INTERNATIONAL JOURNAL OF PURE AND APPLIED RESEARCH IN ENGINEERING AND TECHNOLOGY

A PATH FOR HORIZING YOUR INNOVATIVE WORK

ANALYSIS OF BIG DATA

MS. SHAKEEBA S. KHAN, MS. SAKSHI .S. DESHMUKH

M.E. Scholar, Department of Computer Sci. & Engg., Prof. Ram Meghe Institute of Tech & Research Amravati, India

Accepted Date: 05/03/2015; Published Date: 01/05/2015

Abstract: Big Data is data that either is too large, grows too fast, or does not fit into traditional architectures. Within such data can be valuable information that can be discovered through data analysis [1]. Big data is a collection of complex and large data sets that are difficult to process and mine for patterns and knowledge using traditional database management tools or data processing and mining systems. Big Data is data whose scale, diversity and complexity require new architecture, techniques, algorithms and analytics to manage it and extract value and hidden knowledge from it. Big Data includes structured, semi-structured and unstructured data such as call logs, mobile-banking transactions, online user-generated content such as blog posts and Tweets, online searches, satellite images, etc. As the size of data increases, the amount of irrelevant data usually increases as well and the process becomes impractical. Hence, in such cases, the analyst must be capable of focusing on the informational parts while ignoring the noise data. In this paper, we examine the current trends and characteristics of Big Data, its analysis and challenges.

Keywords: Big Data, Knowledge Discovery, Analytical Challenges, Human Resources

Corresponding Author: MS. SHAKEEBA S. KHAN

PAPER-QR CODE

Access Online On:

www.ijpret.com

How to Cite This Article:

Shakeeba S. Khan, IJPRET, 2015; Volume 3 (9): 328-333

INTRODUCTION

The definition of big data refers to groups of data that are so large and unwieldy that regular database management tools have difficulty in capturing, storing, sharing and managing the information. Big Data refers to the massive amounts of data that collect over time that are difficult to analyze and handle using common database management tools. Big Data includes business transactions, e-mail messages, photos, surveillance videos and activity logs. Big Data also includes unstructured text posted on the Web, such as blogs and social media. Big data refers to a process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach called big data, which uses massive parallelism on readily-available hardware [2, 3].

CHARACTERISTICS OF BIG DATA

Big Data can be described by the following characteristics [4]:

1. Volume (Scale of data) – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered as Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.
2. Variety (Different forms of data) - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data. [5]
3. Velocity (Analysis of streaming data)- The term 'velocity' in this context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.[5]
4. Veracity (Uncertainty of data) - This is a factor which can be a problem for those who analyze the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.[5]
5. Complexity- Data management can become a very complex process, especially when large volumes of data come from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp the information that is supposed to be conveyed by these data. This situation, is therefore, termed as the 'complexity' of Big Data. [5]

DIG DATA ANALYSIS

Data analysis can also be described as knowledge discovery from data. Knowledge discovery is a method where new knowledge is derived from a data set. More accurately, knowledge discovery is a process where different practices of managing and analyzing data are used to extract this new knowledge [6, 7].

- **Data Collection**

The first step in the data processing pipeline is data collection. In this step all data that is to be processed is consolidated for analysis. Difficulties with data collection lie in the different forms that data may have as they arrive from different sources.

- **Data Cleaning**

After collection, data cleaning is performed. There may be data that is either noisy, erroneous or missing values. Data cleaning uses different methods to eliminate this bad data from the dataset. After cleaning, data may need to be transformed as final preparation for analytics.

- **Data Analysis**

After data processing the analysis can begin. In this stage, many different analytic methods and techniques may be performed. These methods and techniques can be broken down into three categories: statistical analysis, data mining and machine learning. Statistical analysis creates models for predication and summarizes datasets. Data mining uses a variety of techniques (clustering, classification, etc.) to discover patterns and models present in the data. Machine learning is used to discover relationships that are present within the data.

CHALLENGES IN BIG DATA

Big data presents a number of challenges; the challenges in Big Data are usually the real implementation hurdles which require immediate attention. Any implementation without handling these challenges may lead to the failure of the technology implementation and some unpleasant results. There are numerous challenges, from privacy and security to access and deployment such as[8,9]:

- **Privacy and Security**

It is the most important challenges with big data which is sensitive and includes conceptual, technical as well as legal significance.

The personal information (e.g. in database of a merchant or social networking website) of a person when combined with external large data sets, leads to the inference of new facts about that person and it's possible that these kinds of facts about the person are secretive and the person might not want the data owner or any person to know about them.Big Data used by law enforcement will increase the chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated.

- **Data Access and Sharing of Information**

If the data in the companies information systems is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete and timely manner. This makes the data management and governance process bit complex adding the necessity to make data open and make it available to government agencies in standardized manner with standardized APIs, metadata and formats thus leading to better decision making, business intelligence and productivity improvements. Sharing of data between companies is awkward because sharing data about their clients and operations threatens the culture of secrecy and competitiveness.

- **Analytical Challenges**

The main challenging questions are as:

- What if data volume gets so large and varied and it is not known how to deal with it?
- Does all data need to be stored?
- Does all data need to be analyzed?
- How to find out which data points are really important?
- How can the data be used to best advantage?

Big data brings along with it some huge analytical challenges. The type of analysis to be done on this huge amount of data which can be unstructured, semi structured or structured requires a large number of advance skills. Moreover the type of analysis which is needed to be done on the data depends highly on the results to be obtained i.e. decision making. This can be done by using one of two techniques: either incorporate massive data volumes in analysis or determine upfront which big data is relevant.

- **Human Resources and Manpower**

Since Big data is at its youth and an emerging technology so it needs to attract organizations and youth with diverse new skill sets. These skills should not be limited to technical ones but also should extend to research, analytical, interpretive and creative ones. These skills need to be developed in individuals hence requires training programs to be held by the organizations.

CONCLUSION

In this universe large amount of information is being handled and exchange. In order to get knowledge of Big Data, analysis of several challenges at the data, model, and system levels are necessary. Here various characteristics, challenges and analysis of Big Data technique for data mining are discussed. Development of safe protocols for management of Big Data is one of the challenging tasks.

REFERENCES

1. State of Big Data Analysis in the cloud: http://dx.doi.org/10.5539/nct.v2nlp62

2. Marr, B. (2013, November 13). The Awesome Ways Big Data is used Today to Change Our World. Retrieved November 14, 2013, from LinkedIn: https://www.linkedin.com/today/post/article/ 20131113065157-64875646-the-awesome-ways-big-data-is-used-today-tochange- our-world

3. http://searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

4. http://venturehire.co

5. Kale Suvarna Vilas, Big Data Mining October 2013,ijcsmr.

6. Agrawal, D., Bernstein, P., Bertino, E., Davidson, S., Dayal, U., Franklin, M. Widom, (2012). Challenges and Opportunities with Big Data. http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf

7. Begoli, E., & Horey, J. (2012). Design Principles for Effective Knowledge Discovery from Big Data. Software Architecture (WICSA) and European Conference on Software Architecture (ECSA), 2012 Joint Working IEEE/IFIP Conference on (pp. 215-218). http://dx.doi.org/10.1109/WICSA-ECSA.212.32

8. Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, Challenges, Tools and Good Practices. IEEE, 404-409.

9. Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013). Big Data: Issues and Challenges Moving Forward. International Conference on System Sciences (pp. 995-1004). Hawaii: IEEE Computer Society.

10. Big Data Analytics-www.datameer.com