A Multi-Agent Framework for Data Extraction,Transformation and Loading in Data Warehouse

Ramzan Talib*, Muhammad Kashif Hanif†, Fakeeha Fatima‡, and Shaeela Ayesha§
Department of Computer Science,
Government College University, Faisalabad, Pakistan

***Abstract*—The rapid growth in size of data sets poses challenge to extract and analyze information in timely manner for better prediction and decision making. Data warehouse is the solution for strategic decision making. Data warehouse serves as a repository to store historical and current data. Extraction, Transformation and Loading (ETL) process gather data from different sources and integrate it into data warehouse. This paper proposes a multi-agent framework that enhance the efficiency of ETL process. Agents perform specific task assigned to them. The identification of errors at different stages of ETL process become easy. This was difficult and time consuming in traditional ETL process. Multi-agent framework identify data sources, extract, integrate, transform, and load data into data warehouse. A monitoring agent remains active during this process and generate alerts if there is issue at any stage.**

***Keywords*—Data Warehouse; Extraction; Loading; Multi-Agent; Operational Data; Transformation**

I. Introduction

In this digital era, data is being generated by different sources at all times. Data can reside on different computers and servers. It is challenging task for organizations to manage and analyze huge volume of data to achieve their goals [1]. Combination of historical and current data is essential to get strategic information. Data warehouse provide the input for strategic decision making. Data warehouse takes data from different sources, process and store in a common repository [2]. Data warehouse is a subject oriented, integrated, time variant and nonvolatile collection of data in support of taking management decision [2]. Data warehouse provide accurate, efficient, and complete view of an organization's operational data to solve complex queries [3].

The most important component of the data warehouse is ETL process. ETL process extract and integrate data from diverse homogeneous and heterogeneous sources. Data sources may contain inconsistent data that can produce incorrect and misleading results. The purpose of ETL process is to extract data from data sources, transform into structured format, and load into target data warehouse [4], [5].

Figure 1 shows the generic architecture of data warehouse. This architecture consists of data source, mapping of ETL process, storage area and analysis layers. The purpose of analysis layer is to mine data for future predication [6], [7].

The data store format in data warehouse is different from operational data sources. Operational systems are essential for day to day operations of any organization. In Online Transaction Processing (OLTP) system, records are stored in flat files using different applications, tools, data formats, and data representational methods [8]. Efficient and effective data extraction methods are essential to make better and reliable strategic decision [9]. A process that converts operational data into analytical data stores in a controlled, secure and suitable format is needed [4]. The ETL process provides data cleaning consistently and reliably with high performance. Different tools can be used at this phase like talend, pentaho, oracle warehouse builder, Microsoft integration services, open text integration services, IBM cognos managers, information builder, SQL server integration tool, and SSIS packages etc. [10], [11].

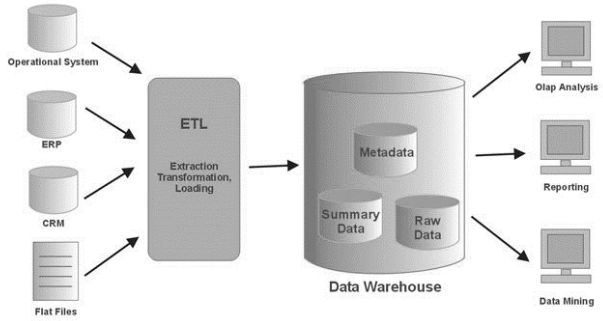

Fig. 1. Architecture of data warehouse [3]

There is need to extract more meaningful, relevant and appropriate data to take reliable, efficient and effective decisions. For this, a multi-agent based framework in ETL process is proposed. The proposed process will make the data extraction, transformation and loading process fast, efficient, and flexible. In this way, efficiency and effectiveness to manage and extraction of data is increased [12].

The remainder of this paper is organized in different sections. In section II, related work is discussed. Section III presents ETL process in data warehouse. Section IV provides a multi-agent framework in ETL process. Section V concludes the paper with future research perspectives.

II. Review of Literature

[5] presented a framework for extraction, transforming and loading data into data warehouse. They have shown extraction is most important phase in ETL process. They also discussed the issues for extracting, transforming and loading

data and their effect on the decision making process. Futher, various quality metrics and ad hoc approaches that enhance the performance of ETL process were proposed. [1] presented different modeling techniques that optimize and enhance the ETL process. Different techniques to design ETL process were discussed in field of academic. These techniques were based on open source and commercial tools. They concluded ETL process is very expensive regarding its cost, time and establishment of data warehouse.

[4] proposed a multi-agent based framework to establish a data warehouse structure for information technology infrastructure library. The use of agents in data warehouse optimizes and enhances the working. ITIL is relatively a complex and large data warehouse infrastructure which manages all IT related fields. By following standards with multi-agent technology help to manage the continuous updation, improve functionality and reduce the chances of risks. [13] presented a multi-agent system used at the data pre-processing stage in e-wedding project. The use of Multi-Agent System (MAS) at earlier stage improve responsiveness and efficiency of the system. A multi-agent system based on Java Agent Development Framework (JADE) is used to cope these issues raised at data pre-processing stage, i.e., handle missing values during data extraction process. JADE support different states of the agents as: agent communication, protocol, behavior, and ontology.

[14] presented novel methods to handle complex data consolidation through multi-agent system in data warehouse. The proposed approach based on more flexible and evaluative architecture in which one can easily add, remove and modify services according to the need. By applying multi-agent based prototype, the integration process done by using UML classes. Two agents the data agent and wrapper agent were used to model the complex data in UML classes. The XML creator agent mapped the UML classes into XML document. [15] discussed how to improve Extraction Transformation and Loading (ETL) process in data warehouse system of higher education system. They presented ETL architecture for HEIS and discussed various issues which arise in development and maintenance of the data model.

[16] discussed the use of agent technology in data warehouse. They have proposed Intelligent Data Warehouse (IDW) model for data extraction, processing and information retrieval optimization. In this model, data is integrated from different sources efficiently. Moreover, data collection process is improved which reduces the extraction time. It incorporates functions that are adaptable and flexible to access the data across the enterprises. [17] presented the workflow process for the refreshment of data. They concluded data updation process should be performed in extraction, transformation and loading phases.

III. Extraction, Transformation and Loading Process

The complexity and disparity of sources is growing with the increased usage of information systems. Data warehouse provides central repository that enable organizations to store all historical data at one place. ETL is a most important phase to extract, clean and transform data in data warehouse. Figure 2 shows data flow in ETL process. ETL process has extraction, transformation, and loading steps (Figure 3).

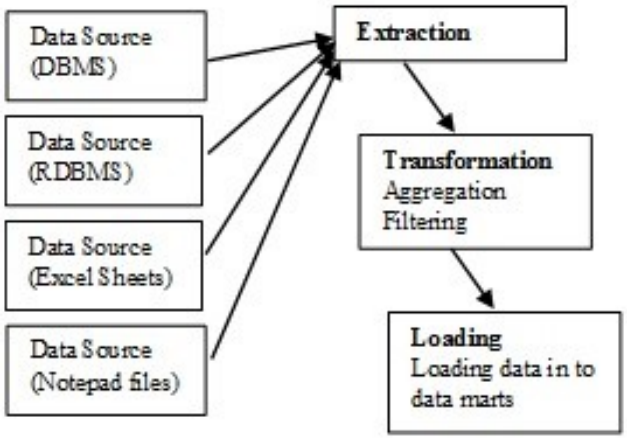

Fig. 2. ETL data flow diagram [5]

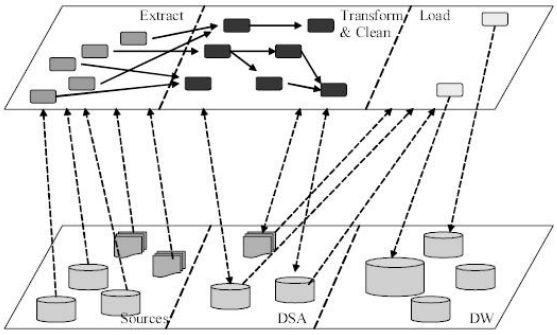

Fig. 3. ETL process [6]

A. Extraction

The first phase of ETL process is data extraction. This phase integrate data from various homogeneous and heterogeneous sources. The source systems can have different data format according to their operational needs [18]. Data validation rules are applied on the operational data according to domain. Validation rules identify whether the data extracted from the sources has the correct values [19]. Following constraints are checked at this stage [10].

- content and meta data
- data object attributes
- extraction mode and protocols to capture data
- monitor the extraction process

B. Transformation

Transformation is a most crucial phase. At the staging area, different mapping functions are performed on the extracted data to remove dirty values, duplications, inconsistencies, and naming conventions [17]. Manipulation operations like cleansing, filtering, enriching, aggregating, sorting, generating surrogate keys, and granularity level are determined to map

the external data source to data warehouse [18], [20]. Set of rules to translate coded values and to derive new values are applied to clean and transmit the data. At this phase, schema and instance level mapping are performed to standardize the data. In addition, data validation and data accuracy constrain are performed [21].

C. Loading

This is a final phase in ETL process. Extracted and transformed data are loaded into targeted data warehouse [14]. Data loading in the data warehouse has its own technical challenges. A major problem is difference between new and existing data at loading time. This step make sure data is converted into targeted data structure of data warehouse rather than source data structures. Moreover, various schema and instance joining and aggregation functions are performed at this phase [15].

IV. A Multi-Agent Framework in ETL Process

Intelligent agents are used now a days in every field of life to solve complex problem by distributing the work. Agents are a software programs that take the autonomous action in different states to attain design objectives. According to [1], responsive, proactive, independent, object oriented and social are important characteristics of agents. In multi-agent based system, agents work collectively and each agent performs specific tasks according to the role assigned [14], [15].

The addition of agents at the data extraction level minimizes the chance of error, increases efficiency and reliability (Figure 4). Moreover, the extraction, transformation and loading time is reduced [22]. Agents invoke messages when any problem arises. An alert is generated for missing value or irrelevant data [10], [16].

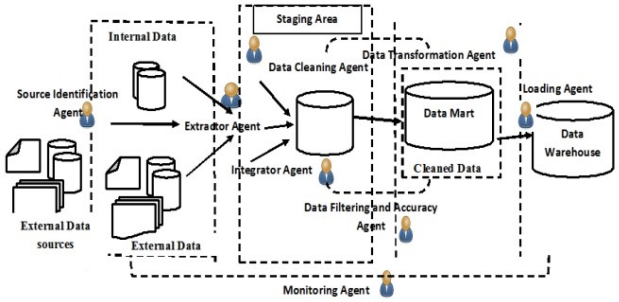

Fig. 4. The Multi-Agents framework for ETL

There is no specific standard and structures for operational data. Discrepancies arise in data formats due to changing characteristics of data [19]. Multi-agents in ETL process helps to reduce the chances of occurrence of errors. Each agent is assigned a specific role by following the standards regarding the semantic and format of data [23]. In this case study, agent based ETL process is analyzed that helps to make the extraction, transformation and loading process efficient, effective, and reliable [21], [24]. Agents are organized into three groups in multi-agent framework.

- Extractor and Integrator Multi-Agent Group
- Transformator and Loader Multi-Agent Group
- Management and Control Agent

A. Extractor and Integrator Multi-Agent Group

Agents in this group extract data from different excel files, flat files, MS access and SQL databases [8]. Agents coordinate with other agents in this group to extract complete, concise, and reliable data (Figure 5). Agents in this group are assigned following roles.

- **Source Identifier Agent (SIA)** identify the data extraction sources.
- **Extractor Agent (EA)** establishes a link with the sources system and extracts data.
- **Data Cleaning Agent (DCA)** is concerned with identifying and eliminating contradictions and inconsistencies. DCA removes duplicate, missing and irrelevant values from data. It is also responsible for the customization and integration of the information from multiple sources [25].
- **Integrator Agent (IA)** integrates and mounts the extracted data in the Data Staging Area (DSA). The extracted records are loaded into data warehouse staging area (Figure 5). At this stage, DCA removes all discrepancies of spelling error, invalid or wrong records to improve the quality and reliability of the data [26].

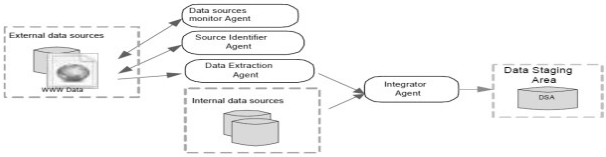

Fig. 5. The Extractor and Integrator Multi-Agent Group

B. Transformator and Loader Multi-Agent Group (TLM)

Transformator and Loader Multi-Agent Group is responsible for data validation, accuracy, consistency, schema and instance related conversion according to semantic rules [25]. TLM follows a work-flow sequence to execute all data transformation in a reliable and efficient way. This multi-agent group is consists of the following agents:

- **Data Validator Agent (DVA)** checks and matches all records of the fact and dimension tables to ensure integrity constraints.
- **Data Filtered and Accuracy Agent (DFAA)** make sure record contains appropriate values and mapped according to data warehouse structure [27].
- **Loader Agent (LA)** is responsible for loading record from logical schema into repository mapped schema. The role of LA is to ensure efficiency and consistency to improve the performance of data warehouse operations and reduce the loading time.

C. Management and Control Agent (MCA)

The purpose of MCA in ETL process is to monitor all the activities of agents. MCA ensure agents are doing work properly and according to the sequence. It also establishes a coordination among agents to enhance the functionality and performance.

V. Conclusion

A multi-agent based ETL framework provide an efficient mechanism to extract, transform and load data in data warehouse. ETL process extract data from homogeneous, heterogeneous, or distributed sources and map in the format according to targeted data warehouse. There exist different methods and tools to enhance the efficiency of ETL process. In this work, an agent based framework is proposed. In proposed framework, agents work collectively to perform tasks according to the roles assigned. The system contains EIM, TLM, and MCA groups of agents to reduce the extraction time and optimize the performance. Research can be carried to design a common model for the meta data of ETL process. Moreover, the implementation of the agent based scenario for analysis purpose in different fields of life can be done.

References

[1] A. Bologa, R. Bologa *et al.*, "Business intelligence using software agents," *Database Systems Journal*, vol. 2, no. 4, pp. 31–42, 2011.

[2] W. H. Inmon, *Building the data warehouse*. John wiley & sons, 2005.

[3] Z. El Akkaoui, E. Zimànyi, J.-N. Mazón, and J. Trujillo, "A model-driven framework for ETL process development," in *Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP*. ACM, 2011, pp. 45–52.

[4] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE transactions on knowledge and data engineering*, vol. 24, no. 9, pp. 1537–1555, 2012.

[5] P. Balaji and D. Srinivasan, "An introduction to multi-agent systems," in *Innovations in multi-agent systems and applications-1*. Springer, 2010, pp. 1–27.

[6] M. Arif and G. Mujtaba, "A survey: Data warehouse architecture," *International Journal of Hybrid Information Technology*, vol. 8, no. 5, pp. 349–356, 2015.

[7] M. Golfarelli and S. Rizzi, *Data warehouse design: Modern principles and methodologies*. McGraw-Hill, Inc., 2009.

[8] V. Gour, S. Sarangdevot, G. S. Tanwar, and A. Sharma, "Improve performance of extract, transform and load (ETL) in data warehouse," *International Journal on Computer Science and Engineering*, vol. 2, no. 3, pp. 786–789, 2010.

[9] S. H. A. El-Sappagh, A. M. A. Hendawi, and A. H. El Bastawissy, "A proposed model for data warehouse ETL processes," *Journal of King Saud University-Computer and Information Sciences*, vol. 23, no. 2, pp. 91–104, 2011.

[10] A. KABIRI and D. CHIADMI, "Survey on ETL processes." *Journal of Theoretical and Applied Information Technology*, vol. 54, no. 2, 2013.

[11] M. Mrunalini, T. S. Kumar, and K. R. Kanth, "Simulating secure data extraction in extraction transformation loading (ETL) processes," in *Computer Modeling and Simulation, 2009. EMS'09. Third UKSim European Symposium on*. IEEE, 2009, pp. 142–147.

[12] M. Singh and S. Jain, "Transformation rules for decomposing heterogeneous data into triples," *Journal of King Saud University-Computer and Information Sciences*, vol. 27, no. 2, pp. 181–192, 2015.

[13] N. Kolsi, A. Abdellatif, and K. Ghedira, "Data warehouse access using multi-agent system," *Distributed and Parallel Databases*, vol. 25, no. 1-2, pp. 29–45, 2009.

[14] A. J. Morais, E. Oliveira, and A. M. Jorge, "A multi-agent recommender system," in *Distributed Computing and Artificial Intelligence*. Springer, 2012, pp. 281–288.

[15] K. Kularbphettong, G. Clayton, and P. Meesad, "A hybrid system based on multi-agent system in the data preprocessing stage," *arXiv preprint arXiv:1003.1792*, 2010.

[16] O. Boussaïd, F. Bentayeb, and J. Darmont, "An mas-based ETL approach for complex data," *arXiv preprint arXiv:0809.2686*, 2008.

[17] M. Wooldridge, *An introduction to multiagent systems*. John Wiley & Sons, 2009.

[18] M. M. Al-Debei, "Data warehouse as a backbone for business intelligence: Issues and challenges," *European Journal of Economics, Finance and Administrative Sciences*, vol. 33, 2011.

[19] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma, "Improving data quality: Consistency and accuracy," in *Proceedings of the 33rd international conference on Very large data bases*. VLDB Endowment, 2007, pp. 315–326.

[20] F. Fatima, M. Javed, F. Amjad, and U. G. Khan, "An approach to enhance quality of the rad model using agents," *The International Journal of Science and Technology*, vol. 5, pp. 2002–2010, 2014.

[21] I. Mekterović, L. Brkić, and M. Baranović, "Improving the ETL process and maintenance of higher education information system data warehouse," *WSEAS transactions on computers*, vol. 8, no. 10, pp. 1681–1690, 2009.

[22] L. Muñoz, J.-N. Mazón, J. Pardillo, and J. Trujillo, "Modelling ETL processes of data warehouses with uml activity diagrams," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2008, pp. 44–53.

[23] R. Gill and J. Singh, "Enactment of medium and small scale enterprise ETL (masseetl)-an open source tool," *International Journal of Computer Science and Information Technologies*, vol. 6, no. 1, pp. 141–147, 2015.

[24] C. A. Moturi and A. Emurugat, "Prototyping an academic data warehouse: Case for a public university in kenya," *British Journal of Applied Science & Technology*, vol. 8, no. 6, 2015.

[25] A. Abello, O. Romero, T. B. Pedersen, R. Berlanga, V. Nebot, M. J. Aramburu, and A. Simitsis, "Using semantic web technologies for exploratory OLAP: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 571–588, 2015.

[26] K. Sivaganesh, P. Srinivasu, and S. C. Satapathy, "Optimization of ETL work flow in data warehouse," *International Journal on Computer Science and Engineering*, vol. 4, no. 9, p. 1579, 2012.

[27] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, *Fundamentals of data warehouses*. Springer Science & Business Media, 2013.