ISSN:0975-9646

ETL Life Cycle

Purnima Bindal , Purnima Khurana

Abstract As the data warehouse is a living IT system, sources and targets might change. Those changes must be maintained and tracked through the lifespan of the system without overwriting or deleting the old information. We need to load data warehouse regularly so that it can serve its purpose of facilitating business analysis and keep updated. The process of extracting data from source systems and bringing it into the data warehouse is commonly called ETL, which stands for extraction, transformation, and loading. The intention of this survey is to present the research work in the field of ETL technology in a structured way. ETL process is described in detail in the paper.

Keywords: extract, transform and loading, data warehouse, online analytical processing, online transaction protocol

1. INTRODUCTION:

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.

1. 1 why do organizations need Data Warehouse?

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.

Well planned, well defined and effective testing scope guarantees smooth conversion of the project to the production. A business gains the real buoyancy once the ETL processes are verified and validated by independent group of experts to make sure that data warehouse is concrete and robust.

1.2 ETL stands for Extract, Transform and Load

Definition - A process is used to enable companies to move data from multiple sources, reformat and cleanse it, and then load the data into another area for analysis or operational system for support of the organizations business process.

a) **Extracts** data from homogeneous or heterogeneous data sources
b) **Transforms** the data for storing it in proper format or structure for querying and analysis purpose
c) **Loads** it into the final target (database, more specifically, **operational data store, data mart, or data warehouse)**[3]

1.3 ETL Testing is categorized into four different Engagements

ETL testing is categorized in the following four types:

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

1. ETL TESTING TECHNIQUES:

- Verify that data is transformed correctly according to various business requirements and rules.
- Make sure that all projected data is loaded into the data warehouse without any data loss and truncation.
- Make sure that ETL application appropriately rejects, replaces with default values and reports invalid data
- Make sure that data is loaded in data warehouse within prescribed and expected time frames to confirm improved performance and scalability.

Apart from these four main ETL testing methods other testing methods like integration testing and user acceptance testing is also carried out to make sure everything is smooth and reliable.

2. ETL TESTING PROCESS:

Similar to any other testing that lies under Independent Verification and Validation, ETL also go through the same phase. Business and Requirement Understanding.

- Validating.
- Test Estimation.
- Test Planning based on the inputs from Test Estimation and Business Requirements.
- Designing Test Cases and Test Scenarios from all the available inputs.
- Once all the Test Cases are ready and are approved, testing team proceed to perform pre-execution check and Test Data preparation for Testing.
- Lastly Execution is performed till Exit Criteria is met.
- Upon successful completion summary report is prepared and closure process is done.

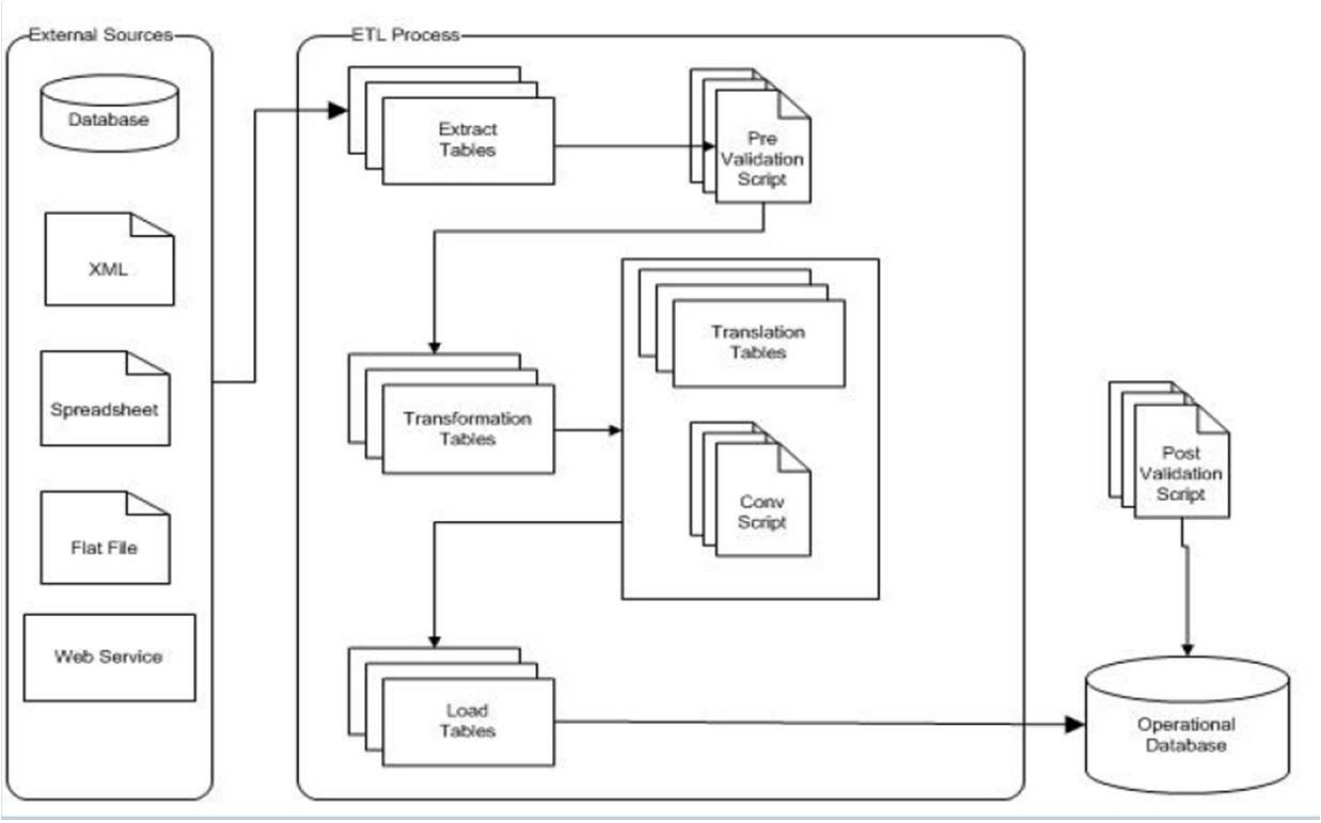
It is necessary to define test strategy which should be mutually accepted by stakeholders before starting actual testing. A well-defined test strategy will make sure that correct approach has been followed meeting the testing aspiration. ETL testing might require writing SQL statements extensively by testing team or may be tailoring the SQL provided by development team. In any case testing team must be aware of the results they are trying to get using those SQL statements.

3. HOW IS TESTING OF DATA WAREHOUSE DIFFERENT FROM TESTING OF TRANSACTIONAL SYSTEMS?

There are several differences when it comes to testing of Data Warehouse (OLAP) and Transactional (OLTP) applications:

- The focus of OLTP application testing is on software code while OLAP application testing is directed at the validation of the correctness of data
- The volume of data involved in OLAP application testing is typically very large when compared to volume of data involved in the testing of OLTP applications
- Data integration projects present different set of challenges for testing of full and incremental loads
- Performance testing of data integration projects presents different set of challenges including the need for large volumes of test data when compared to OLTP applications
- The number of use cases for OLTP applications are finite while the test scenarios for regression and performance testing of OLAP applications can be virtually unlimited[2].

Diagram: ETL Process Overview



4. DATABASE VS. DATA WAREHOUSE TESTING

There is a popular misunderstanding that database testing and data warehouse is similar while the fact is that both hold different direction in testing.

- Database testing is done using smaller scale of data normally with OLTP (Online transaction processing) type of databases while data warehouse testing is done with large volume with data involving OLAP (online analytical processing) databases.
- In database testing normally data is consistently injected from uniform sources while in data warehouse testing most of the data comes from different kind of data sources which are sequentially inconsistent.
- We generally perform only CRUD (Create, read, update and delete) operation in database testing while in data warehouse testing we use read-only (Select) operation.
- Normalized databases are used in DB testing while denormalized DB is used in data warehouse testing.
- There are number of universal verifications that have to be carried out for any kind of data warehouse testing. Below is the list of objects that are treated as essential for validation in ETL testing:

- Verify that expected data is added in target system.
- Verify that data transformation from source to destination works as expected.
- Verify that all DB fields and field data is loaded without any truncation.
- Verify data checksum for record count match.
- Verify that for rejected data proper error logs are generated with all details.
- Verify NULL value fields.
- Verify that duplicate data is not loaded.
- Verify data integrity.

6 ETL LIFE CYCLE

6.1 Initiation Create a conversion working area to use for the transformation process

6.2 **Extract**

- Data is extracted from heterogeneous data sources
- Each data source has its distinct set of characteristics that need to be managed and integrated into the ETL system in order to effectively extract data.
- ETL process needs to effectively integrate systems that have different:
- DBMS
- Operating Systems
- Hardware
- Communication protocols
- Need to have a logical data map before the physical data can be transformed
- The logical data map describes the relationship between the extreme starting points and the extreme ending points of your ETL system usually presented in a table or spreadsheet.
- The content of the logical data mapping document has been proven to be the critical element required to efficiently plan ETL prcesses
- The analysis of the source system is usually broken into two major phases: The data discovery phase and The content analysis phase

6.2.1 The Data Discovery Phase

It is up to the ETL team to drill down further into the data requirements to determine each and every source system, table, and attribute required to load the data warehouse.

- Collecting and Documenting Source Systems
- Keeping trackof source systems
- Determining the System of Record -Point of originating of data
- Definition of the system-of-record is important because in most enterprises
- data is stored redundantlyacross many different systems.
- Enterprises do this to make nonintegrated systems share data. It is very common that the same piece of data is copied, moved, manipulated, transformed, altered, cleansed, or made corrupt throughout the enterprise, resulting in varying versions of the *same* data

6.2.2 The Content Analysis Phase

Understanding the content of the data is crucial for determining the best approach for retrieval-

- **NULL values :** An unhandled NULL value can destroy any ETL process. NULL values pose the biggest risk when they are in foreign key columns. Joining two or more tables based on a column that contains NULL values will cause data loss! Remember, in a relational database NULL is not equal to NULL. That is why those joins fail. Check for NULL values in every foreign key in the source database. When NULL values are present, you must *outer* join the tables-
- **Dates in nondate fields.**Dates are very peculiar elements because they are the only logical elements that can come in various formats, literally containing different values and having the exact same meaning. Fortunately, most database systems support most of the various formats for display purposes but store them in a single standard format
- During the initial load, capturing changes to data content in the source data is unimportant because you are most likely extracting the entire data souce or a potion of it from a predetermined point in time.
- Later the ability to capture data changes in the source system instantly becomes priority
- The ETL team is responsible for capturing data-content changes during the incremental load.

6.3 Transformation

- Main step where the ETL adds value
- Actually changes data and provides guidance whether data can be used for its intended purposes
- Performed in staging area

Data Quality paradigm

- Correct
- Unambiguous
- Consistent
- Complete
- Data quality checks are run at 2 places -after extraction and after cleaning and confirming additional check are run at this point

Cleaning Data

- **Anomaly Detection Data sampling** –count(*) of the rows for a department column
- Column Property Enforcement
  - Null Values in reqd columns
  - Numeric values that fall outside of expected high and lows
  - Cols whose lengths are exceptionally short/long
  - Cols with certain values outside of discrete valid value sets
  - Adherence to a reqd pattern/ member of a set of pattern

Confirming

- Structure Enforcement
  - Tables have proper primary and foreign keys
  - Obey referential integrity
- Data and Rule value enforcement
  - Simple business rules
  - Logical data checks

6.4 Load

Load two things :

Loading Dimensions

Loading Facts

6.4.1 **Loading Dimensions**

- Physically built to have the minimal sets of components

•The primary key is a single field containing meaningless unique integer –Surrogate Keys

•The DW owns these keys and never allows any other entity to assign them

•De-normalized flat tables –all attributes in a dimension must take on a single value in the presence of a dimension primary key.

- Should possess one or more other fields that compose the natural key of the dimension
- The data loading module consists of all the steps required to administer slowly changing dimensions (SCD)and write the dimension to disk as a physical table in the proper dimensional format with correct primary keys, correct natural keys, and final descriptive attributes.
- Creating and assigning the surrogate keysoccur in this module.
- The table is definitely staged, since it is the object to be loaded into the presentation system of the data warehouse.

6.4.2 Loading Facts

- Fact tables hold the measurements of an enterprise. The relationship between fact tables and measurements is extremely simple. If a measurement exists, it can be modeled as a fact table row. If a fact table row exists, it is a measurement.
- When building a fact table, the final ETL step is converting the natural keys in the new input records into the correct, contemporary surrogate keys
- ETL maintains a special surrogate key lookup table for each dimension. This table is updated whenever a new dimension entity is created and whenever a change occurs on an existing dimension entity
- All of the required lookup tables should be pinned in memory so that they can be randomly accessed as each incoming fact recordpresents its natural keys. This is one of the reasons for making the lookup tables separate from the original data warehouse dimension tables.
- Managing Indexes :
  - Performance Killers at load time
  - Drop all indexes in pre-load time
  - Segregate Updates from inserts
  - Load updates
  - Rebuild indexes
- Managing Partitions
  - Partitions allow a table (and its indexes) to be physically divided into *mini tables* for administrative purposes and to improve query performance
  - The most common partitioning strategy on fact tables is to partition the table by the date key. Because the date dimension is preloaded and static, you know exactly what the surrogate keys are
  - Need to partition the fact table on the key that joins to the date dimension for the optimizer to recognize the constraint.
  - The ETL team must be advised of any table partitions that need to be maintained.
- **Outwitting the Rollback Log**
  - The rollback log, also known as the redo log, is invaluable in transaction (OLTP) systems. But in a data warehouse environment where all transactions are managed by the ETL process, the rollback log is a superfluous feature that must be dealt with to achieve optimal load performance. Reasons why the data warehouse does not need rollback logging are:

- All data is entered by a managed process—the ETL system.
- Data is loaded in bulk.

- Data can easily be reloaded if a load process fails.
- Each database management system has different logging features and manages its roll back log differently[4]

7. ETL TESTING CHALLENGES

ETL testing is quite different from Conventional Testing. There are many challenges faced while performing Data Warehouse Testing. Here is the list of few ETL testing challenges:

- 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.
- Faults in Business Process and Procedures.
- Trouble Acquiring and Building Test Data.
- Missing Business Flow Information.

8. ETL TESTING TOOLS AND VENDORS

Tool	Vendor
Oracle Warehouse Builder (OWB)	Oracle
Data Integrator (BODI)	Business Objects
IBM Information Server (Ascential)	IBM
SAS Data Integration Studio	SAS Institute
PowerCenter	Informatica
Oracle Data Integrator (Sunopsis)	Oracle
Data Migrator	Information Builders
Integration Services	Microsoft
Talend Open Studio	Talend
DataFlow	Group 1 Software (Sagent)
Data Integrator	Pervasive
Transformation Server	DataMirror
Transformation Manager	ETL Solutions Ltd.
Data Manager	Cognos
DT/Studio	Embarcadero Technologies
ETL4ALL	IKAN
DB2 Warehouse Edition	IBM
Jitterbit	Jitterbit
Pentaho Data Integration	Pentaho

9. CONCLUSION

Data is important for businesses to make the critical Business Decisions.ETL testing plays a significant role in Validating and Ensuring that the Business Information is Exact, Consistent and Reliable. Also, it minimizes hazard of data loss in production.

REFERENCES

[1] http://www.softwaretestinghelp.com/etl-testing-data-warehouse-testing/
[2] http://www.datagaps.com/etl-testing-challenges
[3] http://en.wikipedia.org/wiki/Extract,_transform,_load
[4] http://www.srmuniv.ac.in/sites/default/files/files/ETL.pdf
[5] http://www.ijarcsse.com/docs/papers/Volume_3/6_June2013/V3I6-0145.pdf
[6] http://docs.oracle.com/cd/B19306_01/server.102/b14223/ettover.htm
[7] http://www.webopedia.com/TERM/E/ETL.html
[8] http://omicsonline.com/open-access/how-is-extraction-important-in-etl-process-2277-1891.1000122.pdf