**ETL Testing**

Everything you need to know about ETL Testing to be successful in improving your data quality.

**What does ETL stand for?**

ETL = Extract, Transform, Load

​“Extract, Transform, Load (ETL) is the general procedure of copying data from one or more data sources into a destination\target system which represents the data differently from the source(s). The ETL process is often used in data warehousing.

* ***Data extraction*** involves extracting data from homogeneous or heterogeneous sources
* ***Data transformation*** processes data by cleaning and transforming them into a proper storage format/structure for the purposes of querying and analysis
* ***Data loading*** describes the insertion of data into the target data store, data mart, data lake or data warehouse.

A properly designed ETL system extracts data from the source systems, enforces data quality and consistency standards, conforms data so that separate sources can be used together, and finally delivers data in a presentation-ready format.“

A hand pointing at a diagram

Description automatically generated

### Why perform an ETL?

To load a data warehouse or data mart regularly (daily/weekly) so that it can serve its purpose of facilitating business analysis. Or move data from files, xml or other sources to a big data lake, data warehouse or data mart.

**What is a Data Warehouse?**

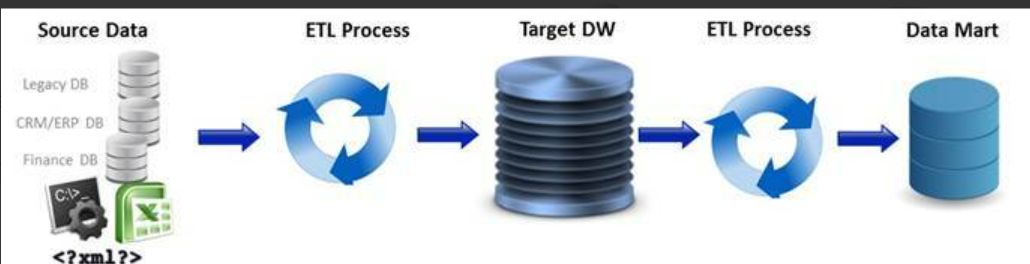
A data warehouse is:

* Typically a relational database that is designed for query and analysis rather than for transaction processing
* A place where historical data is stored for archival, analysis and security purposes.
* contains either raw data or formatted data
* combines data from multiple sources (i.e. Sales data, salaries, operational data , human resource data, inventory data, web logs, social networks, Internet text and docs, other)

**Data Mart**

A data mart is a subset of a data warehouse that has the same characteristics but is usually smaller and is focused on the data for one division or one workgroup within an enterprise.

* A data mart is a subset of a data warehouse focused on a particular line of business, department or subject area.
* Data marts can improve team efficiency, reduce costs and facilitate smarter tactical business decision-making in enterprises.



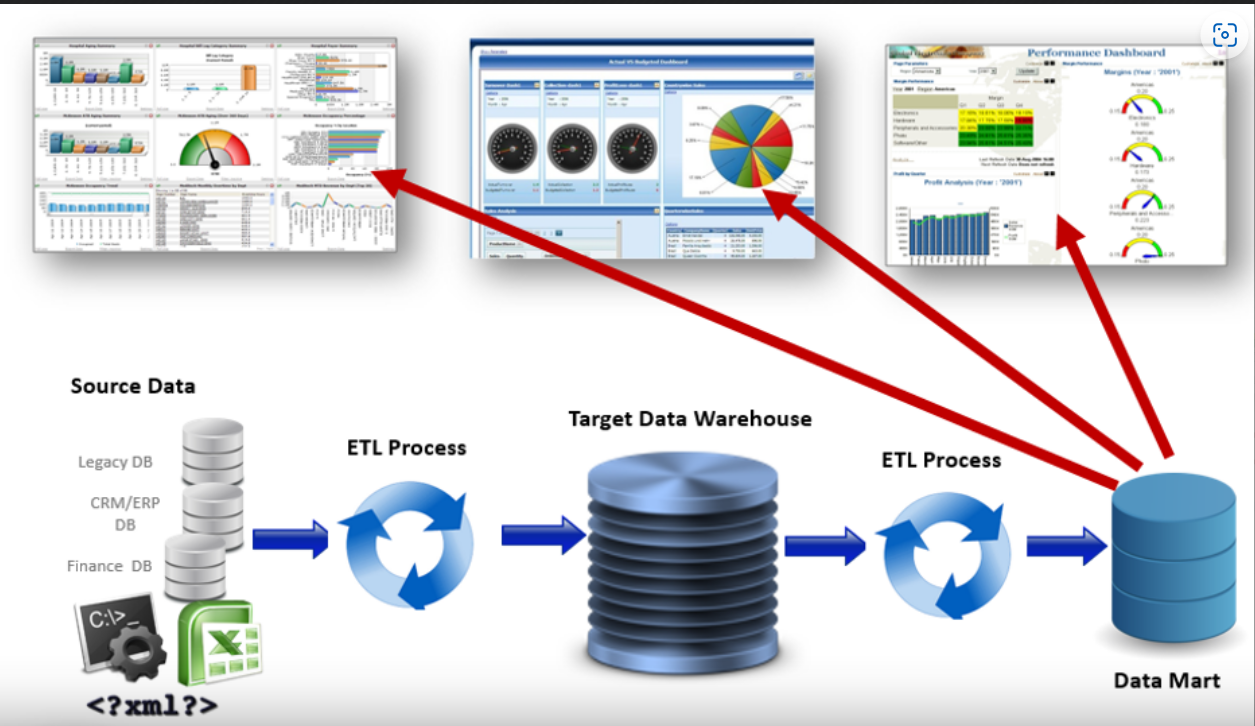
**Why deal with all this data?**

* The benefit to moving data from sources (files, databases, xml, web services, mainframes, external feeds, etc.) is to have it in an optimized target data store to analyse, look for trends and make strategic and tactical decisions based on this data.
* To analyse the data, business intelligence software is connected to the data warehouse or data mart.

### What is Business Intelligence (BI) software?

Business Intelligence (BI) software is used in identifying, analysing business data. BI provides simple access to data which can be used in day-to-day operations and integrates data into logical business areas. It also provides historical, current and predictive views of business operations. BI is made up of several related activities, including data mining, online analytical processing, querying and reporting.

C‑level executives and managers use BI & Analytics software to make critical business decisions based on the underlying data.



**Typical Data Architecture**

Below is a picture that demonstrates a typical data architecture where data may be moving in the following manner:

* From databases, Excel, json, web services or mainframe into a big data lake to a data warehouse to a data mart and being interpreted through business intelligence (BI) solutions
* As a data migration from a data warehouse on premises to a data warehouse in the cloud and interpreted through BI solutions
* From an enterprise application (finance, trading, sales, ERP, research, etc.) into a data warehouse and interpreted through BI solutions

A diagram of data processing

Description automatically generated

### How does the ETL process work?

**Extract** – ETL developers extract data from one or more systems and/or files and copy it into the data warehouse.

**Transform** – They remove inconsistencies, assemble to a common format, adding missing fields, summarizing detailed data and deriving new fields to store calculated data.

**Load** – They map the data to the proper tables and columns, transform and/or load it into the data warehouse.

A diagram of data processing

Description automatically generated

### Who is involved in the ETL process?

There are at least 4 roles involved. They are:

**Data Analyst:** Creates data requirements (source-to-target map or mapping doc)

**Data Architect:** Models and builds data store (Big Data Lake, Data Warehouse, Data Mart, etc.)

**ETL Developer:** Transforms and loads data from sources to target data stores.

**ETL Tester:** Validates the data, based on mappings, as it moves and transforms from sources to targets.

A diagram of a software project

Description automatically generated

### Why perform ETL Testing?

Bad data caused by defects in the ETL process can cause data problems in reporting that can result in poor strategic decision-making.

An example of what bad data can cause:  
A large fast food company depends on business intelligence reports to determine how much raw chicken to order on a monthly basis, by sales region and time of year. If these reports are not correct, then the company could order an incorrect amount which could cost the company millions of dollars in either lost revenue or wasted product.

**How does ETL Testing work?**

ETL Testing is a way to perform validation of the data as it moves from one data store to another. The ETL Tester uses a Mapping Document (if one exists), which is also known as a source-to-target map. This is the critical element required to efficiently plan the target Data Stores. It also defines the Extract, Transform and Load (ETL) process.

The intention is to capture business rules, data flow mapping and data movement requirements. The Mapping Document specifies source input definition, target/output details and business & data transformation rules.

The typical process for ETL Testing is as follows:

**1) Review the Schema and Business Rules / Mappings**

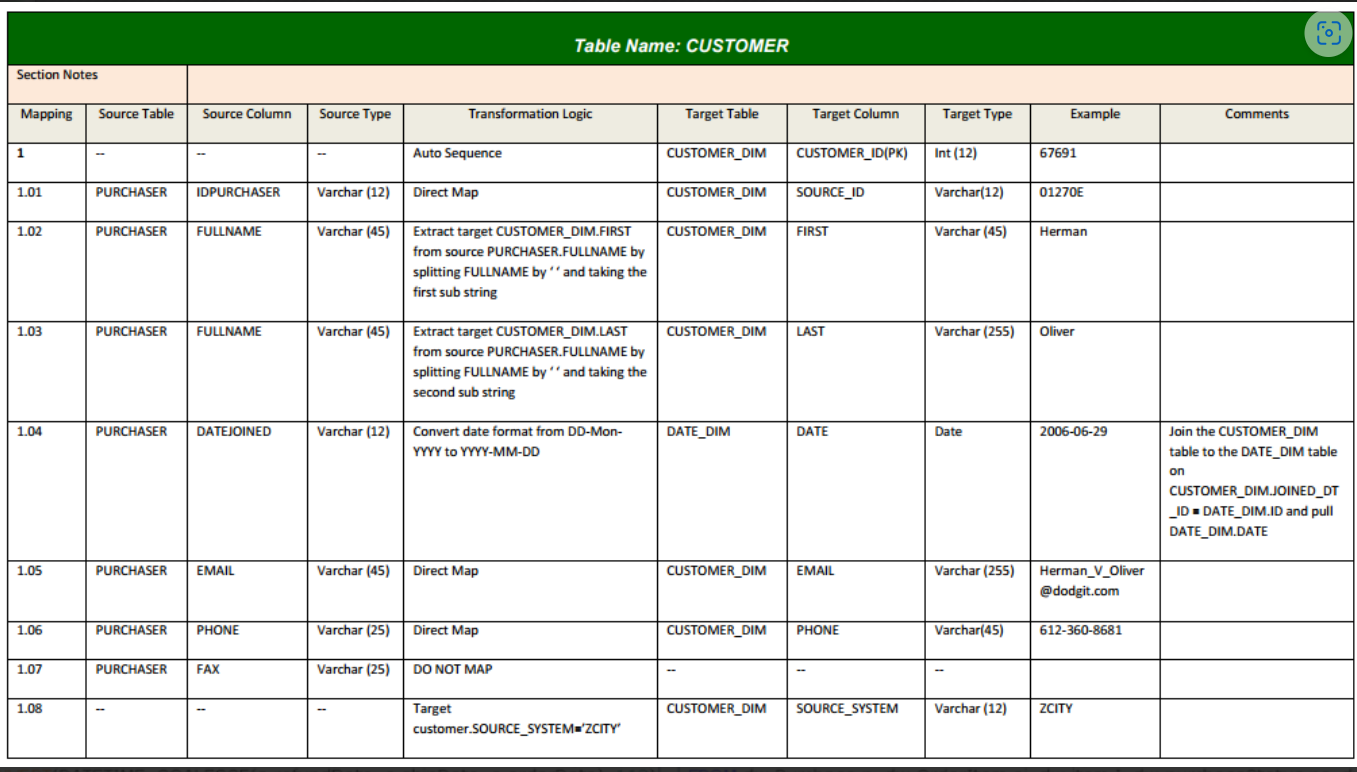
Schemas are ways in which data is organized within a database or data warehouse.

Business Rules are also known as Mappings or Source-to-Target mappings and are typically found in a Mapping Document. The mapping tables in the document are the requirements or rules for extracting, transforming (if at all) and loading (ETL) data from the source database and files into the target data warehouse or big data store. Specifically, the mapping fields show:

* Table names, field names, data types and length of both source and target fields
* How source tables / files should be joined in the new target data set
* Any transformation logic that will be applied
* Any business rules that will be applied

**2) Create Test Cases**

Each Mapping will typically have its own test case. Test Case will typically have two sets of SQL queries. One query will extract data from the sources (flat files, databases, xml, web services, etc.) and the other query will extract data from the target (Data Warehouses or Big Data stores).



**3) Execute Tests, Export Results**

Queries can be executed in SQL developer, Oracle db2 , Toad and other sql editors. Get the results from both the source systems. The test results from the 2 queries are saved into 2 Excel spreadsheets.

**4) Compare Results**

Compare all result sets in the source spreadsheet with the target spreadsheet by eye compare. (also known as ​“Stare & Compare”). Difficult to perform the testing manually using our eye for the large volume of data. So more efforts are required and it could be incorrect.

There are 4 different methods for performing ETL testing:

* Sampling
* Minus Queries
* Using home grown tools, utilities and frameworks
* Using commercial software like Query Surge (Automation tool)

Test cases can be executed in HPALM, other data validation tool, even manual comparison using excel.

### ETL Testing Tools – [Query Surge](https://guru99.link/recommends-querysurge-etl-testing-tool), Informatica Data Validation

### Responsibility of ETL tester –

### An ETL tester’s role is important in safeguarding the business’s data quality. Here are some key responsibilities of an ETL tester:

### Prepare and plan for testing by developing a testing strategy, a test plan, and test cases for the ETL testing process.

### Analyse source data for data quality concerns throughout the ETL testing process.

### Execute test cases to validate the ETL process.

### Identify defects and issues in the ETL process and work with teams to rectify them.

### Communicate testing results with development teams, stakeholders, and other decision-makers.

### Incorporate learnings and best practices to improve the ETL testing process over time.

### What is a Minus Query?

A Minus Query is a query that uses the MINUS operator in SQL to subtract one result set from another result set to evaluate the result set difference. If there is no difference, there is no remaining result set. If there is a difference, the resulting rows will be displayed.

The way to test using Minus Queries is to perform source-minus-target and target-minus-source queries for all data, making sure the extraction process did not provide duplicate data in the source and all unnecessary columns are removed before loading the data for validation.

Challenges during minus query execution

* In the standard minus query implementation, minus queries need to be executed twice (Source-to-Target and Target-to-Source). This doubles execution time and resource utilization.
* Result sets may be inaccurate when duplicate rows of data exist (the minus query may only return 1 row even if there are duplicates)

|  |  |  |
| --- | --- | --- |
| Table Name Customer1 - Source | |  |
|  |  |  |
| First\_Name | Last\_Name | City |
| Thara | Ruth | New Delhi |
| Louis | Millen | Noida |
| Yogi | Anand | Daman |
| Mishra | Das | Gujarat |
| William | Sha | Airoli |
| Peter | Deric | Goa |
| Andy | Jeter | Nizamapur |
| Bala | Joy | Bhayandar |

|  |  |  |
| --- | --- | --- |
| Table Name Customer2 - Target | |  |
|  |  |  |
| First\_Name | Last\_Name | City |
| Thara | Ruth | New Delhi |
| Louis | Millen | Noida |
| Yogi | Anand | Daman |
| Mishra | Das | Gujarat |
| William | Sha | Airoli |
| Peter | Deric | Goa |
| Joe | Mantle | Ulhasnagar |

SELECT FIRST\_NAME,LAST\_NAME,CITY FROM CUTOMER1

MINUS

SELECT FIRST\_NAME,LAST\_NAME,CITY FROM CUTOMER2

|  |  |  |
| --- | --- | --- |
| Andy | Jeter | Nizamapur |
| Bala | Joy | Bhayandar |

SELECT FIRST\_NAME,LAST\_NAME,CITY FROM CUTOMER2

MINUS

SELECT FIRST\_NAME,LAST\_NAME,CITY FROM CUTOMER1

|  |  |  |
| --- | --- | --- |
| Joe | Mantle | Ulhasnagar |

|  |  |  |
| --- | --- | --- |
| Table Name Customer1 - Source | |  |
|  |  |  |
| First\_Name | Last\_Name | City |
| Thara | Ruth | New Delhi |
| Louis | Millen | Noida |
| Yogi | Anand | Daman |
| Mishra | Das | Gujarat |
| William | Sha | Airoli |
| Peter | Deric | Goa |
| Andy | Jeter | Nizamapur |
| Bala | Joy | Bhayandar |
| Joe | Mantle | Ulhasnagar |

|  |  |  |
| --- | --- | --- |
| Table Name Customer2 - Target | |  |
|  |  |  |
| First\_Name | Last\_Name | City |
| Thara | Ruth | New Delhi |
| Louis | Millen | Noida |
| Yogi | Anand | Daman |
| Mishra | Das | Gujarat |
| William | Sha | Airoli |
| Peter | Deric | Goa |
| Andy | Jeter | Nizamapur |
| Bala | Joy | Bhayandar |
| Joe | Mantle | Ulhasnagar |

SELECT FIRST\_NAME,LAST\_NAME,CITY FROM CUTOMER1

MINUS

SELECT FIRST\_NAME,LAST\_NAME,CITY FROM CUTOMER2

**No results will be returned.**

**Challenges Strategies for Testing an ETL**

ETL testing can have many challenges. Identifying them early in the ETL process can prevent bottlenecks and costly delays. Some of the common challenges include:

* **Potential complexity of data transformations.** Transformations of large datasets can be time-consuming and complex.
* **Incorrect data.** Data is often messy and full of errors; ETL testing needs clean, accurate data to have correct results.
* **Data source changes.** Changes to data sources impact the completeness and accuracy of data quality.
* **Complex processes**. Complex data integrations and business processes can cause problems.
* **Slow performance.** Slow processing or slow end-to-end performance caused by massive data volumes can impact data accuracy and completeness.

**OLTP vs OLAP**

**OLTP (Online Transaction Processing) and OLAP (Online Analytical Processing)**are two different types of database systems that are designed to support different types of workloads.

**OLTP** systems are designed to support transactional workloads, which involve inserting, updating, and deleting small amounts of [data](https://nixondata.com/knowledge/apache-spark-for-experts/apache-spark-what-is-data-skew-how-to-identify-data-skew-from-data-skew-impact-fix/) in a database. These systems are optimized for fast read and write performance, and are typically used to support real-time operational systems, such as e-commerce websites, ATM machines, and airline reservation systems, railway reservation systems etc.,

**OLAP** systems, on the other hand, are designed to support analytical workloads, which involve querying large amounts of data to [extract](https://nixondata.com/knowledge/big-data-fundamentals/what-is-etl-what-are-etl-tools-list-of-open-source-etl-tools-etl-tools-available-in-aws/) insights and perform data analysis. These systems are optimized for fast query performance and are typically used to support business intelligence and data warehousing applications.

Here are some examples of the use cases for OLTP and OLAP systems:

* OLTP:
  + E-commerce websites
  + Banking systems – day to day transactions.
  + Inventory management systems
  + Customer relationship management (CRM) systems
* OLAP:
  + Data warehousing and business intelligence
  + Financial analysis
  + Marketing analysis
  + Supply chain analysis

## Difference between OLAP and OLTP

|  |  |  |
| --- | --- | --- |
| **Category** | **OLAP (Online Analytical Processing)** | **OLTP (Online Transaction Processing)** |
| Definition | It is well-known as an online database query management system. | It is well-known as an online database modifying system. |
| Data source | Consists of historical data from various Databases. | Consists of only operational current data. |
| Purpose | It serves the purpose to extract information for analysis and decision-making. | It serves the purpose to Insert, Update, and Delete information from the database. |
| Volume of data | A large amount of data is stored typically in TB, PB | The size of the data is relatively small as the historical data is archived in MB, and GB. |
| Queries | Relatively slow as the amount of data involved is large. Query execution may take hours. | Very Fast as the queries operate on 5% of the data. |
| Processing time | The processing of complex queries can take a lengthy time. | It is comparatively fast in processing because of simple and straightforward queries. |
| Operations | Only read and rarely write operations. | Both read and write operations. |
| Updates | Batch jobs are enabled to refresh the data on a regular basis. | The user manual refreshes the data which are brief and quick. |
| Productivity | Improves the efficiency of business analysts. | Enhances the user’s productivity. |

**2**. **ETL Testing Phases**

**2.1 Introduction to the different Phases of ETL Testing**

The different phases of ETL testing process is as followA diagram of a test

Description automatically generated

* Business and Requirement understanding (BRD and TDD)
* Test planning and Estimation.
* Test case design and Test data preparation.
* Test execution.
* Identifying Buys and reporting bugs
* Test summary report.
* Test closure after retesting.

ETL testing is performed in five stages.

* Identifying different data sources and requirements.
* Data procurement.
* Implement business logics and dimensional Modelling.
* Build and populate data.
* Generate\Build Reports

## ETL Test Life Cycle and Test Artifacts

* **Understand Business Requirements**
* Business requirement document
* Technical specification document
* Design specification document
* Data mapping document(STTM)
* Database structure, application logic, and how data is used in the current system.
* **Validate Business Requirements**
* Cross-check each and every requirement with business analyst and clarify the doubts and issues related to data mapping.
* Cross-check logic of data transformation and validate that there is no gap.
* **Estimate the Testing efforts**
* Estimate how much time is required for test planning, and test artifacts (Test case creation and test execution creation document) creation.
* As per the number of business requirements, high-level test execution efforts are estimated.
* As per the overall project time, budget and team size, no. of testing days are estimated.
* **Test Planning**
* Test methods and techniques used, tools used, and schedule of testing.
* **Creating Test Artifacts**
* It includes test cases, test scenarios preparation, keeping the following key points in the mind.
* The structure of source data should be the same as on the final database.
* All the tables on the target are updated including all referential data.
* The data model is correctly followed in the target.
* Data transformation rules.
* **Test data preparation –** Something developers need to take care. If required, testers will prepare the data for testing.
* **Test cases review and approval.**
* Test cases needs to be reviewed and shared to business for approval
* Once reviews are done on test cases, we can do a dry run for those test cases.
* **Test Execution:**

Test case execution happens as per the schedule. Following are the steps in the ETL test execution phase.

* Execute ETL jobs.
* Check that data load is done successfully on the target and data is valid.
* Check the target and validate the data quality and correctness by executing SQL queries.
* Check ETL tool logs, where it caused error or terminated abnormally.
* Execute the test cases.
* **Bug Reporting Defect closure:**
* If there are any data mismatch identified, it can be reported as a defect.
* Defects are logged and defect closure is done by the development team.
* **Sign off and Reporting:**
* The number of test cases executed, status, and overall status is shared with the team. Sign off can be given if enough testing is done and bug count is close to zero, and no high severity issue is present in the system.

**ETL Testing Types**

ETL testing is categorized in the following types.

* **Testing of New System:** In this, Data is extracted from different types of sources and loaded into the Data warehouse.
* **Structure Validation:** Validate the source and target table structure against corresponding mapping doc.
* **Data Quality check:** Date Check needs to be performed where the format of the date should be same between source and target.
* YYYY-MM-DD
* YYYY-MON-DD
* DD-MM-YYYY
* **Mapping doc validation:** Verify mapping doc whether corresponding ETL information is provided or not. Change log should maintain in every mapping doc.
* **Metadata Testing:** Metadata validation is validating the column names, data types, length of the table, Null columns, Key columns like Primary or foreign key checks.

**Examples : - Desc Table\_Name**

**Select column\_name, data\_type from information\_Schema.columns**

* **Count Validation –** To verify the source and target count is exactly matching.

SELECT COUNT(\*) FROM Customer\_Table ;

* **Duplicate validation –** 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.

To verify there is no duplicate records present in the target table after data loading.

SELECT OrderID, COUNT(OrderID)

FROM Order\_Table

GROUP BY OrderID

HAVING COUNT(OrderID) > 1;

* **Null validation** – To verify the not null columns, are not having any null values.

SELECT Id, Account\_Name, Account\_Type from Account\_Table

Where Id is not null or

Account\_Name is not null or

Account\_Type is not null

* **Workflow execution and monitoring –** To the workflow or the jobs execution where the data is loaded to a target system correctly.

Verify the Job status either success\failure message.

* **Data validation** (Source minus target and Target minus source)
* **Incremental Testing:** This testing is done to check the data integrity of old and new data when the new data added. Incremental testing verifies that the system processes correctly even after the insertion and updating the data during an incremental ETL process.
  + Incremental data can be validated using last\_updated\_date and last\_updated\_by column filter.
  + SELECT \* FROM Table\_Name where last\_updated\_date =’2023-04-01’
  + It can be either inserted data or the updated data.
* **File validation –** File validations can be performed with the below ways.
* The files can be either txt, xlsx, csv formats.
* Structure validation of the file. Like name & number of fields, delimiter, naming convention, Header or Trailer records.
* Check for duplicate records if any.
* Select any one row from the target file and it can be compared with the Source output using the SQL source query.
* Check for data truncation if any in the file.
* **Testing of Reports:** When data is completely loaded into the Data warehouse, then testing of reports is done to check data quality, validity, business logic, and calculations.
* **Data Accuracy Testing:** This testing is done to ensure that the data is accurately loaded and transformed as expected.
* **Migration testing:** To verify the data migration of a source system to the data warehouse with data integrity and no loss of data.
* **Source to target data testing:** To verify that there should not be any data loss or data truncation while loading the data from source to data warehouse.

Data @ Source - Thiruvananthapuram

Data loaded @target - Thiruvanantha

* **Performance testing:** It is done to verify the data loading happens within the expected time and the users of the application can perform their job without any delay.
* **GUI testing:** This testing is done to check reports are showing up correctly to its end users.
* **Tools –** Power BI, Tableau