Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

A WORKFLOW FOR DATA QUALITY MANAGEMENT AND ERROR HANDLING IN ETL PROCESSES

ASIS MOHANTY
Research Scholar, Sri Sri University, Cuttack, Odisha India.
Email: asis.m2021-22ds@srisriuniversity.edu.in

Dr. SUNIL K DHAL
Professor, Sri Sri University, Cuttack, Odisha, India. Email: sunildhal@srisriuniversity.edu.in

Dr. NILAYAM K KAMILA
Senior Lead Software Engineer, Capital One, 802 Delware Avenue, Wilmington Delaware 19801 United States of America. Email: nilayam.kamila@gmail.com

Abstract

Data is important in quality management extracts, transformations, load processes (ETL), which are fundamental to integration and analysis of data from different sources in data warehouses and analysis. Ensuring high data quality and effective error handling is necessary to maintain the reliability of business intelligence and decision-making processes. This study presents a comprehensive workflow to handle data quality and errors in ETL processes, which address general data quality challenges such as discrepancies, lack of data and duplicate records. The proposed workflow includes a series of structured stages: data profiling, definition of data quality rules, detection of automated errors, improvement mechanisms and ongoing monitoring. These stages are supported by specific techniques such as data, data transformation rules, and vigilant mechanisms for data quality deviations, and ensure rapid detection and dissolution of problems. In addition, the workflow is designed to be adaptable in different data communities, with the provisions on integration into the existing ETL framework. This study is provided to reduce data errors, increase data quality and portray the efficiency of the workflow in ETL process efficiency. This research contributes to the field by offering a standardized approach to managing data quality in ETL processes, promoting better data regime and enabling organizations to make more informed, date-driven decisions.

Keywords: Data Quality Management, ETL Process, Error Handling, Data Profiling, Data Governance.

1. INTRODUCTION

In today's data-driven world, companies rely on extracts, transformations, load processes (ETL) to integrate and analyze large versions of data from many sources. However, ensuring high data quality and effective failure management in ETL workflows is an important challenge, as data deviations, lack of values and change errors can come to the decision. Poor data quality can cause incorrect analysis, disability and compliance risk. To solve these challenges, a structured ETL workflow is required that includes strong data quality management and automatic error management mechanisms.

This study proposes a comprehensive workflow to handle data quality and errors in ETL processes. Workflows include checks after loading accountability to ensure automated data verification, classification of good and bad records, mis insulation and data integrity. In addition, decisions such as Machine Learning (ML) technology, such as decisions

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

and random forests, are integrated into workflows to detect errors, predict potential errors and improve the efficiency of poor journal treatment. By taking advantage of these ML models, the ETL process can classify the mail with care, adapt the fault solution and reduce manual intervention.

The proposed workflow ETL provides a scalable and adaptive approach to data quality management, and ensures that businesses can make more accurate, date -driven decisions. By detecting real -time monitoring, automatic error solution and future failure, the purpose of this study is to increase ETL efficiency and reliability. Ultimately, this research data helps to improve data, reduce the risk of operating and ensure high quality data for analysis and commercial intelligence applications.

2. STATEMENT OF THE PROBLEM

Ensuring data quality and errors in ETL procedures is an important challenge for the output organizations on data-driven decisions. Traditional ETL workflows often struggle to handle nonconformities, lack of values and change errors, causing incredible insights and disabilities. Automatic mechanism deficiency to detect and solve errors complicates further data integrity management. This study addresses these challenges by suggesting a structured ETL workflow, including the decision to find out data verification, error classification and future failure. The goal is to improve ETL efficiency, reduce manual intervention and secure, high -quality reliable data for analysis and business intelligence.

3. LITERATURE SURVEY

ETL (Extract, Transform, Load) process is an important component in data integration and analysis, and to ensure data quality management and error management in the ETL workflows have been focused on extensive research. Various studies have proposed frameworks and function to increase data security and optimize ETL performance.

Kim, Park and Lee (2018) emphasized the role of data quality management in ETL processes, which identify major challenges such as inconsistent data format, lack of values and repetition. His study suggested an automated verification mechanism to ensure data integrity. Similarly, Garcia, Perez and Rodriguez (2017) conducted a comprehensive survey on data quality problems, classified general problems and reviewed the existing data profiling and cleaning solutions. Error handling in ETL workflows has also been studied a lot. Singh and Kaur (2019) analysed the mechanism for errors, proposed rules -based and machine learning (ML) - -driven approach to improving RUL time. In a related study, Patel and Shah (2022) introduced an automated data quality assessment system, which includes deviations and verification techniques to improve ETL data.

Machine learning has proved to be a promising approach to increase data quality management in ETL processes. Zhang and Lee (2020) discovered the use of wooden and random forest algorithms to predict and reduce data quality problems, which

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

improved error classification accuracy. Similarly, Huang, Wang and Lee (2015) investigated ETL process adjustment techniques, including work flight automation and parallel processing to improve data reliance and efficiency. Handling large data ETL challenges provide more complications. Chen and XU (2023) discussed scalability issues in large data ETL pipelines, which suggest adaptive error management mechanisms to manage high-step, real-time data. Similarly, Kumar and Srinivasan (2021) focused on strategies to handle real-time errors, and advocated power-based ETL approaches to reduce delay and loss of data.

Other studies have shown specific data stays and frame development techniques. Lopez and Martinez (2014) underwent data methods such as dismissal detections and standardization, while Brown, Wilson and Taylor (2016) suggested a structured ETL structure to ensure continuous data quality verification. In addition, Guayen and Tran (2020) performed a comparative analysis of ETL tools, evaluated their abilities in data profiling, transformation and incorrect handling. Finally, Smith (2012) emphasized the best practice for ETL error management, including strong logging, monitoring and recovery mechanisms.

This literature review emphasizes the increasing importance of automation, machine learning and real -time processing in ETL data quality management and error management. Future research should focus on integrating adaptive teaching techniques and cloud-based ETL solutions to further increase the data's accuracy and operational efficiency.

4. OBJECTIVES

1) To integrate machine learning techniques, especially in the ETL workflow for decisions tree and random forests, automated data quality evaluation, historical error patterns and predetermined data can enable intelligent classification of good and bad records based on quality rules.
2) Decision to predict potential data errors, optimize poor journal treatment and improve the general data integrity through adaptive learning and deviations to decide to detect errors in ETL processes by taking advantage of decision tree and random forest models to detect errors and increase handling.

5. DATA MANAGEMENT

Data follows a structured approach to ensure quality management and errors to ensure ETL workflow reliable data processing. It begins with the water and the quality phase, where the data is read from a source (step 1: Data Reed). If the reading fails, resumes the process or logs in errors. When the data is read, the data undergo the quality control (step 2: source data quality), where the records are classified either as good or bad. In the composition and strain phase, good items continue to step 3: While the data is sent to the poor journal treatment unit (step 3.1) for error handling and potential improvement.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

The data created is then loaded into the target system (Step 4: Computer Management) but is identified and flagged for identifying some unsuccessful mail. The phase of dealing with errors and errors ensures miscarriage and resolution, where the failed record for record items is either restored or stored in S3. Finally, in phase 5: Error catch and after data that uploads purity, which verifies the integrity of the system that uploaded data, ensures perfection and accuracy. It increases workflow quality, automatically to handle errors, and provides a structured recovery mechanism, which makes ETL processes more reliable and scalable to make effective data-driven decisions.

5.1. Data Quality Problem in ETL Process

Problems with data quality in ETL (Extracts, Transforms, Load) processes create important challenges for organizations accurately, continuous and timely data to make decisions. Poor data quality can lead to incorrect insight, operating disabilities and compliance risk. The table below emphasizes the ETL workflows, their causes and the most important data quality problems in possible effects on their causes and data processing and decision -making.

Data Quality Problem	Description	Cause	Impact
Missing and Incomplete Data	Data records contain null or missing values.	Data extraction errors, system failures.	Reduces data reliability and completeness.
Inconsistent Data Formats	Variability in data types, formats, or units.	Different source system standards.	Causes transformation failures and errors.
Duplicate Records	Multiple instances of the same data entry.	Poor deduplication mechanisms.	Leads to redundancy and inaccurate analysis.
Data Anomalies and Outliers	Unusual or incorrect data points.	Human errors, system glitches.	Misleading insights and analytical errors.
Schema Evolution Issues	Structural changes in source databases.	Updates in source systems.	Breaks ETL pipelines, leading to failures.
Data Latency Issues	Delays in data extraction, transformation, or loading.	Network issues, slow processing times.	Affects real-time decision-making.
ETL Process Failures	Errors in extraction, transformation, or loading.	System crashes, incorrect configurations.	Leads to corrupted or incomplete datasets.

To reduce these problems, organizations must discover automatically computer verification in the ETL workflows, detection of machine learning-driven errors and implement a strong error management mechanism.

6. DATA ANALYSIS

The data analysis in this study focuses on evaluating the efficiency of ETL workflows proposed to handle data quality and handle errors in using decision -making wood and random forest machine learning models. These models are used to classify data registrations and classify in incorrect categories and automatically detect errors in the

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

ETL process and improvement in the resolution. The decision provides an explanatory model for identifying patterns in deviations and deviations in data by following a structure of decision tree regulations. However, it is prone to overfit, which may affect the new dataset to influence normalization. Random forest, an outfit method, reduces the problem by making many decisions by making trees and on average their outputs, increasing the strength and accuracy. To assess model performance, four key evaluation metrics are considered:

- ***Accuracy*** measures the overall correctness of classification, indicating the proportion of correctly classified records out of total records.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- ***Precision*** evaluates the proportion of correctly identified erroneous records among all records flagged as errors, reducing false positives.

$$Precision = \frac{TP}{TP + FP}$$

- ***Recall*** measures the model's ability to detect actual errors, ensuring fewer false negatives in data quality validation.

$$Recall = \frac{TP}{TP + FN}$$

- ***F1-score*** provides a balanced measure of precision and recall, offering a comprehensive view of model effectiveness.

$$F1\ Score = 2 \times \frac{Precision\ \times Recall}{Precision + Recall}$$

By taking advantage of these calculations, the study shows that the integration of decision-making wood and random forest model in the ETL process improves the error classification, reduces manual intervention and increases data quality management, and eventually ensures valued and high intelligence data for analyzing and business applications.

6.1 Algorithms of Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.model_selection import train_test_split
import pandas as pd
# Load the dataset (Assume df is a preprocessed dataset)
X = df.drop(columns=['Label'])  # Features (input variables)
```

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

```
y = df['Label']  # Target variable (Valid or Error)
# Split dataset into training (80%) and testing (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Initialize and train Decision Tree Classifier
dt_model = DecisionTreeClassifier(criterion='gini', max_depth=5, random_state=42)
dt_model.fit(X_train, y_train)
# Make predictions
y_pred = dt_model.predict(X_test)
# Calculate Evaluation Metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='binary')  # For binary classification
recall = recall_score(y_test, y_pred, average='binary')
f1 = f1_score(y_test, y_pred, average='binary')
# Print Results
print(f'Accuracy: {accuracy:.2f}')
print(f'Precision: {precision:.2f}')
print(f'Recall: {recall:.2f}')
print(f'F1 Score: {f1:.2f}')
```

6.2 Algorithms of Random Forest

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score
from sklearn.model_selection import train_test_split
import pandas as pd
# Load dataset (Assume df is a preprocessed dataset)
X = df.drop(columns=['Label'])  # Features (input variables)
y = df['Label']  # Target variable (Valid or Error)
# Split dataset into training (80%) and testing (20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Initialize and train Random Forest Classifier
rf_model = RandomForestClassifier(n_estimators=100, criterion='gini', max_depth=10,
random_state=42)
rf_model.fit(X_train, y_train)
# Make predictions
```

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

```
y_pred = rf_model.predict(X_test)
# Calculate Evaluation Metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred, average='binary')  # For binary classification
recall = recall_score(y_test, y_pred, average='binary')
f1 = f1_score(y_test, y_pred, average='binary')
# Print Results
print(f'Accuracy: {accuracy:.2f}')
print(f'Precision: {precision:.2f}')
print(f'Recall: {recall:.2f}')
print(f'F1 Score: {f1:.2f}')
```

6.3. Data Set

The study decision uses several datasets to evaluate data quality management and incorrect handling in ETL processes using wood and random forest models. Each data set has financial, transactions and user-related properties, with the target variable token is reliable, which classifies symbols either reliable (1) or incredible (0).

Dataset -1 contains 200 items, including opening value, closing value, highest and lowest value, trade volume, user followers, user friends and credit rating. It lacks values and includes five float and four integer columns, making it suitable for classification functions.

Dataset -2 contains 500 records, each representing a Blockchain -Token transformation, integrates financial and user -related characteristics to determine reliability based on important affected factors.

Dataset-3 extends the analysis with 1,650 items, including share value, trade volume and user-specific matrix. The Decision Tree model is used to classify symbols reliable by analyzing functions such as user credit assessments, stock trends and trade volumes, measured by means of accuracy, accuracy and F1-score with performance.

The largest dataset, dataset -4, contains 10,000 items, and maintains the same properties as the previous dataset. Decision Tree Model Token ensures a comfortable classification of reliability, without lack of values affecting data integrity. Important viewing measurements are used to validate the model's efficiency to ensure reliable classification results.

These data sets provide a strong foundation for training and evaluating both wooden and random forest models and detecting errors in ETL workflows and increasing data quality management.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

6.4. Results and Discussion

Table 1: Results Comparison

Data	Accuracy		Precession		Recall		F-1Score	
	Decision Tree	*Random Forest*	*Decision Tree*	*Random Forest*	*Decision Tree*	*Random Forest*	*Decision Tree*	*Random Forest*
Set 1	0.95	0.38	0.95	0.43	0.95	0.26	0.95	0.32
Set 2	0.91	0.53	0.95	0.55	0.86	0.62	0.90	0.58
Set 3	0.98	0.50	0.98	0.51	0.98	0.49	0.98	0.50
Set 4	0.96	0.51	0.95	0.49	0.96	0.44	0.96	0.46

The study data in data quality management and error management in the ETL processes evaluates the efficiency of the tree and random forest model. Table 1 presents comparative analysis of both models in four data sets using the main performance measurements: accuracy, accurate, recall and F1-shor.

From the results, the decision trees are continuously high accuracy in all data sets, which have values from 0.91 to 0.98, while random forests with accurate values between 0.38 and 0.53 to a large extent.

Better performance of the decision -making three model in accuracy shows that this data effectively captures the underlying pattern in the quality classification. However, although random forests show low accuracy, it provides better generalization and strength, especially in recall values.

When it comes to accuracy, the decision tree maintains high values (0.95 - 0.98) in all data sets, while random forests have ups and downs, from 0.43 to 0.55. This indicates that the decision is more accurate when it comes to classifying the tree reliable and incredible symbols and lowering false positives. However, Random Forest Set 2 (0.62) shows better recall performance, which highlights its ability to identify the more real wrong items correctly.

The F1 beeps, which balance and remember accuracy, are the highest in the decision tree (0.90 - 0.98), and confirm its general dominance in classification performance. However, F1-score performance set with random forest is relatively better in set 2 (0.58), which strengthens the strength of the handling of unbalanced data.

Key Observations:

- ❖ Decision tree improves general accuracy and classification performance, making it more suitable for accurate data quality management in ETL workflows than random forest.
- ❖ Random forest shows better recall in some cases, indicating the possibility of more efficiently detecting the wrong mail in large and more complex datasets.
- ❖ Display intervals suggest that the decision tree is more effective for wood -structured data, while random forest may require additional parameters setting for optimal results in detecting data errors.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

7. FINDINGS

The study on a workflow for data quality management and decisions when using three- and random forest models provides many major insights into ETL processes:

- ❖ Decision trees work continuously than random forests in terms of accuracy, accurate, recall and F1 score, and receive accuracy values from 0.91 to 0.98, making it a more effective model for ETL data classification.
- ❖ Random forest, although less accurately, shows better recalls in specific data sets, especially in set 2 (0.62 recall), indicates the ability to classify more real wrong items correctly.
- ❖ Excel (0.95 - 0.98) in decision tree accuracy ensures minimum false positive, which is important for maintaining high data integrity in ETL workflows.
- ❖ Random forest struggles with low accuracy (0.38 - 0.53), suggesting that additional setting or enchanted techniques may be necessary to improve efficiency.
- ❖ The F1 score in decision tree remains high (0.90 -0.98), confirms the reliability of balanced accuracy and recalls for effective error handling.
- ❖ Both models ensure reliable data quality management but serve different goals - random forests for accurate classification for normalization in the detection of crucial trees and errors.

8. CONCLUSION

The study suggests that the decision to secure data quality in ETL processes is a wood -proposed model, which provides high accuracy, accuracy and balanced performance in datasets. Although Random Forest provides better recall in some cases, the low general accuracy suggests that it may require further setting for effective applications in ETL error management. Ultimately, the integration of machine learning models in ETL workflows increases data integrity, reduces errors and improves the efficiency of data processing pipes. The decision to adapt both accuracy and memories in future research can detect a hybrid approach to a combination of a combination of wood and random forest models, ensuring a stronger ETL data quality management system.

Reference

1) Brown, T., Wilson, G., & Taylor, R. (2016). ETL Frameworks for Ensuring Data Quality. Journal of Data Engineering, 10(4), 210-225.

2) Chen, H., & Xu, Y. (2023). Data Quality Challenges in Big Data ETL Processes. Big Data Research, 15(2), 50-65.

3) García, M., Pérez, J., & Rodríguez, L. (2017). A Survey on Data Quality Issues in ETL Processes. Information Systems Review, 23(1), 75-92.

4) Huang, Y., Wang, X., & Li, Z. (2015). ETL Process Optimization for Data Quality Improvement. Computing and Informatics, 34(5), 1010-1025.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 59 Issue: 04:2026
DOI: 10.5281/zenodo.20351920

5) Kim, J., Park, S., & Lee, H. (2018). Data Quality Management in ETL Processes. Journal of Data Management, 12(3), 145-160.

6) Kumar, S., & Srinivasan, R. (2021). Real-Time Error Handling in ETL Systems. Journal of Real-Time Data Processing, 7(3), 180-195.

7) Lopez, F., & Martinez, D. (2014). Data Cleansing Techniques for ETL Processes. Data Cleaning Journal, 8(2), 95-110.

8) Nguyen, P., & Tran, L. (2020). A Comparative Study of ETL Tools for Data Quality Management. Software Engineering Journal, 25(3), 300-315.

9) Patel, R., & Shah, M. (2022). Automated Data Quality Assessment in ETL Pipelines. Journal of Big Data, 9(1), 112-130.

10) Singh, A., & Kaur, P. (2019). Error Handling Mechanisms in ETL Workflows. International Journal of Computer Science, 16(2), 89-102.

11) Smith, J. (2012). Best Practices for Error Handling in ETL Processes. Information Technology Journal, 14(1), 50-65.

12) Zhang, L., & Lee, K. (2020). Improving ETL Data Quality with Machine Learning. Data Science Journal, 19(4), 225-240.