International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 01, ISSN No. 2455-2143, Pages 232-236
Published Online May 2023 in IJEAST (http://www.ijeast.com)

EXTRACTION, TRANSFORMATION AND LOADING PROCESS IN THE CLOUD COMPUTING SCENARIO

Dr. Maninti Venkateswarlu, Dr. T. G. Vasista
Ashoka Women's Engineering College, Kurnool

Abstract---**Business Intelligence strategies often need real-time BI and analytics. For this purpose, BI systems focus on data management as a whole and good data and information quality in particular to achieve data consistency and referential integrity across distributed systems. Generating key performance indicators (KPI) of business intelligence might require incorporating the semantic approach of data integration to achieve the consistent and sustainable semantic decisions. So this paper deals with how to improve data and information quality by adopting Extraction-Transformation-Loading (ETL) and Extraction-Loading-Transformation processes towards generating structured data and semantic data with quality towards improving inference capabilities and generating semantic knowledge base for the business domain ontology. However this research did not consider dealing with dynamic or real-time data quality and their role towards producing better business intelligence.**

Keywords---**Business Intelligence, Cloud Data Warehouse, Data Quality, Decision Making, ELT Process, ETL Process, Information Quality**

I. INTRODUCTION

Cloud computing is an efficient way of providing computing resources in the form of outsourcing the data storage, processing and also includes the security aspect (Escobedo-Bailon et al., 2021). Cloud computing offers a service-based pay-as-you-go approach (Ambrust et al., 2009) thus Cloud computing resources involve Infrastructure as a service, platform as a service and software as a service (Escobedo-Bailon et al., 2021). Cloud computing refers to on-demand and integrated computer resources that are made available to those who subscribe for them to use it. These resources include data storage capacity, backup and self-synchronization (Nir, 2010; Zanoon, Al-Haj & Khwaldeh, 2017). Data storage resources of cloud appear either in the form of a cloud storage service or as a version of data centre (Vasconcellos, 2022). Cloud enterprise data warehousing is a top level strategic business and information technology investment initiative to drive the profit and to make it more customer centred (Matthew, 2022). A cloud data warehouse usually appears as software as a service instead of appearing in the form of a physical data warehousing storage. A cloud data warehouse is a managed service in a public cloud. It is optimized for conducting business intelligence and big data analytics based operations (Qlik). The quality of Business Intelligence management is directly proportional to decision making quality of corporate people. Hence, data quality and information quality become critical success factors to achieve comprehensive BI solutions (Wieder & Ossimitz, 2015).

II. RESEARCH OBJECTIVES

Therefore two research objectives are discussed in this paper:

(i) Managing Data Quality in ETL process
(ii) Managing Information Quality in ELT process

III. BUSINESS INTELLIGENCE IN CLOUD COMPUTING

Business intelligence is the process of collecting information from data, so that organizations can be guided towards making business decisions (WinMan) Business Intelligence is a new phenomenon of handling data in business. Cloud computing offers more accessibility options to choose different BI tools (Kasem & Hassanein, 2014) to achieve more profitability (Yiu, Yeung & Cheng, 2020) and competitive advantage (Muntean, 2008). Gartner defined business intelligence as "the use of applications, infrastructure, tools and best practices that enables access to and analysis of information to improve and optimize decision and performance" (Gartner; Kasem & Hassanein, 2014). Cloud Business intelligence is a contemporary revolutionary concept of delivering business intelligence capabilities as a service using cloud architecture in low cost but with faster deployment and flexibility (Kasem & Hassanein, 2014). Business Intelligence relies on a large database called data warehouse. A data warehouse is a collection or large data sets extracted, transformed and loaded from different forms of database (Menon, Rehani & Gund, 2012). A traditional data warehouse usually built as an integrated data set of relational databases. The data in the data warehouse is stored by several operational database systems before loading to data warehouse. The data

warehouse can be viewed as a collection of different data marts. The data warehouse maintains the history of data and integrates to produce a unique data model. The data warehouse can handle queries. ETL is applied on data warehouse and the results are storied in multidimensional arrays and then online analytical processing (OLAP) tools are used for generating the business reports (Nedunchezhain, Moorty & Thurnavukkarasu, 2012).The ETL phase is an essential step in the data warehousing process. It is considered as the most expensive phase in implementing data making systems. Integration of data relies heavily onvolume of data and velocity of data. Use of parallel computing technique enhances the performance of ETL process. While some are suggesting classical ETL process for obtaining contemporary business decision making capabilities (Wang, Hu & Zhou, 2011), some other suggested big data computing environment (Elgendy & Elragal, 2016) in passing (Diouf, Boly & Ndiaye, 2018).

IV. BIG DATA ANALYTICS IN CLOUD COMPUTING

Big data and cloud computing are two main stream technologies of contemporary interest in the Information Technology field. Big data is a concept that deals with storing, processing and analyzing large amounts of data. Big data analytics is believed to reduce the cost of analytical treatment and hence becoming a crucial process in many fields and sectors. In Big data analytics, the ETL: Extraction-Transformation-Loading process takes a paradigm shift into ELT: Extraction-Loading-Transformation (Berisha & Meziu, 2021). For example in Hadoop, ELT process takes place (Kumar, 2020). As data protection, privacy and cyber security are becoming some of the cloud big data analytics challenges. Lack of knowledge and trained profession to work cloud big data is becoming a contemporary big challenge that Information Technology market is facing now (Mohan & Manohar, 2021).

V. MANAGING DATA QUALITY IN CLOUD COMPUTING

According to J. M. Juran, the engineering and management consultant, data is said to be high quality if data can be fit for the intended use of operations, planning and decision-making aspects of business (Chisholm, 2017). The accuracy and relevance of Business Intelligence & Analytics (BI&A) rely on the ability to bring high data quality to the data warehouse from both internal and external sources using ETL process (Souibgui et al., 2019). The nature of data in the data warehouse is different from that of operational data base. While operational database focuses more to contain transactional data supported by having lookup data among normalized data sets, data warehouse data is subject oriented, integrated, time variant, non-volatile, de-normalized and summarized on which OLAP can be performed (El-Sappagh, Hendawi & El Bastawissy, 2011).

Ensuring data quality requires tracking quality defects in the ETL process. Data quality is the reflection of multi dimensional facets and influences (Souibgui et al., 2019). There are three data quality perspectives: (i) data oriented and (ii) process oriented (iii) control oriented. While data oriented approach tackles quality defects at data level, process oriented approach tackles them considering ETL as a process. Control oriented approach focuses on administrative operational aspects such as accounts and permissions. The data level quality relies on features such as amount of data, relevance, completeness and timeliness (Wang & Strong, 1996). The process level quality relies on features such as referential integrity and business semantics. The control level quality relies on accessibility and security features (Souibgui et al., 2019).

A. The ETL Process

Classical ETL process is done in the following three steps (El-Sappagh, Hendawi & El Bastawissy, 2011):

1) Extraction**:** During ETL process, data is extracted from an On-Line Transaction Processing (OLTP) database. The data sources are not only belonging to relational databases but also text files, spreadsheets (e.g. csv files) etc. (Ambika, 2020). Therefore use of ODBC, JDBC drivers become common among ETL team.

2) Transformation**:** The data is transformed to match the data warehouse schema such as mostly star schema but may sometimes have to follow snow flake scheme. Most of the contemporary enterprise data warehouses are also planning to maintain galaxy schema while most of the data marts focus to adopt star schema. The transformation step tends to make some data cleansing too so that correctness, completeness, consistent and unambiguous data can be seen in the data warehouse. A DW schema is composed of fact tables and dimension tables. All the transformations are available in data warehouse as a part of meta-data. A well designed ETL system is expected for ease of modification and generate decision support reports with dashboards of charts and graphs that enable decision making.

Data cleansing**:** Data cleansing is required when data is extracted from source data systems, while loading into staging area or transforming into target data ware area (Ul Haq, 2016). Data cleansing is the process of finding errors in data and correcting them either using manually or automatically. A large part of the cleansing process involves the identification and removing duplicate records (Ridzuan & Zainon, 2019). Other activities include updating missing fields and missing values (Loshin, 2013). Data cleansing is also called data scrubbing. It not only removes errors but also removes inconsistencies from data to improve the data quality. Other data cleansing activities include: Updating nulls, correcting misspellings and illegal values, maintaining consistent data formats, wrong references. Lookups are one way to avoid wrong references. For data warehousing cleaned data is often available from

International Journal of Engineering Applied Sciences and Technology, 2023
Vol. 8, Issue 01, ISSN No. 2455-2143, Pages 232-236
Published Online May 2023 in IJEAST (http://www.ijeast.com)

data staging area. Examples of some of the ETL tools include: CopyManager, DataStage, Extract, PowerMart, DecisionBase, DataTransformationService, MetaSuit, WarehouseAdministrator etc. (Rahm & Do, 2000). For example Informatica cloud data quality simplifies cloud data warehouse management with a single, easy-to-use, self-service data quality tools that operate entirely in the cloud as an economical subscription service and let users focus on building data quality mapping and data profiling based on data quality rules for standardization and cleansing (Informatica, 2019).

3) Loading**:** Loading to target Data warehousing fact and dimension tables and OLAP cube data sets can be achieved through not only using native database drivers but also the using generic ODBC, JDBC drivers.

VI. MANAGING INFORMATION QUALITY IN CLOUD COMPUTING

Turban et al. (2005) view of Information is: Meaningful organization of data the recipient or surprise value of unknown to recipient. The following are the most common information quality dimensions (Al-Hakim, 2007) are: Accessibility, Accuracy, Amount of Information, Believability, Coherence, Capability, compactness, conciseness of representation, consistency of representation, ease of implementation and ease of understanding. Gustavsson & Wanstrom (2009) define information quality as the "ability to satisfy stated and implied needs of the information consumer (p.237) in passing to Alshikhi & Abdullah (2018). The quality of decision improves by improving the information quality towards having knowledge about relationships among problem variables (Raghunathan, 1999).

A. The recent analytics based ELT process

Anwar, Huntz, Koloch & Pitti (2010) argued that there are some problems that arise in integrating data with Data Warehouse approach such as building and maintaining large scale integrated data warehouse and is also difficult and expensive; once it is built it becomes inflexible to make changes/update. If the data is not conducted properly subjected to right ETL process, On-Line Analytical Processing queries cannot be executed conveniently, therefore DWH (Data WareHouse) design becomes more burdened on the warehouse schema and ETL process (Al-Sudairi & Vasista, 2011).

While Levenbach (2013) suggested PEER process as a modeling pathway for demand forecasters/planners, where notations are: P-Preparation of Healthy data, E-Executing Analytic Modeling methodologies, E-Evaluating performance diagnostics and R-Reconciliation of modeling pathways; Vasista (2007) suggested TAMPA methodology where the notations are: T-Transform, A-Analyze, M-Measure, P-Predict and A-Act as an extension to MPA methodology mentioned in Berson, Smith & Thearling (2000) for eCRM solution. TAMPA is a closed loop methodology (Vasista & AlAbdullatif, 2017).

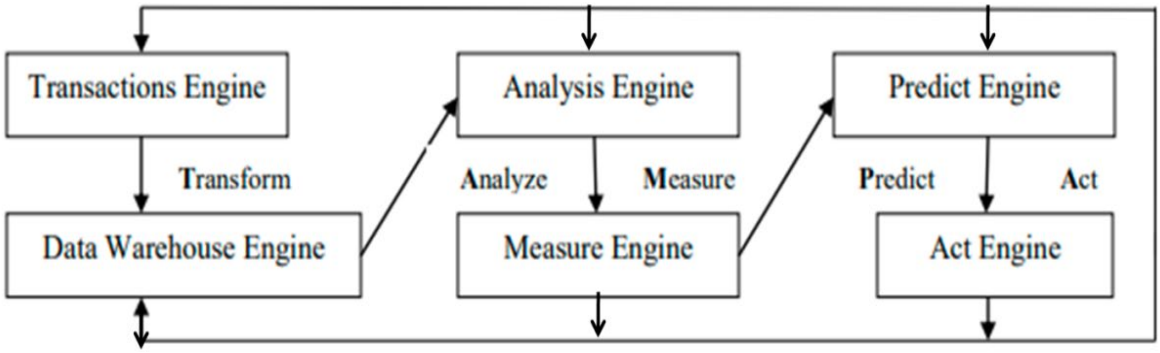

Fig. 1: TAMPA Methodology (Source: Vasista, 2007; Vasista & AlAbdullatif, 2017)

Decision making depends more on the availability of semantic information. Therefore building ontology approach enables true semantic integration across the information sources to draw wider conclusions. Ontology provides efficient information integration. Ontology allows construction of semantic information base. Semantic Information Base (SIB) enables the following features: storing, loading and Data Manipulation Language (DML) access to RDF/OWL data and ontologies; inference using Web Ontology Language (OWL) and Resource Description Framework (RDF) schema semantics and user defined rules; ontology assisted querying of enterprise relational data. RDF meta-data is required to be accessed to retrieve semantic statements and for which what is called triplestore is built. A triplestore is optimized to store and retrieve short statements called triples that appear in the form of subject-predicate-object; hence the name tripelstore is given. It means the relational data is published on the web as RDF and data is linked through SPARQL end point (AlSudairi & Vasista, 2011). To this end, handling and managing the

structured data is done. Then what about dealing with unstructured data? Annotation helps prepare information sets that can be used to train machine learning models for a variety of applications. S (Babovic & Milutinovic, 2013).
The research proposed by Rani et al. (2017) suggested what is called MOUNT framework. It proposed the multi-level semantic annotation and unified data integration using semantic web ontology in big data process. According to which, the heterogeneous collection of big data (which is a combination of structured and unstructured data) will be subjected to two types of annotation: (i) Coarse-grained annotation – in which ontology is created focusing on metadata and (ii) fine-grained annotation – in which a schema of structured information is extracted to be prepared to match the Resource Description Framework format and a schema of unstructured information is extracted to be prepared to match based on lexical and semantic annotation. Then the query processing takes place on the integrated part of both structured and unstructured information. For example in the Smart cities context, data can be collected directly from variety of sensors, smart phones, structured data from RDMBS and unstructured data from text files and other sources are integrated and linked with smart city data repositories to perform analytical check and reasoning so to generate required information as well as new knowledge or inference for decision-making for having better urban governance (Khan, Anjum, Soomro & Tahir, 2015).

VII. CONCLUSION

Data warehouses and data marts are hence called Business Intelligence (BI) enablers. Data integrity is a pre-requisite for data consistency. The main purpose of business intelligence systems in an organization is to increase the quality of data and information, so thatquality of decision making will be improved. For this purpose, Business Intelligence refers to interoperable data infrastructure and standardization of data-related technology, creation of metadata standards for Big Data management and Analytics. Semantics play an important role in harnessing information in business domain in the form of building knowledge graphs along with scalable and flexible data discovery, analysis and dash board based reporting capabilities. But one of the limitations of this research is to see how to consider continuous flow of real-time data and its quality toward making better inference and decision making.

VIII. REFERENCES

[1] Al-Sudairy M. A. T. & Vasista T. G. K. “Semantic Data Integration Approaches for E-Governance”, International Journal of Web & Semantic Technology (IJWEST) 2 (1), 1-12. 2011

[2] Alshikhi O. A. & Abdullah B. M. “Information Quality: Definitions, Measurement, Dimensions and Relationship with Decision Making”. European Journal of Business and Innovation Research, 6 (5), 36-42, 2018

[3] Ambika, P. Chapter Thirteen – Machine learning and deep learning algorithms on the Industrial Internet of Things (IIoT), In Advances in Computers, 117 (1), 321-338, 2020

[4] Ambrust M. et al. Above the clouds: A Berkeley view of cloud computing. Electrical Engineering and Computer Sciences, University of California at Berkeley, 2009

[5] Anwar N., Huntz E., Kolch W., & Pitti A. Semantic Data Integration for Francisella tularensis novicida Proteomic and Genomic Data”. Online Was [Available at] www.cis.strath.ac.uk/~ela/AnwarSWAT4LS_5.pdf retried on Nov 2, 2010.

[6] Berisha B. & Meziu E. Big data Analytics in Cloud computing: An overview, 2021 DOI: 10.13140/RG.2.2.26606.95048.

[7] Bibovic Z. & Milutinovic V. Chapter 2 – Novel system architecture for semantic-based integration of sensor networks. In Advances in Computers, 90, 91-183, 2013.

[8] Chisholm M. Fundamental Concepts of Data Quality, 2017Online [Available at] https://www.firstsanfranciscopartners.com/blog/fundamental-concepts-data-quality/ Retrieved on 10-05-2022.

[9] Diof P. S., Boly A. & Ndiaye S. “Variety of data in the ETL processes in the cloud: State of the art”. In the proceedings of 2018 IEEE International conference on Innovative Research and Development, May 11-12, Bangkok, Thailand, p. 6, 2018

[10] El-Sappagh S. H. A., Hendawi, A. M. A. & El Bastawissy A. H. “A Proposed model for data warehouse ETL processes”. Journal of King Saud University – Computer and Information Systems, 23 (2), 91-104, 2011

[11] Elgendy, N. &Elragal, A. “Big Data Analytics in support of the Decision making process”, Procedia Computer Science, 100, 1071-1084, 2016

[12] Escobedo-Bailon et al. “Cloud Technology as a support for the ETL Process and its influence on decision making”, International Journal of Aquatic Science, 12 (02), 4637-4646, 2021

[13] Gartner, Business Intelligence, [Available at] https://www.gartner.com/en/information-technology/glossary/business-intelligence-bi Rretrieved on 10-05-2022

[14] Informatica. Data Quality in the cloud data warehouse, Informatica Solution Brief, IN17-1119_03762, 2019,Online [Available at] https://www.informatica.com/content/dam/informatica-com/en/collateral/solution-brief/data-quality-in-the-

cloud-data-warehouse_solution-brief_3762en.pdf Retrieved on 10-05-2022.

[15] Kasem M. &Hassanein E.“Cloud Business Intelligence Survey”. International Journal of Computer Applications, 90 (1), 2014

[16] Khan Z., Anjum A., Soomro K., Tahir M. A. “Towards cloud based big data analytics for smart future cities”. Journal of Cloud computing 4 (2), 2015

[17] Kumar, R. Best Approach for your Data Warehouse.2020, Online [Available at] https://www.softwareadvice.com/resources/etl-vs-elt-for-your-data-warehouse/ Retrieved on 15-05-2022.

[18] Levenback H. Prdictive analytics for demand forecasting and planning managers – A Big data challenges. KAIST college of Business, Seoul, Korea, 2013 Online [Available at] https://forecasters.org/wp-content/uploads/gravity_forms/7-2a51b93047891f1ec3608bdbd77ca58d/2013/07/Levenbach_Hans_ISF2013.pdf Retrieved on 10-05-2022

[19] Loshin D. Business Intelligence: The savvy Manager’s Guide, Second Edition, ISBN: 978-0-12-385889-4, Morgan Kaufmann Imprint, 2013

[20] Matthew U. O.“Business demand for a cloud enterprise data warehouse in electronic healthcare computing: Issues and development in e-Healthcare cloud computing”. International Journal of cloud applications and computing, 2022https://doi.org/10.4018/IJCAC.297098

[21] Menon L., Rehani B. &Gund S.“Business Intelligence on the cloud overview, use cases and ROI”. In Proceedings of National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC 2012.

[22] Mohan P. M. & Manohar B. M.. Challenges in Big data analytics and cloud computing. International Business and Management Research, 9(2), 156-161, 2021

[23] Muntean M. “Business Intelligence Solutions for gaining competitive advantage”. In 7th WSEAS international conference on Artificial Intelligence, Knowledge engineering and databases, University of Cambridge, Feb 20-22, UK, 2008

[24] Nedubchezhian P., Moorty V. V. & Thirunavakkarasu P. D.“A Survey on integrating business intelligence with cloud computing”. International Journal of Applied Information Systems, (2), 9-15, 2012

[25] Nir K.“Cloud computing in developing economies”, Computer, 43, 10, 47-55, 2010

[26] Qlik, Cloud Data Warehouse, Online [Available at] https://www.qlik.com/us/cloud-data-migration/cloud-data-warehouse Retrieved on 09-05-2022.

[27] Raghunathan S.” Impact of information quality and decision-maker quality on decision quality: a theoretical model and simulation analysis”, Decision Support System, 26 (4), 275-286, 1999

[28] Rahm E. & Do H. H.“Data Cleansing: problems and Current Approaches”. IEEE Data Engineering Bulletin 23, 4, 11 pages, 2000.

[29] Rani et al. “Multi-level semantic annotation and unified data integration using semantic web ontology in big data process”, Cluster Computing, 1-13, 2017

[30] Ridzuan, F.&Zainon, W. M. N. “A Review on Data Cleansing methods for Big data”.Procedia Computer Science 161, 731-738, 2019.

[31] Souibgui M. et al.“Data quality in ETL process: A preliminary study”. Procedia Computer Science, 159, 676-687, 2019.

[32] Ul Haq Q. S. Data mapping for Data Warehouse Design. https://doi.org/10.1016/C2015-0-04423-9, Morgan Kaufmann Imprint, 2016.

[33] Vasista, T. G. K. “Wise CRM Engine “, Synergy-The Journal of Marketing, 5 (1), 123-127, SIMSR publications, Mumbai, India, 2007.

[34] Vasista T. G. K. & AlAbdullatif A. M. “Role of Electronic Customer Relationship Management in Demand Chain Management: A predictive analytics approach”,International Journal of Information Systems and Supply Chain Management (IJISSCM), 15 Pages,2017 DOI:10.4018/IJISSCM.2017010104.

[35] Wang, R. Y. & Strong, D. M. “Beyond Accuracy: What data quality means to data consumers”, Journal of Management Information Systems, 12 (6), 1996.

[36] Wang, T., Hu, J. & Zhou, H. “Design and Implementation of ETL approach in Business Intelligence Project”. In Practical Applications of Intelligence Systems, 2011, DOI: 10.1007/978-3-642-25658-5_35.

[37] Wieder B. & Ossimitz M-L. “The Impact of Business Intelligence on the Quality Decision”, Procedia Computer Science, 64, 1163-1171., 2015

[38] WinMan, How business intelligence can help you gain competitive advantage. Online [Available at]https://www.winman.com/blog/how-business-intelligence-can-help-you-gain-competitive-advantage Retrieved on 10-05-2022

[39] Yiu L. M. D., Yeung C. L., Cheng T. C. E. “The impact of business intelligence systems on profitability and risks of firms”. International Journal of Production Research, 59 (13), 2019

[40] Zanoon N., Al-Haj, A. & Khwaldeh S. M. “Cloud computing and Big Data is there a relation between the two: A Study”. International journal of Applied Engineering Research, 12 (17), 6970-6982, 2017