International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 131–142
©Cerebration Science Publishing

Article

Toward a New Approach for Real-Time and Semantic Big Data Integration

Hana Mallek [1,*], Leila Bayoudhi [1,2], Faiza Ghozzi [1] and Faiez Gargouri [3]

[1] MIRACL Laboratory, ISIMS, Technology Center of Sfax, Sakiet Ezzit, Sfax 3021, Tunisia; bayoudhi.leila@gmail.com (L.B.); faiza.ghozzi@isims.usf.tn (F.G.)

[2] Faculty of Economic Science and Management of Mahdia, University of Monastir, Hiboon, Mahdia 50000, Tunisia

[3] MIRACL Laboratory, University of Sfax, Airport Road, Sfax 3029, Tunisia; faiez.gargouri@usf.tn

* Correspondence author: mallekhana@gmail.com

Received date: 13 February 2024; Accepted date: 15 May 2024; Published online: 6 January 2025

Abstract: The exponential growth of data derived from the Internet exceeds the social networks plateforms such as Facebook, Twitter and so on. This data includes a wide range of sensor data from a variety of sources, such as IoT devices, satellite data and so on. This massive amount of data led to the emergence of ”Big Data” concept. This new trend has had a considerable impact, particularly in the field of decision support systems. In particular, these impacts are mainly observed in the extraction, transformation and loading processes within business intelligence systems. In this context, three main challenges emerge: dealing with large quantities of data, various types of data and rapid data generation. In this paper, we present and discuss the state of the art of research focused on ETL processes while addressing the challenges of big data. Our study aims to determine the degree to which these works take into account the characteristics of big data in their approaches. Finally, we provide an overview of a new approach to ETL processes, called BRS-ETL, which supports heterogeneous and streaming data.

Keywords: ETL; Big Data; integration; variety; velocity; volume; streaming; heterogeneity

1. Introduction

Currently, the advancement of Internet technology and the proliferation of smart devices and communication systems used by billions of end users has given rise to the concept of “Big Data” [1]. This term refers to various types of data, including images, audio, videos, and text documents, that are generated and consumed at high velocity. Indeed, these data sources need a real-time treatment that allows organizations to extract insights and analyze data as it is generated to provide up-to-date information. Moreover, this timeliness enables decision-makers to respond quickly to changing conditions and identify emerging trends [2].

As a result of this exponential growth of data, many disciplines were influenced by this continuous spread such as sociology, medicine, biology, economics, management, and decision support systems (DSS) [3]. These systems are responsible for enabling business analysts and researchers to process data and extract relevant insights in order to make strategic decisions.

The DSS rely mainly on the Extracting, Transforming, and Loading (ETL) processes. These processes are executed sequentially for extracting data from diverse operational sources, transforming and loading it into a multidimensional Data Warehouse (DW). In this case, ETL processes are no longer adequate to handle such complexity and velocity of information [2,4]. To cope with these new challenges, new technologies were proposed to meet the specific company needs and to adapt the specific characteristics of data to be processed (e.g., SPARK (https://spark.apache.org/ accessed on: 02-01-2024), MapReduce (https://www.ibm.com/topics/mapreduce accessed on: 02-01-2024), KAFKA (https://kafka.apache.org/ accessed on: 02-01-2024), etc.). Furthermore, Big Data storage requires databases that are capable of accommodating the vast variety and volume of information. Indeed, Schema-less databases have emerged as a solution (e.g., MongoDB for document-oriented databases and HBase and Cassandra for column-oriented databases).

So, it is crucial to adapt ETL processes to handle massive data sources in real-time while taking into account data heterogeneity. This allows users to access data quickly to make pertinent decisions. By doing so, adapted ETL processes enable businesses to make faster decisions based on accurate and up-to-date data. Given these challenges, we conducted a comprehensive survey on the integration of massive data, specifically addressing its three key characteristics: Volume, Variety, and Velocity.

In this research paper, our aim is to identify and discuss the numerous challenges associated with integrating vast amounts of data while managing the ETL process. After reviewing several existing works, we will delve into an in-depth discussion to provide insights into an innovative approach to ETL processes called BRS-ETL (Big-Real time-Semantic ETL). Our proposal is based on the limitations found in the literature review to propose a generic approach that covers the 3 Vs of Big Data (Volume, Variety, and Velocity). This approach integrates new big data technologies and considers various characteristics of Big Data and the specific requirements of DSS, ensuring the extraction, transformation, and loading of relevant data at the appropriate time.

The remainder of this paper is structured as follows: Section 2 provides an overview of new technologies within a Big Data environment and offers background information on database technologies. Section 3 outlines various approaches proposed to tackle Big Data challenges while managing the ETL process. Section 4 discusses related work and highlights challenges in Big Data integration. Finally, before concluding, Section 5 presents a novel approach to handling large, heterogeneous, and rapidly generated data.

2. Big Data Technologies

Big data technologies have revolutionized the landscape of handling and storing data, ushering in an era of unprecedented scalability and efficiency. Innovative solutions such as Apache Hadoop (https://hadoop.apache.org/ accessed on:02-01-2024), Apache Spark, and distributed storage systems have emerged, enabling organizations to harness the power of parallel processing and distributed computing to effectively manage vast datasets.

Apache Hadoop stands as an open-source framework tailored for processing and storing large datasets across distributed clusters. At the core of its architecture we find the Hadoop Distributed File System (HDFS), which partitions data into smaller blocks and distributes them across nodes within a cluster. Leveraging the MapReduce [5] programming model for parallel processing, Hadoop Hadoop executes batch processing tasks with remarkable efficiency.

In contrast, Apache Spark is an in-memory data processing engine. It supports a wide range of data processing tasks, including batch processing, real-time stream processing, machine learning and graph processing. Spark's in-memory approach significantly accelerates data processing by minimizing disk I/O operations [6].

When it comes to storing data, organizations have a plethora of options, including distributed storage systems like Amazon S3 (https://aws.amazon.com/fr/s3/ accessed on: 03-02-2024), Google Cloud Storage (https://console.cloud.google.com accessed on: 03-02-2024), and HDFS (https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html accessed on: 03-02-2024). These systems distribute data across multiple servers or nodes in a cluster, ensuring fault tolerance and scalability through data replication mechanisms. As a result, they seamlessly accommodate the immense volumes of data generated in a big data environment while facilitating efficient access and retrieval.

Furthermore, the advent of NoSQL databases, exemplified by MongoDB (https://www.mongodb.com/ accessed on: 03-03-2024) and Cassandra (https://cassandra.apache.org/ accessed on: 03-02-2024), has introduced a paradigm shift from traditional relational databases. NoSQL databases offer flexible, schema-less data storage, with MongoDB specializing in document-oriented storage suited for unstructured and semi-structured data, while Cassandra excels in handling high volumes of write-intensive data with its column-oriented architecture.

Collectively, these technologies address the multifaceted challenges posed by big data [7], providing organizations with the scalability, speed, and adaptability necessary to derive actionable insights and value from massive datasets. Through their synergistic capabilities, they empower enterprises to unlock the full potential of big data and drive innovation in an increasingly data-driven world.

3. Contributions for Big Data Challenges

Processing Big Data in Decisional Systems presents a significant challenge. Ensuring efficient analysis and retrieval of relevant information from vast datasets is a crucial interest in this context.

In the following sections, we meticulously examine papers focusing on ETL processes addressing specific Big Data challenges: volume (see Section 3.1), velocity (see Section 3.2), variety (see Section 3.3) or hybrid a approaches (see Section 3.4).

To facilitate a clear comparison of the state-of-the-art work, each subsection features a table that scrutinizes works based on their approach to coping with structured (S), semi-structured (SS), and unstructured (US) data. Furthermore, each table dissects works focusing on the Extraction (E), Transformation (T), or Loading (L) phases, while also identifying the modeling level considered: Conceptual level (C), Logical level (L), or Physical level (P). Additionally, insights into the application of Big Data (BD) technologies are provided. Finally, each table evaluates whether these works address the multidimensional (MD) structure of a data warehouse or not.

This structured approach allows for a comprehensive understanding of the strategies and methodologies employed in addressing the intricacies of processing Big Data within Decisional Systems.

3.1. ETL-Based Approaches Coping with Volume

The Volume characteristic of Big Data is well-addressed in the literature, within several solutions focusing on ETL processes (see Table 1).

Table 1. A Comparative table of ETL-based approaches focusing on the Volume dimension.

	Data Sources			**Modeling**			**Technology Used**	**ETL Phases**			**MD**
	S	**SS**	**US**	**C**	**L**	**P**		**E**	**T**	**L**	
[8,9]	X					X	MapReduce		X		-
[10,11]	X			X		X	MapReduce		X		-
[12]		X				X	MapReduce	X			-
[13]		X				X	MapReduce		X	X	-
[14,15]		X		X		X	MapReduce	X	X		X
[16,17]	X	X		X		X	-	X			X

The ETLMR Framework [8] adopts the MapReduce programming model to adapt Slowly Changing Dimension (SCD) functionality in ETL processes.

CloudETL [9] uses Hadoop as a platform for running ETL processes and Hive as a Data warehouse. The MapReduce paradigm is also used in the extraction phase with TAREEG Framework [12] for collecting and filterng geographic data.

The proposed approach addresses several aspects that enhance decisional systems.

Moreover, Gupta et al. [13] introduced a framework called ”Web ETL” designed to handle web data. The aim of this work is to optimize the transformation and loading phases of the ETL process. The transformation phase is responsible for filtering erroneous and redundant data using data mining techniques. Subsequently, the loading phase is responsible for loading the filtered data into the data warehouse through the Hadoop ecosystem.

On the other hand, the authors proposed BigDimETL [14] to adapt typical ETL processes with the MapReduce paradigm. This adaptation aims to merge the parallelism aspect with the specific requirements of decisional systems. The proposal involves formalizing and implementing ETL basic operations (i.e., select, project, and join) during the Transformation phase. Following this, the authors [16] proposed a solution for adapting the extraction phase into the context of big data. They achieved this by formalizing and implementing a conversion algorithm to transform JSON file formats into a column-oriented NoSQL database.

These solutions improve undoubtedly decisional systems. However, the physical-level process modeling makes their usage more complex.

At the conceptual level, we find the P-ETL Framework [10,11] uses the MapReduce paradigm with Changing Data Capture (CDC) functionality. The proposed framework is developed within the Apache Hadoop environment to adapt the classical schema of ETL processes with the Map-Reduce paradigm. This work is significant in the decision-making context. It demonstrates the reliability of the Map-Reduce paradigm in accelerating the CDC functionality of the ETL process, but it does not take into account basic operations (selection, projection, conversion, etc.) that are responsible for filtering and cleaning the data to be loaded into a multidimensional Data Warehouse (DW).

Additionally, the conceptual level is addressed in BigDimETL [15,17] through an activity diagram template. In [17], the authors presented a conceptual modeling for the extraction phase aimed at ensuring the conversion and vertical partitioning operations. In [15], the authors presented a conceptual modeling for the transformation phase.

By reviewing Table 1, it becomes evident that all the referenced works have adopted the MapReduce paradigm to construct a physical solution, with a primary emphasis on the Transformation phase within ETL processes. This underscores the importance placed on efficient data processing and manipulation during the extraction, transformation, and loading stages.

At the modeling level, there is a clear trend towards prioritizing the physical implementation aspect, which involves the actual execution of tasks within the computing environment. However, it is noteworthy that only a few works address the conceptual level, which encompasses the high-level design and structure of the data integration process [10,11,15,16].

Despite the relevance of these works in advancing ETL processes, it is crucial to recognize their limitations. One notable aspect is the omission of considerations for handling unstructured data and accommodating the multidimensional structure during data integration. Only the work by [16] addressed the NoSQL model with the multidimensional structure. Neglecting these factors can impede the effectiveness of decision-making processes and limit the insights derived from the data. Future research should strive to address these gaps in order to develop more comprehensive and robust solutions for handling big data in decisional systems.

3.2. ETL-Based Approaches Coping with Variety

Several works addressed the variety feature in the ETL processes (see Table 2). For example, Bimonte et al. [18,19] coped with the data Variety through a multimodel Star Schema (Graph-oriented, document-oriented, etc.) to store data according to their native models in order to preserve data variety.

Table 2. A Comparative table of ETL-based approaches focusing on the Variety dimension.

	Data Sources			**Modeling Level**			**BD and Semantic Technology Used**	**ETL Phases**			**MD**
	S	**SS**	**US**	**C**	**L**	**P**		**E**	**T**	**L**	
[18,19]			X		X	X	-		X	X	X
[20]		X				X	Hive Ontology	X	X	X	-
[21]		X				X	Ontology	X	X	X	-
[22]		X		X		X	Web services Ontology	X	X	X	-
[23]			X			X	MapReduce HBase	X	X		-
[24]	X	X	X	X		X	Open linked data RDF OWL SPARQL		X		-
[25]	X	X	X	X		X	Open linked data RDF SPARQL		X		-

The objective of this work is to study the performance of a multi-model DBMS when used to store multidimensional data for OLAP-type analyses. On the other hand, Hilali et al. [20] addressed ELT basic operations and introduced Mapping-ELT (M-ELT), wherein the authors leveraged Hive to enhance data warehousing capabilities. They also utilized ontology to tackle the issue of semantic heterogeneity.

Moreover, Boulahia et al. [21] proposed a theoretical semantic solution by using ontologies to annotate text in the Extraction, Transformation and Loading phases. In addition, Guo et al. [23] proposed the SHMR system for storing and retrieving semantic information from heterogeneous multimedia data sources, to resolve a variety problems.

Furthermore, Mahmoud et al. [24] implemented a semantic ETL model that uses semantic web technologies (i.e., linked data as RDF) for aggregating, integrating, and representing data. Regarding the transformation phase, it is responsible for converting structured, semi-structured, or unstructured data into RDF format.

Regarding [22], the authors used web services for orchestrating ETL flows. Indeed, they proposed a generalization of all schemas of data sources (relational, semantic and graph) and ETL operators through Data-Driven Engineering.

Moreover, Bensal et al. [25] proposed a semantic ETL framework which generates a semantic model of the datasets to be integrated. Then, it generates a semantic linked data (as RDF triples) that can be queried using SPARQL (http://www.w3.org/TR/rdf-sparql-query/ accessed on: 02-01-2024). This work used a semantic approach only on the transform step of the ETL process.

Based on the findings presented in Table 2, it is evident that all research efforts addressing data variety have introduced physical solutions that incorporate novel semantic or Big Data technologies. However, it is worth noting that these solutions did not account for the multidimensional structure, with the exception of the works by Bimonte et al. [18,19].

This indicates a significant gap in the current research landscape, as overlooking the multidimensional structure could limit the effectiveness and relevance of data integration efforts. Addressing this aspect is crucial for developing more comprehensive and robust solutions that can handle diverse data types while aligning with the complex structures often found in decisional systems.

3.3. ETL-Based Approaches Coping with Velocity

Regarding the velocity feature of big data, several approaches have been proposed to address real-time data processing, aiming to overcome the limitations of the MapReduce paradigm. These approaches utilize in-memory processing, particularly leveraging SPARK technology, to efficiently handle the high velocity of data from various sources [26].

In this context, Machado et al. [27] proposed the Distributed On-Demand ETL (DOD-ETL) tool, which targets the bottleneck of ETL processes to achieve near real-time Change Data Capture (CDC) ETL functionality.

Mehmood et al. [28] implemented a semi-stream join process for NoSQL data streams within a real-time data warehousing scenario. This process involves extracting data streams from two distinct sources, performing real-time rearrangement and filtering according to the data warehousing format, and subsequently joining them.

Furthermore, Mehmood et al. [29] utilized in-memory database techniques to reduce frequent disk access in the data ETL process for join operations.

Zdravevski et al. [30] proposed a cloud-based architecture for efficient Big Data ETL, where Spark is employed for the extract and transform phases, followed by loading the results into a data warehouse using distributed load agents (DLAs). This approach optimizes resource utilization by offloading processing tasks to cluster slaves (edge nodes), thereby reducing the workload on the database server.

Additionally, Biswas et al. [31] introduced a formal modeling approach for real-time ETL, integrating machine learning models to automate data integration processes. This approach has been applied in credit risk assessment to mitigate risks and enhance revenue generation.

All the solutions listed in Table 3 advocate for a physical solution using Spark technology, indicating a clear trend towards leveraging the capabilities of Spark for big data processing. On the other hand, it is notable that a NoSQL document-oriented database, especially MongoDB, is utilized with [28,29], helping to provide real and semi-stream data in a timely manner.

Table 3. A Comparative table of ETL-based approaches focusing on the Velocity dimension.

Approach	Data Sources			Modeling Level			Used	ETL Phases			MD
	S	SS	US	C	L	P	Technology	E	T	L	
[29]						X				X	-
[30]	X	X				X	Spark	X	X		-
[27]	X	X		X		X	Kafka			X	-
[31]	X	X			X	X	MongoDB	X	X		X
[28]	X	X				X	MongoDB		X		-

3.4. Hybrid Approaches

This subsection sheds light on hybrid approaches where the works take into consideration more than one characteristic of big data such as variety-velocity, volume-variety etc. (see Table 4).

Table 4. A Comparative table of ETL-based approaches focusing on two or three dimensions of Big Data.

	Data Sources			BD Technology Used	Covered BD 3 Vs			ETL Phases			MD
	S	SS	US		Vol	Var	Vel	E	T	L	
[32]				MapReduce						X	-
[33]	X	X	X	Neo4j HBase		X	X	X	X	X	-
[34]			X	MapReduce	X	X		X			-
[35]				Hadoop Flink Hive HBase	X		X	X		X	-
[36]		X		OWL mongoDB	X	X	X	X		X	-

Indeed, both variety and velocity are addressed in [33], where an open-source graph database called pandaDB is introduced. PandaDB serves as an extension of Neo4j, specifically designed to facilitate the processing of unstructured data. The authors have also optimized the processing of query data within the graph, enabling a deeper understanding of the semantic information extracted from the unstructured data stored within it.

The variety and velocity problems are studied and surveyed by [37], where the authors present a systematic literature review about the integration of multimedia data sources through ETL processes in real time. The conclusion of this study is that the majority of researchers primarily focus on multimedia data, but only a few have specifically explored the concept of Multimedia Big Data in real time, particularly in a decisional context. Additionally, among the works, only a few considered the conceptual level in their approach to handling Multimedia Big Data.

Moreover, Sujatha et al. [34] presented a solution to accelerate multimedia retrieval through the use of the MapReduce paradigm. The approach primarily focuses on the extraction phase to semantically filter pertinent data.

Additionally, Balti et al. [35] considered both the volume and velocity characteristics of multimedia big data. The authors proposed a multidimensional framework for Big Earth Observation (EO) data warehousing. It consists of three main parts: data collection and preprocessing, data loading and storage, and visualization for spatio-temporal analysis.

To speed up the execution of the ETL processes and to give very fast streams, Berkani et al. [32] proposed a near real-time Data Warehouse design (NRTDW) that deals with ETL processes and selected materialized views. At runtime, ETL transformations are analyzed to see whether they can be executed using the allocated memory and the selected materialized views. Indeed, the use of views minimizes the graph comparison against the RDF quads set.

Furthermore, Abbess et al. [36] took into consideration the Big Data 3Vs namely, variety, volume and velocity. Their approach is made up of three steps. The first one consists in wrapping data sources to MongoDB databases using transformation rules. The second step consists in local OWL ontologies generation (i.e., modules construction). Regarding the third step, it consists in merging ontology modules into a global ontology [38,39]. To address the velocity feature of Big Data, the authors proposed distributed processing of data without implementing the solution. For the variety issue, the authors proposed to copy all data sources to a common representation.

4. Discussion and Synthesis

After reviewing the state of the art presented in the previous section, it can be concluded that significant efforts have been made in data integration to tackle the challenges posed by the substantial data volume.

These approaches have harnessed parallelism and used the MapReduce paradigm to enhance efficiency [8–15,17]. Other researchers have focused on tackling data variety challenges by leveraging semantic aspects [20–25]. Other works studies have dedicated efforts to addressing real-time data challenges using in-memory technology [27,29,30]. Additionally, hybrid approaches aim to address a combination of challenges, such as variety and volume [34], volume and velocity [35], or velocity and variety [32,33].

While these studies are relevant, they have highlighted a wide range of challenges associated especially with decision support systems.

These challenges are manifested at different levels, each of which is developed below.

- **Modeling Level** where the Big Data integration faces significant challenges due to the limited attention given to the conceptual modeling of different ETL phases. Conceptual modeling is used to model the different ETL operations required to handle the Big variety and velocity of data.

  It also provides an abstract view of how data is extracted, transformed, and loaded, which can help minimize the DSS cost. In fact, 70% of DW projects are reserved for ETL process development, making it crucial to address the conceptual level of ETL processes [40].

  However, the majority of works tend to focus on technical implementation rather than conceptual presentation, leading to a notable gap in addressing the conceptual level of ETL processes, while covering the variety and velocity characteristics. Therefore, it is essential to give more attention to the conceptual modeling of ETL processes to ensure its ability to handle the Big variety and velocity of data.

- **Decision support system specifities** where the multidimensional data modeling is the elementary structure of loaded data into the DW. This structure is an effective model for organizing and structuring data as Facts and Dimensions for analysis and reporting purposes.

  However, the studied approaches do not take sufficient account of these specific requirements. This limits the effectiveness of data integration for decision support systems. Hence, meeting this challenge means adapting data integration techniques to accommodate the multidimensional structure of data and optimizing its use in decisional environments [14].

- **ETL processes bottleneck** where the excessive focus on either the Extraction or Transformation phases of ETL processes can lead to increased pressure within these particular phases, and to potentially hinder the overall ETL process [10].

  These bottlenecks can result in inefficiencies, increased processing times and resource restrictions, which impact the ability to maintain timely and accurate decision support. To overcome this problem, it is essential to balance the allocation of tasks and resources between all phases of the ETL process. In this case, implementing a more holistic approach to ETL ensures a better, more efficient data integration process.

- **Covering Big data characteristics**, where none of the existing work suggests a complete solution that addresses the three critical characteristics of Big Data: volume, variety and velocity. Indeed, each one presents its own set of challenges and solve them individually. This may not be enough to exploit the full potential of Big Data for DSS. In fact, we need a more integrated approach that takes all three characteristics into account simultaneously. Research should focus on developing strategies and frameworks that are capable of effectively managing data volumes, types, and speeds in the context of DSS.

5. New Approach: BRS-ETL

In this section, and in order to resolve the majority of data integration problems presented in the preveous sections, we will present our new approach called BRS-ETL (Big, Real time and Semantic ETL), which is a novel framework designed to address the challenges posed by big volume, real-time, and heterogeneous data.

Traditional ETL processes often struggle to cope with the velocity and variety of data streams, leading to bottlenecks in data processing pipelines and hindering timely decision-making.

The BRS-ETL framework aims to overcome these challenges by leveraging advanced techniques and methodologies tailored specifically for handling big volume, real-time, and heterogeneous data. By integrating cutting-edge technologies such as distributed computing, stream processing, and semantic analysis, BRS-ETL offers a comprehensive solution for efficiently ingesting, transforming, and loading data from diverse sources in a timely manner.

One of the key features of the BRS-ETL framework is its ability to seamlessly scale to accommodate fluctuating data volumes and processing demands.

Furthermore, BRS-ETL incorporates semantic-enhanced capabilities, allowing organizations to extract valuable insights from the rich semantic information embedded within their data sources. By leveraging semantic analysis techniques, BRS-ETL enables organizations to uncover hidden relationships, patterns, and trends within their data, empowering more informed decision-making.

Overall, the BRS-ETL framework represents a significant advancement in data integration and processing, offering organizations a powerful tool for unlocking the full potential of their big volume, real-time, and heterogeneous data assets.

5.1. BRS-ETL Architecture

The BRS-ETL framework comprises several important components that enable it to effectively handle big volume, real-time, and heterogeneous data (see Figure 1). Some of the key components include:

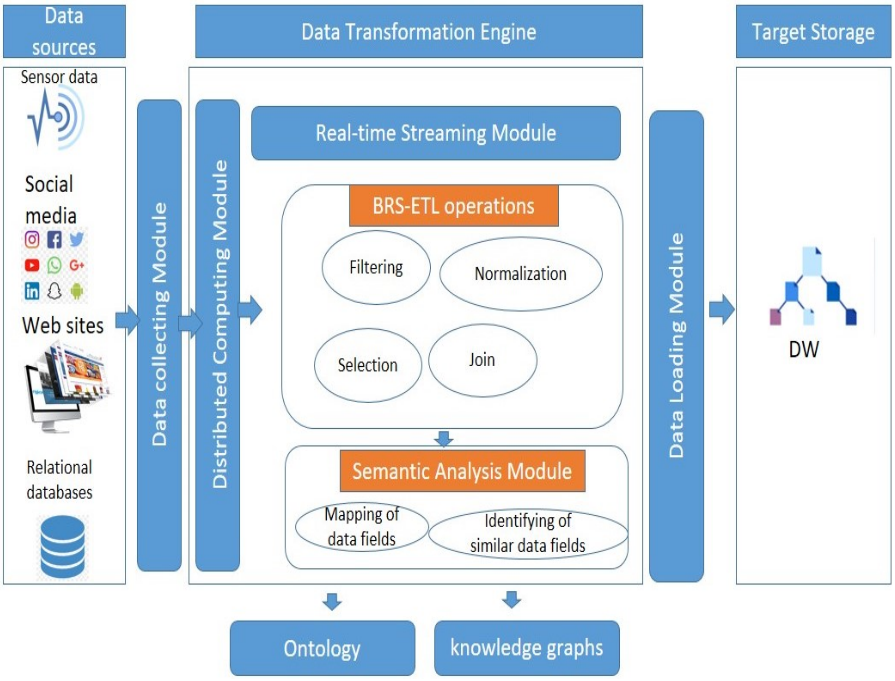

Figure 1. BRS-ETL Framework.

Data Collecting Module: This component is responsible for continuous data collection from various sources, including databases, streaming platforms, websites, social media, sensor data, etc.

Data Transformation Engine: The transformation engine performs adapted typical ETL operations within the distributed real-time environment, such as filtering, cleansing, and aggregating to prepare the data for analysis. It leverages advanced algorithms and techniques to handle large volumes of data and accommodate real-time processing requirements. This module incorporates three sub-modules:

- **Distributed Computing Module:** BRS-ETL relies on distributed computing infrastructure to enable parallel processing and scalability.
- **Real-time Streaming Module:** For real-time data processing, BRS-ETL includes a streaming module that supports continuous data processing by invoking BRS-ETL operations for filtering, joining, and cleaning data in real-time.
- **Semantic Analysis Module:** This module serves as a metadata layer which helps to analyze and extract semantic content from the data, enabling the extraction of meaningful insights and relationships. It employs natural language processing (NLP) techniques, ontology modeling [41,42], and knowledge graphs to enhance the understanding of the data and facilitate more informed decision-making. Operations such as mapping data fields and identifying similar data fields are performed within this module.

Data Loading Module: The loading module is responsible for storing processed data into a multidimensional data warehouse for further analysis and reporting. It ensures data integrity, reliability, and consistency while accommodating diverse data formats and structures.

5.2. Toward Implementation

Several new technologies play a crucial role in empowering the BRS-ETL modules to effectively handle big volume, real-time, and heterogeneous data. In Figure 2, we illustrate the key big data technologies that each module can leverage to achieve its objectives.

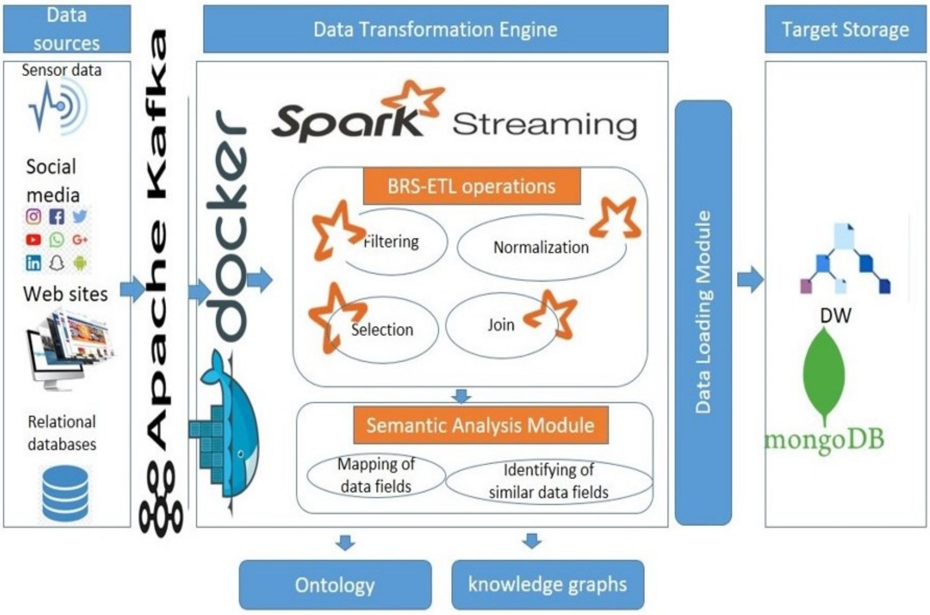

Figure 2. BRS-ETL Framework Technologies.

For **the Data Collecting Module:** Apache Kafka [43] stands out as a distributed streaming platform enabling high-throughput, real-time data ingestion from various sources.

In the **Data Transformation Engine:** We consider Docker [44] for the Distributed Computing Module. Docker, an open-source container orchestration platform, automates the deployment, scaling, and management of containerized applications. It provides a flexible and scalable infrastructure for distributed computing.

Within **the Real-time Streaming Module:** Apache Spark is chosen as a distributed computing framework offering in-memory processing capabilities for large-scale data transformation and analysis.

For **the Semantic Analysis Module:** Natural Language Processing (NLP) Libraries (e.g., NLTK, SpaCy) are utilized. These libraries provide tools and algorithms for processing and analyzing human language data, enabling semantic understanding and extraction. Additionally, Knowledge Graphs (e.g., Neo4j) are employed as Graph databases that represent knowledge in a structured format, facilitating semantic analysis and relationship discovery.

Regarding **the Data Loading Module:** Apache MongoDB is introduced as a distributed document-oriented Data Warehouse that supports real-time storage capabilities, ensuring efficient and reliable storage of processed data.

This carefully selected suite of technologies empowers each module within the BRS-ETL framework to tackle the challenges of big volume, real-time, and heterogeneous data effectively, ensuring seamless data integration and processing.

6. Conclusions

The rapid proliferation of data originating from Internet sources and social media platforms and sensor data has given rise to the concept of Big Data. This phenomenon has produced wide-ranging impacts, notably within the domain of DSS. In particular, these impacts are mainly noticeable in the ETL phases.

This research paper offers a comprehensive survey of data integration techniques that aim at tackling various challenges posed by big data. Upon analyzing this study, it becomes clear that existing research frequently neglects several critical constraints, including the modeling level, consideration of multidimensional structures, coverage of important big data characteristics, and potential bottlenecks.

To address these issues, we proposed a new approach called BRS-ETL that covers all three primary big data characteristics by integrating additional modules into ETL processes to handle big, heterogeneous, and rapidly changing data. Our approach, enhances the capabilities of traditional ETL processes to effectively manage the complexities of modern data environments.

Moving forward, there are several promising future research directions to explore. One area of focus is ontology-based data integration implementation, in order to involve enhancing semantic enrichment for large-scale and heterogeneous data with diverse schemas. By leveraging ontologies and semantic technologies, organizations can improve the accuracy and efficiency of data integration processes, enabling better decision-making and insights extraction.

Additionally, it is essential to consider the multidimensional structure when dealing with semantic and real-time extracted data. By incorporating multidimensional modeling techniques into data integration frameworks, organizations can better capture the complex relationships and hierarchies present in their data, leading to more comprehensive and meaningful analyses.

Overall, addressing these challenges and exploring new research directions will contribute to advancing the field of data integration and improving the effectiveness of data management practices in the era of big data. The proposed BRS-ETL framework represents a significant step towards achieving these goals and unlocking the full potential of data-driven decision-making.

Author Contributions

Conceptualization, H.M. and L.B.; methodology, H.M.; validation, H.M., L.B. and F.G.; investigation, H.M.; resources, F.G.; writing—original draft preparation,all manuscript's authors; writing—review and editing, F.G.; supervision, F.G. All authors have read and agreed to the published version of the manuscript.

Funding

The authors declare that they have no funding for the present study.

Conflicts of Interest

Author declares no conflict of interest.

Data Availability Statement

Not applicable.

References

1. Oussous, A., Benjelloun, F. Z., Lahcen, A. A., and Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, 30(4), 431-448.
2. Mehmood, Erum, and Tayyaba Anees. ”Challenges and solutions for processing real-time big data stream: a systematic literature review.” *IEEE Access* 8 (2020): 119123-119143.

3. Rinaldi, Antonio M., and Cristiano Russo. ”A semantic-based model to represent multimedia big data.” In *Proceedings of the 10th International Conference on Management of Digital Ecosystems*. 2018.
4. Boulahia, C., Behja, H., Chbihi Louhdi, M. R., and Boulahia, Z. (2024). The multi-criteria evaluation of research efforts based on ETL software: from business intelligence approach to big data and semantic approaches. *Evolutionary Intelligence*, 1-26.
5. Hedayati, S., Maleki, N., Olsson, T., Ahlgren, F., Seyednezhad, M.and Berahmand, K. (2023). MapReduce scheduling algorithms in Hadoop: a systematic study. *Journal of Cloud Computing*, 12(1), 143.
6. Ibtisum, S., Bazgir, E., Rahman, S. A.and Hossain, S. S. (2023). A comparative analysis of big data processing paradigms: Mapreduce vs. apache spark. *World Journal of Advanced Research and Reviews*, 20(1), 1089-1098.
7. Challal, Z., Bala, W., Mokeddem, H., Boukhalfa, K., Boussaid, O.and Benkhelifa, E. (2019, October). Document-oriented versus column-oriented data storage for social graph data warehouse. In *Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 242-247). IEEE.
8. X., Liu, C., Thomsen and T. B., Pedersen (2013). ETLMR: a highly scalable dimensional ETL framework based on mapreduce. Transactions on Large-Scale Data-and Knowledge-Centered Systems VIII: Special Issue on Advances in Data Warehousing and Knowledge Discovery, 1-31.
9. X., Liu, C., Thomsen and T. B., Pedersen (2014, July). CloudETL: scalable dimensional ETL for hive. In *Proceedings of the 18th International Database Engineering and Applications Symposium* (pp. 195-206).
10. M., Bala, O., Boussaid and Z. Alimazighi (2014, November). P-ETL: Parallel-ETL based on the MapReduce paradigm. In *Proceedings of the 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)* (pp. 42-49). IEEE.
11. M., Bala, O., Boussaid and Z. Alimazighi (2017). A Fine-Grained Distribution Approach for ETL Processes in Big Data Environments. *Data and Knowledge Engineering*, 111, 114-136.
12. Alarabi, L., Eldawy, A., Alghamdi, R., and Mokbel, M. F. (2014, November). TAREEG: A MapReduce-based system for extracting spatial data from OpenStreetMap. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 83-92).
13. G., Gupta, N., Kumar, and I., Chhabra (2020). Optimised transformation algorithm for hadoop data loading in web ETL framework. *EAI Endorsed Transactions on Scalable Information Systems*, 7(25), e6-e6.
14. Mallek, H., Ghozzi, F., and Gargouri, F. (2020). Towards extract-transform-load operations in a big data context. *International Journal of Sociotechnology and Knowledge Development (IJSKD)*, 12(2), 77-95.
15. Mallek, H., Ghozzi, F., and Gargouri, F.(2023) Conceptual modeling of Big Data SPJ operations with Twitter social medium. *Social Network Analysis and Mining*, 1-31.
16. H., Mallek, F., Ghozzi, F. Gargouri, Conversion operation: from semi-structured collection of documents to Column-oriented structure. In *Proceedings of the 22nd International Conference on Hybrid Intelligent Systems* (HIS 2022).
17. Mallek, H., Ghozzi, F., and Gargouri, F. Conceptual modeling of Big Data extraction phase. *International Journal of Hybrid Intelligent Systems*, 1-16.
18. Bimonte, S., Hifdi, Y., Maliari, M., Marcel, P., and Rizzi, S. (2020, March). To each his own: Accommodating data variety by a multimodel star schema. In *Proceedings of the 22nd International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data Co-Located with EDBT/ICDT 2020 Joint Conference (EDBT/ICDT 2020).*
19. Bimonte, S., Gallinucci, E., Marcel, P.and Rizzi, S. (2022). Data variety, come as you are in multi-model data warehouses. *Information Systems*, 104, 101734.
20. Hilali, I., Arfaoui, N.and Ejbali, R. (2022, March). A new approach for integrating data into big data warehouse. In *Proceedings of the Fourteenth International Conference on Machine Vision (ICMV 2021)* (Vol. 12084, pp. 475-480). SPIE.
21. Boulahia, C., Behja, H.and Louhdi, M. R. C. (2021, June). Towards Semantic ETL for integration of textual scientific documents in a Big Data environment: A theoretical approach. In *Proceedings of the 2020 6th IEEE Congress on Information Science and Technology (CiSt)* (pp. 133-138). IEEE.
22. Berkani, N., Bellatreche, L., and Guittet, L. (2018). ETL processes in the era of variety. *Transactions on Large-Scale Data-and Knowledge-Centered Systems: Special Issue on Database-and Expert-Systems Applications*, 98-129.
23. Guo, Kehua, et al. ”An effective and economical architecture for semantic-based heterogeneous multimedia big data retrieval.” Journal of Systems and Software 102 (2015): 207-216.

24. Mahmoud, A. , Mahmoud, Y. Shams, Elzeki, O. M. and Awad, N.A. (2021). Using Semantic Web Technologies to Improve the Extract Transform Load Model . *Computers, Materials & Continua*, 68(2), 2711-2726.
25. Bansal, S.K.and Kagemann, S. (2015). Integrating Big Data: A Semantic Extract- Transform-Load Framework. Computer, 48(3), 42-50.
26. Biswas, N., Biswas, S., Mondal, K. C., and Maiti, S. (2024). Challenges and Solutions of Real-Time Data Integration Techniques by ETL Application. *Big Data Analytics Techniques for Market Intelligence*, 348-371.
27. Machado, Gustavo V., et al. ”DOD-ETL: distributed on-demand ETL for near real-time business intelligence.” *Journal of Internet Services and Applications* 10 (2019): 1-15.
28. Mehmood, Erum, and Tayyaba Anees. ”Performance analysis of not only SQL semi-stream join using MongoDB for real-time data warehousing.” *IEEE Access* 7 (2019): 134215-134225.
29. Mehmood, Erum, and Tayyaba Anees. ”Distributed real-time ETL architecture for unstructured big data.” *Knowledge and Information Systems* 64.12 (2022): 3419-3445.
30. Zdravevski, E., Lameski, P., Apanowicz, C., and Śl zak, D. (2020). From Big Data to Business Analytics: The case study of churn prediction. *Applied Soft Computing*, 90, 106164.
31. Biswas, N., Mondal, A. S., Kusumastuti, A., Saha, S.and Mondal, K. C. (2022). Automated credit assessment framework using ETL process and machine learning. *Innovations in Systems and Software Engineering*, 1-14.
32. Berkani, N., Bellatreche, L., and Ordonez, C. (2018, May). ETL-aware materialized view selection in semantic data stream warehouses. In *Proceedings of the 2018 12th International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-11). IEEE.
33. Zhao, Zihao, et al. ”PandaDB: An AI-Native Graph Database for Unified Managing Structured and Unstructured Data.” *International Conference on Database Systems for Advanced Applications*. Cham, Switzerland: Springer Nature, 2023.
34. Sujatha, D., M. Subramaniam, and Chinnanadar Ramachandran Rene Robin. ”A new design of multimedia big data retrieval enabled by deep feature learning and Adaptive Semantic Similarity Function.” *Multimedia Systems* 28.3 (2022): 1039-1058.
35. Balti, Hanen, et al. ”Multidimensional architecture using a massive and heterogeneous data: Application to drought monitoring.” *Future Generation Computer Systems* 136 (2022): 1-14.
36. Abbes, H. and Gargouri, F. (2018). MongoDB-Based Modular Ontology Building for Big Data Integration. *Journal on Data Semantics*, 7, 1-27.
37. Mallek, H., Ghozzi, F., and Gargouri, F. Real-Time ETL for multimedia sources A systematic literature review. In *Proceedings of the 23nd International Conference on Intelligent Systems Design and Applications (ISDA 2023)*.
38. Bayoudhi, L., Sassi, N., and Jaziri, W. (2017). A hybrid storage strategy to manage the evolution of an OWL 2 DL domain ontology. *Procedia Computer Science*, 112, 574-583.
39. Bayoudhi, L., Sassi, N., Jaziri, W. (2019). Efficient management and storage of a multiversion OWL 2 DL domain ontology. *Expert Syst.* 36.
40. El Akkaoui, Z., Zimányi, E., Mazón, J. N.and Trujillo, J. (2011, October). A model-driven framework for ETL process development. In *Proceedings of the ACM 14th International Workshop on Data Warehousing and OLAP* (pp. 45-52).
41. Bayoudhi, L., Sassi, N., and Jaziri, W. (2021). Towards a semantic querying approach for a multi-version OWL 2 DL ontology. *International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM)*, 13, 80-90.
42. Bayoudhi, L., Sassi, N., and Jaziri, W. (2018). How to repair inconsistency in OWL 2 DL ontology versions? *Data and Knowledge Engineering*, 116, 138–158.
43. Raptis, T. P.and Passarella, A. (2023). A survey on networked data streaming with Apache Kafka. IEEE access.
44. He, Q., Zhang, F., Bian, G., Zhang, W., Li, Z.and Duan, D. (2023). Real-time network virtualization based on SDN and Docker container. *Cluster Computing*, 26(3), 2069-2083.