**The ETL Testing Process Checklist**

**Source Data:**

The nature of a Data Warehouse project is that data is collected from various sources and also in various formats. So one of the most essential things is to define strict contracts for the interfaces thru which the data is collected.

* Where does the data come from?
* Which data formats are used?
* How is the data formatted?
* Which Information is necessary to handle the data correctly?
* How are different data types formatted? (number, date, time, ...)

Content

* Is there one record in the source for one observation fact? (for example one row of diagnosis contains primary and secondary diagnosis)
* How are deleted records marked?
* Is it a complete set of data? (e.g. all observations or just a set of changes)
* Are trailing whitespaces part of the data?

Structure

* Data source (data files in directory or on a share, database tables or views, web services, …)
* Data format
* Comma separated file, EBCDIC encoded files, etc.

Encoding

* How are the files encoded (ASCII, UTF-8, ISO 8859-1, …)
* Destination Database (is important for dates)
* Language settings
* Date Formats
* Number Formats

*ETL* is a process of *extracting* data from source system(s), *transformed* into a format that can be

**ETL Vs ETL Testing :**

*ETL* is a process of *extracting* data from source system(s), *transformed* into a format that can be analyzed and finally *loaded* into a data warehouse.

ETL Testing is a process of checking whether the ETL process has been successfully performed or not as expected. It follows a data centric approach to validate that the data has been transformed and loaded into the target as expected.

**ETL Testing Techniques:**

Passage of data from extraction to loading could result in system errors which would result in poor information transferred across the organization. ETL testing ensures that such errors do not occur, and eliminates the bugs/errors based on the following strategies:

1. **Data Completeness:** ensures that all the data from the source are loaded into the target destination.
2. **Data Correctness:** ensures that all the data is accurately transformed and loaded from the source to the target destination.
3. **Performance Testing**: ensures if an ETL system can handle an expected load of multiple users and transactions.
4. **Metadata Testing:** checks whether data retains its integrity up to the metadata level, It involves validating the source and the target table structure w.r.t. the mapping document. The mapping document contains all the fields that define the structure of tables in the source and the target systems like length, data type, index etc.
5. **Syntax Testing:** checks for poor data due to invalid characters and incorrect character cases.
6. **Data Validation:** checks whether the values of the data post-transformation are the same as their expected values with respect to the source values.

**ETL Testing Process :**

ETL Testing process consists of 4 steps namely, Test Planning, Test Design, Execution and Test Closure. Let’s have a look on each step one-by-one:

1. **Test Planning:** This step is based on the business requirement. We look out for dependencies, risks and mitigation plans in this phase.
2. **Test Design:** Draft a test scenarios from all the available inputs. Also design mapping document and SQL scripts.
3. **Execution:** Execution is performed till exit criteria are met. It includes running ETL jobs, monitoring job runs, SQL script execution, defect logging, defect retesting and regression testing.
4. **Test Closure:** In this step a summary report is prepared and closure process is done. In this phase sign off is given to the next phase to promote the code.

**Checklist for ETL Testing:**

Following are the points keep in mind while performing ETL testing:

1. **Data transformation from source to destination:** The primary goal of ETL Testing is to ensure that data is moved from source to destination as expected so that it can be used to validate the accuracy of reports.

**2. Ensure that all DB fields and field data is loaded without any truncation.**

**3. Check for record count match:** Consider that Source file contains 200 records, Out of which 180 was moved to staging table using business filters. Once the development team loads the data in the target table. There can be 3 possibilities:

* If the target tables have 180 records, it means no records are rejected.
* If the target table have say 160 records, then 20 records should have been rejected and should be available in reject file/Error Table.
* If the target tables have say 185 records, which is fundamentally incorrect.

**4. Rejected data proper error logs:** Check for the Data Types/Size of the Columns that are available in the Target table to avoid unnecessary rejections due to size constraints.

**5. Duplicate data is not loaded:** Ensure that the unique key, primary key and any other column should be unique as per the i2b2/tranSMART schema.

**6. Maintaining data integrity:** Ensure all the fields that define the structure of tables in the source and the target systems like length, data type, index etc. must be same.