IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12136

Data Science and Quality Management

Aneesh Vishnu[1]

Student, BCom, Jain (Deemed-to-be-University), Bangalore, India[1]

Abstract: Our civilization has become much more computerised, which has greatly improved our ability to generate and gather data from a variety of sources. Almost every element of our lives has been inundated with an enormous amount of data. It is necessary to convert the enormous volume of data into knowledge and relevant information. Data mining is a promising and flourishing field of computer science as a result of this. Data mining is the automated or practical extraction of patterns that represent implicitly stored or recorded knowledge from huge information repositories, such as databases, data warehouses, the web, and other big information repositories or data streams. Any type of data can be used for data mining as long as it has value for the intended application. In this paper, we discuss in detail data warehouse and data warehouse data, which is almost basic form of data for data science applications. We also present to you a typical framework of a data warehouse and data pre-processing techniques. Additionally, we talk about OLAP (Online Analytical Processing) Data Marts, a subset of an organisational data store that is typically focused on a single objective or key data area and can be disseminated to meet business requirements.

Keywords: computerization, data science, databases, data warehouses, data pre-processing techniques, OLAP (Online Analytical Processing) Data Marts, Total Quality Management.

I. INTRODUCTION

The development of databases and information technology from simple file processing systems has been extensive. New and potent data models are incorporated into modern database systems, data warehousing and mining [1] for advanced data analysis, and web-based databases [2]. Data mining and analytics have been sparked by the consistent and brilliant advancement of computer hardware, powerful processing, reasonably priced data collection tools, and storage media. The data warehouse is a new type of data repository design. This is a single-site repository for numerous heterogeneous data sources that are arranged according to a single schema to aid in management decision-making. Following are the data processing techniques:

- Data cleaning
- Data integration
- Data reduction
- Data transformation
- Load and refresh

II. DATA WAREHOUSE

Data warehouses generalize and consolidate data in multi-dimensional spaces. The construction of data warehouses involves data cleaning [3], data integration [4] and data transformation [5] and the data can be viewed as an important step for data mining. Data warehouse provides architectures and tools for business executives to systematically organize, understand and use their data to make strategic decisions. Figure 1 shows data transformation and its movement. The information gathered in a warehouse can be used in any of the following domains:

- Tuning production strategies
- Customer analysis/customer behavior
- Operations analysis

A. Process flow in data warehouse

There are four major processes that contribute to a data warehouse:

- Extract and load the data
- Cleaning and transforming the data
- Backup and archive the data
- Managing queries and directing them to the appropriate sources

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023

DOI: 10.17148/IJARCCE.2023.12136

In order to recover the data in the event of data loss, software failure or hardware failure, it is necessary to keep regular backups. Archiving involves removing the old data from the system in a format that allows it to be quickly restored whenever required. Query and analysis tools are used in query management process.

This process performs the following functions:

- Manages the queries
- Helps speed up the execution time of the queries
- Directs the queries to the most effective system sources.
- Ensures that all system sources are used in the most effective way

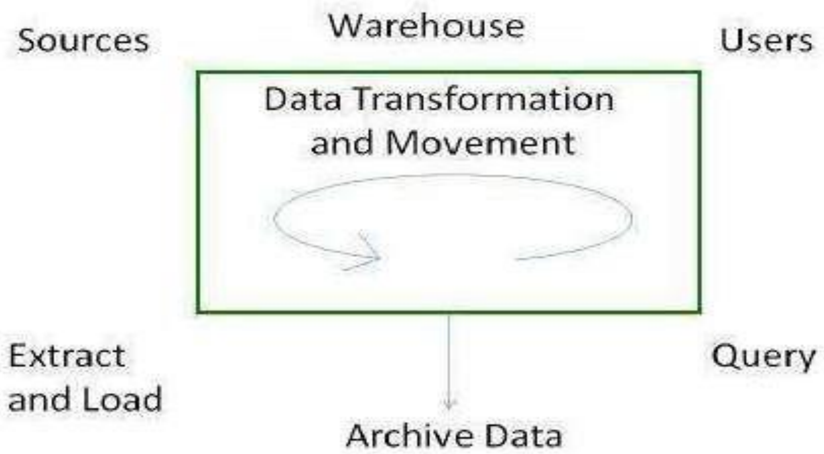

Figure 1 Data transformation and its movement.

III. DATA SCIENCE

Data science is a field that primarily developed out of necessity, as opposed to being an area that was being explored. It has evolved over time from being used in the relatively narrow realm of measurements and research to being a ubiquitous presence in every sphere of science and business. In this section, we examine some of the key research and application areas where data science is now in use and on the cutting edge of development.[6-7]

Figure 2 shows the process of Data Science

The three components of data science are data organisation, data assembly, and data dissemination. At any rate, data wrangling, which combines data assembling and orchestrating, includes assembling as a crucial component. In other words, knowledge of the what, why, and how is what distinguishes data science from other disciplines. A data expert must fundamentally acknowledge the identity of the overall populace involved in the profit-making venture.[7]

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⋇ Impact Factor 7.918 ⋇ Vol. 12, Issue 1, January 2023
DOI: 10.17148/IJARCCE.2023.12136

IV. DATA SCIENCE IN QUALITY MANAGEMENT

Using Data Science and machine learning techniques, the industrial process data may be split into label data and no label data. When MES and numerous other systems are used, some data is linked to the appropriate alert, management, and other events, which is evidently reflected in the relational database. This kind of data can be treated as labeled data and has certain distinct categorization criteria. Although we can classify the majority of the process data further since it is stored in the database as normal data, there are no explicit classification criteria in the database, so the type of data is processed as unlabeled data. In the unlabeled data, the majority of deep-level fault issues are frequently mixed together.[8-9]

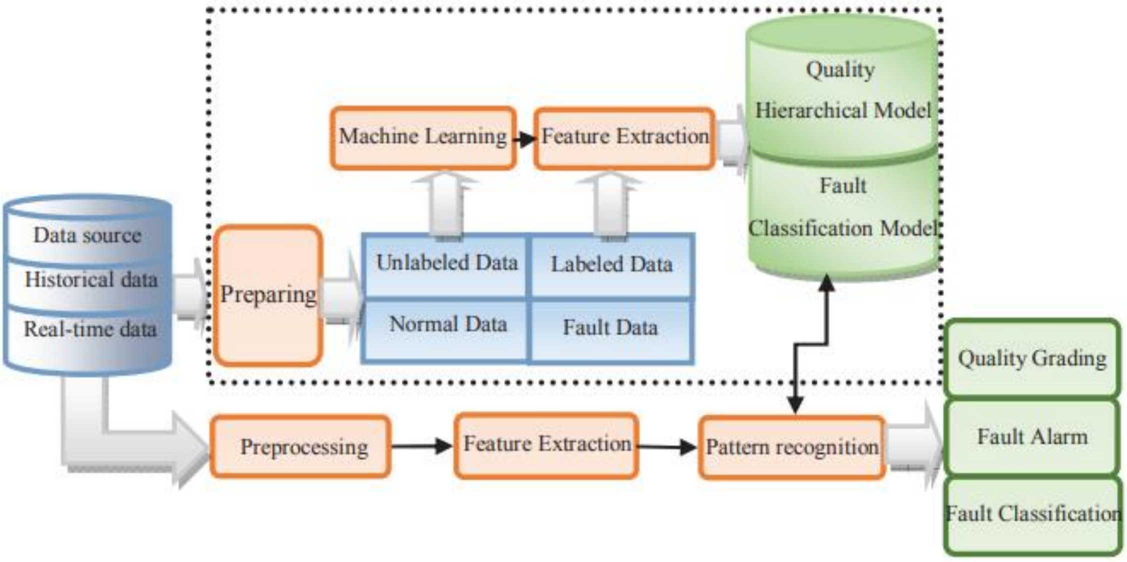

Figure 3 shows the process of ML after Data Science methods.

V. EDA

Data Scientists widely use EDA to understand datasets for decision-making and data cleaning processes. EDA reveals crucial information about the data, such as hidden patterns, outliers, variance, covariance, correlations between features. The information is essential for the hypothesis's design and creating better-performing models.

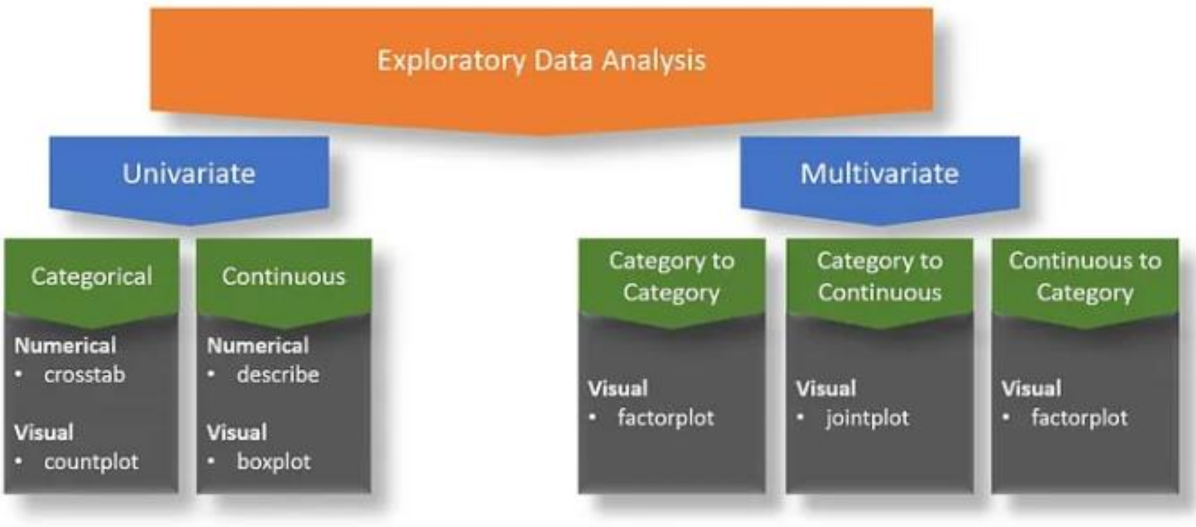

Figure 4 shows some EDA Plots with conditions

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.918 ⁜ Vol. 12, Issue 1, January 2023
DOI: 10.17148/IJARCCE.2023.12136

VI. TOTAL QUALITY MANAGEMENT

Process P1 refers to the requirement to understand data definition in order to gauge the quality of database information. It also refers to the assessment of the data meaning and data structures. Assessing information quality is done through Process P2, which manages information as an asset. Process P3 detects business performance metrics, computes the cost of associated information and the cost of missing information quality, computes the value for the customer, determines customer segmentation, and computes the value of the information. Perform data re-engineering and cleaning as part of Process P4 to enhance the information product by addressing the signs of a lacking quality. The flawed data is transformed into data with a sufficient level of quality during this process. Process P5: Enhance the information processing quality. Establish an environment for high-quality information (Process P6) represents the guidelines and cultural prerequisites necessary to maintain an environment for information quality that is always improving. As a result, this serves as the foundation for all other procedures and establishes the implementation strategy for DQ.[9]

VII. CONCLUSIONS

Data Science includes analysis, measurement, and process improvement. These comprise the procedures required for data collection and measurement in order to analyse performance and enhance effectiveness and efficiency. They comprise processes for measuring, monitoring, auditing, performance analysis, and improvement (for corrective and preventative actions, for example). While analysis and improvement processes are frequently viewed as autonomous processes that interact with other processes, receive input from measurement results, and send outputs for the improvement of those processes, measurement processes are frequently documented as an integral part of the management, resource, and realisation processes. TQM is closely related to outcomes of EDA through data analysis.

REFERENCES

[1] Fu D, Peng X, Yang Y. "Unbalanced tree-formed verification data for trusted platforms," Security and Communication Networks, 2016, 9(7), pp. 622-633.
[2] Zhang Y, Hepner G F. "The Dynamic-Time-Warping-based k-means plus plus clustering and its application in phenoregion delineation," International Journal of Remote Sensing, 2017, 38(6), pp. 1720-1736.
[3] Li Z, Huang Q, Carbone G J, et al. "A high-performance query analytical framework for supporting data-intensive climate studies," Computers Environment and Urban systems, 2017, 62, pp. 210-221.
[4] Khamphakdee N, Benjamas N, Saiyod S. "Performance Evaluation of Big Data Technology on Designing Big Network Traffic Data Analysis System," 2016 Joint 8th International Conference on soft computing and Intelligent Systems (SCIS) and 17th International Symposium on Advanced Intelligent Systems (ISIS), 2016, pp. 454-459.
[5] Park K, Baek C, Peng L M. "A Development of Streaming Big Data Analysis System Using In-memory Cluster Computing Framework: Spark," Park J, Jin H, Jeong Y S, et al. Lecture Notes in Electrical Engineering. NEW YORK: SPRINGER, 2016, pp. 157-163.
[6] Schaetzle A, Przyjaciel-Zablocki M, Skilevic S, et al. "S2RDF: RDF Querying with SPARQL on Spark," Proceedings of the VLDB Endowment, 2016,9(10), pp. 804-815.
[7] Joy R, Sherly K K. "Parallel Frequent Itemset Mining with Spark RDD Framework for Disease Prediction," Proceedings of IEEE International Conference on Circuit, power and Computing technologies (ICCPCT 2016), 2016.
[8] Shafer J, Rixner S, Cox A L. "The Hadoop distributed filesystem: Balancing portability and performance," IEEE International Symposium on PERFORMANCE Analysis of Systems & Software. IEEE, 2010, pp. 122-133.
[9] Chen H, Lu K, Sun M M, et al. "Enumeration system on HBase for lowlatency" IEEE-ACM International Symposium on Cluster Cloud and Grid Computing. NEW YORK: IEEE, 2015, pp. 1185-1188.

GUIDE

VISHESH S born on 13th June 1992, hails from Bangalore (Karnataka) and has completed B.E in Telecommunication Engineering from VTU, Belgaum, Karnataka in 2015. He has also completed his MBA in e-Business and PG Diploma in International Business. He also worked as an intern under Dr. Shivananju BN, former Research Scholar, Department of Instrumentation, IISc, Bangalore. His research interests include Embedded Systems, Wireless Communication, BAN and Medical Electronics. He is also the Founder and Managing Director of the corporate company Konigtronics Private Limited. He has guided over a thousand students/interns/professionals in their research work and projects. He is also the co-author of many International Research Papers. Many international students (from more than 12 countries) are also working for his research projects.