Data Warehouse Modernization Using Document-Oriented ETL Framework for Real Time Analytics

Monika Patel and Dhiren B. Patel

Abstract For past several years Relational Databases performing and providing services to many applications. As data is growing, the necessity for a new kind of database which can handle such a huge amount of data in semistructured and unstructured form has also increased. NoSQL databases are able to manage complexity of data structure as well as they can handle such a huge amount of data. Analysis is again a challenging process of such semi-structured and unstructured data. Traditional data warehouses are incapable of analyzing such schema-less data for decision making as relational databases need to know schema in advance. ETL (Extract-Transform-Load) is the process used to collect data from source, process and transform them and storing into a data warehouse for further analysis. The paper presents a framework for an ETL process for document-oriented data warehouse which provides real-time analytics using classification approach at the data warehouse stage. The proposed framework is designed and verified to enhance execution time for real-time analytics.

Keywords Extract-transform-load framework · Data warehouse · MongoDB · Real-time analytics · NoSQL/SQL

1 Introduction

For analysis and storing data, data warehouses are used from decades in the industries. In the scenario of Big Data DWs become more necessitate to handle semistructured and unstructured data. But traditional Data Warehouses have limitations related to structure of data. NoSQL databases can overcome the limitations of traditional databases and allows schema-free data storage and also provides horizontal scalability. As the data increases substantially, it is important to integrate the data to

M. Patel (✉) · D. B. Patel
Department of Computer Science, Gujarat Vidyapith, Ahmedabad, Gujarat, India
e-mail: monapatel18r@gmail.com

D. B. Patel
e-mail: dhiren_b_patel@gujaratvidyapith.org

V. S. Rathore et al. (eds.), *Rising Threats in Expert Applications and Solutions*,
Lecture Notes in Networks and Systems 434,
https://doi.org/10.1007/978-981-19-1122-4_5

extract the vision for business expansion and policy formation for most businesses [1, 2]. Extract-Transform-Load process is used to handle structured, semi-structured and unstructured data for Data Warehouse [3].

1.1 Relational Data Warehouse

Businesses require to store their data into data warehouses for analytics to take crucial decisions. Typically, data warehouses are designed and managed using relational databases for knowledge discovery using relational queries. The extensive growth of unstructured data needs expansion of a data warehouses that can manage data in an unstructured way, with each record having different attributes. Non-Relational technologies have ability to accumulate and process such schema free data and provides scope to design new age data warehouse [4].

1.2 ETL

For creation of a Data Warehouse, ETL (Extract-Transform-Load) is the primary process for collecting and organizing data for business decisions. Extracting, Transforming and Loading is cumbersome process in the designing of any data warehouse. An Extraction is the process of collecting data from sources and temporary lodging into the stage area for further transformation. Transformation includes cleaning, splitting, joining, and aggregating data. Lodging process loads data into the designated Data Warehouse [5, 6].

1.3 NoSQL Based Data Warehousing

NoSQL databases are new age databases and has ability to manage huge amount of data in unstructured form. Researchers have identified and pinpointed many advantages of NoSQL databases over relational databases and emphasized the scope of data warehouse development using non-relational databases. Some of the foremost benefit of NoSQL databases are schema-free data structure and scalability [7].

1.4 Real-Time Analytics

Real-time analytics is the need for current businesses to make quick decisions to cop-up the competition. Many researchers are working with the real-time analytics using NoSQL technologies [8–10]. Papers deals with real-time stream processing using

stream join algorithms to improve latency. In another paper, authors have designed a framework for handling structured and unstructured data using NoSQL technologies for data analytics.

The proposed paper provides the modern framework of handling structured and unstructured data using modern technologies (NoSQL) for data analytics [4, 11]. Prototyping of the proposed framework with execution time and memory usage will be analysed to check the performance.

The paper organizes into sections, where Sect. 2 represents related work done in the proposed research area. Section 3 shows the Proposed Framework, Sect. 4 represents methodologies and process will be used to execute the proposed framework. Section 5 covers Results and Discussion of the proposed framework and Sect. 6 Concludes the paper.

2 Related Work

The authors have analyzed 15 most prevalent ETL tools and carried out main features an ETL tool can provide. The paper also suggests the appropriate ETL tool for a specific type of organization. Further, it has highlights the limitations of most ETL tools such as horizontal scalability, real-time processing and automation [1]. The researchers analyzed the performance of incremental lodging in ETL using codebased ETL tools. Incremental loading is done by snapshot-based Change Data Capture, which will record the changes in the data while loading them into the data warehouse, then dimension and fact processing will be performed. Furthermore, authors have developed a code-based ETL framework to supports RDBMS and NoSQL database with rich transformation techniques [12].

In this paper, a novel data warehouses approach is presented using document-oriented databases. The approach consists of five steps such as choosing NoSQL database, extracting schema, identifying structure graph, multidimensional schema and designing the data warehouse schema [13]. In this research, authors have provided a data warehouse solution for NoSQL and SQL data sources for improved stream processing using MongoDB. The analysis between structured and unstructured semi-stream join processing using MongoDB has been made while implementing Extract-Transform-load phase. CPU and memory usage is analyzed while executing using synthetic and real datasets. Finally, the authors have guided for choosing appropriate memory size to reduce disk I/O cost for effective implementation of real-time data warehouse [14]. This paper includes has reviewed 74 articles for real-time stream processing systems in systematic way, which provide guidance to researchers for implementation of a robust real-time stream processing framework for real-time Data Warehouse. This study pinpoints a problem of less research done for ETL for unstructured data [8].

The authors have designed an ETL based platform for transformation of multi-dimensional concept-based model using document-oriented database for NoSQL

warehousing. The Business Process Modeling Notation is used for modeling transformation rules and then the model is evaluated for read and write request latency using TPC-DS benchmark. Based on the results researchers have come to the conclusion that NoSQL data warehouse manage data volume and queries without affecting performance [15]. This research proposes BigDimETL with multidimensional structure for the ETL process to achieve knowledge discovery. To develop a data warehouse MapReduce and Hbase are used to store data into facts and dimensions. Three queries are used to analyze the execution time, queries can be run using different factors such as table size, joins, aggregations and concurrent queries [16]. The researchers have highlighted the need of tools and techniques for unstructured data, as available tools are able to manage structured data only. The paper designed the method for handling unstructured data using NoSQL platform like MongoDB. Researchers have focused on features of MongoDB, which can analyze data of any structure without using classy data warehouse workloads [17].

3 Proposed Work

Prevailing ETL frameworks are usually transform data into structure form before storing them into Data Warehouse for analysis. The proposed framework has been designed based on literature review done from similar papers of ETL frameworks [1]. Transforming semi-structured and unstructured data into structured data to create a data warehouse is inefficient and needs the schema to be known in beforehand.

The present Semi-structured and unstructured data gathered from various applications makes it hard to transform data into the structure form for analysis [18].

The proposed ETL framework is designed using a Non-Relational DocumentOriented Database for modern Data Warehouse. Figure 1 represents proposed ETL framework. Very less work done in the field of ETL process to store and analyze documents (JSON) in data warehouse.

Every business in the market is facing blast of data because of wide use of internet and online platforms. It is very important to store and analyze products, sales and customer data to understand the pattern and to increase growth of businesses. The data generated from online platforms have no structure or have document like structure with no schema defined. NoSQL databases are becoming popular to manage such schema-free data with and also provides horizontal scalability. To encounter the organizational requests of handling semi-structured data for analysis the proposed ETL framework is designed. The proposed framework challenges typical ETL processes for storage, retrieval, scalability and real-time analytics [19, 20]. The proposed framework receives input from source (JSON documents) into MongoDB for transformation. MongoDB's aggregation framework creates pipeline for data aggregation. Transformed documents will be stored using Amazon Web Services as a Data Warehouse [21]. To enhance real-time analytics, data pre-processing will be performed on transformed data for on-demand analytics using data classification approach.

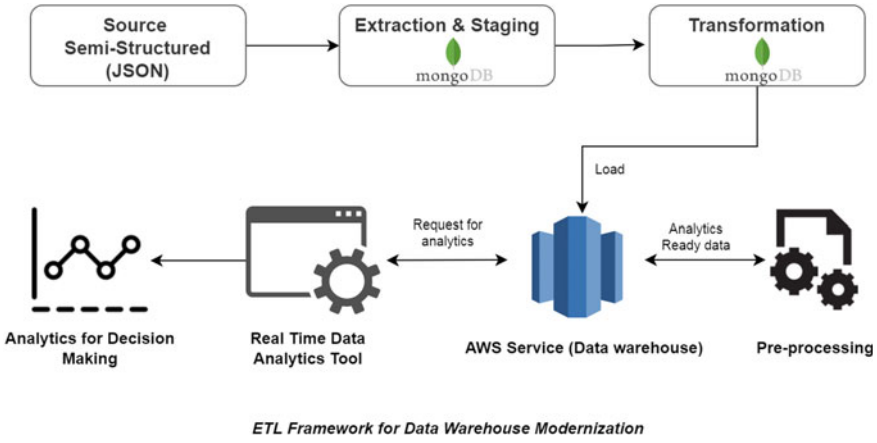

Fig. 1 Proposed ETL Framework

Selected classification method will pre-process the data into analytics-ready data for quicker results [11]. Apache Spark data stream processing engine can be used to speed up the process and to identify the essential components while analytics [22, 23]. Pre-processed data will reduce the analysis time and provides real-time analytics. An ETL processing tool HCI integrates health care databases into knowledgebase using AWS cloud services [24].

4 Methodology

Figure 2 shows the method of implementation stated in ETL framework.

Step 1 Extraction of JSON documents from source system. GSTR1 and GSTR 2A files of various clients are extracted from Goods and Service Tax to analyze the sales and purchase over the time.

Step 2 Staging into MongoDB for further process. In MongoDB JSON documents will be stored as collections.

Step 3 Data Cleaning, Filtering, attribute selection. This step selects the required attributes for further analysis.

Step 4 Transformation of documents using MongoDB aggregation Framework. Using Aggregation pipeline data are transformed and grouped using "$unwind", "$group" command for data aggregation.

Step 5 Loading into Data Warehouse using Amazon Web Services.

Step 6 To Speed up the analysis, Pre-processing is performed using classification approach to classify the data into categories and understand the relationship between variables.

Step 7 Processed data will be stored into Data Warehouse to fasten the analysis.

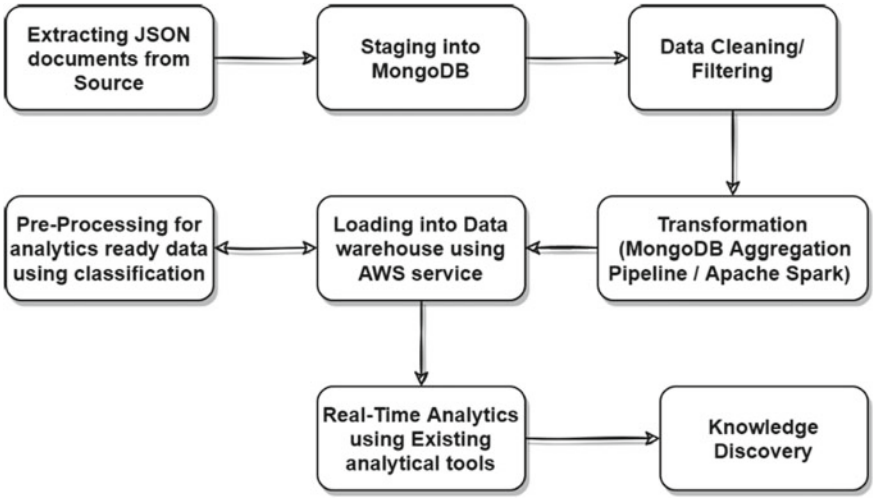

Fig. 2 Flow of proposed ETL Framework

Step 8 Testing Execution time using Existing Real-Time analytics tools.
Step 9 Knowledge Discovery and decision-making using analytics.

5 Results and Discussion

As can be seen, the approaches and technologies used to manage unstructured data is limited and they only provide solution till one or more extent of the ETL and Data Warehouse process (Table 1).

The studied approaches can provide transformation of unstructured data into structured and vice versa to design a NoSQL based Data Warehouse for Real-Time Analytics. Analyzing unstructured data from the Data Warehouse can be more complicated, thus it requires analytical techniques to speed up the process. The proposed work covers major aspects of designing, implementing and providing real-time analysis using classification approach at the data warehouse for unstructured data.

Figure 3 shows the technologies used for designing NoSQL based data warehousing using structured/unstructured data and points out the performance at the specific phase of implementation. Which leads to further research and development in the area using other technologies. Here majorly used technologies are evaluated based on capacity and performance for structured data handling, unstructured data handling, real-time analytics, execution time and response time. Proposed work overcomes the limitations in the work done till the date with new trend of technologies with added data mining approach at the data warehouse level for real time analytics.

Table 1 Technological trend in the development of NoSQL based Data Warehouse

References number	Input data	Technology	Scope of the research
[12]	Database tables, CSV files	Postgres SQL Anaconda Python	Modeling of a near real time process
[13]	DBLP, LINKEDIN	Hadoop MapReduce	Design a data warehouse Scheme from unstructured data
[14]	100YWeathe rSmall	MongoDB Atlas	Semi-stream join processing at ETL phase
[15]	Multi-dimensional schema	Talend for big data, Java routines	Transformation rules to transform multi-dimensional schema into document-oriented system
[16]	Twitter	Hbase, Oracle SQL, Hadoop hortonworks	Adapting ETL process with big data for decision making
[17]	CSV, MP3 file	MongoDB, SQL	Query processing comparison in SQL and MongoDB
Proposed work	GST—JSON files	MongoDB, Amazon red shift	ETL framework for document oriented data warehouse with real-time analytics using classification approach

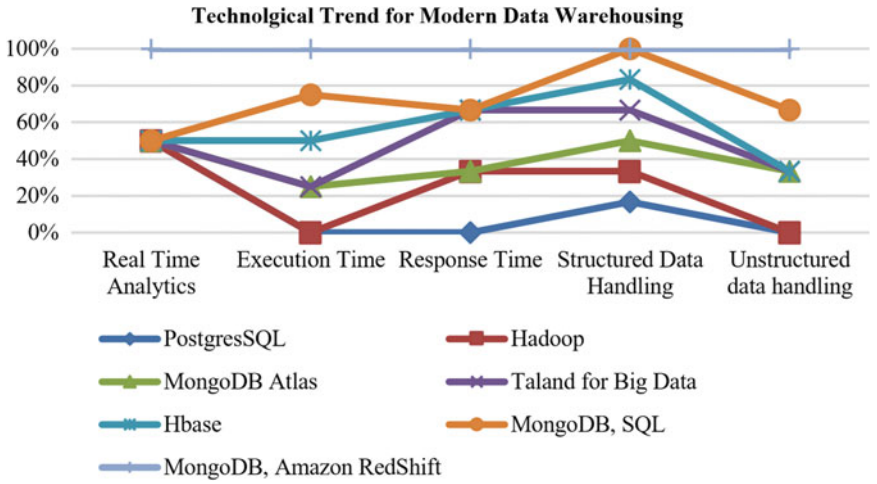

Fig. 3 Technological Trend for Modern Data Warehousing

Appling Data mining on the data warehouse compliments each other by improving real-time analytics.

6 Conclusion

This paper proposed an ETL framework for data warehousing using NoSQL databases. The framework will improve execution time by pre-processing data priorly for better performance in the era of big data. This newly designed framework using NoSQL technologies seems to offer promising results for handling huge amount of semi-structure data and real-time analytics. The study of latest technologies has been done to identify the superlative tools which can improve performance in every aspect of developing and executing document-oriented data warehousing. The researcher has analyzed and determined that the working on semi-structured data directly will save time of transformation into structured form. As future work, we will implement the proposed ETL framework with defined methodologies which will establish transformation at the data warehouse stage using appropriate classification approach before analytics.

References

1. M. Patel, D.B. Patel, *Progressive Growth of ETL Tools: A Literature Review of Past to Equip Future*, vol 1187 (Springer Singapore, 2021)
2. F. Halawa, H. Dauod, I.G. Lee, Y. Li, S.W. Yoon, S.H. Chung, Introduction of a real time location system to enhance the warehouse safety and operational efficiency. Int. J. Prod. Econ. **224** (2020)
3. M.Y. Santos, C. Costa, C. Andrade, *Enhancing Big Data Warehousing for Efficient, Integrated and Advanced Analytics*: Visionary Paper The final authenticated version is available online at https://doi.org/10.1007/978-3-030-21297-1_19 (June, 2019)
4. A. Sabtu et al., The challenges of extract, transform and loading (ETL) system implementation for near real-time environment. Int. Conf. Res. Innov. Inf. Syst. ICRIIS, 3–7 (2017)
5. S. Gorhe, *ETL in Near-Real Time Environment: Challenges and Opportunities*
6. R. Bagave, *Enhancing Extraction in ETL flow by modifying as P-ECTL based on Spark Model Msc in Cloud Computing Rahul Bagave Supervisor*
7. S. Muhammad, F. Ali, *Next-generation ETL Framework to address the challenges posed by Big Data*
8. E. Mehmood, T. Anees, Challenges and solutions for processing real-time big data stream: a systematic literature review. IEEE Access **8**, 119,123–119,143 (2020)
9. X. Li, Y. Mao, Real-time data ETL framework for big real-time data analysis, in *2015 IEEE International Conference on Information and Automation ICIA 2015—Conjunction with 2015 IEEE International Conference on Automation and Logistics*, Aug (2015), pp. 1289–1294
10. R.J. Santos, J. Bernardino, M. Vieira, 24/7 real-time data warehousing: A tool for continuous actionable knowledge, in *2011 IEEE 35th Annual Computer Software and Applications Conference* (July) (2011), pp. 279–288
11. E. Zdravevski, P. Lameski, A. DImitrievski, M. Grzegorowski, C. Apanowicz, Cluster-size optimization within a cloud-based ETL framework for Big Data, in 2019 IEEE International Conference on Big Data (Big Data) (2019), pp. 3754–3763

12. N. Biswas, A. Sarkar, K.C. Mondal, Efficient incremental loading in ETL processing for real-time data integration. Innov. Syst. Softw. Eng. **16**(1), 53–61 (2020)
13. S. Bouaziz, A. Nabli, F. Gargouri, Design a data warehouse schema from document-oriented database. Procedia Comput. Sci. **159**, 221–230 (2019)
14. E. Mehmood, T. Anees, Performance analysis of not only SQL semi-stream join using MongoDB for real-time data warehousing. IEEE Access **7**, 134215–134225 (2019)
15. R. Yangui, A. Nabli, F. Gargouri, *ETL Based Framework for NoSQL Warehousing*, pp. 0–14, Aug (2017)
16. H. Mallek, F. Ghozzi, O. Teste, F. Gargouri, BigDimETL with NoSQL Database. Procedia Comput. Sci. **126**, 798–807 (2018)
17. R. Aluvalu, M.A. Jabbar, Handling data analytics on unstructured data using MongoDB. IET Conf. Publ. **2018** (CP747) (2018)
18. D. Prakash, *NOSOLAP: Moving from Data Warehouse Requirements to NoSQL Databases*, no. Enase (2019), pp. 452–458
19. R. Indrakumari, T. Poongodi, P. Suresh, B. Balamurugan, The growing role of integrated and insightful big and real-time data analytics platforms, 1st ed., vol. 117, no. 1. Elsevier Inc. (2020)
20. A. Wibowo, Problems and available solutions on the stage of extract, transform, and loading in near real-time data warehousing (a literature study), in *2015 International Seminar on Intelligent Technology and Its Applications (ISITIA) 2015* (2015), pp. 345–349
21. N. Fikri, M. Rida, N. Abghour, K. Moussaid, A. El Omri, An adaptive and realtime based architecture for financial data integration. J. Big Data **6** (1) (2019)
22. F.K. Putri, J. Kwon, A distributed system for finding high profit areas over big taxi trip data with MognoDB and Spark, in 2017 IEEE 6th International Congress on Big Data (BigData Congress) 2017 (2017), pp. 533–536
23. A. Priya, K.R. #1, S. Rani, *A Survey Paper on the Comparison of NOSQL Engines (Mongo dB vs. Cassandra) using Spark*, vol. 6, no. 5 (2019), pp. 233–236
24. S.A. Trivedi, M. Patel, S. Patel, Health care cube integrator for health care databases. Web Semant. Acad. Press. (2020), pp. 129–151

Sources

ISSN | Enter ISSN or ISSNs | Find sources

ISSN: 2367-3370 ×

Improved Citescore

We have updated the CiteScore methodology to ensure a more robust, stable and comprehensive metric which provides an indication of research impact, earlier. The updated methodology will be applied to the calculation of CiteScore, as well as retroactively for all previous CiteScore years (ie. 2018, 2017, 2016...). The previous CiteScore values have been removed and are no longer available.

View CiteScore methodology. >

Filter refine list

Apply Clear filters

Display options

- Display only Open Access journals

Counts for 4-year timeframe

- No minimum selected
- Minimum citations
- Minimum documents

Citescore highest quartile

- Show only titles in top 10 percent
- 1st quartile
- 2nd quartile
- 3rd quartile
- 4th quartile

Source type

- Journals
- Book Series
- Conference Proceedings
- Trade Publications

Apply Clear filters

1 result

Download Scopus Source List | Learn more about Scopus Source List

All | Export to Excel | Save to source list

View metrics for year: 2021

	Source title	CiteScore	Highest percentile	Citations 2018-21	Documents 2018-21	% Cited
1	Lecture Notes in Networks and Systems	0.7	12% 236/270 Control and Systems Engineering	9,307	14,098	26

Top of page

About Scopus

What is Scopus

Content coverage

Scopus blog

Scopus API

Privacy matters

Language

日本語版を表示する

查看简体中文版本

查看繁體中文版本

Просмотр версии на русском языке

Customer Service

Help

Tutorials

Contact us