ETL TESTING IN DATAWAREHOUSING

K.VISHNU VANDANA1, V.SUJATHA2
1Assistant Professor, Brindavan Institute of Technology & Science, Kurnool
2Assistant Professor, Kottam College of Engineering, Kurnool

ABSTRACT
It has been observed that Independent Verification and Validation is gaining huge market potential and many companies are now seeing this as prospective business gain. Customers have been offered different range of products in terms of service offerings, distributed in many areas based on technology, process and solutions. ETL or data warehouse is one of the offerings which are developing rapidly and successfully. During the development of the data warehouse (DW), too much data is transformed, integrated, structured, cleansed, and grouped in a single structure. These various types of changes could lead to data corruption or data manipulation. Therefore, DW testing is a very critical stage in the DW development process. A number of attempts were made to describe how the testing process should take place in the DW environment. In this paper, I will concentrate on ETL testing techniques, the four different engagements existed in ETL Afterwards, I will highlight the weakness points that exist in the available DW testing approaches.

Keywords: Data Warehouse, ETL Testing, Data Migration.

1 INTRODUCTION
“Data Warehouse is relational database which is subject oriented, integrated, time-variant and non-volatile collection of data used to support strategic decision process.” Exhaustive testing [2] of a Data Warehouse during its design and on an ongoing basis (for the incremental activities) comprises Data Warehouse testing. This type of testing is getting common today. The reasons for this are manifold, prominent ones being increase in Enterprise Mergers & Acquisitions, Data Center Migrations, Increased Compliance regulations, Senior Management’s increased focus on data and data driven decision makings. Data Storage has become cheaper and easier compared to the file-storage days.[1] Hardware has become cheaper and Data Mining solutions are easily available. Data driven decision has proved to be more accurate. In this context testing the data warehouse implementations are also of utmost significance. Organization decisions depend entirely on the Enterprise data and the data has to be of utmost quality. Complex Business rules and transformation logic implementations mandates a diligent and thorough testing.

Why a Data Warehouse?
A Data warehouse is a composite and collaborated data model that capture[3]s the entire data of an organization. It brings together data from heterogeneous sources into one single destination. It is not just bringing together .Data is Extracted, Transformed and loaded(ETL) into the Data warehouse. This processing of the data is usually done in what is known as “Staging area”. The data need not be normalized, as it will be mainly for read-only purposes or basically querying and analytical purposes.

Why Data warehouses?
Organizations with organized IT practices are looking forward to create a next level of technology transformation. They are now trying to make themselves much more operational with easy-to-interoperate data. Having said that data is most important part of any organization, it may be everyday data or historical data. Data is backbone of any report and reports are the baseline on which all the vital management decisions are taken. Most of the companies are taking a step forward for constructing their data warehouse to store and monitor real time data as well as historical data. Crafting an efficient data warehouse is not an easy job. Many organizations have distributed departments with different applications running on distributed technology. ETL tool is employed in order to make a flawless integration between different data sources from different departments. ETL tool will work as an integrator, extracting data from different sources; transforming it in preferred format based on the business transformation rules and loading it in cohesive DB known are Data Warehouse.

Architecture of Data warehousing
Listed below are the some of the main reasons why organizations go in for Data warehouse project. Business Mandated: Mergers and Acquisitions are very common in today’s business arena. Every day, there are talks going on between retail giants in huge number of domains. Is it simple Extraction, Transformation and Loading of voluminous data? There is a huge amount of organizational data movement, so ETL testing, to ensure if the data movement and the data transformations is a key to ensure a seamless merger/acquisition. Data Mart: Data Mart can be described as a specialized subset of the Data warehouse. It caters to the data requirements of a specific group like Finance, HR and Marketing. It contains data which assimilated, processed and is in decipherable format for the end-user. Dependent data marts are those that derive the data from the underlying Enterprise data marts.
**Decision Support:** In any Organization for Policy Making senior Management relies on Decision Support systems. Hence the Decision Making support systems should be very intelligent to monitor the ideas and conclusions. If there is any bug in the data model the data warehousing implementation or the ETL process should be capable of translating into disastrous decision by the organization.

![Data Warehouse Architecture](image)

**Figure 1** Architecture of Data Warehouse

**2 RELATED WORK**

**Data warehouse testing**

A good understanding of data modeling and the need for the data warehouse will equip the test analyst with guidelines for coming up with an apt testing strategy. Hence, it is very important that during the Requirement Analysis phase, importance must be given in understanding the Data warehouse implementation to the maximum possible extent\[10\]. Data warehouse testing strategies will, in most case be a consortium of several smaller strategies. This is due to the nature of Data warehouse testing the test execution does not start at the end of implementation. In short, test execution itself has multiple phases and is staggered throughout the life cycle of the Data warehouse implementation.

Listed below are the main phases of Data warehouse testing and the different types of testing required in each of the phases\[7\].

![Data Warehouse Testing Types](image)

**Figure 2** Data Warehouse Testing Types

As depicted in the figure above, different types of testing are required throughout the lifecycle of a Data warehouse implementation.

During ETL phase of a Data warehouse implementation, Data quality testing is of utmost importance. Any defect slippage in this phase will be very costly to rectify later. Functional testing needs to be carried out to validate the Transformation logic.\[3\]

**3 ETL TESTING**

Building a data warehouse is a major undertaking that's expected to yield substantial business benefits in order to justify the cost and effort\[4\]. To ensure that your ETL data warehouse project lives up to expectations, ETL testing must be front and center, particularly in the all-important early stages of the project. ETL testing seeks to ensure that transfer of data from heterogeneous sources to the central warehouse occurs without data loss, with strict adherence to transformation rules, and in compliance with all validity checks. Effective ETL testing should detect early on any problems with the source data as well as any inconsistencies or ambiguities in the business rules that are intended to guide data transformation and integration. ETL testing should continue as production gets underway, to ensure that the extraction, transformation, and loading processes are going as planned and that the result is warehoused data that's consistent, accurate, and complete.

ETL stands for Extract, Transform and Load.
Extract - It can be defined as extracting the data from numerous heterogeneous systems.

Transform - Applying the business logics as specified by the business on the data derived from sources. The transform logic also includes the following:
- Cleaning (e.g., mapping NULL to 0 or "Male" to "M" and "Female" to "F" etc.)
- Filtering (e.g., selecting only certain columns to load)
- Splitting a column into multiple columns and vice versa
- Joining together data from multiple sources (e.g., lookup, merge)
- Transposing rows and columns

Load - Loading the data into the final warehouse after completing the above two process.

Very well planned and defined only guarantees smooth conversion of the project to the production. If and only if independent group of experts validates and verifies the processes, the Data warehouse can be said concrete and robust.

ETL or Data warehouse testing is categorized into four different engagements irrespective of technology or ETL tools used:

New Data Warehouse Testing – New DW is built and verified from scratch. Data input is taken from customer requirements and different data sources and new data warehouse is build and verified with the help of ETL tools.

Migration Testing – In this type of project customer will have an existing DW and ETL performing the job but they are looking to bag new tool in order to improve efficiency.

Change Request – In this type of project new data is added from different sources to an existing DW[9]. Also, there might be a condition where customer needs to change their existing business rule or they might integrate the new rule.

Report Testing – Report are the end result of any Data Warehouse and the basic propose for which DW is build. Report must be tested by validating layout, data in the report and calculation.

Some Levels of testing are outlined below, but this should not be considered as the exhaustive list of testing activities[6].

Requirements Testing
- Are the requirements complete?
- Are the requirements testable?
- Are the requirements clear (is there any ambiguity)?

Data Validation Testing
- Compare record counts between data sources
- Ensure that the ETL application properly rejects, replaces with default values and reports invalid data
- Verify that data is transformed correctly according to system requirements and business rules
- Compare unique values of key fields between source data and warehouse data
- Ensure that all projected data is loaded into the data warehouse without any data loss or truncation
- Test the boundaries of each field to find any database limitations

Integration Testing
- Verify the sequence and outcome of ETL batch jobs
- Verify that ETL processes function with upstream and downstream processes
- Verify the initial load of records on data warehouse
- Verify any incremental loading of records at a later date for newly inserted or updated data
- Test the rejected records that fail ETL rules
- Test error log generation

Report Testing
- Verify report data with the data source
- Create SQL queries to verify source/target data
- Verify field-level data
- User Acceptance Testing
- Verify that the business rules have been met
- Confirm that the system is acceptable to the client

Performance Testing
- Verify that data loads and queries are executed within anticipated timeframes
- Verify that maximum anticipated volume of data is loaded within an acceptable timeframe
- Verify load times with various amounts of data to predict scalability

Regression Testing
Ensure that current functionality stays intact whenever new code is released.
4 ETL TESTING: Challenges
ETL testing is quite different from conventional testing. There are many challenges we faced while performing data warehouse testing. Here is the list of few ETL testing challenges[9]

- Incompatible and duplicate data.
- Loss of data during ETL process.
- Unavailability of inclusive test bed.
- Testers have no privileges to execute ETL jobs by their own.
- Volume and complexity of data is very huge.
- Fault in business process and procedures.
- Trouble acquiring and building test data.
- Missing business flow information.

CONCLUSIONS
Even though data warehouses are used since the early 1990s, the design and modeling of ETL processes is still accomplished in a vendor-dependent way by using tools that allow to specify them according to the schemas represented in concrete platforms. Therefore, their design, implementation and even more their maintenance must be done depending on the target platform. In this paper, we have provided, to the best of our knowledge, the first approach for the specification of ETL processes in a vendor-independent way and the automatic generation of its corresponding code in commercial platforms. Clearly, the importance, complexity and criticality of such an near real time are housing a significant topic of research and practice; therefore, we conclude with the hope that the above mentioned issues will be addressed in a principled manner in the future by both the industry and the academia.

7. REFERENCES