Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Modeling and Analyzing a Comprehensive Framework of Big Data Value Chain

Riyam Qasim Mubarak Salih

Oil Products Distribution Company, department of Central Matching, Baghdad, Iraq .Email: riyam.qasim2022@gmail.com

ARTICLE INFO

Article history:
Received: 12 /1/2025
Rrevised form: 8 /2/2025
Accepted : 9 /3/2025
Available online: 30 /3/2025

Keywords:

Big data

Value chain

Data collection

Data cleaning

Data analyzing

ABSTRACT

The rapid growth of big data has introduced significant challenges in storing, processing, and extracting actionable insights from diverse datasets. **I**n this paper, presenting and addressing the problem statements of this paper is how to model and analyses the Big Data Value Chain (BDVC) systematically in order to improve the efficiency and how to use to enable the data-driven of the decision make. We propose an integrated framework that leverages a novel algorithmic approach for data ingestion, storage, processing, cleaning, and analysis. The proposed system combines predictive analytics with state-of-the-art preprocessing techniques to transform raw data into valuable business insights. The framework is validated using three different datasets from the UCI Machine Learning repository, Kaggle, and AWS Public Datasets. Our results demonstrate improvements in data quality and predictive performance, thereby facilitating better compliance with regulatory standards and enhancing overall decision support.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.11977

1. Introduction

Big Data is a very large and complex amount of data that is difficult or impossible to store, analyze and manage using traditional storage and processing tools and techniques or traditional data management applications [1].

Therefore, data is classified into in-memory data and out-of-memory data. In-memory data is data that is processed and analyzed using simple tools, while out-of-memory data is data that is difficult to process and analyze using in-memory data. Therefore, it is necessary to find a method or a set of combined methods to deal with big data in order to extract what is intended from it [2, 3].

Big data is a set of data that, if existing programs want to process and manage it within an acceptable period of time, requires a larger available volume, which is difficult to provide. Big data could be initiated from different sources

∗Corresponding author Riyam Qasim Mubark salih

Email addresses: *:* riyam.qasim2022@gmail.com

Communicated by 'sub etitor'

including Internet, cloud-computing, social-media platforms, variety sensors, medical device(s), scientific laboratories(s), and more [4].

One of the most important features of big data is its large size, as well as its rapid change, in addition to its diversity and complexity at the same time. Therefore, it requires many tools and technologies that play an effective and special role in carrying out the storage, analysis and extraction of the required values [5]. When dealing with big data by institutions and companies, it provides predictive information regarding customer trends, tendencies and behavior, thus facilitating the analysis process, decision-making, and extracting the required values [6, 7].

The semi-structured data contains some organized elements and others that are unorganized, and it is data that has partially organized structures, such as an email that contains a set of organized fields such as (sender, recipient, subject) but may contain unorganized text in the content of the email. While organizations and companies which are dealing with large-amount(s) of data in various industries face several obstacles and difficulties, which are: data analysis, data storage, data quality, security, temporal analysis, data reliance, data flow control, privacy and security problem, cost problem, data protection laws problem [8 – 10].

A value chain is a set of interconnected resources and processes that begin with the creation of a dataset or directly dealing with an existing and available dataset, primarily electronically, and extend to analyzing and extracting the desired set of values from the data. The concept of the value chain was developed by Porter in the 1980s [11].

Furthermore, the data value chain (DVC) could be expressed as a mechanism that defines a large number of processes that can be repeated in order to gradually obtain the value of data from its creation stage to its processing and transfer (if required) and its flexible provision to the relevant user. The data value chain consists of five main steps: data collection, data storage, data processing, data cleaning, and data analysis [12, 13].

The main objective of the article is to address the challenges by dealing with three different and diverse sources of datasets to implement digital solutions using Big Data and data analytics. Thus, modeling an analytical system based on the principles of BDVC. Data is generated (data recording and capture), collected (data validation and storage), analyzed (data processing and analysis for generating fresh knowledge) and exploited (results for further utilizing) and thus, heterogeneous data needs are linked to generate insights [14, 15].

2. Literature review

Faroukhi et al. [16], reviews the evolution of Big Data Value Chains (BDVCs) and their potential for data monetization. It presents an exhaustive end-to-end BDVC framework and discusses approaches to monetize data within organizations. The study emphasizes the need for tailored monetization models for big data environments.

Kim et al. [17], explores how data-driven approaches complement expert-based analyses in value chain mapping, using business transaction records for comparative analysis. Findings suggest data-driven mappings enhance accuracy but require significant preprocessing and algorithmic effort. Expert validation remains vital for flexible interpretations.

Lamba et al. [18], identifies 14 enablers for successful big data initiatives in operations and supply chain management (OSCM) using MCDM techniques. Key enablers include top management commitment and big data quality management. A hierarchical and causal model highlights interrelationship among these enablers.

Aydin et al. [19], develops a model of the BDVC highlighting technologies for data acquisition, storage, and analysis. It reviews big data applications and provides a framework to guide technology selection for specific use cases. The study aims to align big data characteristics with technology capabilities.

Jain et al. [20], addresses implementing big data analytics (BDA) in sustainable supply chain management (SSCM) using the PESTEL framework. A structural model identifies factors like IT policy and environmental optimization as critical for sustainability. It highlights BDA's role in eco-efficiency and strategic decision-making.

Table (1) shows a comprehensive comparison between the analytical model in this article and the studied literature reviews studies in term of design approach, findings, limitations and practical implications.

Table 1 - A comprehensive comparisonof the literature reviews.

Model	Design Approach	Findings	Limitations	Practical Implication
Faroukhy et al. [16]	construct an end-to-end BDVC and explore data monetization approaches	facilitate data monetization, offering pathways for profit generation through direct or indirect means	Limited focus on empirical validation and industry-specific applications	Highlights the necessity of tailored monetization models and comprehensive BDVC
Kim et al. [17]	Experimental comparison of expert-based and data-driven value chain mapping approaches	Data-driven mappings are accurate but rigid; expert insights enhance flexibility	High algorithm development effort and dependency on quality preprocessing for data use	Promotes a hybrid approach combining data-driven and expert insights for realistic mappings
Lamba et al. [18]	MCDM techniques (ISM, Fuzzy-TISM, DEMATEL) to model interrelationships among big data enablers	Top management commitment and quality management are key drivers in OSCM big data initiatives	Inputs limited to a small group of experts; broader validation needed	Guides prioritization of enablers to enhance big data implementation in OSCM
Aydin et al. [19]	Development of a BDVC model linking big data characteristics to technologies for specific stages.	Provides a comprehensive technology framework for BDVC stages and practical application examples.	Does not deeply address implementation challenges across industries or data types.	Assists in selecting appropriate technologies for big data processes in various use cases.
Jain et al. [20]	PESTEL framework and TISM methodology to identify and structure SSCM factors for BDA implementation.	Policy support, IT culture, and environmental optimization are critical for sustainable supply chains.	Focused on manufacturing in emerging economies; less generalizable to other sectors or regions.	Informs strategic planning for eco-efficient supply chains using big data analytics.

3. Methodology

The important and basic thing in the process of dealing with big data lies in two stages, the first of which is the decision about what data you want to deal with, or in other words, what are the sources of big data you want to deal with, and the second part includes the steps, processes and stages that will be applied to the big data that were identified in the first part above. On this basis, this section was divided into two parts, which are the data set and the analytical model, as shown below:

3.1 Datasets

First thing it should be done is that using different datasets and applying the comprehensive steps of BDVC. In this article three different datasets are used for exploring BDVC, including collection, storage, processing, cleaning, and analysis: (UCI Machine Learning repository, Kaggle Datasets, AWS Public Datasets). The UCI repository provides a collection of datasets for machine learning and data analysis tasks. Kaggle hosts a variety of datasets for beginners and advanced users. AWS (Amazon Web Services) offers a registry of public datasets such as satellite imagery, genomics data, and climate observations. Each of these sources provides datasets that align well with the value chain steps (collection, storage, processing, cleaning, and analysis). In this article, three used datasets are namely Source 1, Source 2, Source 3 respectively.

3.2Analytical model

To deal with proposed model for analyzing Big Data as BDVC an effective procedure is initiated to meet the main objectives in this article. The built model in this article consists of many stages, in this article the most important five stages (i.e., data collection, data storage, data processing, data cleaning, data analysis) are modelled. Big data analysis relies on systematic steps, including data collection, storage, processing, cleaning, and analysis. Below are the mathematical formulas or expressions associated with each step:

A) Data Collection

The process of data collection (or data gathering) can be expressed using the union of collected datasets:

$$D = \bigcup_{i=1}^{n} d_i \qquad (1)$$

Where:

D: The complete dataset.

di: An individual source (dataset) collected from a specific source (dataset).

n: The number of sources (datasets).

B) Data Storage

To calculate the total required storage size:

$$S = \sum_{i=1}^{n} S_i \qquad (2)$$

Where:

S: total storage size needed.

s_i: size of the data from source i (dataset i).

C) Data Processing

Processing involves transformations and filtering, which can be represented as:

$$D\acute{} = f(D) \qquad (3)$$

Where:

D: raw data.

D′: processed data.

f: transformation or processing function (e.g., filtering, merging, or aggregating).

D) Data Cleaning

$$D' = \frac{D''}{E} \tag{4}$$

Where:

D′′: cleaned data.
D′: processed data.
E: set of outliers or error data.

E) Data Analysis

The formulas depend on the purpose of the analysis. Here, the analysis consists of both mean and variance:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{5}$$

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{6}$$

Where:

μ: mean value.
x_i: value of the data from source i (dataset i).
n: the number of sources (datasets).
σ^2: the variance value.

The key steps that were completed are illustrated in Figure 1, which may be seen here. In order to achieve the objective that this article intends to achieve, these stages are derived from the additional BDVC core five processes, and they involve the combination of a new sophisticated algorithm.

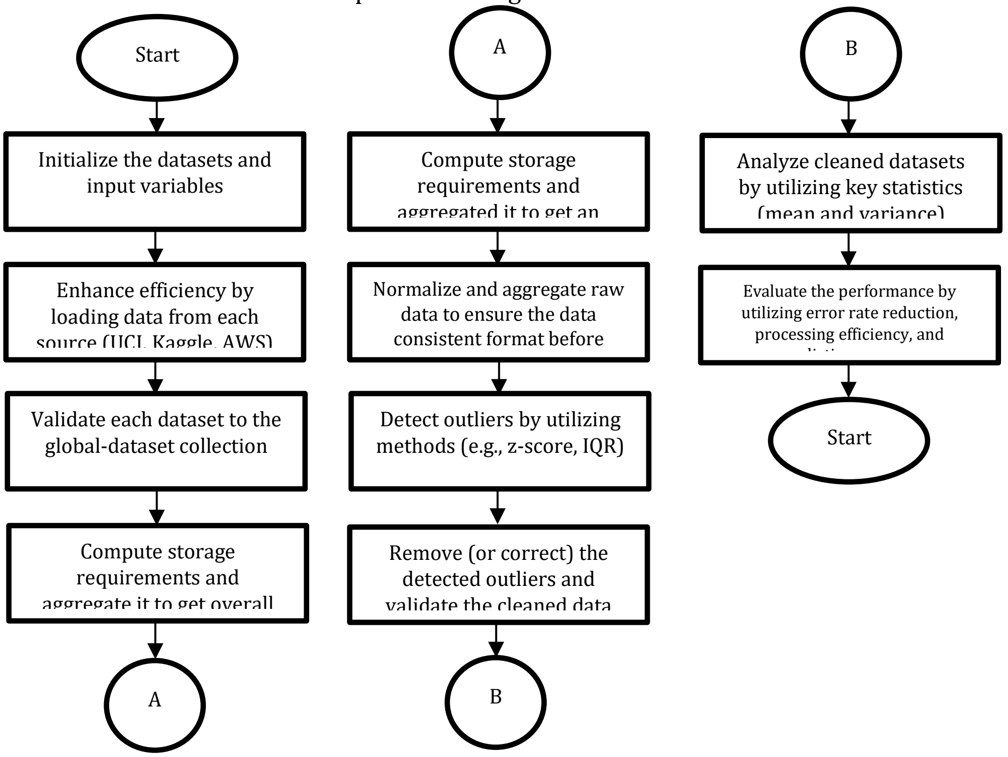
Start
Initialize the datasets and input variables
Enhance efficiency by loading data from each source (UCI, Kaggle, AWS)
Validate each dataset to the global-dataset collection
Compute storage requirements and aggregate it to get overall
A
A
Compute storage requirements and aggregated it to get an
Normalize and aggregate raw data to ensure the data consistent format before
Detect outliers by utilizing methods (e.g., z-score, IQR)
Remove (or correct) the detected outliers and validate the cleaned data
B
B
Analyze cleaned datasets by utilizing key statistics (mean and variance)
Evaluate the performance by utilizing error rate reduction, processing efficiency, and
Start

Fig. 1. Proposed framework.

4. Results and discussion

To deep understanding the model objectives, the main five steps are modeled based on formulas (1), (2), (3), (4) and (5) respectively as shown in the following steps:

Step 1(Data Collection): Figure (2) illustrates a 3D waste plot of generating data points, which is mainly representing by data collected from three different sources. Each axis in this figure represents the corresponding source, while the points characterize the individual data entries. This phase is responsible for gathering the data from various sources or datasets or systems. It ensures that a diverse and effective dataset is available for the sub-sequent steps.

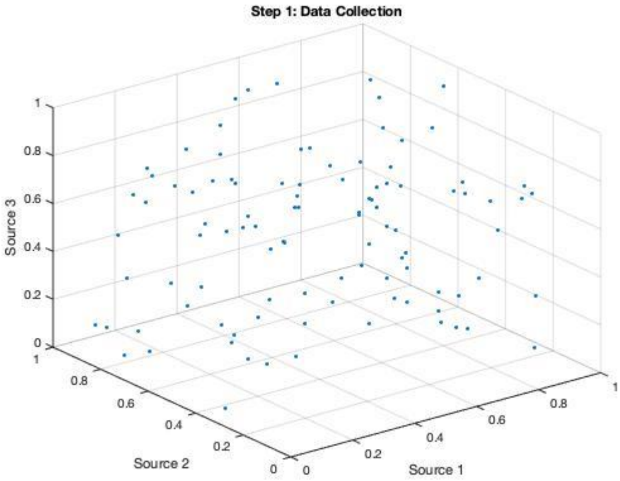

Fig. 2. Data Collection.

Step 2 (Data Storage): Figure 3 illustrates the size of the three different datasets in unit of megabytes (MB). Each bar represents the storage requirement for each dataset, which is allowing to show an effective comparison among the three various datasets. After collection stage, data is needing to be stored in an efficient way for ease access and processing as well. The requirements in this step are expressed crucial, especially for large-scale datasets. This step involves calculating and visualizing the size of each dataset to ensure adequate storage allocation.

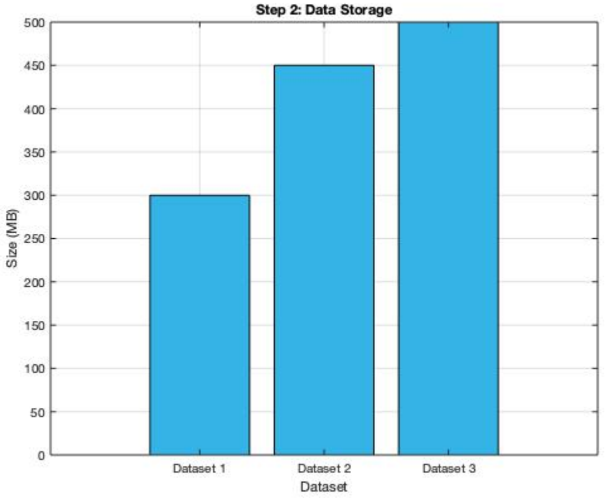

Fig. 3. Data Storage.

Step 3 (Data Processing): Figure (4) illustrates the mean value which is calculated for each data entry the across all sources. In the data processing, it is used to make transformation the raw of data into a readable, useful and usable format by applying the operations of aggregation and normalization. In this figure, the mean value is calculated for each entry across all sources, providing a simplified representation of the data while preserving essential patterns.

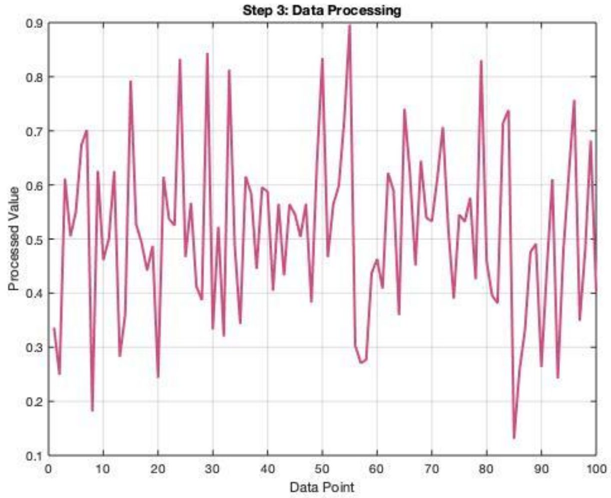

Fig. 4. Data Processing.

Step 4 (Data Cleaning): Figure (5) illustrates the comprehensive comparison of the noisy data (as shown in this figure as an orange circles) and the cleaned data (as shown in this figure as a green stars). Noisy or invalid data points are identified and then removed, to keep only the cleaned dataset for further analysis. Cleaning used to remove irrelevant information, unusable or errors from the data. The figure shows how the cleaning used to eliminate outliers to enhance the data integrity.

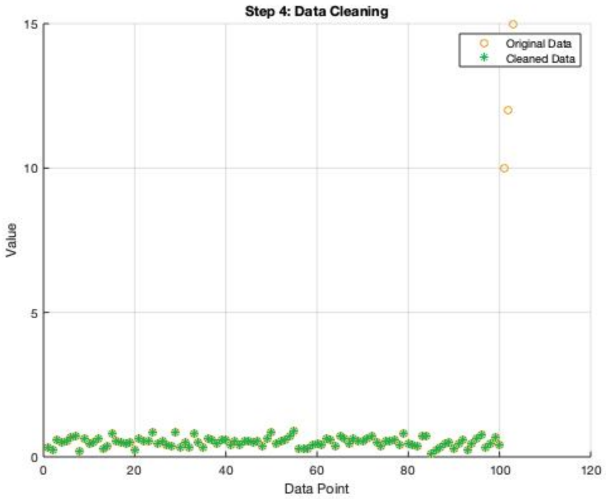

Fig. 5. Data Cleaning.

Step 5 (Data Analysis): The last analytical figure for the last step in the BDVC as shown in figure (6). Which is including the last phase (cleaned-data) distribution of the frequency. Generally, the x-axis is expressing data values while the y-axis shows the number of event processes related to the dataset. This phase provides visualization attribute of the distribution of values, give a highlight for trending.

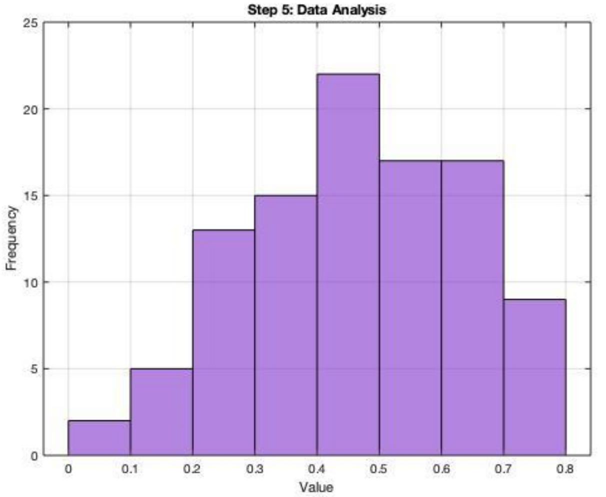

Fig. 6. Data Analysis.

A supplementary evaluation phase has been incorporated in order to simulate an evaluation metric (for instance, the Error Reduction Percentage) for each of the five processing steps in their corresponding order. Figure 6 contains a bar chart that illustrates how the quality of the data improves as it moves through the processes of collecting, storing, processing, cleaning, and analyzing the data. This chart is included in this figure because it provides an illustration of how the data will improve.

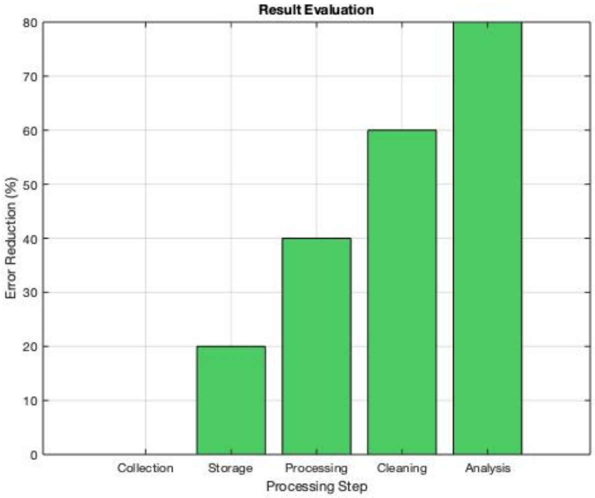

Fig. 6. Result Evaluation.

5. Conclusions

The Big Data Value Chain model that was created and worked on in this article represents an integrated model in all aspects, whether computational or engineering. It performs all engineering and mathematical operations related to Big Data, starting from the creation stage, through the stage of extracting the desired and relevant value, to analyzing it and then dealing with it after it is ready.

One of the most prominent points that show the importance of working in this article is creating an analytical study of the Big Data Value Chain by creating a framework that provides assistance to companies and institutions that have a close relationship with working with Big Data in a way that leads to achieving two main goals: improving efficiency in addition to making data-based decisions using predictive analytics.

As is known, the main stages related to managing and processing Big Data include the data creation stage as well as the stages of storing, processing, analyzing and visualizing it, by using three data sets: UCI, Kaggle and AWS. The results showed that the Big Data Value Chain was improved to obtain, and actionable business insights were obtained.

As future research directions, some of the following functionalities may be added to improve efficiency such as extend the proposed framework, enhance proposed algorithm by integrating multilevel deep learning algorithm, scalabil the related studies, and utilize user-centric evaluation for more accurate evaluation results

Acknowledgements

I would like to thank the Ministry of Higher Education and Scientific Research in Iraq for their courage, help and support during my study and work in this article.

References

[1] Sagiroglu, Seref, and Duygu Sinanc. "Big data: A review." In *2013 international conference on collaboration technologies and systems (CTS)*, pp. 42-47. IEEE, 2013.

[2] Chakraborty, Pranjal, Naser Ezzati-Jivan, Vahid Azhari, and François Tetreault. "AltOOM: A Data-driven Out of Memory Root Cause Identification Strategy." In *2023 IEEE International Conference on Big Data (BigData)*, pp. 1637-1646. IEEE, 2023.

[3] Leis, Viktor, Michael Haubenschild, Alfons Kemper, and Thomas Neumann. "LeanStore: In-memory data management beyond main memory." In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pp. 185-196. IEEE, 2018.

[4] Leskovec, Jure, Anand Rajaraman, and Jeffrey David Ullman. *Mining of massive data sets*. Cambridge university press, 2020.

[5] Flyverbom, Mikkel, Ronald Deibert, and Dirk Matten. "The governance of digital technology, big data, and the internet: New roles and responsibilities for business." *Business & Society* 58, no. 1 (2019): 3-19.

[6] Okorie, Gold Nmesoma, Zainab Efe Egieya, Uneku Ikwue, Chioma Ann Udeh, Ejuma Martha Adaga, Obinna Donald DaraOjimba, and Osato Itohan Oriekhoe. "Leveraging big data for personalized marketing campaigns: a review." *International Journal of Management & Entrepreneurship Research* 6, no. 1 (2024): 216-242.

[7] Theodorakopoulos, Leonidas, and Alexandra Theodoropoulou. "Leveraging Big Data Analytics for Understanding Consumer Behavior in Digital Marketing: A Systematic Review." *Human Behavior and Emerging Technologies* 2024, no. 1 (2024): 3641502.

[8] Azad, Poopak, Nima Jafari Navimipour, Amir Masoud Rahmani, and Arash Sharifi. "The role of structured and unstructured data managing mechanisms in the Internet of things." *Cluster computing* 23 (2020): 1185-1198.

[9] Strekalova, Yulia A., and Mustapha Bouakkaz. "Semi-structured data." In *Encyclopedia of Big Data*, pp. 816-819. Cham: Springer International Publishing, 2022.

[10] Zhang, Lijun, Ning Li, and Zhanhuai Li. "An overview on supervised semi-structured data classification." In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1-10. IEEE, 2021.

[11] Feller, Andrew, Dan Shunk, and Tom Callarman. "Value chains versus supply chains." *BP trends* 1, no. 3 (2006): 165-173.

[12] Shankar, Shashi Kant, María Jesús Rodríguez-Triana, Adolfo Ruiz-Calleja, Luis P. Prieto, Pankaj Chejara, and Alejandra Martínez-Monés. "Multimodal data value chain (m-dvc): A conceptual tool to support the development of multimodal learning analytics solutions." *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje* 15, no. 2 (2020): 113-122.

[13] De Simone, Cristina, Federica Ceci, and Cristina Alaimo. "Data ecosystem and data value chain: An exploration of drones technology applications." In *Sustainable Digital Transformation: Paving the Way Towards Smart Organizations and Societies*, pp. 203-218. Cham: Springer International Publishing, 2022.

[14] Gervasi, Massimiliano, Nicolò Gianmauro Totaro, Giorgia Specchia, and Maria Elena Latino. "Unveiling the Roots of Big Data Project Failure: a Critical Analysis of the Distinguishing Features and Uncertainties in Evaluating Big Data Potential Value." In *itaDATA*. 2023.

[15] Faroukhi, Abou Zakaria, Imane El Alaoui, Youssef Gahi, and Aouatif Amine. "An adaptable big data value chain framework for end-to-end big data monetization." *Big Data and Cognitive Computing* 4, no. 4 (2020): 34.

[16] Faroukhi, Abou Zakaria, Imane El Alaoui, Youssef Gahi, and Aouatif Amine. "Big data monetization throughout Big Data Value Chain: a comprehensive review." *Journal of Big Data* 7 (2020): 1-22.

[17] Kim, Kyungtae, and Sungjoo Lee. "How can big data complement expert analysis? A value chain case study." *Sustainability* 10, no. 3 (2018): 709.

[18] Lamba, Kuldeep, and Surya Prakash Singh. "Modeling big data enablers for operations and supply chain management." *The International Journal of Logistics Management* 29, no. 2 (2018): 629-658.

[19] Aydin, Ahmet Arif. "A comparative perspective on technologies of Big Data value chain." *IEEE Access* (2023).

[20] Jain, Prashant, Dhanraj P. Tambuskar, and Vaibhav Narwane. "Identification of critical factors for big data analytics implementation in sustainable supply chain in emerging economies." *Journal of Engineering, Design and Technology* 22, no. 3 (2024): 926-968.