THOMSON REUTERS ENDNOTE

[NC-Rase 18]
DOI: 10.5281/zenodo.1488685

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

A COMPREHENSIVE SURVEY ON BIG DATA ANALYTICS

Dr. Sumita U. Sharma
Dept. of Electronics, DCPE, HVPM, Amravati

ABSTRACT

In the past few years, there has been a rapid growth in innovative ideas and developments in the field of technology. The world has ushered into many new computing paradigms like pervasive computing, cloud computing, embedded programming, Internet of Things and many. But amongst all of them Data Analytics enjoys a special place. The term big data and data analytics has become the modern buzz word. Not only there are many myths and hypes about this new technology, it poses many challenges and issues. The current paper presents a detailed and comprehensive study on the data analytics, the techniques used with the challenges and the future of this multifaceted modern technology.

Keywords: *Zinc oxide, thin film, solar cell, photoconductivity.*

I. INTRODUCTION

The data has become the most important entity in the modern age of internet and communication technology. Enormous amounts of data is generated and uploaded on the web, every day and every hour. The data is generated from social networking applications, ecommerce applications, and web applications and by many private and government organizations. This massive explosion of data in terms of volumes, velocity and variety leads to a new business model [1]. Data storing, processing and applying various analytical and mining algorithms is the modern driving force for developing decision support systems and business intelligence to enhance the productivity and the business of organizations. Scores of data mining and analytics tools are available and are being used for the purpose of extracting productive knowledge from the large data stores. Appropriate analytic tools used on the datasets provide efficient means to arriving at the error free results [2].

II. THE BIG DATA

The era that we are living in, provides us with massive amounts of data that is readily available on hand to the data scientists. The term big data, in general, refers to datasets which are not only big in size but are also generated at a very high velocity and comes with large variety. Storing and organizing such data makes is also a complex task and cannot be handled using conventional tools and techniques. The question is how big should be the size of data to call it big data. The big data sizes are continuously increasing, and today they range from dozen terabytes (TB) to petabytes (PB) in a single data set. Furthermore, there are many difficulties related to big data like capturing data, data storage, search techniques, sharing of data, and the analytics[3]. The big data includes a very high volumes of data generated at high velocity and it can be of variety of forms. The data stored and managed as big data can be of three types

1. Structured Data – Relational Database
2. Semi structured data – XML data
3. Unstructured Data – Word, Pdf, excel, files, images, and other kinds of multimedia data

Software tools like Hadoop and its related tools have been emerged to handle the big data in a distributed computing environment. Hadoop is as an Open Source software framework developed in Java by Apache Software Foundation for data storage, batch processing and distributed processing of big data applications. Hadoop comes with a set of core elements viz Hadoop Common, MapReduce, Yarn and HDFS. The commercially available Hadoop framework bundles with large number of components each used for a specific functionality on the bid data. The commonly used Hadoop components are Pig, Hive and Spark that are popularly used.

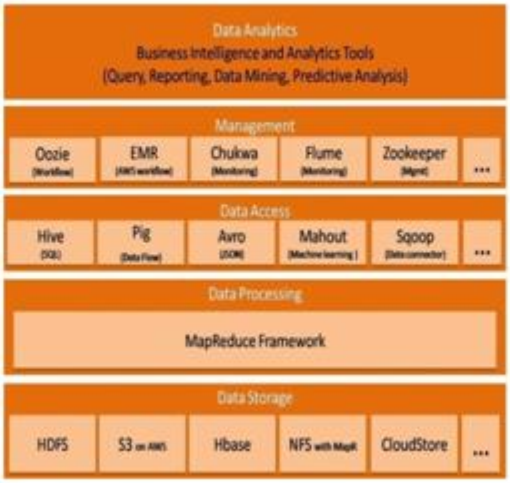

Fig.1. Hadoop Component Ecosystem

III. BIG DATA ANALYTICS

The big data analytics aims at the processing unstructured data generated from application like web and mobile-banking transactions, social networking data contents like blog posts, tweets, online searches, and images which can be transformed into valuable business information using computational methods to disclose trends and patterns between datasets[4].

Types of data analytics

a. Descriptive analytics
The descriptive analytics is an early step of data processing that generates a summary of the historical data to yield useful information and possibly prepare the data for future analysis. It answers the question of what happened. Descriptive analytics mines the raw data from multiple data sources to provide valuable insights into the past. The results thus generated just indicate that there is something is wrong or right but does not explain why. The companies that deal with large amount of historical data do not restrict themselves with descriptive analytics but prefer combining it with other types of data analytics.

b. Diagnostic analytics
The second type of analytics known as diagnostic analytics is a an advanced analytics which examines data or contents to answer the question "Why did it happen?". It is characterized by techniques such as data discovery, data mining and correlations. Diagnostic analytics takes a deeper look at data to attempt to understand the causes of events and behaviors.

c. Predictive analytics
Predictive Analytics makes use the of historical data to forecast future consumer behavior and trends. It is the use of historical data to predict future trends. Statistical models and machine learning algorithms are extensively used in this type of analytics to recognize patterns and learn from the past data. Predictive analysis is also process that employs machine learning to analyze data and make future predictions. Most of businesses prefer the use of predictive analytics to develop strategic marketing campaigns for future.

d. Prescriptive analytics
The main idea of prescriptive analytics is to actually prescribe what action needs to be taken to reduce and rectify a future problems. It also takes a full advantage of a promising trend. An example of this analytics is to identify opportunities for repeat purchases based on customer analytics and sales history.

IV. DATA ANALYTICS TOOLS

The data mining is a process of mining and analyzing huge amounts of data from the datasets and extracting useful information. The data mining techniques provides answers to many business questions using various algorithms which would have taken long time. The data mining can be applied to various business segments like recognizing the customers buying products, online fraud detection, market trends analysis. Main steps involved in the mining process are as follows.

a. Data Pre-processing
The first step of data mining involves activities like removing noise and anomalies from the dataset. It also makes sure to fill the missing values in the datasets and normalizing the data.

b. Clustering of data
This step breaks huge data into smaller groups or the sub classes.

c. Classification of data
At this stage the data is tagged or classified into user defined categories based on the specific needs of the application and the type of dataset.

d. Outlier analysis
It helps in identifying the items which deviate from rest of the items in the dataset. It also detects anomalies in the data.

e. Associative analysis
It helps in building the relationship between the items of the dataset. It also helps in predicting the occurrences of the items in the dataset.

f. Regression
Regression employs method for predicting the values of a dependent items by constructing a model or a mathematical function out of independent items.

g. Summarization
It summarizes the process of data mining and yields a compact model of the large dataset.

There are large number of ready to use tools available for data mining and executing various mining algorithms on the datasets. These algorithms include pattern recognition, statistical methods, and machine learning techniques. Some of the commonly used software tools used for data mining and mostly used by data scientists and researchers are Weka, RapidMiner, KMine, SAS, Orange, Apache Mahout. Using these tools, a user can process large dataset by applying various mining algorithms without actually writing programs, since these tools like Weka comes with a large numbers of in-built mining algorithms. Almost all the data mining tools come under the category of open source software and can be freely downloaded and used for mining process. There are many programming languages also available used in the fields of data analytics and mining like Java, Python, R, SCALA, and most modern Julia.

V. CONCLUSION

It is expected that the techniques of data analytics radically change the way we live in the modern age and do business in the future. Today, we make use of analytics for generating results and provide us with right decision support system in our lives and the businesses. It is also expected that data analytics will make what is impossible, possible. But the data analytics has just begun and long way to go as we are currently at an early stages of the data era.

ISSN 2348 – 8034
Impact Factor- 5.070

REFERENCES

1. *ChetanaTukkoji , Dr. Seetharam, Assistant Professor, Dept of CSE, GITAM University, Bangalore, A Comprehensive Survey on Big-Data Issues, Challenges and Management Approaches on Cloud Environment, International Journal of Advanced Research in Computer and Communication Engineering ISO 3297:2007 Certified Vol. 6, Issue 2, February 2017.*
2. *Vijayaraj ; R. Saravanan ; P. Victer Paul ; R. Raju, A comprehensive survey on big data analytics tools, 2016 Online International Conference on Green Engineering and Technologies (IC-GET), 978-1-5090-4556-3.*
3. *Nada Elgendy and Ahmed ElragalNadaElgendy and Ahmed Elragal, Big Data Analytics: A Literature Review Paper, Department of Business Informatics & Operations, German University in Cairo (GUC), Cairo, Egypt {nada.el-gendy,ahmed.elragal}@guc.edu.eg, August 2014, DOI 10.1007/978-3-319-08976-8_16ISSN 0302-9743*
4. *Jasmine Zakir, Tom Seymour, Kristi Berg, BIG DATA ANALYTICS, Issues in Information Systems Volume 16, Issue II, pp. 81-90, 2015.*
5. *LONGBING CAO, University of Technology Sydney, Australia, Data Science: A Comprehensive Overview, Longbing Cao. 2017. Data science: A comprehensive overview. ACM Comput. Surv. 50, 3, Article 43 (June*
6. *2017), 42 pages. DOI: http://dx.doi.org/10.1145/3076253*
7. *KautkarRohit, M. Tech Scholar, Computer science and technology, Maharashtra Institute of Technology (MIT), Aurangabad, Maharashtra, India , A COMPREHENSIVE SURVEY ON DATA MINING, IJRET: International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308*