An Overview on Data Mining and Data Fusion

Vinayak Jain

***Abstract:** Strong adoption of Internet and Communication technologies across industries in the last two decades has led to large-scale digitization of business processes. While this has helped in the instant availability of information, over the period, the source and amount of this information have increased multi-fold giving rise to Big Data. With the increase in volume, the relevance of data in its raw format continues to decrease over time. According to HACE Theorem, Big Data has autonomous sources being distributed and decentralized data in a complex relationship with each other. Making sense of this ever-growing large pool of data has become increasingly difficult and has created a new problem waning the initial gains made via the digitization of systems and processes. This gave rise to the evolution of multiple Data Mining techniques that have helped to classify large volumes of data into relevant segments and drive value to help provide meaningful information. To extract and discover knowledge from data, Knowledge Discovering Databases (KDD) help in the refining of data. This paper discusses various data mining techniques that help to identify patterns and relationships to help make business decisions using data analysis. Furthermore, the Data Fusion method is reviewed which deals with joint analysis of multiple inter-related datasets providing multiple complementary views to help further with precise decision-making.*

***Keywords:** Big Data, Data Mining, Data Fusion, KDD, HACE Theorem*

I. INTRODUCTION

Big Data refers to too large data sets, size in terabytes and petabytes. The National Institute of Standards and Technology (NIST), the report defined big data as consisting of "extensive datasets—primarily in characteristics of volume, velocity, variability, variety —that require a scalable architecture for efficient storage, manipulation, and analysis." Big data is also defined as the amount of data that even exceed a petabyte, i.e., one million gigabytes (GB) [1]. Exponential growth and availability of data in our world is also a definition of big data.

Doug Laney the big data Industry analyst defined the “Three Vs” of big data:

1. *Volume:* It is the quality of data generated through various online and offline means; it is estimated that 2.5 quintillion bytes of data are generated every day dealing with that the challenge with such huge volumes of data is to extract relevant data from such a high volume of data.
2. *Velocity:* It is defined as the speed at which data is generated, as data is generated at an exponential rate. Every minute, 3.8 million search queries are received by Google. 156 million messages are being sent by email users. 243,000 photos uploaded on Facebook. Finding ways to collect, process and make insight out of huge volumes of data is the biggest challenge for Industry experts and data scientists.
3. Variety Big Data analyse demands a large amount of data and ifferences in its variety. As data is consist of two types, structured and unstructured data, where structured data uses XML format which is easy to enter, store, query, and analyse. Whereas unstructured data is more heterogeneous and difficult to extract value from it. Such as audio, video, images, emails, social media content, web pages etc [1].

   Beyond the Big Three Vs, big-data practitioners and thought leaders have proposed additional Vs:
4. *Veracity:* It is important to know the relevance of data in big data to avoid future trouble. If the collected data is not valid and has a questionable source, it becomes imperative that the institute is to able trust the quality of data which invalids our result.
5. *Variability:* There are many examples in which the same word has different meanings. This makes the job of processing the language very difficult for a computer, so data scientists must keep a check on this phenomenon by creating programs that could understand meaning and context.
6. *Visualization:* Visualization is the way to make the data understandable for investors, decision-makers and nontechnical people. It tells the data scientist's story through graphs, models, and flowcharts, it converts the data into information, then the information into insights, insights to knowledge and the create advantage from the knowledge.

There are a lot of Vs in big data, big data is growing exponentially, and the organization needs to make insights into it and take advantage of big data.

A. Types of Big Data

Two types of big data: structured and unstructured.

1. *Structured data:* It consists of data that consists of numbers and words, basically labelled data that is easily analysed and categorized [2][5]. These data come from network sensors, smartphones, and GPS devices. It also includes sales figures, account balances and transaction data. The tool that can be used for structured data is SQL (structured query language).

Manuscript received on 12 December 2022 | Revised Manuscript received on 22 December 2022 | Manuscript Accepted on 15 May 2023 | Manuscript published on 30 May 2023.
*Correspondence Author(s)

Vinayak Jain*, Department of Computer Science, SRM University, NCR-Delhi, Sonepat (Haryana), India. E-mail: vjain8250@gmail.com, ORCID ID: https://orcid.org/0000-0002-7632-5368

Retrieval Number:100.1/ijdm.A1624053123
DOI:10.54105/ijdm.A1624.053123
Journal Website: www.ijdm.latticescipub.com

Published By:
Lattice Science Publication (LSP)
© Copyright: All rights reserved.

2. *Unstructured data:* It includes more complex information. It consists of unlabelled data like customer reviews, Images, and videos. It is hard to categorise and analyse this type of data [5].

II. DATA MINING

Organizations use the data mining process for converting raw data into information, it is a process of extraction and pattern discovery in large data sets, it involves methods intersection with machine learning, statistics, and database systems, etc [2]. The data mining goal is to extract the information with intelligent methods and then transform it for further use. It is an analysis step of the knowledge discovery in databases (KDD) process. Data management, data pre-processing, complexity considerations, post-processing of discovered structures, visualization, and online updating are also coming under data mining.

It can be applied to any form of large-scale data or information processing, and also on any application of computer decision support system (DSS), including AI, i.e., artificial intelligence (e.g., machine learning) and business intelligence.

III. DATA FUSION

It is the process of extracting and integrating multiple data sources to finalize the most consistent, precise and the most accurate answer in improvement for better decision-making than for any answer in an individual data source [3]. The result of this is data are better generated as low, intermediate, or high-level acc. to the processing stage which can be stored, manipulated, and analysed.

According to Researchers of the Joint Directors of Laboratories (JDL) workshop, the definition of data fusion: is "A multi-level process dealing with the association, correlation, combination of data and information from single and multiple sources to achieve refined position, identify estimates and complete and timely assessments of situations, threats and their significance" [4]. And Hall and Llinas: "data fusion techniques combine data from multiple sensors and related information from associated databases to achieve improved accuracy and more specific inferences than could be achieved by the use of a single sensor alone."

Thus, we define data fusion as an average of all multiple sources which gives core information with higher quality in nature, more validation and less cost.

IV. HACE THEOREM

Big Data is the sum of a huge volume of data, autonomous sources having distributed and decentralized control and complex and evolving relationships binding the data [5].

These factor sums up to be a difficult job to extract useful information from Big Data.

To make things clear, let us assume some blindfolded men are made to conclude what they think of a giant statue which is Big Data here, by touching it and recording their conclusion independently. After giving them time with the statue, we find that to some people the statue feels like a wall and to some, it's like a car depending on the person's view limited to the region they are in contact with different people describe their view differently.

Big Data is comparable to collecting information from different men to draw the best possible picture of the statue in real time. The task of drawing a picture of the statue is not as simple as it looks as there might be people who hesitate to speak of their observations or some people might not describe their observations properly and they may even have privacy concerns about the messages they want to deliver and several difficulties can come across the process of picturing the statue, so it's not as simple as asking to a person to explain about his observations, same goes with the Big Data.

To define Big Data in a proper fashion HACE Theorem suggests the characteristics of the Big Data are:

1. *Huge with heterogeneous and diverse data sources* - One of the most important characteristics of Big Data is the huge volume it is present in, and it has multiple dimensions and is heterogeneous. This kind of data comes from various sites like Twitter, Instagram, Facebook, and LinkedIn etc [5].
2. *Decentralized control* - Autonomous data sources with distributed and decentralized controls are one of the main characteristics of Big Data. Being autonomous, each data source generates and collects information without involving any centralized control. This is like the World Wide Web set where each web server provides information, and each server can fully function without reliance on another server.
3. *Complex and evolving data* - Data collected from multiple sources which are not structured could be termed as complex data Examples of complex data would be videos and images, documents, and a web server. Evolving data can be described as the evolution of humans and other species as time passes new parameters are being added to the data we generate.

V. DATA MINING AND BIG DATA

Generally, data mining (sometimes called data or knowledge discovery) is the process of analysing data from different angles and scenarios and compacting it into useful information - information that can be used to increase income, cuts costs, or decision making. Technically, data mining is the process of noting correlations or patterns among dozens of fields in a large relational database.

A. Data mining techniques

1. *Classification* - It is a complex data mining technique, that uses data attributes to shift them into indistinct categories, to help us in further conclusion [2, 6]. Smartphone companies use data mining to classify customers buying products like camera-specific, performance-specific and budget-specific phones, etc. These classifications help companies to learn more about their customer.

2. *Clustering* - This technique is very alike to classification, grouping data together based on their comparability. Cluster groups are more unstructured, making it a simpler option for data mining. In the supermarket example, a simple cluster group could be veg. and non-veg items instead of the specific classes.
3. *Association rules* - Association in data mining is all about detecting patterns, specifically based on linked variables. It's like a machine learning training model, where it helps us recommend the next thing related to primary things such as YouTube recommendation or movie recommendation which uses tags and genres for its purpose.
4. *Regression analysis* - It is used for strategy and model; it helps in identifying the relationship and effects of variables in a set. The Organization may be projecting the price based on availability, consumer demand, and competition.
5. *Anomaly/outlier detection* - An Odd one out is what you call an outlier. Data needs to be able to identify and understand the outliers which can give error indications and differentiate data points in your data as well. For example, if there is an unidentified cell in human blood, you'll want to investigate that outlier and understand the reason for its existence.

The real data mining task is the semi-automatic analysis of large quantities of data to draw out previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining) using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the given data and may be used in more analysis as in machine learning, predictive analytics, or data mining step to further identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system (DSS). Neither the data collection, preparation, result in interpretation and reporting are part of the data mining step but do come under the KDD process as necessary steps [7].

VI. KNOWLEDGE DISCOVERY IN DATABASES (KDD)

It is a process in which we focus on extraction and discovering knowledge from data. It is a process to redefine and enhance data mining.

Steps Involved in KDD Process:

1. *Data Cleaning* **-** It is defined as the removal of null values, noisy and irrelevant data that has no value or effect on our insight, from the data set.
2. *Data Integration* **-** It is defined as combining heterogeneous data from different sources into, Data warehouse, i.e., a common source. Data Synchronization tools.
   - Data Migration tools.
   - Extract Load Transformation (ELT) process.
3. *Data Selection* **-** Collecting relevant and valid data which can enhance its value without going any further redundancy and concurrency of data. It can be achieved with given techniques such as -
   - *Neural network*.
   - *Decision Trees*.
   - *Naive Bayes*.
   - *Clustering*, *Regression*, etc.
4. *Data Transformation* **-** It is defined as the process of converting data into a desirable form that is required for mining.
   It is a two-step process:
   - Data Mapping: To capture transformation we assign source base elements to the destination.
   - Code generation: The actual transformation program is created in this step.
5. *Data Mining* **-** Data mining is defined as the integrated average of all datasets to maximise its relevant Transforms task-relevant data into patterns.
6. *Pattern Evaluation* **-** It is defined as examining knowledge's strictly increasing pattern on a given measure.
   - Find an interestingness score from the pattern.
   - Uses summarization and Visualization for a better understanding ability for the user.
7. *Knowledge representation***:** It is defined as the visualization tool for knowledge discovery to visualize data mining results.
   - Reports.
   - Tables.
   - Graphs
   - Models
   - Flowcharts
   - Discriminant rules, classification rules, characterization rules, etc.

VII. CLASSIFICATION OF DATA FUSION

Data Fusion plays a key role in compacting the huge amount of data from multiple sources to a central warehouse, helping large organisations greatly benefit from it [9].

The key problems of assimilation of all these data are as follows:

Data Fusion is working on these challenges by making it easy to move data around, with two main focuses:

1. Build data pipeline without writing any code **-** With a few clicks, we can use open source CDAP project, to connect more than 100 connectors between a source and link forming a pipeline.

2. Do transformation without writing any code **-** Data Fusion comes with a set of built-in transformations that you can effectively apply to your data.

A. Based on the type of architecture

1. *Centralized architecture:* The information from all the input sources will be received in the central processor through the fusion node as an observational measurement. Then the fusion process will obtain in the central process. There are some limitations of a centralized architecture,

to the amount of bandwidth requires sending the raw data. Also, a buffer can occur according to the data transferring will affect the results of data [7][8].

2. *Decentralized architecture:* The decentralized architecture is consisting of N number of nodes; each node has a specific processor and the data fusion not occurring in a specific node. Therefore, each node fuses its current data with the data received from its peer. The increasing numbers of nodes cause an increment in the communication cost and that is one of the limitations of decentralized architecture.
3. *Distributed architecture:* In this architecture, each input source must process the measurements before being transmitted to the fusion node that's mean the completion of the state estimation and the data is associated with the source node before communicating it to the fusion node.

Then, the estimation state for the object will provide by each node depending on the local view of that object. This type of architecture provides different options and variations that range from only one fusion node to several intermediate fusion nodes.

4. *Hierarchical architecture:* It is a set of the above two architectures where these two architectures are difficult to implement independently. However, the requirements, demand, data availability and organization of the data fusion system will be responsible for selecting the best architecture.

B. Other features of Data Fusion are:

1. *Open source:* It's built on top of CDAP with big community that keep on developing new connectors.
2. *Accessible:* due to the user interface you can start Data fusion with novice background.
3. *Metadata:* search integrated datasets by technical and business metadata. Track lineage for integrated datasets in the dataset & field level.
4. *Flexible:* without UI, Data Fusion is extensible, and you can add code to automate it.

VIII. INDUSTRIAL TOOLS

A. Tool of Big Data Analytics

1. *Hadoop:* Apache Hadoop is an open-source software library framework for big data. It works on map-reduce architecture, i.e., the master node will subdivide the job into two main tasks: mapping and reducing [9].
2. Qu bole: Qu bole is a platform for Big Data management which is Autonomous. It's an open-source tool which manages and optimizes itself and focuses on business outcomes.
3. *Kaggle:* It is one of the world's largest big data communities. It is used by organizations and researchers to post their data & statistics. Data is seamlessly analysed here.

B. Data mining tools for businesses.

1. *Data Melt:* This tool helps with mathematics, statistics, calculations, data analysis, and visualization. It can combine with scripting languages like Python, Ruby, and many Java packages.
2. *ELKI Data Mining Framework It is designed to be easy to use for students, organizations, and researchers. It focuses on algorithms that impact outlier systems and unsupervised clusters.*
3. *The R Project:* It is used for Statistical Computing and graphics, and it is supported on many operating Systems.

C. Data Fusion tools for businesses

1. *Xplenty:* It is an Extract, Transform, Load or ETL tool that provides a method with very less code or zero code for automating the data flows in an organization [10]. It makes data access easy through visual data pipeline creation. The user can easily clean data, normalize data and transform data to a destination.
2. *Jitterbit:* It is an Application Programming Interface integration platform which is designed to streamline the connection of cloud, on-premise, and SaaS applications. It could also offer a way to add Artificial Intelligent technology into your applications and could help you develop various innovative solutions using AI tech.
3. *IBM App Connect*: This is a tool for different types of applications which instantly connects applications and data from existing systems and modern technologies across all environments and it offers real-time and batches support. You can expose data as REST APIs and take advantage of hundreds and hundreds of pre-built connectors included in this platform.

IX. CONCLUSION

This paper reviews the techniques for performing big data, as big data means the collection of a complex and huge amount of data. Big data is exponentially growing and the need for it is rising in business, technology, medical and engineering domains. Big data technologies and some of the most widely used architectures for the implementation of data fusion solutions to problems from different industrial tools. Data mining is the analytical process of exploring data in search of patterns and then getting insights from the data. Data mining is used to create DSS (Decision Support System) that helps in creating knowledge-driven models for an advisory system, classification, configuration, diagnosis, and interpretation. Data Fusion technology is used by a large organisation to extract a large amount of data from multiple sources to a central warehouse.

DECLARATION

Funding/ Grants/ Financial Support	No, I did not receive.
Conflicts of Interest/ Competing Interests	The article bears No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material/ Data Access Statement	Please refer to reference.
Authors Contributions	I am only the sole author of the article.

REFERENCES

1. ‘’What is Big Data?’’ retrieved from https://datasciencedegree.wisconsin.edu/data-science/what-is-big-data/ (2017)
2. Wikipedia, Data Mining, retrieved from https://en.wikipedia.org/wiki/Data_mining (2022)
3. Wikipedia, Data Fusion, retrieved from https://en.wikipedia.org/wiki/Data_fusion (2022)
4. Castanedo Fedrico, A Review of Data Fusion Techniques (2013) [CrossRef]
5. Prof. Khan MD Sameeruddin, Prof. Subbarao GVNKV, GNV Reddy Vibhav, Hace Theorem based Data Mining using Big Data(2016)
6. Information Technology, Data mining in business analytics, retrieved from https://www.wgu.edu/blog/data-mining-business-analytics2005.html(2005)
7. Geeksforgeeks, KDD Process in Data Mining https://www.geeksforgeeks.org/kdd-process-in-data-mining/ (2022)
8. Odaudu Ngbede Salefu, Ime Jarlath Umoh, Emmanuel Adewale Adedokun, Francis Franklin Marshall, Donald Etim Ikpe, Big Data Fusion and Emerging Technologies. 2019 IEEE 1st International Conference on Mechatronics, Automation and Cyber-Physical Computer System (2019)
9. David Taylor, ‘’ Top 15 Big Data Tools and Software (Open Source) 2022’’ https://www.guru99.com/big-data-tools.html(2022)
10. Abe Dearmer,“ 17 Best Data Integration Platforms’’ https://www.xplenty.com/blog/17-best-data-integration-platforms/ (2021)

AUTHORS PROFILE

Vinayak Jain has completed his B.Tech CSE in Artificial Intelligence and Data Science from SRM University NCR-Delhi, Sonepat. He has a working experience of 6 months as a Data Analyst at Ingenious E-brain Solution Pvt. Ltd. where he was responsible for study and analysis of literature & technical research of current technology trends in the markets and helping the clients with their intended requests, His areas of interest are Artificial Intelligence, Machine Learning, Big Data, Computer Vision and Extended Reality. Currently working as a Freelance Web Designer on a prototype of the website and mobile application using the graphic tool Figma and Abode XD.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.