ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

SCIENTIFIC Research and Community

Review Article

Open Access

Internal Controls in the ETL Process

Sneha Dingre

Data Analyst/ Modeler, Miami, FL, USA

ABSTRACT

The paper explores the significance of internal audit controls at each phase in the ETL process. The subsequent sections delve into internal controls during the extract, transform, and load phases, addressing key aspects such as data accuracy, completeness, consistency, error handling, logging, and metadata management. The paper concludes by stressing the need for controls during the load phase to ensure optimal loading procedures and data integrity.

***Corresponding author**

Sneha Dingre, Data Analyst/ Modeler, Miami, FL, USA.

Received: June 03, 2023; **Accepted:** June 05, 2023; **Published:** June 20, 2023

Keywords: Extract, Transform, Load, Data, Accuracy, Internal, Audit

Introduction to ETL Process and the Need for Internal Audit Controls

Organizations often employ extensive Extract, Transform, Load (ETL) processes, handling data ingestion from numerous sources into their data warehouse or facilitating data migration between systems. This intricate journey involves the transformation of data, application of business logic, and subsequent ingestion into the data warehouse, tailored to meet the requirements of the target destination [1]. The ETL process comprises three fundamental phases: extraction, transformation, and loading. In the extraction phase, the focus is on identifying and capturing pertinent data for ingestion into the warehouse. The subsequent transformation phase involves the application of business rules, constraints, and checks on the data. The final loading phase completes the cycle. The significance of data quality in ETL processes is underscored by the authors, emphasizing that the way ETL processes are developed profoundly impacts data quality [2]. Given their inherent complexity, improperly established ETL processes can introduce errors, potentially leading to the incorporation of erroneous data into the data warehouse.

High-quality data, defined as complete, consistent, valid, conformed, accurate, and timely, is essential for informed decision-making. During the extraction phase, data from diverse sources is typically copied to a staging area. Any inaccuracies in this process pose a risk of feeding incorrect data into the warehouse [3]. Potential reasons for incorrect data in the staging area include errors in change-data-capture (CDC) parameters, resulting in missing data or data duplication. Consequently, robust data audits and internal controls are imperative to ensure the ingestion of accurate and authentic data initially, thereby upholding data accuracy and authenticity. The author identifies four main dimensions for assessing data quality during the ETL process: data relevance, data soundness, data process, and data infrastructure [4]. These dimensions serve as a comprehensive framework for evaluating and ensuring the quality of data throughout the intricate stages of ETL processes.

Internal Controls During the Extract Phase

The ETL (Extract, Transform, Load) process initiates with the extraction phase, where data is gathered from diverse sources to facilitate subsequent transformations. Within this phase, several crucial activities take place. The following chart illustrates the sequential steps involved in the extraction process. To ensure data integrity and accuracy, internal controls are essential at each step of the extraction phase. For every control implemented at a specific step, a monitoring control is established to oversee its effectiveness in the transition to the subsequent step. This cyclical pattern is integral to guaranteeing the seamless operation of controls throughout the entire process.

Establishing proper access controls for data source identification is imperative, aligning with the responsibilities associated with each role. Once authentication and authorization controls are in place, the focus shifts to ensuring the quality of the extracted data. Recommended controls encompass addressing issues such as missing data values, empty fields, and incorrect data formats. Following data extraction, a pivotal step involves implementing metadata management controls. These controls serve to meticulously align metadata with the actual characteristics of the extracted data, ensuring accurate representation throughout the process. In terms of data validation during the extract phase of ETL, there are several controls that can be put into place to ensure data accuracy, completeness, and consistency.

Ensuring Data Accuracy and Completeness

Data sources come in different formats-structured, unstructured, or semi-structured. Each of these formats has distinct data validation techniques that can be employed to ensure that the data extracted during the extract phase meets predefined format requirements. Depending on the source being extracted, enforcing data validations, such as checks for data field types, date formats, and textual data validations, can assist engineers in identifying data inconsistencies in the early stages. Many times, data types differ between source systems and target systems, and these discrepancies can lead to issues such as data loss or truncation. Implementing checks for data type verification, data format conformation, and textual validation can guarantee that the extracted fields are accurately represented. Having these checks in place is crucial for maintaining data integrity throughout the ETL process.

Maintaining Logs and Audits
During the extraction phase, it is crucial to maintain error logs and audit logs to facilitate troubleshooting and maintenance activities in case anything goes wrong. Keeping key information, such as timestamps, process details, and trigger events, can aid in troubleshooting when pipeline failures occur. Additionally, as data sources might change over time, it is crucial to keep a log of changes and implement a version control process that tracks the modifications in extraction processes.

Metadata Management
Metadata for objects in a data warehouse usually holds information about a view's or table's definition, staging table's definition, data about logical and or physical data models. It is crucial to cleanse and standardize this metadata after extraction process is done [4]. This metadata allows for impact analysis and data lineage showing how data moved through the data warehouse. Figure 1 shows an approach to metadata management using warehouse architecture [4].

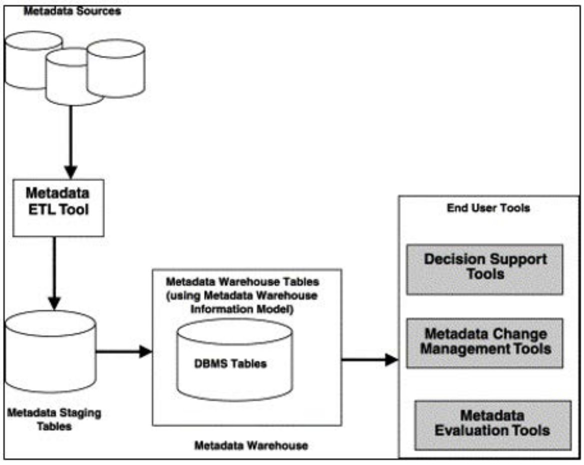
Metadata Sources
Metadata ETL Tool
Metadata Staging Tables
Metadata Warehouse Tables (using Metadata Warehouse Information Model)
DBMS Tables
Metadata Warehouse
End User Tools
Decision Support Tools
Metadata Change Management Tools
Metadata Evaluation Tools

Figure 1: The Metadata Warehouse Architecture [4]

Internal Controls During the Transform Phase
The Transform phase in the ETL process is dedicated to reshaping the extracted data to meet specific requirements. This involves comprehensive data cleaning and transformative actions such as aggregation and filtering. The ensuing chart elucidates the sequential flow of activities within the transformation phase, portraying the orchestrated steps involved in molding the data to its desired state.

Internal controls established during the transformation phase begin with data cleaning. While cleaning the data, it is crucial to implement validation controls over the cleaning process. These validation controls ensure that the cleaned data adheres to the specified requirements. After cleaning, the data undergoes various transformational activities, including filtering and the addition of transformative data elements. Implementing controls over each data transformation guarantees that the data is transformed according to the designed requirements. It is also essential to establish time-bound controls to track the duration of various transformation activities. Following data transformations, the data needs to be formatted into a unified structure easy to load into the designed data model of the data warehouse. Controls over formatting and ensuring that data from multiple sources is formatted into a unified data format ensure smooth aggregation of data from various sources. Post data formatting, the data enters the aggregation phase, ready to be loaded into the data warehouse. In the process of aggregating data, data quality will be affected. Hence, it is essential to have robust controls during this phase [5].

Data Validation Rules
Objective of the data validation is to understand the accuracy, completeness and consistency of the transformed data. Some of the activities during this is to define data validation rules to implement checks for data accuracy, completeness and consistency. Validate the data against predefined rules such as data format, type and range requirements. Conduct consistency checks to maintain data integrity.

Error Handling and Logging
It is crucial to identify, address, and monitor errors encountered during the transformation process, establish mechanisms for capturing errors and implementing error-handling procedures. Detailed information about errors can be logged for audit and troubleshooting purposes. Data quality metrics can be monitored and automated processes can be set up for continuous monitoring. Generate reports on key data quality issues and trends.

Documenting Enhancements and Transformation Logic
To continuously improve the transformation processes, optimize performance, and ensure governance, it is important to maintain clear documentation for transformation logic. Implement version control for tracking changes to transformation rules. Establishing controls for metadata management can help with data lineage and transparency.

Internal Controls During the Load Phase
The Load phase of ETL involves loading the transformed data into a designated database. The chart below illustrates the sequential steps in the load process of ETL. Throughout the load phase, it is crucial to define and comprehend the targeted schema and structure of the data warehouse. Following the understanding of the schema, all transformational activities can be scheduled and loaded into the data warehouse.

Internal controls during the load phase revolve around the successful loading of data at optimal speed and time. Establishing controls includes verifying if the data is loaded in the correct format and structure, and tracking timelines for each load to ensure optimal loading procedures are followed. As mentioned by the authors, maintaining minimal execution time is crucial. Therefore, controls over timelines are essential for optimization in [6]. Implementing controls for comparing data sources with final data warehouse tables will also prove useful. Additional controls encompass data encryption in the data warehouse and maintaining data logs, among others.

To ensure the accuracy, completeness, and integrity of the loaded data, referential integrity checks during the loading process can be implemented. It is also crucial to aapplying data validation rules along with conducting checks for data types, formats, and range requirements to validate the accuracy and completeness of loaded data. Lastly, loaded data can be reconciled with the source to ensure accuracy.

Conclusion

Robust controls are necessary for maintaining data accuracy and integrity throughout the ETL process. This paper emphasizes the critical importance of these controls during extraction, transformation, and loading phases to mitigate errors and ensure the reliability of organizational data. Implementing stringent measures, including data validation, error handling, and metadata management, is essential for organizations seeking to enhance the effectiveness and trustworthiness of their ETL processes.

References

1. Radhakrishna V, Sravan Kiran V, Ravikiran K (2012) Automating ETL process with scripting technology. 2012 Nirma University International Conference on Engineering (NUiCONE), Ahmedabad, India 1-4.
2. Hamed I, Ghozzi F (2015) A knowledge-based approach for quality-aware ETL process. 2015 6th International Conference on Information Systems and Economic Intelligence (SIIE), Hammamet, Tunisia 104-112.
3. Saranya N, Brindha R, Aishwariya N, Kokila R, Matheswaran P, et al. (2021) Data Migration using ETL Workflow. 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India 1661-1664.
4. Munawar (2021) Extract Transform Loading (ETL) Based Data Quality for Data Warehouse Development. 2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI), Jakarta, Indonesia 373-378.
5. Xp W, Li J (2023) Design of Data Quality control system based on ETL. Journal of Physics: Conference Series 2476: 012083.
6. Panos Vassiliadis (2009) A survey of Extract-Transform-Load Technology. International Journal of Data Warehousing and Mining 5: 1-27.