ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/hari-prasad-bomma/480

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 1 & Iss: 2

Research Article

Daily Regression Suite – DRS A Framework to Optimize Data Quality

Hari Prasad Bomma*

Citation: Bomma HP. Daily Regression Suite – DRS A Framework to Optimize Data Quality. *J Artif Intell Mach Learn & Data Sci 2023*, 1(2), 2198-2200. DOI: doi.org/10.51219/JAIMLD/hari-prasad-bomma/480

Received: 02 May, 2023; **Accepted:** 18 May, 2023; **Published:** 20 May, 2023

***Corresponding author:** Hari Prasad Bomma, Data Engineer, USA, E-mail: haribomma2007@gmail.com

ABSTRACT

Data quality testing is an important and integral part of ETL to make sure certain data quality standards are met. Testing makes sure the data qualities such as Accuracy, Reliability, Timelines and Relevance are maintained. In this article we will review the current methodology, architecture and a data quality framework Daily Regression Suite (DRS) that can help curtail data issues in production systems.

Keywords: Data quality, Data warehouse, Regression suite, ETL testing, Data anomalies, Integrations, ETL Optimization, Framework

1. Introduction

Data quality is defined as the degree to which data meets a company's expectations of accuracy, validity, completeness and consistency. Data quality is critical for analysis, reporting and decision making. High quality data helps in informed decision making and efficient business operations. On the other hand quality issues in data can have a huge impact on business and user experience. Poor quality data leads to all kinds of problems like improper financial, inaccurate reporting, dissatisfied customer, missed opportunities and flawed decision making.

Despite all the checks in place, data issues can arise over time due to various factors such as job failures, absence of data from the source, invalid data from different systems, irregular data volumes and data extraction from non quality systems which may cause processing failures. These issues can lead to data inconsistency, data loss, duplicity and other complications.

Implementing an inbuilt regression testing suite DRS that runs daily in the production environment can help mitigate these challenges. This suite ensures that data quality is consistently maintained and promptly notifies any discrepancies, saving time and providing an effective solution for reducing data issues.

2. Research Background

Traditionally, the ETL process plays a crucial role in populating data warehouse. This includes the historical load during the initial setup and subsequent incremental updates carried out on regular intervals, ensuring that the data warehouse stays current and functional.

A typical data flow commences with the extraction of data from diverse source systems. This raw data is subsequently transformed in accordance with specific business requirements and logic. Post transformation, the refined data is loaded into target tables or databases.

Generally, when historical data load happens for the first time or an ETL is built for first time, data undergoes thorough quality testing to ensure its accuracy, reliability and consistency. Data quality testing includes accuracy check to ensure the data is correct and free from errors. Completeness check to guarantee all required information is present. Consistency checks to ensure data is uniform and without conflicts across different sources.

2.1. Data quality testing process

To ensure the highest quality of data organizations follow a

comprehensive testing life cycle. This life cycle encompasses several critical stages:

- **Requirement analysis:** Identify and document the data quality requirements based on business needs and objectives.
- **Test planning:** Develop a detailed test plan outlining the scope, approach, resources and schedule of testing activities.
- **Test design:** Create test cases and scenarios that cover all aspects of data quality, such as accuracy, completeness, consistency and timeliness.
- **Test execution:** Conduct the tests according to the plan, executing test cases and recording the results.
- **Defect management:** Track and manage any data quality issues detected during testing, ensuring they are addressed and resolved.
- **Test closure:** Review the entire testing process, documenting lessons learned and best practices for future projects.

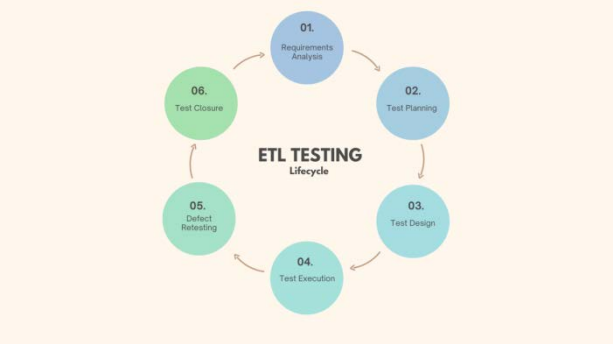

Figure 1: Traditional ETL Testing Cycle.

Once deployment is complete, testing vigilance often wanes until new updates are introduced. This delay in ongoing scrutiny can hinder the early detection of data issues and data anomalies caused by various factors. To mitigate such discrepancies, an inbuilt regression testing suite DRS can be deployed along with critical tables to the production environment.

3. Methodology

Implementing an inbuilt Daily Regression Suite that runs daily/regular intervals in the production environment involves several key steps and prerequisites.

3.1. Requirement and setup

Clearly outline the goals of the regression testing suite, such as ensuring data consistency, preventing data loss and identifying anomalies. Precise data quality rules and criteria a regression testing suite will validate needs to be established. Suitable data quality and regression testing tools that integrate seamlessly with existing ETL processes help in a smooth transition.

3.2. Development

Write automated test scripts using the chosen tools to perform daily regression tests. These scripts should be capable of validating data against the predefined quality rules. Implement the regression testing suite in an iterative manner, allowing for continuous improvement and adaptation based on feedback. A custom table can be built with columns such as Rule number, Rule description, Source query, Source connection, target query, target connection, expected count, actual count, range min, range max, last executed at, loaded by, status and active indicator etc. This table captures data such as number of records received, load date and time, range of data etc to compare and notify if an anomaly is detected.

Once the regression testing suite is fully integrated into the ETL workflow, ensure to run tests automatically after daily data load completes. Implement real time monitoring and alerting systems to notify stakeholders immediately when discrepancies are detected. Deploy the regression testing suite in the production environment, configuring it to run at specified intervals (e.g., daily). Build notification mechanism to review the test results and validate them against the expected outcomes. This involves analyzing test logs and reports to identify and address any issues. Regularly update the test scripts to reflect changes in business requirements and data structures. Continuously refine the regression testing suite based on feedback and new data quality challenges that arise.

3.3. Reporting

After implementing the regression testing suite, the next step is to generate reports that provide insights into data quality and any detected issues. Choosing the appropriate reporting tools that integrate well with existing Data Quality and ETL workflows is also a key task.

Reporting metrics includes identifying the key metrics that needs to be tracked and report on, such as: Data Quality Scores which aggregate measures of data quality based on predefined rules. Error rates, frequency and types of data quality issues detected. Trend Analysis changes in data quality over time.

Automated tools can generate these reports at scheduled intervals, ensuring timely insights. Creating interactive dashboards allows stakeholders to drill down into specific data quality issues and track progress over time. By connecting data sources, transforming data, creating visualizations and scheduling automated report generation and distribution, user friendly reporting tools can enhance transparency, improve decision making and track improvements. Incorporating automated reporting into the data quality framework provides clear, accessible insights into data quality status, timely information for informed decision making and continuous monitoring for ongoing refinement.

3.4. Plan for scalability

Ensure that all relevant stakeholders are trained on the importance of data quality and how to use the regression testing suite. Involving them in the testing process and encouraging feedbacks improve the system. Design regression testing suite to handle growing data volumes and complexities. Planning for scalability ensures that the suite can accommodate future data quality requirements. This foresight prevents future bottlenecks and ensures that the suite remains effective as the organization's data needs expand.

4. Justification of Chosen Methods

This kind of data quality framework reduces the manual effort required and increase the efficiency of identifying data issues early in the ETL process.

Automated scripts also ensure that tests are executed regularly and reliably, minimizing human error. Integrating the regression testing suite with ETL processes ensures that data quality checks are an inherent part of the data pipeline. This integration helps catch issues at various stages of data transformation and loading,

preventing poor quality data from reaching the final destination. It also streamlines the process by making data quality checks automatic and continuous.

Figure 2: Regression suite enabled ETL Cycle.

Continuous monitoring and alerting systems are crucial for real time detection of data discrepancies. They provide immediate notifications, allowing prompt action to rectify issues. This proactive approach helps maintain data integrity and reduces the impact of data quality issues on business operations.

Detailed reports and interactive dashboards provide clear insights into data quality status. They help stakeholders quickly identify and understand data issues, track progress over time and make informed decisions. Visualizations and dashboards make data quality metrics accessible and actionable.

Regularly updating test cases and scripts ensures that the regression testing suite remains relevant and effective as business requirements and data structures evolve. Continuous improvement helps in adapting to new data quality challenges, ensuring that the suite remains robust and comprehensive.

Training stakeholders on the importance of data quality and the regression testing suite ensures that everyone involved understands their roles and responsibilities. Involving stakeholders fosters a culture of data quality and encourages feedback, which is vital for continuous improvement and buy in.

5. Conclusion

Implementing a robust Daily Regression testing Suite is essential for maintaining high data quality within production environments. Organizations can proactively address data issues by automating test scripts, integrating with ETL processes, continuously monitoring data and generating detailed reports. The benefits of such a suite include enhanced transparency, improved decision making and efficient tracking of data quality improvements. Ensuring stakeholder involvement and planning for scalability further contribute to its effectiveness. Ultimately, a thorough regression testing suite not only supports data integrity but also empowers businesses to make more informed and reliable decisions.

6. References

1. Everett Gerald D and Raymond McLeod Jr. Software testing: testing across the entire software development life cycle. John Wiley and Sons, 2007.
2. Bashir Imran and Amrit L Goel. Testing object-oriented software: life cycle Solutions. Springer Science and Business Media, 2012.
3. Kimball Ralph and Joe Caserta. The data warehouse ETL toolkit. John Wiley and Sons, 2004.
4. Matteo Golfarelli and Stefano Rizzi. “Data Warehouse Testing,” International Journal of Data Warehousing and Mining (IJDWM), 2011;7: 26-43.
5. Sara B Dakrory, Tarek M Mahmoud and Abdelmgeid A Ali, “Automated ETL Testing on the Data Quality of a Data Warehouse,” International Journal of Computer Applications, 2015;131: 9-16.