps to perform data preparation, validation, and optimization. Makes use of graphical user interface to execute various workflow. Best suited for business and commercial applications.

5. Rattle, Orange

Rattle is developed using the R Programming language. It has the ability to run on various operating system such as Linux, Windows and Mac OS.

To deal with data science along with python programing ‘orange’ plays a vital role. It can deal with gadgets where we can drag and drop also it can be connected to different widgets. It uses the concept of machine learning algorithm to preprocess, clustering the data.

CONCLUSION

By using the tools of big data, we can analyze various type of data that helps to enhance different types of business. It plays a major role in the Banking, health care, Fast Moving Consumer Goods (FMCG) Products Sales, Fraud Detection, Customer response in online sales, etc. Hence study, analysis, and implementation of big data analytics has become a mandate which is possible with the help of above mentioned big data tools.

REFERENCES

[1] Mariam Adedoyin-Olowe1 et.al “A Survey of Data Mining Techniques for Social Media Analysis “

[2] Chu, Cheng, et al. "Map-reduce for machine learning on multicore." Advances in neural information processing systems 19 (2007)

[3] Groves, Peter, Basel Kayyali, David Knott, and Steve Van Kuiken. "The 'big data' revolution in healthcare." McKinsey Quarterly 2013.

[4] Shang, Weiyi, Zhen Ming Jiang, HadiHemmati, Bram Adams, Ahmed E. Hassan, and Patrick Martin. "Assisting developers of big data analytics applications when deploying on Hadoop clouds." In Proceedings of the 2013 International Conference on Software Engineering, pp. 402-411. IEEE Press, 2013.

[5] http://data-infomed.com/why-more-data-and simple algorithms-beat-complex-analytics-models/