Data Validation using ETL – A Theoretical Perspective

[1]Sanjay Kumar S, [2]Roshaan J S, [3]Surya V, [4]Srinivas S and [5]Sreemathy J
[1,2,3,4,5] Bachelor of Engineering, Sri Eshwar college of Engineering, Coimbatore, Tamil Nadu
[1]sanjaykumarcsk2002@gmail.com, [2]roshaan.j.s2019cse@sece.ac.in, [3]surya.v2019cse@sece.ac.in,
[4]srinivassooriyakumar@gmail.com, [5]sreemathy.j@sece.ac.in

Article Info
A. Haldorai et al. (eds.), *2nd International Conference on Materials Science and Sustainable Manufacturing Technology,* Advances in Computational Intelligence in Materials Science.
Doi: https://doi.org/10.53759/acims/978-9914-9946-9-8_2

Abstract - Data is more readily available than ever, but when it is erroneous or insufficient, it can be challenging to interpret. As a result, data validation is essential to improving the quality of data for sound decision-making. The authors have discussed some of the most important concepts and challenges in data validation. It is obvious that human oversight cannot be completely removed from this process. Information priceless human qualities that cannot be taught. Humans are still cautious to take action on decisions that have not been validated by another person, despite today's highly advanced data validation technology or automated approaches. A data validation dashboard can be used by an expert data practitioner to monitor the complete data analysis procedure. The dashboard may make it easier for teams or project managers to assign tasks and resources while also more efficiently monitoring the progress and success of their work. This paper offers insightful discussions of the fundamental ideas, key points, and validation procedure for data validation and quality assurance. Additionally, the article compares several data validation technologies, and several significant industry players are explored. Additionally, the main problems, difficulties, and requirements are explored.

Keywords - Data Validation, Component, ETL, Test, Data Warehouse.

I. INTRODUCTION

Due to the growing amount of data being produced, big data analytics has garnered a lot of interest from academics and industry professionals. To maximize the value of the data warehouse and the data stored inside, data validation must be done. In essence, the Extract, Transform, and Load (ETL) procedure, which entails moving the source database to the target data warehouse, includes a stage called data validation. To make sure the data they are dealing with is authentic, it enables users to undertake end-to-end testing, including testing for data accuracy, data completeness, and data quality. The validation messages aid in the improvement of reporting performance, query optimization, stuck thread reduction, and data model correction. Validated data can be used by businesses for demand planning and financial forecasts. For instance, developing and evaluating demand prediction models can increase the forecasting's accuracy.

II. DATA WAREHOUSE

The load and analytical process of turning data into useful information for top management are currently too much for data warehouse tools and technologies to handle. Business intelligence (BI) activities, notably analytics, are designed to be facilitated and supported by a specific type of data management system termed a "data warehouse". Bill Inmon initially used the phrase "Data Warehouse" in 1990.

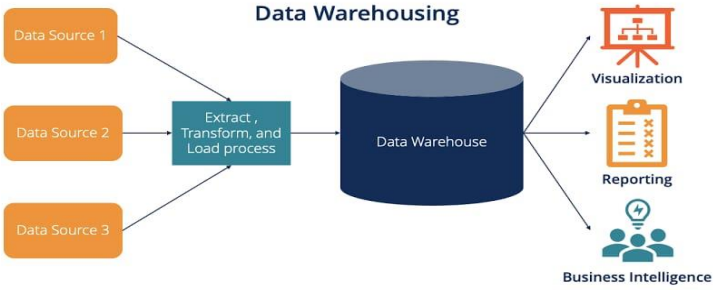

Fig 1. Data Warehouse

Inmon defines data warehouses as groups of data that are non-volatile, subject-oriented, integrated, and time-variant. Businesses store and analyze data using a combination of databases, data lakes, and data warehouses. A data warehouse serves as a central store for data that is gathered from several data sources. Several relational databases and the transactional system feed data into a data warehouse. **Fig 1** shows data warehouse.

III. DATA BASE VS DATAWAREHOUSE

Despite certain similarities, a data warehouse and a standard database are not the same thing. The primary distinction is that databases collect data for a range of transactional uses. A data warehouse is an example of an OLAP system, also referred to as an online database query response system. An OLTP system is an online database editing system, similar to an ATM. Learn more about how OLTP and OLAP differ from one another.

IV. WORKING OF ETL

Using a software called extraction-transformation-loading (ETL) technologies, data extraction, cleansing, customization, integration, reformatting, and loading into a data warehouse are all accomplished. Five steps make up the ETL process: extract, clean, transform, load, and analyze. The extract, convert, and load process steps are the three most important ones. Before being merged and converted by the ETL component and fed into the data warehouse, data is often extracted from the source systems (data that is stored in databases, flat files, web services, etc.). At data marts or streams created for reporting and analytics on top of the data warehouse, the data stream comes to a stop. The creation of the ETL process may be one of the most difficult parts of creating a warehouse due to its complexity, time-intensive nature, and utilization of the bulk of the resources, costs, and implementation time. An in-depth comprehension of the source area, destination area, and mapping area is necessary Currently, the mapping area lacks standard models, but the source and destination sections do. Two examples of source area standard models are entity relationship diagrams and star schemas. **Fig 2** shows ETL.

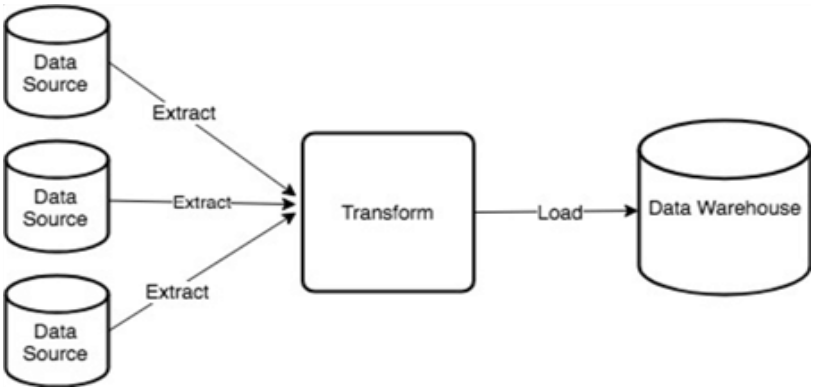

Fig 2. ETL

The top ETL testing tools are listed below:

- RightData
- Integrate.io
- iCEDQ
- BiG EVAL
- Informatica Data Validation
- QuerySurge
- Datagaps ETL Validator
- QualiDI
- Talend Open Studio for Data Integration
- Codoid's ETL Testing Services
- Data Centric Testing
- SSISTester
- TestBench
- DataQ

V. REASON FOR USE OF TALEND OPEN STUDIO OVER OTHER TOOLS

The most widely used open source ETL software is Talend's ETL tool. Instead of running the configurations of an ETL engine through a pipeline, Open Studio creates Java code for ETL pipelines. It has certain performance benefits as a result of this strategy. The most widely used open source ETL software is Talend's ETL tool. Instead of running the

configurations of an ETL engine through a pipeline, Open Studio creates Java code for ETL pipelines. It has certain performance benefits as a result of this strategy. The key characteristics of Talend open studio include: Any kind of relational database, flat files, etc. are all supported by Talend Data Integration. An integrated GUI that makes it easier to design and build ETL procedures. Over 900 components make up the built-in data connections in Talend Data Integration. It immediately recognizes business uncertainty and consistency issues in transformation rules. It enables the execution of distant jobs. It provides early defect detection to cut expenses. Based on ETL best practices. It gives both quantitative and qualitative measurements. Between the ETL development, ETL testing, and ETL production environments, context change is available. Tracking of data flow in real-time and thorough execution statistics.

VI. DATA VALIDATION

Verifying and validating data before usage is the process of validating data that has been acquired. Data validation is a necessary step in every form of data handling operation, including data collection, analysis, and presentation of the results. Since validation takes time, it can be tempting to ignore it occasionally. But doing so is crucial if you want to achieve the finest results. Organizations benefit from data validation because it eliminates problems that could arise from deteriorating data. It doesn't provide a comprehensive solution, but it helps businesses look for data that is missing, incomplete, erroneous, or inconsistent that could unintentionally alter the results for which the data is meant. A system has several built-in checks to make sure the data being entered and stored is logically consistent. The digital revolution has made data validation considerably quicker now. The majority of data integration platforms automate and incorporate the data validation process, making it part of the workflow as a whole rather than an extra step. It rarely necessitates human interaction in such automatic systems. Because bad data might lead to problems later on and because data cleansing is more expensive if done later in the process, data validation becomes crucial.

VII. DATA VALIDATION TESTS

Data Uniformity

To ensure that the entity's real value matches exactly in all locations, data uniformity tests are carried out. Two different tests are conceivable in this situation:

(i) Checks performed on the same schema: Within the same schema, the data entity could be present in two tables (either source system or target system)

(ii) Cross-schema checks: The data entity may be migrated into the target schema in its current state, meaning that both the source system and the destination system have it.

Entity Presence

In this kind of test, we must confirm that all the entities (Tables and Fields) between the source and target are compatible. According to the Data Model architecture, there are two possibilities: an entity might be present or could be absent. Verify that all of the tables (and columns) that are present in both the source and the target are identical. We perform a text comparison after pulling a list of all Tables (and columns). Only when the same entity names are used throughout does this sanity check succeed. Sometimes multiple table names are used, making it impossible to compare directly. It's possible that we'll need to map this information in the Data Mapping sheet and check it for errors.

Data Reliability Test

We validate the data's logical accuracy, as the term implies. This kind of test falls into two groups. This allows the tester to identify problems with data quality even in the source system.

(i) Non-numerical type: For this category, we check the non-numerical content's accuracy. Examples include genuine phone numbers, pin codes, and emails.

(ii) Domain analysis: For this kind of test, we choose certain data domains and check them for flaws.

Validation of Metadata

When validating metadata, we ensure that the target's table and column data type definitions are created appropriately and implemented in accordance with the requirements of the data model architecture.

Here, there are two groups:

Metadata design: The first inspection entails verifying that the data model is correctly created in accordance with the operational requirements for the target tables. When designing the destination system, data architects may move schema entities or make adjustments. Verifying that the appropriate scripts were produced utilising the data models should be the next check. In order to determine whether the tables and field definitions for each category below, we first confirm that the metadata defined for the target system satisfies the business requirement. Here are a few examples of the metadata checks:

(i) Check the data type: For instance, will the Decimal (8, 16, or 20 bytes) or Double type of Total Sales function correctly?

(ii) Data Length Check: For instance, will 500 characters be enough for the Address field's data length? It may be the case that data migration is carried out as the organisation expands into new geographies. The addresses of

the new geography might have an excessively long format, and if the length is maintained, a use case might be incorrect.

(iii) Example of an index check: Does the target system have an index for the OrderId column? What if a company merger occurred, necessitating data transfer, and the Orders table increased in size by 100 times in the target system?

(iv) Checking Metadata Across Environments: With this check, make sure the Production environment's metadata matches the QA test environments. The QA environment is one place where tests may succeed while failing in other places.

Delta change: These tests find errors that occur while a project is ongoing and changes to the metadata of the source system occur in the middle of the project but are not implemented in the target systems. Example: The Customer table in the source system received a new field called CSI (Customer Satisfaction Index), but the destination system did not receive it.

Data Completeness

Once automated, these sanity checks can be repeated often and reveal any missing record or row counts between the source and target tables.

There are two distinct test types:

(i) Record count: In this section, we contrast the number of records in matching tables in the source and destination systems. This is a short sanity check to confirm the ETL or migration job's post-running. Whenever the counts do not agree, we have a flaw. Occasionally records are discarded throughout the job run. Some of these could be true. However, we provide a case study for this as testers.

(ii) Column data profiling: When record counts are high, this kind of sanity test is useful. Here, we build logical data sets that lower the number of records before doing a source-to-target comparison. Filter all unique values in a column wherever possible; for instance, the OrderDetails table may contain numerous instances of the ProductID. Select a distinct list for ProductID from the source and target databases, then validate it. As a result, record counts are drastically decreased, and sanity tests are completed more quickly. Similar to the tests mentioned above, we can also select all the important columns and determine whether the KPIs (minimum, maximum, average, maximum or minimum length, etc.) in the source and target tables match.

Advantages

By halving the time spent handling data, Talend Open Studio lowers developer rates. It is very trustworthy and effective when dealing with large datasets. Functional mistakes also occur far less frequently than they do with manual ETL. The numerous Talend users can help the software's developers identify any mistakes that might have happened when designing the ETL method. It provides users with a selection of open-source integration solutions for no cost. Data validation makes sure that the dataset is precise, clean, and complete as well as that the data is not corrupted by eliminating data errors from all projects. Eliminating data mistakes from all projects ensures that the dataset is accurate, clean, and comprehensive as well as that the data is not destroyed. A technique or gadget is validated when there is established scientific proof that it can reliably provide high-quality results. Talend offers pre-built widgets to link with databases, web services, and FTP, for example, saving time and quality issues in constructing the fundamental functionalities required to create a complete workflow. Additionally, it promotes a consistent mentality and approach to processing stages rather than reasoning that is entirely free-form and dependent on the developer's mood. Even if you work with a variety of procedures, this makes it simpler to comprehend the workflow that is applied.

Disadvantages

When utilizing Talend, common source control procedures do not function as you may anticipate. You must take extra precautions while managing your workspace directory since the intermediary items Talend generates in the workspace are not just a collection of.java files. This means that before updating the version of Talend you are actively developing with, you must make a backup of your whole workspace directory. Even while we should do this even when upgrading a basic Eclipse setup, it is far simpler to get back to where you were if something goes wrong with Eclipse than with Talend. This is a major worry for developers, and they are reluctant to spend the time necessary to become comfortable with a single tool in order to address it. The team and enterprise versions of the program include tools to aid with this, but they also have their own adoption, financial, and team consensus challenges.

VIII. CONCLUSION

A crucial step in reducing the size of the datasets and increasing process effectiveness is data validation. But every method or procedure has advantages and disadvantages, therefore it's important to fully acknowledge both. The ideal work process can be provided by data validation, which can increase quality and accuracy. We've covered a few of the crucial elements in this article that might help you understand data validation. If an analyst adopts this strategy with the proper process, data validation can produce the best results for big data, making data management easier. The data validation process is an essential step in data and analytics processes to filter quality data and improve process efficiency.

It facilitates data handling and generates data that is trustworthy, consistent, and accurate. In order to develop validation tests that are simple to conduct and compliant with current criteria, businesses are looking into a variety of methods, including automation. In contrast, Talend Open Studio can save a lot of time. By eliminating duplicates and loading the resulting files in a cloud environment, our project assists in validating the datasets.

References

[1]. Koren and M. Jurcevic, "Concept-Level Model of Integrated Syntax and Semantic Validation for Internet of Medical Things Data," 2021 IEEE 15th International Conference on Semantic Computing (ICSC), Jan. 2021, doi: 10.1109/icsc50631.2021.000

[2]. K. Oyoo, "Automatic Data Validation and Testing for Enterprise Asset Management in the Power and Utilities industry," SoutheastCon 2021, Mar. 2021, doi: 10.1109/southeastcon45413.2021.9401912.

[3]. P. Yang et al., "Lifelogging Data Validation Model for Internet of Things Enabled Personalized Healthcare," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 48, no. 1, pp. 50–64, Jan. 2018, doi: 10.1109/tsmc.2016.2586075.

[4]. M. Kim and J. Yun, "Data Reliability Enhancement Method through Data Validation in Crowdsensing System," 2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN), Jul. 2019, doi: 10.1109/icufn.2019.8806104.

[5]. G. Zhang, "A data traceability method to improve data quality in a big data environment," 2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC), Jul. 2020, doi: 10.1109/dsc50466.2020.00051.

[6]. J. Ding, X.-H. Hu, and V. Gudivada, "A Machine Learning Based Framework for Verification and Validation of Massive Scale Image Data," IEEE Transactions on Big Data, vol. 7, no. 2, pp. 451–467, Jun. 2021, doi: 10.1109/tbdata.2017.2680460.

[7]. Y. Chung, T. Kraska, N. Polyzotis, K. H. Tae, and S. E. Whang, "Slice Finder: Automated Data Slicing for Model Validation," 2019 IEEE 35th International Conference on Data Engineering (ICDE), Apr. 2019, doi: 10.1109/icde.2019.00139.

[8]. Ibrahim, Ermatita, Saparudin, and Z. Adetya, "Analysis of weakness of data validation from social CRM," 2017 International Conference on Data and Software Engineering (ICoDSE), Nov. 2017, doi: 10.1109/icodse.2017.8285849.

[9]. J. Gao, C. Xie, and C. Tao, "Big Data Validation and Quality Assurance -- Issuses, Challenges, and Needs," 2016 IEEE Symposium on Service-Oriented System Engineering (SOSE), Mar. 2016, doi: 10.1109/sose.2016.63.

[10]. Y. Yang, S. Kang, and J. Seo, "Improved Machine Reading Comprehension Using Data Validation for Weakly Labeled Data," IEEE Access, vol. 8, pp. 5667–5677, 2020, doi: 10.1109/access.2019.2963569.