Big Data ETL Implementation Approaches: A Systematic Literature Review

Joshua C. Nwokeji*, Faisal Aqlan†, Anugu Apoorva*, and Ayodele Olagunju†
*Comp. & Info. Sys. Dept. Gannon Uni. † Indus.Engr., Dept., Penn. State Uni. ‡ Uni., of Saskatchewan;
Email: Nwokeji001@gannon.edu

***Abstract*—Extract, transform, load (ETL) is an essential technique for integrating data from multiple sources into a data warehouse. ETL is applicable to data warehousing, big data, and business intelligence. Through a systematic literature review of 97 papers, this research identifies and evaluates the current approaches used to implement existing ETL solutions. We found that conceptual modeling such as UML, BPMN, and MDA is the most popular approach used to implement ETL solutions. However, innovative approaches such as machine learning, artificial intelligence, and robotics are either under-utilized or not used at all to develop ETL solutions. Additionally, we discuss the implications of these to ETL research and practice.**

I. Introduction

Organizations are rapidly generating 'big data' from various sources, e.g., social media and e-commerce systems. The term big data is used to define data that are too voluminous and complex to be processed by traditional data processing systems [1]. Big data becomes more meaningful to organizations when analyzed to derive business intelligence that support decision making. Normally, big data analytics starts with integrating the generated data into a data warehouse using various techniques; ETL (extract, transform, load) is a popular technique used for this purpose [1].

A typical ETL process is carried out in three steps. In the first step, data in various formats (e.g., txt, csv, xls) are extracted from different data sources. The second step involves applying transformation techniques such as normalization, filtering, and sorting to clean these data. Finally, the cleaned data are migrated (loaded) into a data warehouse to be processed and analyzed to derive inteliigence, knoweledge, and wisdom [1]. In this era of big data, ETL research is becoming increasingly important and a means to developing new approaches to solve the growing challenges of data integration in organizations.

As ETL evolve, researchers and practitioners should be aware of the current techniques used to implement ETL solutions, and the implications of these to research and practice. Hence, this paper identifies and evaluates existing ETL implementation techniques from 97 papers selected using systematic literature review (SLR) method. In sections II, III, and IV we respectively discus SLR method, present and discuss result, and conclude our research.

DOI reference number: 10.18293/SEKE2018-152

II. Method

To meet the objectives of this research, we adopt the method for conducting SLR in software engineering proposed by Kitchenham and Charters [2]. In the paragraphs that follow, we discuss how this research conforms to this method.

a) Define Review Objectives and Question: The main objective of this review is to identify and evaluate approaches used in implementing existing ETL solutions. We identified a review question that closely align with this objective:

RQ: What approaches are currently used to implement ETL solutions?

b) Develop Search Strategy: This include selecting data sources, identifying search keywords, and conducting the search [2]. We selected the following data sources based on their relevance to software engineering and computer science: IEEE Xplore, ACM Digital Library, Science Direct, ProQuest, and Google Scholar. After a series of pilot searches, we identified and selected the keywords shown in Table I. To conduct the search, we combined each keywords in concept [A] with the keywords in concepts [B], and [C] using boolean operators (and/or).

TABLE I: SLR Keywords

Main Concept	Corresponding Keywords
[A] Extract Transform Load	[A1] Extract*, Transform*, Load*; [A2] ETL
[B] Approach	[B1] Framework; [B2] Models; [B3] Tools; [4] Technology; [B5] Software
[C] Quality	[C1] Quality Attributes; [C2] Quality Measures; [C3] Quality Features; [C4] Quality Characteristics; [C5] QoX

c) Specify Selection Criteria:: After developing the search strategy, the authors discussed and agreed on a set of criteria for including and excluding papers from review. These criteria are described in Table II.

TABLE II: Selection Criteria

Criteria	Include	Exclude
Year	Papers must be published between 2006 and 2016	Papers published before 2006
Relevance	Papers whose title and abstracts are relevant to ETL and its related concepts, as well as the review objectives and questions.	Papers that do not relate to the ETL and other key concepts of this research
Quality	Peer reviewed papers in journals, conferences, and workshops.	Non-peer reviewed papers. Also keynotes and presentations.
Rigor	Papers that demonstrate rigor through the use of appropriate scientific research and validation methods.	Papers that do not use appropriate scientific research and validation methods

d) Extract Data:: We searched the data sources with the keywords shown in Table I, this returns a total of 865 papers. Then, we removed duplicate papers and we excluded papers that do not meet the year criterion described in Table II. Afterwards, we applied the relevance criteria to exclude papers whose titles and abstracts are not related to the main concepts of the review. Finally, we selected a total of 97 papers as primary studies, after we applied the quality and rigor criteria. Please note that we did not show these primary studies due to space restrictions.

III. Result and Discussion

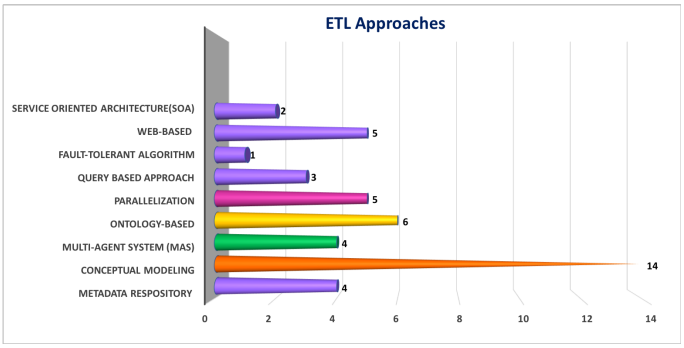

Fig. 1: Approaches used to Implement ETL Solutions

a) **RQ: What approaches are currently used to implement ETL solutions?:** We found nine (9) popular implementation approaches from selected papers (primary studies). As shown in Figure 1, these approaches are plotted in the vertical axis, while the horizontal axis shows the number of papers that reported them. For clarity, we also listed them here as follows : (i) Service oriented architecture (SOA) (ii) Web-based technologies, e.g., semantic web (iii) Fault-tolerant algorithm. (iv) Structured query languages (SQL). (v) Parallelization (parallel computing paradigm), e.g., Map Reduce. (vi) Domain ontology (vii) Multi-agent system (MAS).(viii) Conceptual modeling e.g., unified modeling language (UML) and business process modeling (BPMN) (ix) Meta-data repository.

b) **Variety of Approaches:** The results from this systematic literature review reveal that a variety of approaches are used to develop ETL solutions, see Figure 1. The leading among these is conceptual modeling approaches, such as the unified modeling language (UML), business process model and notation (BPMN), and model driven architecture (MDA) or model driven development (MDD). Conceptual models tend to represent real world concepts at level of abstraction that facilitates easy comprehension and analysis [3].

Although the application of conceptual modeling to ETL offers some advantages such as automatic code generation [4]; the overly emphasis on this, at the expense of other approaches, calls for concerns. For instance, while conceptual modeling approaches appear useful in present times; there is no clear research direction and indication of how conceptual modeling approaches can be applied to develop effective solutions to address future exponential increase in data complexity, volume, and heterogeneity.

More so, the focus on conceptual modeling appears to overshadow the application of new and innovative approaches such as artificial intelligence, machine learning, and robotics to developing ETL solutions. It is clear if or how these innovative can be applied to develop ETL solutions. This calls for research questions that should be focus of current studies in ETL. For instance, can artificial intelligence be used to automate data extraction from heterogeneous sources, or can machine learning and robotics be applied to transform and load data? Hence, we expect the future directions of ETL research to be in the area of developing new approaches based on current innovative technologies such as machine learning and artificial intelligence. These new approaches will be driven by the new data requirements, and the technology advancements.

IV. Conclusion and Future Work

This paper presents a systematic literature review, conducted to identify and discuss current approaches used to implement ETL solution; quality attributes to be considered when selecting ETL approach; and prevailing challenges in ETL research. In addition, we identify and discuss current trends in ETL research focusing on application domain, frequency or articles published per year and number of articles published per region. Based on the results identified from 96 papers, published between 2006 and 2016, we found that approaches for implementing ETL solutions overly focus on conceptual modeling, neglecting emerging and innovative approaches such as machine learning.

These challenges should call for concerns and questions among big data researchers and practitioners. The concern, for instance, would be if we (humans and machines) will ever catch up with the demands, and solve the problems, of big data integration and management. The questions are: Will big data eventually over grow human intelligence and machine capacity? Can we invent newer approaches (beyond what we currently have) to dealing with big data management? certainly these questions can be answered through intensified concerted effort by ETL researchers and practitioners. These should be the focus of future work.

References

[1] V. N. Gudivada, R. A. Baeza-Yates, and V. V. Raghavan, "Big data: Promises and problems." *IEEE Computer*, vol. 48, no. 3, pp. 20–23, 2015.

[2] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature reviews in Software Engineering Version 2.3," *Engineering*, vol. 45, no. 4ve, p. 1051, 2007.

[3] J. C. Nwokeji, F. Aqlan, B. Barn, T. Clark, and V. Kulkarni, "A modelling technique for enterprise agility," *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018.

[4] Z. El Akkaoui, E. Zimanyi, J. N. Mazon Lopez, J. C. Trujillo Mondejar *et al.*, "A bpmn-based design and maintenance framework for etl processes," *International Journal of Data Warehousing and Mining*, 2013.