2023

BIG DATA & DATA ANALYSIS

Written by:

Mohammed Al-Hosni | 1615 | 8A

Abdulrahman Al Ghammari | 1654 | 8A

Al Mulham Waldwadi | 1630 | 8A

Contents

Report aim

The following report discusses technological advancements of *Big Data* and *Data Analysis* from the perspective of both theoretical aspects from their respective characteristics and the methodologies utilized in employing them. The report also aims to go over as their practical implementation and usage for in enhancing both economical and corporate opportunities.

Table of contents

Big data

This section of the report will discuss Big Data in terms of its structures, characteristics, and the methods and tools used to employ Big Data while going over the technology's importance.

Definition and structure:

Definition of Big Data: Big data is a huge collection of varying datasets relating to a given topic ranging into the Petabyte or even Exabyte range. Due to both the variety and size, particular challenges arise in terms of capturing, storing, and analyzing said data leading to the use of specialized tools and techniques being necessary. The type of data stored in Big Data clusters is generally categorized into three types:

Structured Big Data: Data cluster including quantitative data that can be easily organized into tables with immediate context such as names, phone numbers, bank statements.

Unstructured Big Data: Qualitative data clusters that cannot be organized meaningfully into tables and require conversion into understandable computer formats (e.g., png, mp4, etc.) or require context the addition of context to provide meaningful information. Examples of such Big Data includes video, audio recordings, and photographs.

Semi-structured Big Data: Semi-structured data is the Big Data is in between the two extremes just stated, where it is data that is somewhat quantitative (structured) but also includes some qualitive (unstructured) values that require context (e.g., Folders organized by a given topic)

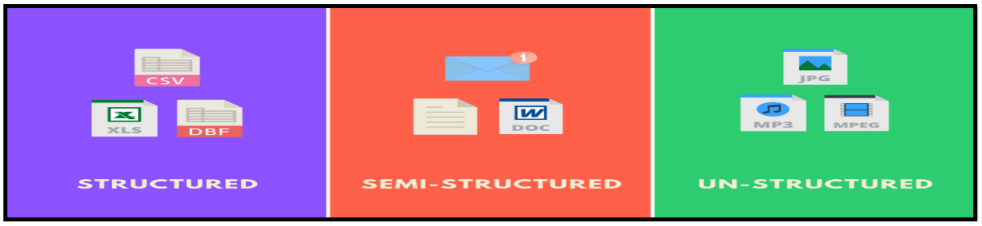

Figure 1 - Examples of the stated structures

Characteristics of a Big Data cluster:

The "Seven Vs" are characteristics that indicate significant information for any given Big Data cluster, knowing said characteristics allows analysts to know the requirements of the data cluster and the optimal methodologies to analyze the data.

Volume	The size of the data in terms of storage requirement.
Velocity	As stated by (IT Chronicles, n.d.), velocity refers to both the rate at which new information is added but also how fast the information needs to be processed.
Variety	Number of unique datatypes (text, photos, videos, etc.,) in the data cluster. For example, an unstructured Big Data cluster would be relatively high in variety.
Variability	How much does a single entry vary. A single photo could have multiple versions with different implications for each one.
Veracity	The overall tolerance of inaccuracy and how trustworthy the data is.
Value	The Return-on-Investment relative to resources expended analysing the data.
Visualization	How easily can the data be visualized into a relevant format.

Biggest Big Data challenges:

How Big Data is stored: The main complication of Big Data relates to its storage due to its size. For that – according to (Google Cloud, n.d.) – digital storage environments known as "Data Lakes" are utilized. Unlike Databases, Data Lakes store data in its raw format without tables organization, enhancing storage efficiency, as simply removing these aspects significantly reduces hardware requirements. Moreover, the popular Big Data framework Hadoop uses a process where the Data Lake is split into files and spread throughout several machines, reducing processing stress.

Processing big data: This involves conversion of unstructured data or editing information. Said tasks are usually light on hardware but due to the size, processing power demands are significant.

How data can be turned into meaning: Due to the vastitude of Big Data, analysis through conventional computational methods is insufficient as the processing strength of such techniques is not sufficient to handle such amount. This is where advanced Data analysis such as machine learning is used and is what will be discussed in the following section of the report.

Security concerns: Since Big Data captures and analyzes data potentially relating to personal aspects of people (in the context of a customer, patient, citizen etc.), a huge concern is being generated against private companies owning such data due to the potential for data misuse or leaks.

Data Analysis

Further on, another aspect commonly related to data in general and especially in the case of Big Data is Data Analysis, and for that the subsequent section of the report will go over what Data Analysis is and its relation to Big Data and the techniques utilized for analyzing Big Data.

By definition, Data Analysis is a Disciplinary field that encompasses the inspection, organisation, interpretation and visualisation of data. Its end goal is to gain better understanding of datasets through computation as to uncover patterns that otherwise would be too complex. This increases accuracy of modern decision-making tools such as AI and enhances forecasting and predictive ability by learning from historical data as to form a coherent conclusion and actionable steps for potential growth.

The relations between Big Data and Data Analysis:

Giving meaning to raw Big Data: Big Data is the process of storing, maintaining, and organizing data, all of which is done for the sole purpose of analyzing it, as to eventually give analysts an idea of future trends based on the historical data, hence turning data amounting to nothing in its raw form into information that can lead decision-makers towards a more accurate path. The terminology utilized when combining the two technologies is Big Data Analytics.

The Data Analysis process: As stated by (Kelley, 2023), The Data Analysis process is a five-step procedure that raw data needs to take in order to be turned into useful information, with said steps being:

Figure 2 – The stages of the Data Analysis process

The first step involves knowing purpose and business expectations from using Data Analysis. While both the second and third steps of Big Data involve storing and cleaning (processing) the information. With steps four and five being specific to analysis. As such, it can be noticed that Big Data has no use without Data Analysis and vice versa.

Data Analysis techniques for Big Data:

Analysis of Big Data is a slow process and increasing its rate is through expending resources on processing power. As such, it needs to be used by analysis techniques that benefit from vast data.

Deep Learning: Deep Learning (DL) – a subset of AI and Machine Learning – is the process of creating neural network algorithms that take data as inputs and through a process of initial trial and error, learns to reach a point where the values it provides as outputs are as accurate. Due to said trial-and-error process, DL thrives on large subsets of data, a perfect situation for Big Data Analytics. After processing, DL is utilized for pattern recognition and data validation. Applications that have recently grown significantly in popularity are automated driving and ChatGPT, which use pattern recognition to output their signals (steering and replies respectively).

Data mining: Data Mining takes large sets of information and uses them for pattern recognition, with the difference between it and ML being that ML teaches a computer to interpret data automatically, while human interaction is still needed in interpreting and taking action based on the results of Data Mining. A major example of practical Data Mining as mentioned by (Twin, 2022), is in behavioral analysis for customer shopping patterns and which items are frequently sold together. Which is used to create dedicated offers and discounts. Meaning that the larger the data the more accurate, inclusive, and representative of the average customer the results are.

Regression Analysis: Regression analysis is the process of checking how much the change of one variable affects the other. As mentioned by (Beers, 2022), the analysis can be on the combined effect of multiple variables on one and is where Big Data is demanded for regression analysis. Since – in large enough scale – this could give analysts a clear idea of future prospects on the price of a given stock or commodity based on current trends of related variables.

Application of Big Data Analytics

The following section will discuss the potential practical application of Big Data Analytics both from a corporate and economic standpoint. Additionally, this could be related to Oman's 2040 vision in both the innovation and competitive index requirements.

Corporate applications:

Pricing selection: Companies can understand future trends of operational costs such as materials and in turn price their products adequately based on fluctuations of relative variables.

Client behavior: Companies can gain feedback from client behavioral analysis and improve on weaker aspects of the business. Through knowing the average customer's behavior and identity, companies can precisely target potential clients and dedicate items/services for their current ones.

Economical:

Improved education curriculums: Through tracking student data and performance, governments can track exact deficiencies in the education system that need to be improved while on the other hand also know which subjects or course materials have seen the most success. ***(Replace it all pls).***

Crime management: Big Data Analytics in surveillance applications can help governments detect crime faster via automated surveillance feed analysis, in turn reducing crime rate and severity.

Traffic control: Through storing and in-depth analysis of traffic data, governments can detect the source of congestion, and in turn dedicate decision making efforts towards the main issue.

Common corporate and economical applications:

Risk management: Although risk management is a broad term, Big Data Analytics can detect any disturbances or unlikely pattern immediately. This in turn helps both companies and countries detect risks such as fraud before they manifest into larger issues.

Competitive intelligence advantage: By utilizing Big Data, companies and governments have a much wider range of intelligence and sources of information that potential competitors do not have. This allows them to avoid issues, innovate, and cease opportunities as soon as possible.

Prospects for Big Data Analytics

Positive trends:

Full real-time automated analytics: As automated Big Data Analytics technologies become increasingly more sophisticated, the rate at which said data is processed and analyzed will also increase, leading to a world where new analytics can be accessed immediately without delay.

Access to valuable analytics on the personal scale: As more and more Big Data clusters become popular, the demand for open-sourced data will also rise, leading to big data clusters that can be accessed by the public. Meaning, anyone could have access to valuable analytics for personal use.

Improvements in related processes: As the demand for Big Data increases, this also means that relevant technologies such as processors or storage techniques would need to keep up, therefore a rise in innovation would be seen in sectors that could have potentially been considered as stagnant.

Potential concerns:

Confidentiality and security concerns: The concern grows over both corporate and client data from potential Big Data breaches. Since the fact that a single breach into the cluster from malicious attackers could lead to leaks and ransomware in sizes that have not been seen before due to the vast amount of information found in said Big Data clusters. However, as mentioned, since security is a big part of the Big Data and is a relevant technology, it is also likely that security measures would increase to accommodate for the increased risk.

The chip shortage: The ongoing chip shortage crisis is significantly effecting all areas of technological hardware production, including those that Big Data utilized. As such advancements in Big Data technology would be hindered up until positive developments are seen in the situation.

References

Beers, B. (2022, August 18). *What is Regression? Definition, Calculation, and Example*. Retrieved from investopedia: https://www.investopedia.com/terms/r/regression.asp

Google Cloud. (n.d.). *What is a Data Lake?* . Retrieved from google: https://cloud.google.com/learn/what-is-a-data-lake

IT Chronicles. (n.d.). *What is Big Data*. Retrieved from itchronicles.

Kelley, K. (2023, February 7). *What is Data Analysis? Methods, Process and Types Explained*. Retrieved from simplilearn: https://www.simplilearn.com/data-analysis-methods-process-types-article

Twin, A. (2022, August 2). *What Is Data Mining? How It Works, Benefits, Techniques, and Examples*. Retrieved from investopedia: https://www.investopedia.com/terms/d/datamining.asp

"The government's data policy that does not comply with the changes of the times is acting as a stumbling block to the spread of the Fourth Industrial Revolution. A paradigm shift in data use is required to strengthen the competitiveness of the nascent big data industry."

- Chang Byeong-kyu

Chairman of The Fourth Industrial Revolution Committee