Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2020, 7(10):101-105

Research Article ISSN: 2394 - 658X

Data Quality Assurance in Big Data and ETL Processes

Ravi Shankar Koppula

Satsyil Corp, Herndon, VA, USA
Ravikoppula100@gmail.com

ABSTRACT

The rapid growth of data across various industries necessitates efficient and effective methods for Extract, Transform, and Load (ETL) processes. This paper delves into the critical aspect of data quality assurance within the context of big data and ETL workflows. By exploring the unique challenges posed by large-scale data, this study evaluates existing quality assurance techniques and their applicability to big data environments. It emphasizes the significance of maintaining high data quality to ensure reliable decision support and analytics. The paper also presents design patterns for incorporating data cleaning techniques within ETL solutions, highlighting the importance of data profiling and cleansing in maintaining data integrity. Through a comprehensive analysis, the paper aims to provide insights and best practices for ensuring data quality in big data systems, thus enhancing the overall effectiveness of ETL processes.

Key words: Data Quality Assurance, Big Data, ETL Processes, Data Profiling, Data Cleansing, Data Transformation, Data Integration, Data Integrity, Quality Assurance Techniques, Data Warehousing

---

INTRODUCTION

Data quality has attracted considerable interest over the last few years, as dirty data in the data warehouse can lead to incorrect or misleading information for decision support. Many quality assurance techniques have been developed for conventional datasets, but it is not obvious that they can be used in the context of big data, as very large datasets introduce new challenges that did not exist before [1]. The purpose of this study is to use existing techniques to improve data quality in the context of big data within the extract-transform-load (ETL) process of data warehousing systems. There are two goals in particular: (1) Provide insight into the challenges regarding data cleaning in the context of big data, and (2) Present design patterns for incorporating existing data cleaning techniques in ETL solutions.

With data warehouses becoming an essential part of decision support systems, data warehouse administrators need to ensure that the information in the data warehouse is trustworthy. The extraction of data from source systems into the data warehouse is not always a straightforward process. Data can be moved from many different sources, similar data can be stored in different ways, and information that is untrustworthy or inaccurate may be replicated in the data warehouse [2]. Data quality has attracted considerable interest over the last few years, often referred to as "the whole truth and nothing but the truth." Dirty data in the data warehouse can lead to incorrect or misleading information for decision support, and many quality assurance techniques have been developed for conventional datasets.

Overview of Big Data and ETL Processes

Big data is defined as data that is too big, fast, or complex for a business to process and analyze using traditional methods [2]. Big data comes from various sources: sensors, social networks, the web, images, and videos, just to name a few. This data can come from satellites in space orbiting the earth, sensors placed in cars, social network's status updates in one second during high impact events, and diagnostic tests that can indicate health problems of patients. Voyager 2, an unmanned space probe, has been sending data to NASA for over 30 years. The data sent back every day by Voyager 1 is approximately 6400 bits. It is 0.15 images from that spacecraft. Therefore, it can be argued that it is not only the data itself that makes it big. Rather, it is the volume of the data that leads to it being classified as big [3]. The term ETL (Extract-Transform-Load) refers to a data processing sequence. First, the extract phase collects data from several heterogeneous sources. Second, the transform phase

applies several transformations to the extracted data, such as cleaning or reformatting. Finally, the load phase stores the transformed data in the same warehouse.

Over the past decade, virtualization, parallelism, and cloud computing have appeared as disruptive technologies that profoundly changed the landscape of computing systems. Recently, research emerged to address new challenges posed by the scale, complexity, and heterogeneity of big data. Unfortunately, conventional ETL systems were not designed with these challenges in mind. ETL tools were purposefully designed to operate a relatively small and static set of sources/targets with well-defined schemas. Moreover, the actual ETL logic was embedded in the specific ETL tool and could hardly be reused if the ETL tool was changed.

IMPORTANCE OF DATA QUALITY ASSURANCE

Quality has a significant and direct effect on the usability of data. For proper utilization and reliance on that data, there is a need for awareness on quality and its influences on availability, adequacy, relevancy, and accuracy of specific data [4]. Despite the extensive implementation of big data analytics applications in several industrial sectors, an efficient solution to big data quality assurance issues is still lacking. Moreover, there is no common understanding of the terms quality assurance and quality challenges. Nevertheless, data quality assurance (DQA) is viewed as the most important factor in ensuring big data analytics reliability and usability [2]. Conversely, ignoring data quality assurance could negatively affect the result and efficiency of data mining and big data applications for individual organization operation. Due to the unregulated accessibility of different types of data sources, the oncoming issues in big data analytics settings have been discussed.

Data quality remains an essential but challenging research area. Although substantial effort has been made and several quality measures were established, continuous efforts are still warranted. Quality issues have gotten more attention over the past few years, especially in ETL processes subdivided as data cleansing and sophisticated record linkage. In addition, while establishing quality measures and analysis of upstream sources is necessary, a scheme or framework for systematic post-process supervision and auditing of data quality over time is still missing. Due to the inherent characteristics of big data, including high volume, velocity, variety, and uncertainty, new types of quality issues have emerged, including noise, inconsistency, incompleteness, and more. Furthermore, despite the intensive research in the past few years, there are substantial gaps in today's knowledge. More specifically, either complete or tractable computational complexity remains largely unexplored. Alternatively, currently, there is no developed solution applicable to big data computation.

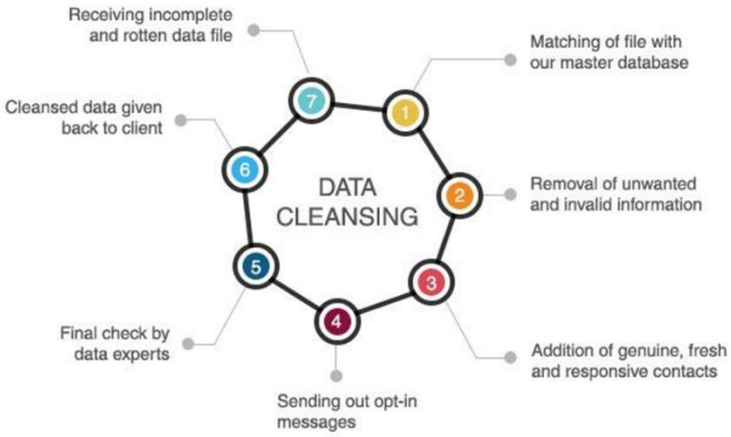
Impact of Poor Data Quality

Historically, data was structured and generally collected through well-defined models and standards [5]. In terms of data quality, these models and standards were also a type of guarantee that the collected data would be in accordance with its purpose making it able to deliver reliable information. Ironically, the emergence of new data sources, such as the web and social networks offering new insight behaviors, was accompanied by the skepticism regarding the quality of these data. Due to artificial intelligence and the massive computational power, it was possible to reconsider and effectively leverage these data in relatively new ways.

Today, we are in the era of big data. With humongous amounts of data being generated and stored, companies across multiple areas are striving to make sense of big data for improved decision making and customer satisfaction [2]. The data collected is mainly semi-structured and unstructured. New methodologies and mechanisms are being developed in order to overcome the traditional mindset of collecting clean, structured and analyzed beforehand data prior to processing and mine these massive datasets regardless of their size, structure and quality. However, the new data paradigm, although offering unlimited opportunities, comes with its own challenges. The data processed, either unstructured or semi-structured, stem from a multitude of sources, each one with its own idiosyncrasies, and have been collected with totally different aims and in various timeframes. Complicating the matters further, in the case of an organization, the data coming from social networks are user-generated and publicly available, thus they are noisy, misused, incomplete, duplicate and of unknown reliability. All the aforementioned issues are tied to the quality of the data fed in the processes since, as the saying goes, "garbage in, garbage out".

KEY COMPONENTS OF DATA QUALITY ASSURANCE

Data profiling and cleansing are examined, focusing on their roles in maintaining data quality. Data profiling involves analyzing data to uncover anomalies and inconsistencies, while data cleansing aims to repair or validate affected data. The importance of these processes in ensuring compliance with data quality standards and improving overall data quality is discussed. There is an emphasis on profiling data at the source at high velocity to reduce costs for defect data, ensuring consistency, detecting anomalies, and cleansing data as it is loaded into the target repository. Data cleansing steps are iterative as shown in below figure.



All companies are trying to understand the value of data and take advantage of it. However, a huge amount of data is stored that does not comply with Furthermore, it is essential to understand the complexity and challenges behind maintaining data the fundamental requirements and causes challenges. To take full advantage of big data, data quality (DQ) mitigation becomes necessary to guarantee the needed level of DQ until the data is consumed. The understanding of the generalized Data Warehouse architecture is the basis for understanding the fundamental components of an ETL process quality across the Big Data Warehouse architecture. Three methodologies to do profiling and subsequently cleanse the big data finding will be explored. The focus will be on the off-the-shelf technologies and solutions that are widely used nowadays at companies like Google, Facebook, and LinkedIn, and that support handling the scale of petabytes of data [6]. Profiling is the examination of a particular dataset to describe, summarize and understand its content. Profiling is a foundational step to cleanse a dataset, as the profiling outcomes and findings are subsequently used to identify the cleansing tasks to execute on the dataset.

BEST PRACTICES IN DATA QUALITY ASSURANCE

At the heart of a successful quality assurance strategy is implementing automated testing and monitoring of Data Quality (DQ) Assessments in Data Lineage, Extraction, Transformation, and Loading (ETL) processes. With an explosion of data, the data training pipelines are increasing in complexity, diversity of data types/sources, and frequency of ingestion. Therefore, manual auditing/testing approaches become impossible to handle. Only continuous monitoring, with the possibility of alerts, and automated retraining, can help detect issues and fix/retrain the processes in a reasonable manner [2].

The implementation of automated execution in real-time (i.e., monitoring pipelines) and daily batch processes can be achieved via scheduling tools such as Apache Airflow or TensorFlow Extended (TFX). For both of these tools, a variety of 0nDQ tests have been coded and integrated, and in parallel, alerts are being produced for key tests. Alerts are designed to inform data engineers/data scientists/ML engineers about issues in variables that require their attention [7].

On the subject of test libraries, a significant effort and collaboration across teams have been invested in creating a centralized library of DQ tests. One of the issues has been the negotiation of a taxonomy of different types of tests - including: 1-D tests that look at one variable, 2-D tests looking at two variables, tests that apply to just gold standard datasets, etc. Jupyter notebooks implementing 0nD Tests have then been coded prior to their deployment in SQL/PySpark/Apache PIG scripts or executed with Python. As testified by the evaluation section of this paper, successful adoption depends also on the context and needs of different teams across the organization. Multi-level testing - both spotting issues across a data lineage and at the level of individual operations/transformations - is also an important topic.

Automated Testing and Monitoring

Automating User-Centered Design of Data-Intensive Processes

Quality assurance (QA) is a system development activity that is responsible for identifying, fixing, and preventing data and process inconsistencies. It comprises a collection of tools and techniques to perform different tests on the process, data, and varied semantic models involved in the Business Intelligence (BI) solution, focusing on the desired quality characteristics of ETL processes in terms of performance, anomalies, and compliance [2].

Automated testing tools are responsible for the continuous execution of preset tests in a BI solution whenever a change occurs. In the context of data design, it should automatically query different data pipelines and models and check for the consistency of their results and metadata. This technique identifies the root cause of changes and failures, thus reducing the time needed to identify new bugs in the system and ensuring a certain level of

maintained quality [7]. In a broader view, monitoring can be seen as an extension of the testing mechanism. Regular queries are executed in the system to check if the data and process flow satisfy predefined quality conditions in respect to completeness, freshness, believability, or any other desired characteristics. Anomalies detected during the monitoring process can trigger alerts to notify the corresponding technical or end users.

CHALLENGES AND SOLUTIONS IN DATA QUALITY ASSURANCE

The integrity of data and rationality of data choice and use to yield useful knowledge from data are the preconditions for a successful deployment of big data technologies. Assurance of data quality has become nonetheless an indispensable task for organizations opting to utilize big data. In this pursuit, the innovative design and application of big data technologies together with conventional data quality methods have been the focus of research efforts in academia and industrial domains [2]. Data quality assurance in big data thus demands a comprehensive approach that integrates scalable big data technologies and conventional methods, with the intention to tackle the challenges arising from the unique characteristics of big data.

The assurance of data quality in ETL processes in ongoing data streams has not been comprehensively studied. It is unclear how to monitor emerging data streams in terms of data quality, detect data quality problems, and take corrective actions on problematic data, such as filtering, fixing, and enriching. To guarantee the quality of the data used in knowledge discovery tasks, existing data quality measures involving completeness, accuracy, consistency, timeliness, uniqueness, validity, and others adaptable to streaming data need to be adapted to non-streaming data. New data quality measures tailored to both big data and ETL processes need to be designed [3]. There are numerous obstacles across organizations that hinder data quality assurance in big data (research gaps). Inputs and schemata of available big data technologies are heterogeneous. The application of such technologies in combination with complex data quality methods often involves administrative and technological obstacles. For instance, the adaptation of traditional data quality methods that involve extensive human monitoring, integrity constraints, expensive cleaning and integration algorithms, and other expensive technologies, often does not scale and is infeasible in big data.

Handling Unstructured Data

Big Data and the accompanying ETL processes create challenges concerning the immense volume of data and data diversity, as there is a constant rise of new types of data from various sources. Oncologic imaging, eHealth records, genomic sequences, social media, and financial records are some particular types of datasets needed to be taken into consideration in the decision-making processes of healthcare providers as they contain valuable information. That said, several scientific and engineering fields constantly augment their datasets, therefore posing a challenge of diverse data integration techniques. The rapid development of new sensors and devices as well as data-generating web applications, such as social networks, have contributed to the huge volume of data generated every day [2]. In 2022, the estimation was that every individual produced over 1.7 Mb of data every second. In turn, this data growth creates the challenge of great dataset variety. There are many different types of data collected, from text files to images, XML documents to audio files, well-defined tabular data to geospatial records. In Big Data, specialized software is needed to manage well-defined data types. When data is not well-defined, the data must be organized, firstly after being read, before discarding the records with poor data quality. Unstructured data is all data that does not follow a strict schema already known at the moment of data acquisition. This type of data can be found, e. g., in emails, documents, text messages, and social posts. The need for unstructured data quality assurance arises because of the fact that poorly quality data can decrease the value of the data-driven information, thus leading to wrong and, therefore, costly decisions. The most common cognitional data filtering and cleaning operations on unstructured textual data are the following: creation of a stopword list (removal of common words like “the”, “and”, “is”, etc.); spelling correction; stemming, which means unifying the word variations (e. g., “houses” transforms into “hous”); lemmatization which means bringing the word back to the dictionary form (e. g., “going” would be “go”); removal of words with too high or too few frequencies (removal of exceedingly repetitive words, like “the”, or removal of domain-specific words that appear only in one document); and removal of offending words. Such offending expressions usually contain swear words, or are considered rude or hurtful [5]. Once the unstructured textual data is filtered and cleaned, the data quality might be ascertained by computing the number of rejected records or comparing the dataset with o factually known ‘ground truth’ dataset.

CONCLUSION

In conclusion, the importance of data quality assurance within big data and ETL processes cannot be overstated. As organizations increasingly rely on data-driven decision-making, the integrity and reliability of their data become paramount. This paper has examined the challenges and strategies associated with ensuring data quality in large-scale data environments.

The unique characteristics of big data—volume, velocity, variety, and veracity—introduce complexities that traditional data quality approaches may not fully address. Therefore, specialized techniques and tools are

necessary to maintain high standards of data quality. Key strategies include robust data profiling, comprehensive data cleansing, and continuous monitoring to detect and rectify anomalies promptly.

Incorporating data quality assurance into the ETL process ensures that data transformations and integrations do not compromise the data's accuracy, completeness, and consistency. This integration is essential for maintaining the trustworthiness of the data used in analytics and decision support systems. Best practices such as implementing automated data validation checks, employing scalable data quality tools, and fostering a culture of data stewardship are vital.

This study highlights the critical role of design patterns in embedding data quality checks within ETL pipelines, demonstrating their effectiveness in maintaining data integrity. By adopting these practices, organizations can significantly enhance the reliability of their big data systems, ultimately leading to more informed and accurate business decisions.

Future research should focus on developing more advanced, AI-driven data quality assurance techniques that can adapt to the evolving landscape of big data. Additionally, there is a need for standardization in data quality metrics and benchmarks to provide a consistent framework for evaluating data quality across different domains and industries.

In summary, ensuring data quality in big data and ETL processes is a continuous and evolving challenge that requires dedicated resources, innovative tools, and a proactive approach. By prioritizing data quality, organizations can unlock the full potential of their data assets, driving better outcomes and gaining a competitive edge in the data-driven world.

REFERENCES

[1]. M. Merino, J. G. Blas, D. Vazquez, R. S. Montero, and I. M. Llorente, "Improving the energy efficiency of large-scale distributed storage systems," Future Generation Computer Systems, vol. 55, pp. 100-111, Mar. 2016

[2]. V. Theodorou, "Automating User-Centered Design of Data-Intensive Processes," 2017.

[3]. V. Manickam and M. Rajasekaran Indra, "Dynamic multi-variant relational scheme-based intelligent ETL framework for healthcare management," 2020.

[4]. S. Juddoo and C. George, "Discovering the most important data quality dimensions in health big data using latent semantic analysis," 2018.

[5]. S. Anstiss, "Understanding data quality issues in dynamic organizational environments – a literature review," 2012.

[6]. X. Chu, "Scalable and Holistic Qualitative Data Cleaning," 2017.

[7]. C. N Williams, "Testing Data Vault-Based Data Warehouse," 2015.