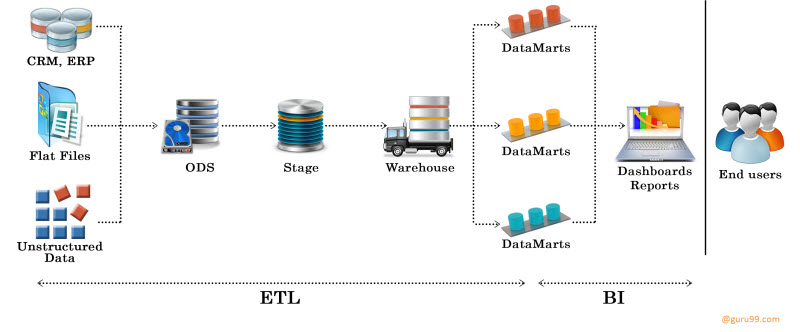
### What is BI?

Business Intelligence is the process of collecting raw data or business data and turning it into information that is useful and more meaningful.  The raw data is the records of the daily transaction of an organization such as interactions with customers, administration of finance, and management of employee and so on.  These data’s will be used for “Reporting, Analysis, Data mining, Data quality and Interpretation, Predictive Analysis”.

### What is Data Warehouse?

A data warehouse is a database that is designed for query and analysis rather than for transaction processing. The data warehouse is constructed by integrating the data from multiple heterogeneous sources.It enables the company or organization to consolidate data from several sources and separates analysis workload from transaction workload.  Data is turned into high quality information to meet all enterprise reporting requirements for all levels of users.

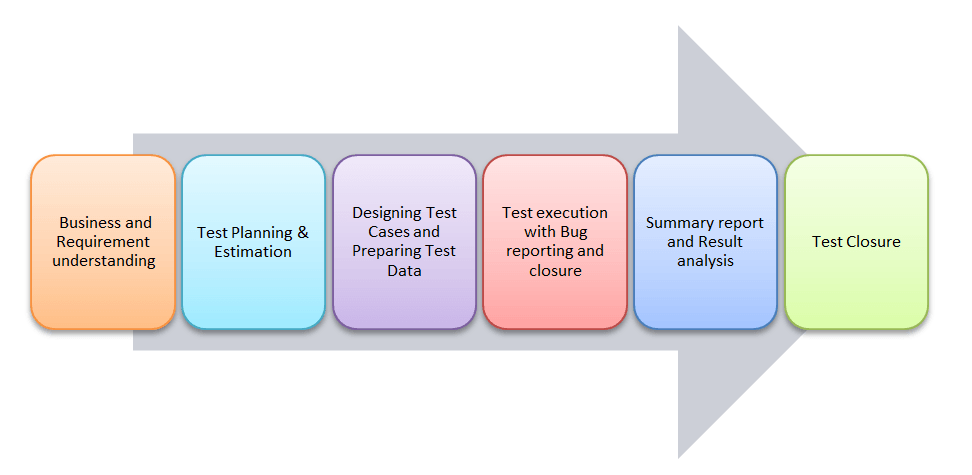


### What is ETL?

ETL stands for Extract-Transform-Load and it is a process of how data is loaded from the source system to the data warehouse.  Data is extracted from an OLTP database, transformed to match the data warehouse schema and loaded into the data warehouse database.  Many data warehouses also incorporate data from non-OLTP systems such as text files, legacy systems and spreadsheets.

### ETL Testing Process

Similar to other Testing Process, ETL also go through different phases. The different phases of ETL testing process is as follows

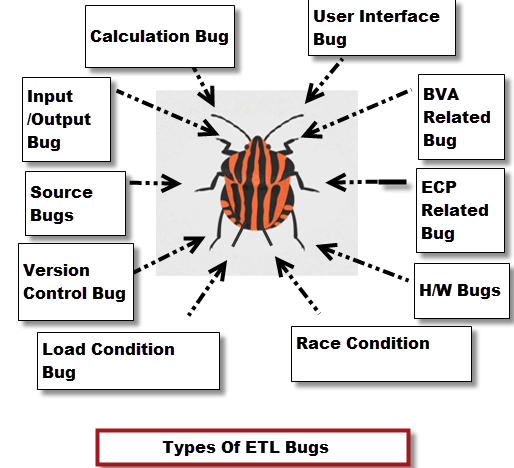


ETL testing is performed in five stages

1. Identifying data sources and requirements
2. Data acquisition
3. Implement business logics and dimensional Modelling
4. Build and populate data
5. Build Reports

|  |  |
| --- | --- |
| **Test Scenario** | **Test Cases** |
| Mapping doc validation | Verify mapping doc whether corresponding ETL information is provided or not.  Change log should maintain in every mapping doc. |
| Validation | 1. Validate the source and target table structure against corresponding mapping doc. 2. Source data type and target data type should be same 3. Length of data types in both source and target should be equal 4. Verify that data field types and formats are specified 5. Source data type length should not less than the target data type length 6. Validate the name of columns in the table against mapping doc. |
| Constraint Validation | Ensure the constraints are defined for specific table as expected |
| Data consistency issues | 1. The data type and length for a particular attribute may vary in files or tables though the semantic definition is the same. 2. Misuse of integrity constraints |
| Completeness Issues | 1. Ensure that all expected data is loaded into target table. 2. Compare record counts between source and target. 3. Check for any rejected records 4. Check data should not be truncated in the column of target tables 5. Check boundary value analysis 6. Compares unique values of key fields between data loaded to WH and source data |
| Correctness Issues | 1. Data that is misspelled or inaccurately recorded 2. Null, non-unique or out of range data |
| Transformation | Transformation |
| Data Quality | 1. Number check: Need to number check and validate it 2. Date Check: They have to follow date format and it should be same across all records 3. Precision Check 4. Data check 5. Null check |
| Null Validate | Verify the null values, where “Not Null” specified for a specific column. |
| Duplicate Check | 1. Needs to validate the unique key, primary key and any other column should be unique as per the business requirements are having any duplicate rows 2. Check if any duplicate values exist in any column which is extracting from multiple columns in source and combining into one column 3. As per the client requirements, needs to be ensure that no duplicates in combination of multiple columns within target only |
| Date Validation | Date values are using many areas in ETL development for   1. To know the row creation date 2. Identify active records as per the ETL development perspective 3. Identify active records as per the business requirements perspective 4. Sometimes based on the date values the updates and inserts are generated. |
| Complete Data Validation | 1. To validate the complete data set in source and target table minus a query in a best solution 2. We need to source minus target and target minus source 3. If minus query returns any value those should be considered as mismatching rows 4. Needs to matching rows among source and target using intersect statement 5. The count returned by intersect should match with individual counts of source and target tables 6. If minus query returns of rows and count intersect is less than source count or target table then we can consider as duplicate rows are existed. |
| Data Cleanness | Unnecessary columns should be deleted before loading into the staging area. |

### Types of ETL Bugs



|  |  |
| --- | --- |
| **Type of Bugs** | **Description** |
| User interface bugs/cosmetic bugs | * Related to GUI of application * Font style, font size, colors, alignment, spelling mistakes, navigation and so on |
| Boundary Value Analysis (BVA) related bug | * Minimum and maximum values |
| Equivalence Class Partitioning (ECP) related bug | * Valid and invalid type |
| Input/Output bugs | * Valid values not accepted * Invalid values accepted |
| Calculation bugs | * Mathematical errors * Final output is wrong |
| Load Condition bugs | * Does not allows multiple users * Does not allows customer expected load |
| Race Condition bugs | * System crash & hang * System cannot run client platforms |
| Version control bugs | * No logo matching * No version information available * This occurs usually in regression testing |
| H/W bugs | * Device is not responding to the application |
| Help Source bugs | * Mistakes in help documents |

### Difference between Database testing and ETL testing

|  |  |
| --- | --- |
| **ETL Testing** | **Data Base Testing** |
| Verifies whether data is moved as expected | The primary goal is to check if the data is following the rules/ standards defined in the Data Model |
| Verifies whether counts in the source and target are matching Verifies whether the data transformed is as per expectation | Verify that there are no orphan records and foreign-primary key relations are maintained |
| Verifies that the foreign primary key relations are preserved during the ETL | Verifies that there are no redundant tables and database is optimally normalized |
| Verifies for duplication in loaded data | Verify if data is missing in columns where required |

### Responsibilities of an ETL tester

Key responsibilities of an ETL tester are segregated into three categories

* Stage table/ SFS or MFS
* Business transformation logic applied
* Target table loading from stage file or table after applying atransformation.

Some of the responsibilities of an ETL tester are

* Test ETL software
* Test components of  ETL datawarehouse
* Execute backend data-driven test
* Create, design and execute test cases, test plans and test harness
* Identify the problem and provide solutions for potential issues
* Approve requirements and design specifications
* Data transfers and Test flat file
* Writing SQL queries3 for various scenarios like count test

### ETL Performance Testing and Tuning

ETL performance testingis a confirmation test to ensure that an ETL system can handle the load of multiple users and transactions.  The goal of performance tuning is to optimize session performance by eliminating performance bottlenecks. To tune or improve the performance of the session, you have to identify performance bottlenecks and eliminate it. Performance bottlenecks can be found in source and target databases, the mapping, the session and the system. One of the best tools used for performance testing is Informatica.

### Automation of ETL Testing

The general methodology of ETL testing is to use SQL scripting or do “eyeballing” of data.. These approaches to ETL testing are time-consuming, error-prone and seldom provide complete test coverage. To accelerate, improve coverage, reduce costs, improve defect detection ration of ETL testing in production and development environments, automation is the need of the hour. One such tool is Informatica.

### Best Practices for ETL Testing

1. Make sure data is transformed correctly
2. Without any data loss and truncation projected data should be loaded into the data warehouse
3. Ensure that ETL application appropriately rejects and replaces with default values and reports invalid data
4. Need to ensure that the data loaded in data warehouse within prescribed and expected time frames to confirm scalability and performance
5. All methods should have appropriate unit tests regardless of visibility
6. To measure their effectiveness all unit tests should use appropriate coverage techniques
7. Strive for one assertion per test case
8. Create unit tests that target exceptions