Big Data: Characteristics, Challenges and Data Mining

Walunj Swapnil K.
Research Scholar, MCA
Thakur Institute of Management Studies, Career Development and Research (TIMSCDR)

Yadav Anil H.
Research Scholar, MCA
Thakur Institute of Management Studies, Career Development and Research (TIMSCDR)

Sonu Gupta
Assistant Professor
Thakur Institute of Management Studies, Career Development and Research (TIMSCDR)

ABSTRACT

Big Data refers to the data or sets of records that are too large in volume to be operated using the existing database management tools and techniques. They are produced in many important applications, such as search engines, business informatics, social networks, social media, genomics, meteorology, and weather forecast. Big data presents a big challenge for database and data investigative research. The main objective of this paper is to give a brief introduction of Big Data, its architecture, characteristics and challenges.

Keywords

Acquisition, Modeling, Mining, Static Data, Dynamic Data.

1. INTRODUCTION

Recent advancement in technology has led to generation of a great quantity of data from distinctive domains over the past 20 years. Big data is a broad term for data sets so great in volume or complicated that traditional data processing applications are inadequate.[1] Although the big data have large amount of data or volume, it also processes the number of unique characteristics unlike traditional data. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set. For example, big data is usually unstructured and requires more time for analysis and processing. This development calls for new system architectures for data acquisition, transmission, storage, and large-scale data processing mechanisms.

Big Data is data that are enormous in size and exceeds the processing capacity of regular or traditional database systems. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The data is so enormous and are generated so fast that it doesn't fit the structures of normal or regular database architecture. To analyze the data new alternative way must be used to process it.

The hot IT buzzword of 2012, big data has become viable as cost-effective approaches have emerged to tame the volume, velocity and variability of massive data. Inside the data, valuable patterns and confidential information were kept hidden because of extensive amount of work to be done to extract the data. To leading corporations, such as Walmart or Google, this power has been in reach for some time, but at ridiculous amount of cost. Today's available hardware, cloud architectures and open source software makes big data available to everyone. Big data processing is extremely feasible and available for even the small garage startups, who can affordably and easily rent server time in the cloud.

The value of big data to an organization falls into two categories: analytical use, and enabling new products. Big data analytics can show insights kept confidential previously by data that cannot be afforded to be processed, such as peer influence among customers, revealed by analyzing shoppers' transactions, social and geographical data. If necessary in order to process every record of data in reasonable time it eliminates the unnecessary need for sampling and elevates an investigative approach to data, in contrast to the somewhat static nature of running predetermined records. Accuracy in big data may lead to more confident decision making.

The number of successful web startups in the last decade are great examples of big data being used as an enabler of new products and services. For example, by combining a large number of signals from a user's actions and those of their friends, Facebook has been able to craft a highly personalized user experience and create a new kind of advertising business. It's no coincidence that the major share of ideas and tools underpinning big data have emerged from Google, Yahoo, Amazon and Facebook. And better decisions can mean greater operational efficiency, cost reduction and reduced risk.

The entrance of big data into the new startups of e commerce brings with it a valuable counterpart: agility. Successfully making full use of big data and derive benefits from it the records in big data requires experimentation and extensive searching for the data. Whether it is about making the new products or searching for new and alternative ways to take competitive advantage, the job calls for strong desire and an entrepreneurial outlook.

2. IMPORTANCE OF BIG DATA

In August 2010, Office of Management and Budget, and Office of Science and Technology Policy of White House gave a statement that Big Data is a national challenge and is at the same priority as of healthcare and national security. The National Science Foundation, the National Institutes of Health, the U.S. Geological Survey, the Departments of Defence and Energy, and the Defence Advanced Research Projects Agency affirmed a joint R&D initiative in March 2012 that planned to invest more than $200 million to build new big data tools and techniques. Its aim is "…realizing of the technologies needed to perform operation and mine huge amounts of information; apply that skills and information to other scientific areas as well as address the national aims in the fields of health energy defence, education and researcher". [2]

3. DATA MINING

Usually, data mining (sometimes called data or knowledge discovery) is the process of analysing data from different vision and abstracting it into useful information - information that can be used to generate more profit, reduce costs, or both. Data mining software comprises of analytical tools for analysing and processing data. It allows users to analyse data from many different perspective or angles, categorize it, and make an abstract of relationships identified. Technically, data mining is the process of finding dependency and overlapping or patterns among dozens of fields in large volume of relational databases.

3.1 Challenges of Big Data Mining

For an intelligent working database system to operate Big Data, the important factor is to measure up to the exceptionally huge volume of data and provide treatment for the characteristics featured by the HACE theorem (explained in Section 6). A conceptual view of the Big Data operating framework, includes three tiers from inside out with considerations on data accessing and computing (Tier I), data privacy and domain knowledge (Tier II), and Big Data mining algorithms (Tier III) (discussed in section 7).

3.2 Why old Data Mining Techniques are not sufficient

There are three fundamental issue fields that need to be pointed in dealing with big data: storage issues, management issues, and processing issues respectively. Each of these represents a huge set of technical research problems in its own privileges. [3]

3.2.1 Storage and Transport Issues

The quantity of data gets overloaded each time we have discovered a new storage medium. What is different about the most recent explosion – due to huge generation of data in social media – is that there has been no recent or latest storage medium. Moreover, data is being generated by each and every one and by everything (e.g., devices, etc…) – not just, as so far, by professionals such as scientist, journalists, writers, etc.

3.2.2 Management Issues

Management will, perhaps, be the most complicated issue to with big data. This issue first showed up a decade ago in the UK e-Science initiatives where data was segregated geographically and "owned" and "managed" by numbers of entities. Resolving problems of access, metadata, utilization, updating, governance, and reference (in publications) have proven to be major wobbling blocks.

3.2.3 Processing Issues

Assume that an exabyte of data needs to be processed completely. For simplicity, assume the data is chunked into blocks of 8 words, so 1 exabyte =1K petabytes. Assuming a processor expends 100 instructions on one block at 5 gigahertz, the time needed for end-to-end processing would be 20 nanoseconds. To process 1K petabytes will need a total end-to-end processing time of roughly 635 years. Thus, effective processing of exabytes of data will require extensive parallel processing and new analytics algorithms in order to provide timely and action able knowledge.

4. ARCHITECTURE OF BIG DATA

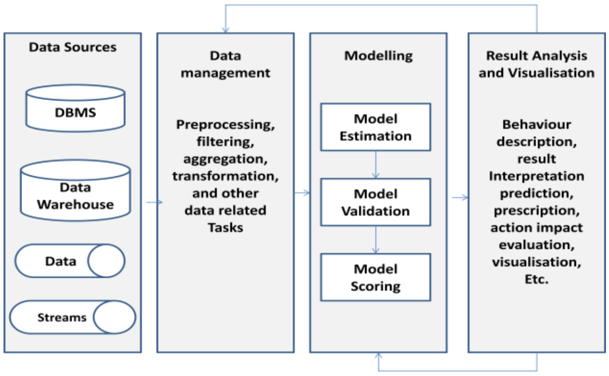
Data Sources
DBMS
Data Warehouse
Data
Streams
Data management
Preprocessing, filtering, aggregation, transformation, and other data related Tasks
Modelling
Model Estimation
Model Validation
Model Scoring
Result Analysis and Visualisation
Behaviour description, result Interpretation prediction, prescription, action impact evaluation, visualisation, Etc.

Fig 1 Overview of the analytics workflow for Big Data. [4]

One of the most time-consuming and extensive work tasks of analytics and investigative approach is preparation of data for analysis and processing; a problem often made worse by Big Data as it already stretches infrastructure to its limits. Performing analytics on huge volumes of data records requires efficient methods to perform operations on the data. Cloud analytics solutions need to consider the multiple Cloud deployment models adopted by enterprises, where clouds can be of certain type. [5]

Private: deployed on a private network, managed by the organization or by an external firm. A private Cloud is suitable for businesses that need the highest level of control of security and data confidentiality.

Public: deployed off-site over the Internet and available to the common people. Public Cloud offers high efficiency and shared resources at cheap rate.

Hybrid: joins both Clouds where additional resources from a public Cloud can be provided as a requirement to a private Cloud. Considering the Cloud deployments, the following scenarios are usually envisioned considering the availability of data and analytics models: (i) data and schema are private; (ii) data is public, structures are private; (iii) data and schema are public; and (iv) data is private, structures are public.

5. CHARACTERISTICS OF BIG DATA

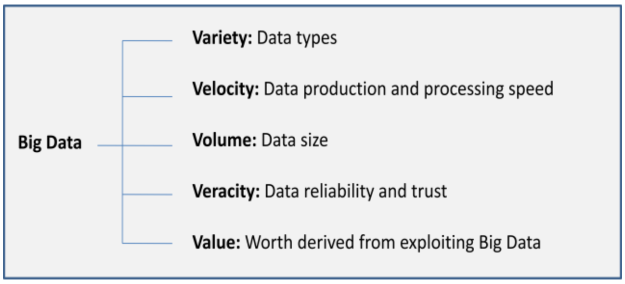
Big Data
Variety: Data types
Velocity: Data production and processing speed
Volume: Data size
Veracity: Data reliability and trust
Value: Worth derived from exploiting Big Data

Fig. 2 Characteristic of Big Data [4]

Characteristics of Big Data by what is usually referred to as a multi V model, is shown in Fig. 2. Variety represents the types of records in data, velocity refers to the rate at which the specific amount of data is generated and analyzed, and volume defines the amount or number of records of data. Veracity means how much amount of the data can be trusted given the reliability of its source.

Data Volume: Data volume defines the measures of amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it. As amount of data volume increases, the value of different data records will decrease in proportion to age, type, richness, and quantity among all other factors.

Data Velocity: Data velocity is a mean to measure the speed of data generation, streaming, and arithmetic operations. E-Commerce and other start-ups have rapidly increased the speed and richness of data used for different business transactions (for instance, web-site clicks). Managing the Data velocity is much more and bigger than a band width issue; it is also an ingest issue (extract transform-load).

Data Variety: Data variety is a measure of the richness of the data representation of the different types of data stored in the database – text, images video, audio, etc. From an analytic perspective, it is probably the biggest obstacle to effectively use large volumes of data. Incompatible data formats, incomplete data, non-aligned data structures, and inconsistent data semantics represents significant challenges that can lead to analytic spread out over a large area in an untidy or irregular way.

Data Value: Data value measures the usefulness of data in making decisions. It has been noted that "the purpose of computing is insight, not numbers". Data science is exploratory and useful in getting to know the data, but "analytic science" encompasses the predictive power of big data.

Complexity: Complexity measures the amount of interconnectedness (possibly very large) and interdependence and overlapping of data in big data structures such that even a slight change (or combination of small changes) in one or a few elements can affect very large changes or a small change that ripple across or cascade through the system and substantially affect its behavior, or no change at all.

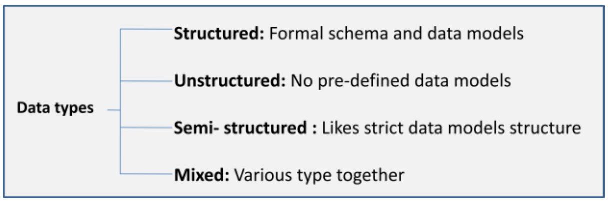
Data types
Structured: Formal schema and data models
Unstructured: No pre-defined data models
Semi- structured : Likes strict data models structure
Mixed: Various type together

Fig. 3 Variety of Big Data [4]

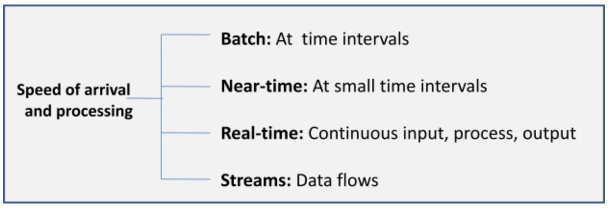
Speed of arrival and processing
Batch: At time intervals
Near-time: At small time intervals
Real-time: Continuous input, process, output
Streams: Data flows

Fig. 4 Velocity of Big Data [4]

Considering data velocity [4], it is considered that, to complicate matters further, arrival of data and processing or analyzing data are performed at different speeds, as illustrated in Fig. 4. Whilst for some applications, the arrival and processing of data can be performed in a block, other analytics applications require continuous and real-time analyses sometimes require immediate action upon processing of incoming data streams i.e. the action is taken when the data flows continuously.

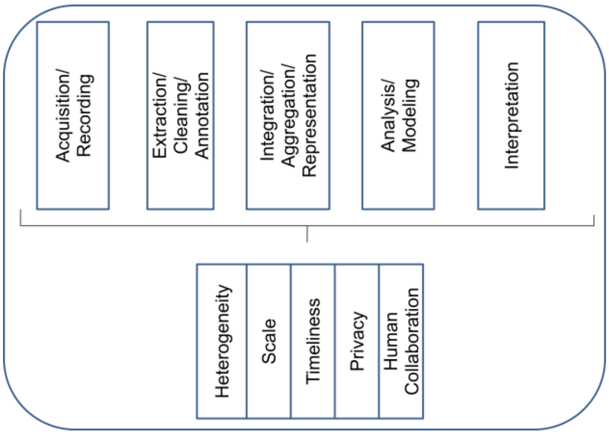

Fig 5: Big Data Pipeline[6]

5.1 Data Acquisition and Recording

Big Data does not falls out of sky or neither is it grown on tree: the generated data is recorded from some data generating source. Another biggest challenge is to automatically produce the correct metadata to describe what data is recorded and how it is recorded and measured. Another important issue here is data provenance. Recording information about the data at its generation is not usefulness. This information can be understood completely and flowed along through the data analysis pipeline.

5.2 Information Extraction and Cleaning

Frequently, the data generated and stored will not support the format which needs to be ready for analysis. We cannot use the unsupported data in this form and still effectively analyze it. To do the analysis, we need a process to extract the information that selects the required information from the underlying unstructured data and expresses it in a structured form suitable for analysis and process. Analyzing and processing the data correctly and completely is a continuously a technical challenge faced.

5.3 Data Integration, Aggregation, and Representation

Data stored in databases are not similar, they are heterogeneous data i.e. Variety of data. It is not enough simply to record the data and store it into a repository. Data analyzing and processing is considerably more challenging than simply locating, identifying, understanding, and citing data.

5.4 Query Processing, Data Modeling, and Analysis

Methods for querying and mining the records of Big Data are fundamentally different from that of traditional statistical analysis on small samples like tables, views, cursor etc. Big Data is often noisy, dynamic, heterogeneous, inter-related interdependent, overlapping and untrustworthy.

5.5 Interpretation

Having the ability to analyze or process Big Data is off the limits value if users cannot understand or interpret the analysis. Ultimately, a decision-maker, provided with the result of analysis, has to interpret these results. This interpretation does not occur out of nowhere in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis.

6. HACE THEOREM

HACE Theorem: Big Data initiates with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. The upper defined characteristics make big data an extreme challenge for retrieving and selecting useful records from the Big Data. [7]

6.1 Huge Data with Heterogeneous and Diverse Dimensionality

One of the fundamental characteristics of the Big Data is the huge amount of data that is varied represented by heterogeneous and diverse dimensionalities. This is because different information collectors and users use their own schema for storing the data and data recording, as well the nature of different applications used to select or retrieve the data also results in diverse representations of the data.

6.2 Autonomous Sources with Distributed and Decentralized Control

Autonomous data sources with distributed and decentralized controls are a main characteristic of Big Data applications. Since the big data is distributed it makes the accessing, selecting and retrieving of data easy. Being autonomous, each data sources is able to generate the records and collect information from other cloud without involving (or relying on) any centralized control.

6.3 Complex and Evolving Relationships

As the volume of the Big Data increases, so do thecomplexity and the relationships underneath the data. Interdependency and overlapping of the data is common disadvantage in big data. In an early stage of data centralized information systems where data is stored in one location, where the focus is on identifying, selecting and retrieving best feature values to represent each observation related to the record to be fetched. This is similar for instance to use a number of data fields, such as age, gender, income, education background etc., to characterize each individual.

7. BIG DATA MINING

The Biggest fundamental challenge for the Big Data applications and databases is to explore and search for the specified records in the large volumes of data and extract useful information or knowledge or data records for future actions to be performed on the data.

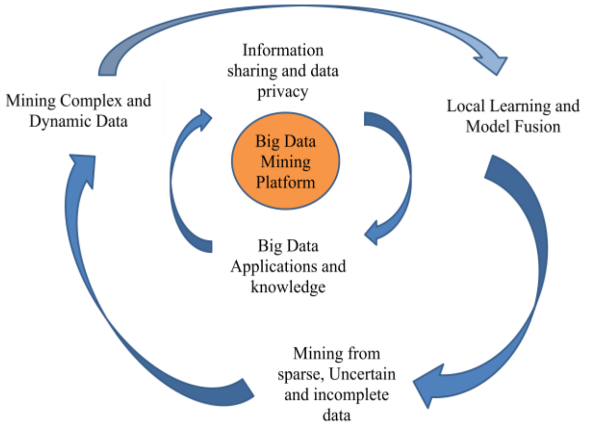
Information sharing and data privacy
Mining Complex and Dynamic Data
Local Learning and Model Fusion
Big Data Mining Platform
Big Data Applications and knowledge
Mining from sparse, Uncertain and incomplete data

Fig 6: Big Data Processing Framework [8]

7.1 Tier I: Big Data Mining Platform

In usual data mining systems, the mining procedures often require too much of intensive computing units for data analysis, processing and comparisons for the suitable records. User needs the computing platform so that user can have efficient access to, at least, two types of resources that are data records and computing processors for all computational problems. For small scale data mining tasks, a single desktop computer, which contains hard disk to store the records and CPU processors to compute and process all the data records, is sufficient to fulfill the aim of data mining. There are numbers of algorithms written for data mining to handle this type of problem settings. For medium scale data mining tasks, data are typically large than the small scale or garage setup (and possibly distributed) so it cannot be fitted in to the hard drive of the computer.

7.2 Tier II: Big Data Semantics and Application Knowledge

Manually written syntaxes, semantics and grammar and application knowledge in Big Data refer to numbers of aspects related to the regulations, restrictions, rules, policies, user knowledge of using application, and domain information. The two most crucial issues at Tier II include sharing data and privacy or confidentiality and domain and application knowledge.

7.3 Tier III: Big Data Mining Algorithms

Local Learning and Model Fusion for Multiple Information Sources: The Big Data applications developed are heavily featured with autonomous sources and decentralized controls for easy accessing data, arithmetic distributed data sources to a centralized site for data records mining is systematically prohibitive due to the transmission cost and confidentiality concerns.

Mining from Sparse, Uncertain, and Incomplete Data:

Useless i.e. remained, inappropriate, and incomplete data are defining features for Big Data applications. Being few number of data points cannot draw the reliable conclusions. This is typically a complicated data dimensionality issue, where data in a high dimensional space does not give the hint of clear trends or distributions. For most machine languages and data mining algorithms, high dimensional unused data significantly degrades the difficulty and the reliability of the structural models derived from the data records.

Mining Complex and Dynamic Data:

The rise of Big Data is driven by the increasing speed of complicated data and their changes in quantity and in nature. While complex dependency schema beneath the data raise the complications for learning machines, they also offer exciting chances that simple records representations are incapable of achieving.

8. CONCLUSION

Big data is able to process and store that data and probably in bulk of amount in soon future. Hopefully, technology will get better. New technologies and tools that have ability to record, monitor measure and merge all kinds of data surrounding us, needs to be introduced very soon. Industries need new technologies and tools for anonymzing data, analysis, tracking and inspecting information, sharing and maintaining, private data in future. So many aspects of life which generates the big data on daily basis that manages big data world need to be shined as possible.

There are too much of future important challenges in Big Data management and analytics that arise from the nature of data: large, diverse, and evolving.

Today, technologies successfully proved importance of big data, learned issues of big data, discovered new ways for data mining. Learned Characteristics of big data and Three tiers of Big Data Mining.

9. REFERENCES

[1] Wei Fan, Albert Bifet. Mining big data: current status, and forecast to the future, ACM SIGKDD Explorations Newsletter, Volume 14 Issue 2, December 2012

[2] Sneha Gupta, Manoj S. Chaudhari, Big Data Issues and Challenges, nternational Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 3, Issue: 2

[3] Stephen Kaisler ; Frank Armour ; J. Alberto Espinosa ; William Money. Big Data: Issues and Challenges Moving Forward, 46th Hawaii International Conference on System Sciences (HICSS), 2013, ISSN :1530-1605

[4] Big Data computing and clouds: Trends and future direction by Rajkumar Buyya

[5] Marcos D. Assunção et.al. Big Data computing and clouds: Trends and future directions, Journal of Parallel and Distributed Computing Volumes 79–80, May 2015

[6] Xindong Wu, , Xingquan Zhu, Gong-Qing Wu, Wei Ding, Data Mining with Big Data, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014

[7] Deepak S. Tamhane, Sultana N. Sayyad. Big Data Analysis Using Hace Theorem, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 4, Issue 1, January 2015

[8] Kale Suvarna Vilas. Big Data Mining. International Journal of Computer Science and Management Research eETECME October 2013, ISSN 2278-733X