Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 9, September 2014, pg.56 – 68

RESEARCH ARTICLE

ANALYSIS OF BIG DATA

[1]Anshul Sharma, [2]Preeti Gulia

[1]M.Tech Scholar, CSA Department, Maharshi Dayanand University, Rohtak

[2]Assistant Professor, CSA Department, Maharshi Dayanand University, Rohtak

anshu.sharma215@yahoo.com; preetigulai81@gmail.com

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. These kinds of difficulties complicate the analysis of multichannel data as compared to the analysis of single-channel data. In this paper, we examine the current trends and characteristics of Big Data, its analysis and how these are presenting challenges in data collection, storage and management [1], HACE theorem that characterizes the features of the Big Data revolution.

I. 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].

II. CHARACTERISTICS OF BIG DATA

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

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.

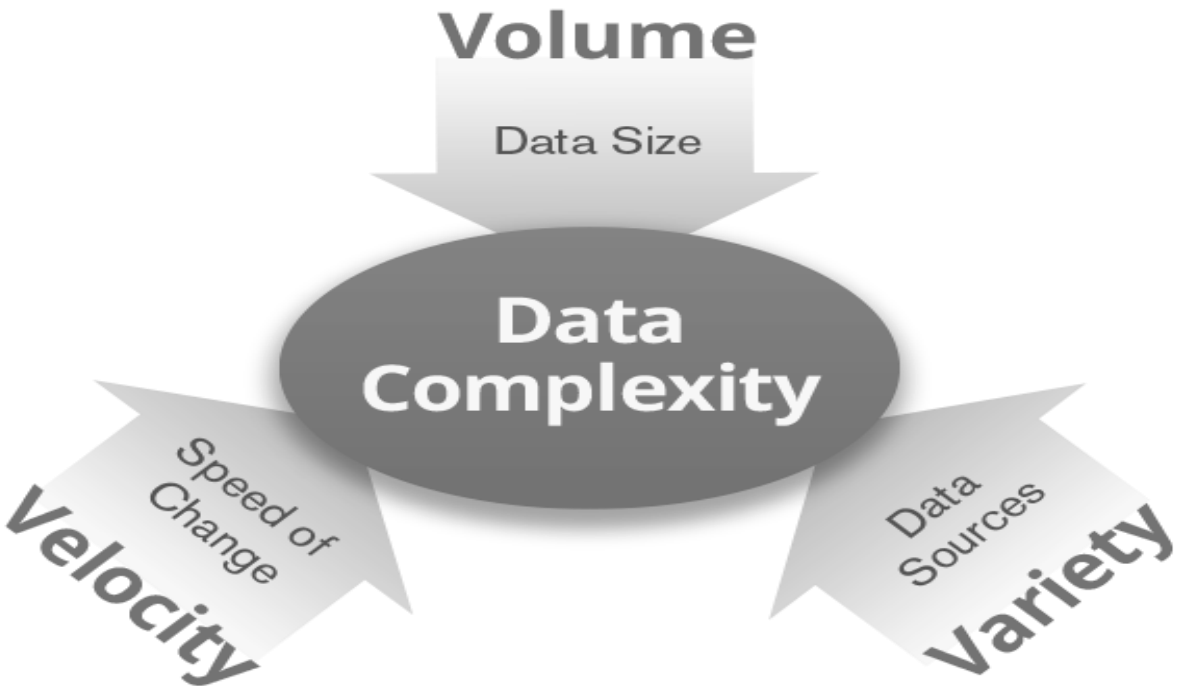

Fig. 3.1 Three Important Characteristics of BIG DATA [10]

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]

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]

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]

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]

III. THE SEVEN STEPS OF BIG DATA DELIVERY

- Collect: Data is collected from the data sources and distributed across multiple nodes – often a grid – each of which processes a subset of data in parallel.
- Process: The system then uses that same high-powered parallelism to perform fast computations against the data on each node. Next, the nodes reduce the resulting data findings into more consumable data sets to be used by either a human being (in the case of analytics) or machine (in the case of large-scale interpretation of results).

- Manage: Often the big data being processed is heterogeneous, originated from different transactional systems. Nearly all of that data needs to be understood, defined, annotated, cleansed and audited for security purposes.

- Measure: Companies will often measure the rate at which data can be integrated with other customer behaviors or records, and whether the rate of integration or correction is increasing over time. Business requirements should determine the type of measurement and the ongoing tracking.

- Consume: The resulting use of the data should fit in with the original requirement for the processing. For instance, if bringing in a few hundred terabytes of social media interactions demonstrates whether and how social media data delivers additional product purchases, then there should be rules for how social media data is accessed and updated. This is equally important for machine-to-machine data access.

- Store: As the "data-as-a-service" trend takes shape, increasingly the data stays in a single location, while the programs that access it move around. Whether the data is stored for short-term batch processing or longer-term retention, storage solutions should be deliberately addressed.

- Govern: Data governance encompasses the policies and oversight of data from a business perspective. As defined, data governance applies to each of the six preceding stages of big data delivery.

IV. HACE THEOREM

This is theorem to model Big Data characteristics. According to it Big Data have the following characteristics:

Heterogeneous, **A**utonomous, **C**omplex & **E**volving

Big data starts with large volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. These characteristics make it an extreme challenge for discovering useful knowledge from the Big Data[5].

A. Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the huge volume of data represented by heterogeneous and diverse dimensionalities. This is because different information collectors use their own schemata for data recording, and the nature of different applications also results in diverse representations of the data. For example, each single human being in a bio-medical world can be represented by using simple demographic information such as gender, age, family disease history etc. For X-ray examination and CT scan of each individual, images or videos are used to represent the results because they provide visual information for doctors to carry detailed examinations. For a DNA or genomic related test, microarray expression images and sequences are used to represent the genetic code information because

this is the way that our current techniques acquire the data. Under such circumstances, the heterogeneous features refer to the different types of representations for the same individuals, and the diverse features refer to the variety of the features involved to represent each single observation. Different organizations (or health practitioners) may have their own schemata to represent each patient, the data heterogeneity and diverse dimensionality issues become major challenges if we are trying to enable data aggregation by combining data from all sources.

B. Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Being autonomous, each data source is able to generate and collect information without involving (or relying on) any centralized control. This is similar to the World Wide Web (WWW) setting where each web server provides a certain amount of information and each server is able to fully function without necessarily relying on other servers. On the other hand, the enormous volumes of the data also make an application vulnerable to attacks or malfunctions, if the whole system has to rely on any centralized control unit. For major Big Data related applications, such as Google, Flicker, Facebook, and Walmart, a large number of server farms are deployed all over the world to ensure nonstop services and quick responses for local markets. Such autonomous sources are not only the solutions of the technical designs, but also the results of the legislation and the regulation rules in different countries/regions. For example, Asian markets of Walmart are inherently different from its North American markets in terms of seasonal promotions, top sell items, and customer behaviors. More specifically, the local government regulations also impact on the wholesale management process and eventually result in data representations and data warehouses for local markets.

C. Complex and Evolving Relationships

While the volume of the Big Data increases, so do the complexity and the relationships underneath the data. In an early stage of data centralized information systems, the focus is on finding best feature values to represent each observation. This is similar to using a number of data fields, such as age, gender, income, education background etc., to characterize each individual. This type of sample-feature representation inherently treats each individual as an independent entity without considering their social connections which is one of the most important factors of the human society. Such social connections commonly exist not only in our daily activities, but also are very popular in virtual worlds. For example, major social

network sites, such as Facebook or Twitter, are mainly characterized by social functions such as friend-connections and followers (in Twitter). The correlations between individuals inherently complicate the whole data representation and any reasoning process. In the sample-feature representation, individuals are regarded similar if they share similar feature values, whereas in the sample-feature-relationship representation, two individuals can be linked together (through their social connections) even though they might share nothing in common in the feature domains at all. In a dynamic world, the features used to represent the individuals and the social ties used to represent our connections may also evolve with respect to temporal, spatial, and other factors. Such a complication is becoming part of the reality for Big Data applications, where the key is to take complex (non-linear, many-to-many) data relationships, along with the evolving changes, into consideration, to discover useful patterns from Big Data collections.

V. BIG 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 integration is later performed to keep data as cohesive as possible.

Data Cleansing

After collection, data cleansing or 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.

VI. 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]:

(i) 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.

- Information regarding the people is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of.

- Another important consequence arising would be Social stratification where a literate person would be taking advantages of the Big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse.

- 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.

(ii) 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.

(iii) 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.

(iv) 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. Moreover the Universities need to introduce curriculum on big data to produce skilled employees in this expertise.

(v) Technical Challenges

***Fault Tolerance*:** With the incoming of new technologies like Cloud computing and big data it is always intended that whenever the failure occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch. Fault-tolerant computing is extremely hard, involving intricate algorithms. It is simply not possible to devise absolutely foolproof, 100% reliable fault tolerant machines or software. Thus the main task is to reduce the probability of failure to an "acceptable" level. Unfortunately, the more we strive to reduce this probability, the higher the cost. Two methods which seem to increase the fault tolerance in big data are as:

First is to divide the whole computation being done into tasks and assign these tasks to different nodes for computation.

Second is, one node is assigned the work of observing that these nodes are working properly.

If something happens that particular task is restarted. But sometimes it's quite possible that that the whole computation can't be divided into such independent tasks. There could be some tasks which might be recursive in nature and the output of the previous computation of task is the input to the next computation. Thus restarting the whole computation becomes cumbersome process. This can be avoided by applying Checkpoints which keeps the state of the system at certain intervals of the time. In case of any failure, the computation can restart from last checkpoint maintained.

Scalability: The scalability issue of big data requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goal of each workload cost effectively. It also requires dealing with the system failures in an efficient manner which occurs more frequently if operating on large clusters. There has been a huge shift in the technologies being used. Hard Disk Drives (HDD) are being replaced by the solid state Drives and Phase Change technology which are not having the same performance between sequential and random data transfer. Thus, what kinds of storage devices are to be used; is again a big question for data storage.

Quality of Data: Collection of huge amount of data and its storage comes at a cost. More data if used for decision making or for predictive analysis in business will definitely lead to better results. Business Leaders will always want more and more data storage whereas the IT Leaders will take all technical

aspects in mind before storing all the data. Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results and conclusions can be drawn. This further leads to various questions like how it can be ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate or not to draw conclusions from it etc.

Heterogeneous Data: Unstructured data represents almost every kind of data being produced like social media interactions to recorded meetings to handling of PDF documents, fax transfers, to emails and more. Working with unstructured data is cumbersome and of course costly too. Converting all this unstructured data into structured one is also not feasible.

Structured data is always organized into highly mechanized and manageable way. It shows well integration with database but unstructured data is completely raw and unorganized.

VII. BIG DATA ADVANTAGES

The Big Data has numerous advantages on society, science and technology. It is onto the way that how it is used for the human beings. Some of the advantages are described below[8]:

Understanding Customers

This is one of the biggest and most publicized areas of big data use today. Here, big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browse logs as well as text analytics and sensor data to get a more complete picture of their customers. The big objective, in many cases, is to create predictive models.

Understanding and Optimizing Business Process

Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts. HR business processes are also being improved using big data analytics. This includes the optimization of talent acquisition, as well as the measurement of company culture and staff engagement using big data tool.

Improving Science and Research

Science and research is currently being transformed by the new possibilities big data brings.

For example, CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator. Experiments to unlock the secrets of our universe – how it started and works - generate huge amounts of data. The CERN data centre has 65,000 processors to analyse its 30 petabytes of data. However, it uses the computing powers of thousands of computers distributed across 150 data centers worldwide to analyze the data. Such computing powers can be leveraged to transform so many other areas of science and research.

Improving Healthcare and Public Health

The computing power of big data analytics enables us to decode entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns. The clinical trials of the future won't be limited by small sample sizes but could potentially include everyone.

Optimizing Machine and Device Performance

Big data analytics help machines and devices become smarter and more autonomous. For example, big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings. Big data tools are also used to optimize energy grids using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouses.

Financial Trading

High Frequency Trading (HFT) is an area where big data finds a lot of use today. Here, big data algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make buy and sell decisions in split seconds.

Improving Security and Law Enforcement

Big data is applied heavily in improving security and enabling law enforcement. The revelations are that the National Security Agency (NSA) in the U.S. uses big data analytics to foil terrorist plots (and maybe spy on us). Others use big data techniques to detect and prevent cyber-attacks. Police forces use big data tools to catch criminals and even predict criminal activity, credit card companies use big data to detect fraudulent transactions.

VIII. CONCLUSION

Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. Big data can be characterized by 3Vs: the extreme volume of data, the wide variety of types of data and the velocity at which the data must be must processed. Although big data doesn't refer to any specific quantity, the term is often used when speaking about petabytes and exabytes of data, much of which cannot be integrated easily. Because big data takes too much time and costs too much money to load into a traditional relational database for analysis, new approaches to storing and analyzing data have emerged that rely less on data schema and data quality. Instead, raw data with extended metadata is aggregated in a data lake and machine learning and artificial intelligence (AI) programs use complex algorithms to look for repeatable patterns[2,3].

Advances in data storage and mining technologies make it possible to preserve increasing amounts of data generated directly or indirectly by users and analyze it to yield valuable new insights. For example, companies can study consumer purchasing trends to better target marketing. In addition, near-real-time data from mobile phones could provide detailed characteristics about shoppers that help reveal their complex decision-making processes as they walk through malls.

Big data can expose people's hidden behavioral patterns and even shed light on their intentions. More precisely, it can bridge the gap between what people want to do and what they actually do as well as how they interact with others and their environment. This information is useful to government agencies as well as private companies to support decision making in areas ranging from law enforcement to social services to homeland security.

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

