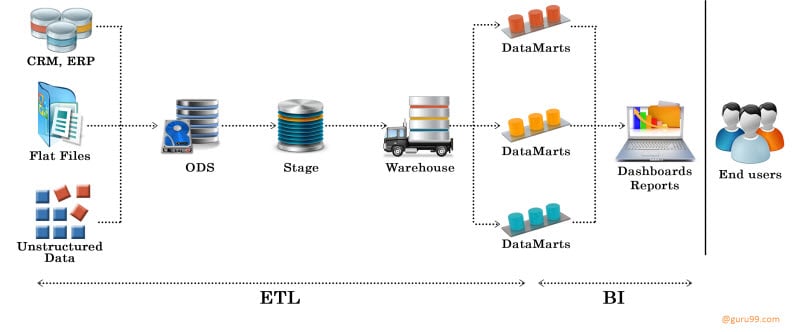
**In this ETL Testing tutorial, you will learn:**

* [What is ETL Testing?](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#what-is-etl-testing)
* [Data Warehouse Testing](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#data-warehouse-testing)
* [What is ETL?](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#what-is-etl)
* [ETL Testing Process](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#etl-testing-process)
* [Types of ETL Testing](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#types-of-etl-testing)
* [How to Create ETL Test Case](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#how-to-create-etl-test-case)
* [ETL Test Scenarios and Test Cases](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#etl-test-scenarios-and-test-cases)
* [Types of ETL Bugs](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#types-of-etl-bugs)
* [Difference between Database Testing and ETL Testing](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#difference-between-database-testing-and-etl-testing)
* [Responsibilities of an ETL Tester](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#responsibilities-of-an-etl-tester)
* [Performance Testing in ETL](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#performance-testing-in-etl)
* [Automation of ETL Testing](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#automation-of-etl-testing)
* [Best Practices for ETL Testing](https://www.guru99.com/utlimate-guide-etl-datawarehouse-testing.html#best-practices-for-etl-testing)

**What is ETL Testing?**

ETL testing is done to ensure that the data that has been loaded from a source to the destination after business transformation is accurate.

It also involves the verification of data at various middle stages that are being used between source and destination. ETL stands for Extract-Transform-Load.



**Data Warehouse Testing** is a testing method in which the data inside a data warehouse is tested for integrity, reliability, accuracy and consistency in order to comply with the company’s data framework.

The main purpose of data warehouse testing is to ensure that the integrated data inside the data warehouse is reliable enough for a company to make decisions on.

## 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.

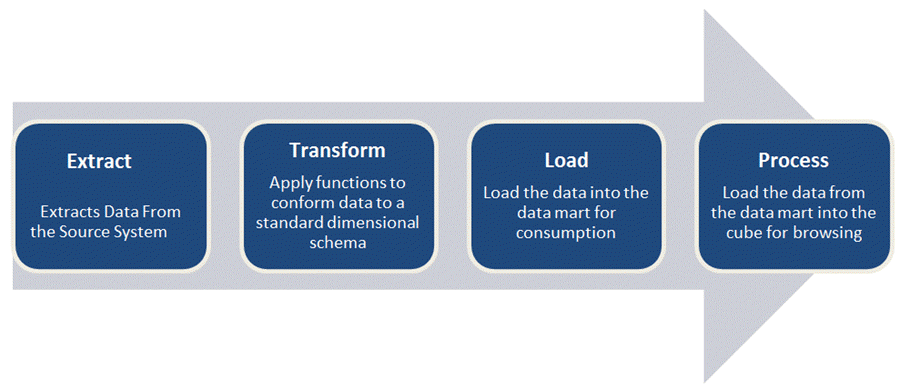
**Let see how it works**

For example, there is a retail store which has different departments like sales, marketing, logistics etc. Each of them is handling the customer information independently, and the way they store that data is quite different. The sales department have stored it by customer’s name, while marketing department by customer id.

Now if they want to check the history of the customer and want to know what the different products he/she bought owing to different marketing campaigns; it would be very tedious.

The solution is to use a [data warehouse](https://www.guru99.com/data-warehousing.html) to store information from different sources in a uniform structure using ETL. ETL can transform dissimilar data sets into an unified structure. Later use BI tools to derive meaningful insights and reports from this data.

The following diagram in this ETL testing tutorial gives you the ROAD MAP of the ETL Testing process flow and various ETL testing concepts:



1. **Extract**

* Extract relevant data

1. **Transform**

* Transform data to DW (Data Warehouse) format
* Build keys – A key is one or more data attributes that uniquely identify an entity. Various [types of keys](https://www.guru99.com/dbms-keys.html) are primary key, alternate key, foreign key, composite key, surrogate key. The datawarehouse owns these keys and never allows any other entity to assign them.
* Cleansing of data :After the data is extracted, it will move into the next phase, of cleaning and conforming of data. Cleaning does the omission in the data as well as identifying and fixing the errors. Conforming means resolving the conflicts between those data’s that is incompatible, so that they can be used in an enterprise data warehouse. In addition to these, this system creates meta-data that is used to diagnose source system problems and improves data quality.

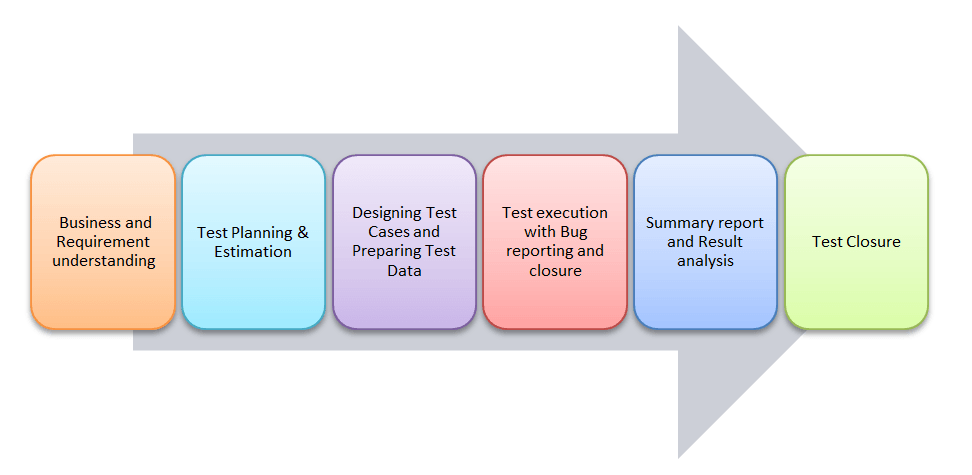
1. **Load**

Load data into DW (Data Warehouse)

Build aggregates – Creating an aggregate is summarizing and storing data which is available in [fact table](https://www.guru99.com/fact-table-vs-dimension-table.html) in order to improve the performance of end-user queries.

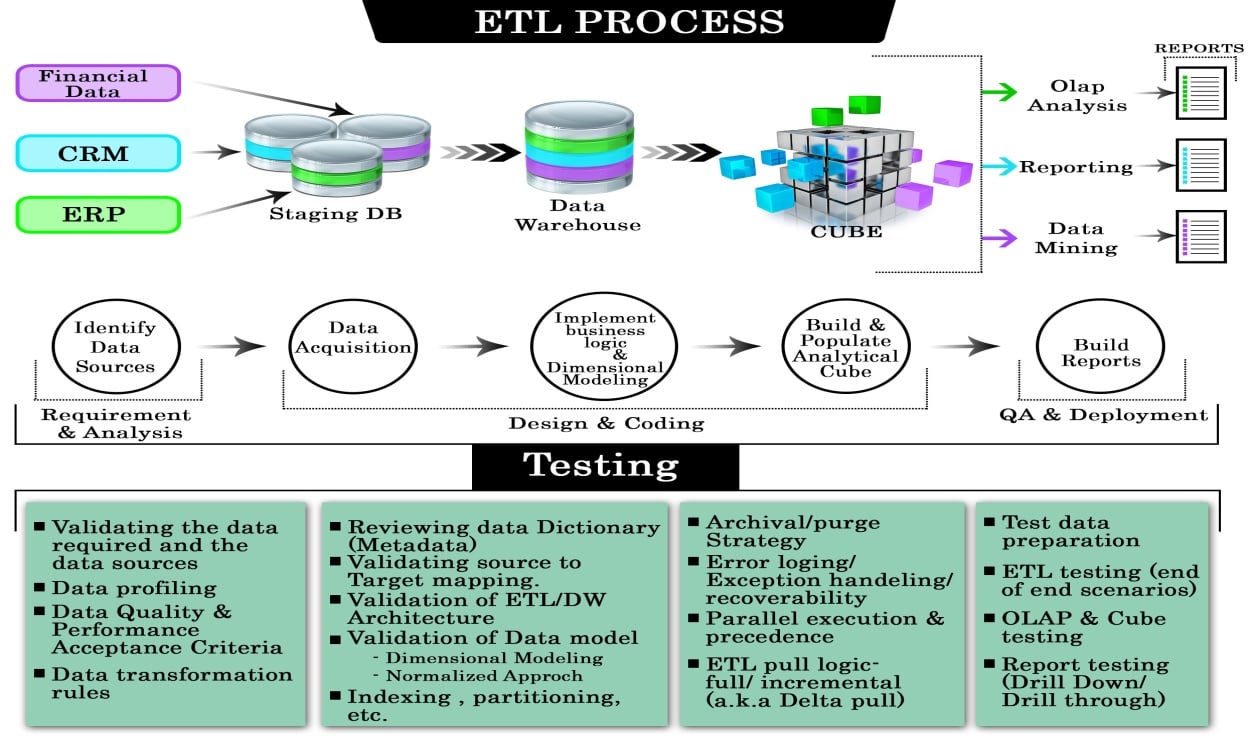
**ETL Testing Process**

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



| **Types Of Testing** | **Testing Process** |
| --- | --- |
| **Production Validation Testing** | “Table balancing” or “production reconciliation” this type of ETL testing is done on data as it is being moved into production systems.  To support your business decision, the data in your production systems has to be in the correct order.  [Informatica](https://www.guru99.com/informatica-tutorials.html)Data Validation Option provides the ETL testing automation and management capabilities to ensure that production systems are not compromised by the data. |
| **Source to Target Testing (Validation Testing)** | Such type of testing is carried out to validate whether the data values transformed are the expected data values. |
| **Application Upgrades** | Such type of ETL testing can be automatically generated, saving substantial test development time.  This type of testing checks whether the data extracted from an older application or repository are exactly same as the data in a repository or new application. |
| **Metadata Testing** | Metadata testing includes testing of data type check, data length check and index/constraint check. |
| **Data Completeness Testing** | To verify that all the expected data is loaded in target from the source, data completeness testing is done.  Some of the tests that can be run are compare and validate counts, aggregates and actual data between the source and target for columns with simple transformation or no transformation. |
| **Data Accuracy Testing** | This testing is done to ensure that the data is accurately loaded and transformed as expected. |
| **Data Transformation Testing** | Testing data transformation is done as in many cases it cannot be achieved by writing one source[SQL](https://www.guru99.com/sql.html)query and comparing the output with the target.  Multiple SQL queries may need to be run for each row to verify the transformation rules. |
| **Data Quality Testing** | Data Quality Tests includes syntax and reference tests. In order to avoid any error due to date or order number during business process Data Quality testing is done.  **Syntax Tests**: It will report dirty data, based on invalid characters, character pattern, incorrect upper or lower case order etc.  **Reference Tests**: It will check the data according to the data model. For example: Customer ID  Data quality testing includes number check, date check, precision check, data check , null check etc. |
| **Incremental ETL testing** | This testing is done to check the data integrity of old and new data with the addition of new data.  Incremental testing verifies that the inserts and updates are getting processed as expected during incremental ETL process. |
| **GUI/Navigation Testing** | This testing is done to check the navigation or GUI aspects of the front end reports. |

**How to Create ETL Test Case**

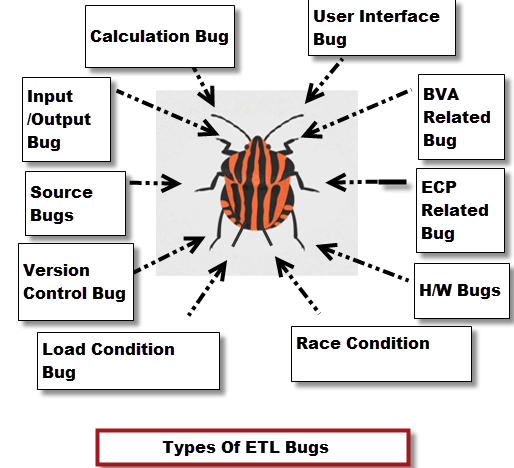
ETL testing is a concept which can be applied to different tools and databases in information management industry. **The objective of ETL testing is to assure that the data that has been loaded from a source to destination after business transformation is accurate.** It also involves the verification of data at various middle stages that are being used between source and destination.

While performing ETL testing, two documents that will always be used by an ETL tester are

1. **ETL mapping sheets :**An ETL mapping sheets contain all the information of source and destination tables including each and every column and their look-up in reference tables. An ETL testers need to be comfortable with SQL queries as ETL testing may involve writing big queries with multiple joins to validate data at any stage of ETL. ETL mapping sheets provide a significant help while writing queries for data verification.
2. **DB Schema of Source, Target:**It should be kept handy to verify any detail in mapping sheets.

| **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 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](https://www.guru99.com/regression-testing.html) |
| **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 a transformation.

**Some of the responsibilities of an ETL tester are**

* Test ETL software
* Test components of ETL data warehouse
* Execute backend data-driven test
* Create, design and execute [test cases](https://www.guru99.com/test-case.html), 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

**Performance Testing in ETL**

Performance Testing in ETL is a testing technique to ensure that an ETL system can handle load of multiple users and transactions.

The primary goal of ETL [Performance Testing](https://www.guru99.com/performance-testing.html) is to optimize and improve session performance by identification and elimination of performance bottlenecks. The source and target databases, mappings, sessions and the system possibly have performance bottlenecks.

One of the best tools used for Performance Testing/Tuning 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](https://www.guru99.com/test-coverage-in-software-testing.html).

To accelerate, improve coverage, reduce costs, improve[Defect](https://www.guru99.com/defect-management-process.html)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](https://www.guru99.com/unit-testing-guide.html) that target exceptions

# **Top 25 ETL Testing Interview Questions & Answers:**

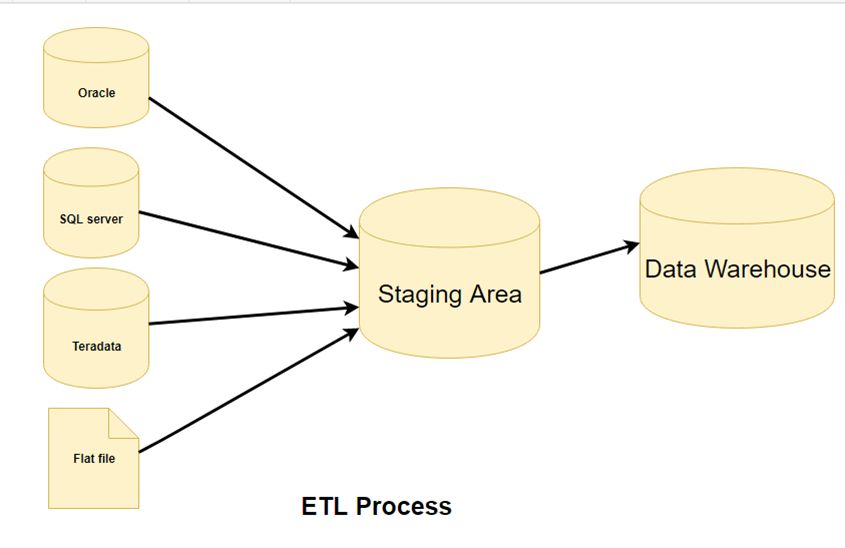
### **1) What is ETL?**

In data warehousing architecture, ETL is an important component, which manages the data for any business process. ETL stands for **Extract, Transform** and **Load**. Extract does the process of reading data from a database. Transform does the converting of data into a format that could be appropriate for reporting and analysis. While, load does the process of writing the data into the target database.

### **2) Explain what are the ETL testing operations includes?**

ETL testing includes:

* Verify whether the data is transforming correctly according to business requirements
* Verify that the projected data is loaded into the data warehouse without any truncation and data loss
* Make sure that ETL application reports invalid data and replaces with default values
* Make sure that data loads at expected time frame to improve scalability and performance



### 3) **Mention what are the types of data warehouse applications and what is the difference between data mining and data warehousing?**

The types of data warehouse applications are

* Info Processing
* Analytical Processing
* Data Mining

[**Data mining**](https://www.guru99.com/data-mining-tutorial.html)can be define as the process of extracting hidden predictive information from large databases and interpret the data while data warehousing may make use of a data mine for analytical processing of the data in a faster way.

[**Data warehousing**](https://www.guru99.com/data-warehousing.html) is the process of aggregating data from multiple sources into one common repository

### **4) What are the various tools used in ETL?**

* Cognos Decision Stream
* Oracle Warehouse Builder
* Business Objects XI
* SAS business warehouse
* SAS Enterprise ETL server

### **5) What is fact? What are the types of facts?**

It is a central component of a multi-dimensional model which contains the measures to be analysed. Facts are related to dimensions.

**Types of facts are**

* Additive Facts
* Semi-additive Facts
* Non-additive Facts

### **6) Explain what are Cubes and OLAP Cubes?**

Cubes are data processing units comprised of fact tables and dimensions from the data warehouse. It provides multi-dimensional analysis.

OLAP stands for **Online Analytics Processing**, and OLAP cube stores large data in muti-dimensional form for reporting purposes. It consists of facts called as measures categorized by dimensions.

### **7) Explain what is tracing level and what are the types?**

Tracing level is the amount of data stored in the log files. Tracing level can be classified in two Normal and Verbose.

**Normal level** explains the tracing level in a detailed manner while **verbose** explains the tracing levels at each and every row.

**8) Explain what is Grain of Fact?**

Grain fact can be defined as the level at which the fact information is stored. It is also known as **Fact Granularity**

**9) Explain what factless fact schema is and what is Measures?**

A fact table without measures is known as Factless fact table. It can view the number of occurring events. For example, it is used to record an event such as employee count in a company.

The numeric data based on columns in a fact table is known as Measures

**10) Explain what is transformation?**

A transformation is a repository object which generates, modifies or passes data. Transformation are of two **types Active and Passive**

**11) Explain the use of Lookup Transformation?**

The Lookup Transformation is useful for

* Getting a related value from a table using a column value
* Update slowly changing dimension table
* Verify whether records already exist in the table

**12) Explain what is partitioning, hash partitioning and round robin partitioning?**

To improve performance, transactions are sub divided, this is called as Partitioning. Partioning enables[Informatica](https://www.guru99.com/informatica-tutorials.html)Server for creating of multiple connection to various sources

**The types of partitions are**

**Round-Robin Partitioning:**

* By informatica data is distributed evenly among all partitions
* In each partition where the number of rows to process are approximately same this partioning is applicable

**Hash Partitioning:**

* For the purpose of partitioning keys to group data among partitions Informatica server applies a hash function
* It is used when ensuring the processes groups of rows with the same partitioning key in the same partition need to be ensured

**13) Mention what is the advantage of using DataReader Destination Adapter?**

The advantage of using the DataReader Destination Adapter is that it populates an **ADO recordset**(consist of records and columns) in memory and exposes the data from the DataFlow task by implementing the DataReader interface, so that other application can consume the data.

14) **Using SSIS (SQL Server Integration Service) what are the possible ways to update table?**

To update table using SSIS the possible ways are:

* Use a[SQL](https://www.guru99.com/sql.html)command
* Use a staging table
* Use Cache
* Use the Script Task
* Use full database name for updating if MSSQL is used

**15) In case you have non-OLEDB (Object Linking and Embedding Database) source for the lookup what would you do?**

In case if you have non-OLEBD source for the lookup then you have to use Cache to load data and use it as source

### **16) In what case do you use dynamic cache and static cache in connected and unconnected transformations?**

* Dynamic cache is used when you have to update master table and slowly changing dimensions (SCD) type 1
* For flat files Static cache is used

### **17) Explain what are the differences between Unconnected and Connected lookup?**

|  |  |
| --- | --- |
| **Connected Lookup** | **Unconnected Lookup** |
| Connected lookup participates in mapping | It is used when lookup function is used instead of an expression transformation while mapping |
| Multiple values can be returned | Only returns one output port |
| It can be connected to another transformations and returns a value | Another transformation cannot be connected |
| Static or dynamic cache can be used for connected Lookup | Unconnected as only static cache |
| Connected lookup supports user defined default values | Unconnected look up does not support user defined default values |
| In Connected Lookup multiple column can be return from the same row or insert into dynamic lookup cache | Unconnected lookup designate one return port and returns one column from each row |

### **18) Explain what is data source view?**

A data source view allows to define the relational schema which will be used in the analysis services databases. Rather than directly from data source objects, dimensions and cubes are created from data source views.

**19) Explain what is the difference between OLAP tools and ETL tools ?**

The difference between ETL and OLAP tool is that

[ETL tool](https://www.guru99.com/best-etl-tools.html) is meant for the extraction of data from the legacy systems and load into specified data base with some process of cleansing data.

**Example:** Data stage, Informatica etc.

While OLAP is meant for reporting purpose in OLAP data available in multi-directional model.

**Example:** Business Objects, Cognos etc.

### **20) How you can extract SAP data using Informatica?**

* With the power connect option you extract SAP data using informatica
* Install and configure the PowerConnect tool
* Import the source into the Source Analyzer. Between Informatica and SAP Powerconnect act as a gateaway. The next step is to generate the ABAP code for the mapping then only informatica can pull data from SAP
* To connect and import sources from external systems Power Connect is used

### **21) Mention what is the difference between Power Mart and Power Center?**

|  |  |
| --- | --- |
| **Power Center** | **Power Mart** |
| Suppose to process huge volume of data | Suppose to process low volume of data |
| It supports ERP sources such as SAP, people soft etc. | It does not support ERP sources |
| It supports local and global repository | It supports local repository |
| It converts local into global repository | It has no specification to convert local into global repository |

### **22) Explain what staging area is and what is the purpose of a staging area?**

Data staging is an area where you hold the data temporary on data warehouse server. Data staging includes following steps

* Source data extraction and data transformation (restructuring)
* Data transformation (data cleansing, value transformation)
* Surrogate key assignments

**23) What is Bus Schema?**

For the various business process to identify the common dimensions, BUS schema is used. It comes with a conformed dimensions along with a standardized definition of information

**24) Explain what is data purging?**

Data purging is a process of deleting data from data warehouse. It deletes junk data’s like rows with null values or extra spaces.

### **25) Explain what are Schema Objects?**

Schema objects are the logical structure that directly refer to the databases data. Schema objects includes tables, views, sequence synonyms, indexes, clusters, functions packages and database links

**26) Explain these terms Session, Worklet, Mapplet and Workflow ?**

* **Mapplet:** It arranges or creates sets of transformation
* **Worklet:** It represents a specific set of tasks given
* **Workflow:** It’s a set of instructions that tell the server how to execute tasks
* **Session:** It is a set of parameters that tells the server how to move data from sources to target

**What is Data wareHouse**

1. It is a database that is designed for querying and analysis rather than for transaction Processing.

2. It separates analysis workload from transaction system.

3. This helps in:

i. Maintaining historical records

ii. Analyzing the data to gain a better understanding of the business and to improve the Business.

4. Data warehouse is a subject-oriented, integrated, time varying, non-volatile collection of data in support of the management's decision-making process.

i. **Subject** Oriented: This is used to analyze particular subject area.

ii. **Integrated**: This shows that integrates data from different sources.

iii. **Time** **variant**: Historical data is usually maintained in a Data warehouse, i.e. retrieval can be for any period. In transactional system only the most recent/current data is maintained. But in the Data warehouse recent/current and the previous/historical data is maintained.

iv. **Non**-**Volatile**: Once the data is placed in the data warehouse, it cannot be changed, which means we will never be able to change the data.

**What is ETL?**

1. ETL stands for Extract-Transform-Load.

• **Extract** is the process of reading data from a source database/ transactional system.

• **Transform** is the process of converting the extracted data to required from.

• **Load** is the process of writing the data into the target database/ analytical system.

2. It is a process which defines how data is loaded from the source system to the target system (data warehouse).

**Data Ware House Architecture**

There are four layers in DWH architecture:

 Data Source Layer

 Data Staging Area

 Data Storage Layer

 Reporting Layer

**Source Layer** --First layer is the Data Source layer, which refers to various data stores in multiple formats like relational database, Flat Files, Excel files, Xml Files etc.

-- These data stores business data like Sales, Customer, Finance, Product etc.

**Staging Layer**--After that the next step is Extract, where the required data from data source layer is extracted and put into the data staging area.

-- Data Staging area is intermediate layer between Data Source Layer and Data Storage Layer used for processing data during the ETL process.

-- Basically needs staging area to hold the data and to perform data

Transformations, before loading the data into warehouse.

--Actual transformation transactional data into analytical data is done in data staging area.

**Storage Layer**--And finally, we have the Data Storage layer i.e. data warehouse, the place where the successfully cleaned, integrated, transformed and ordered data is stored in a multi-dimensional environment. Now, the data is available for analysis and query purposes.

**Reporting Layer**-In reporting layer, data in data storage layer is used to create various type of management reports from where user can take business decisions for planning, designing, forecasting etc.

**-- Meta data** is nothing but the data about data.

--**Meta data repository** is used to store meta data of data which is actually

present in data warehouse i.e. Data storage layer

--**Data mart** can be defined as the subset of data warehouse.

A data mart is focused on a single functional area e.g. product, customers, employees, sales etc. It is a **subject-oriented database** and is also known as **High Performance Query Structures (HPQS).**

**OLTP (Online Transaction Processing System):**

1. OLTP is nothing but a database which actually stores the daily transactions which are created from one and more applications.

2. Data in OLTP is called as the current data.

3. Mostly normalized data is used in OLTP system.

**OLAP (Online Analytical Processing System) :**

1. OLAP is use to store analytical data

2. It deals with analyzing the data for decision making and planning, designing etc.

3. Data in OLAP is called as the Historical data.

4. Mostly DE normalized data is used in OLAP system.

**Normalization**

* Normalization is the process of efficiently organizing the data in the database.
* Normalization is used to minimize the redundancy.
* Normalization divides the larger table into the smaller table and links them using relationship.

**Data Models**

* Data model tells how the logical structure of a database is modeled/designed.
* Data models define how data is connected to each other and how it will be processed and stored inside the system.

---Types of Data Models:

i. Conceptual Data Model

ii. Logical Data Model

iii. Physical Data Model

1. **Conceptual Data Model**

-- A conceptual data model is high level design of database.

--Features of conceptual data model include:

1. Displays the important entities and the relationships among them.

2. No attribute is specified.

3. No primary key is specified.

1. **Logical Data Model**

-- Logical Data Model defines the data as much as possible, to show how they can be physically implemented in the database.

i. Includes all entities and relationships among them.

ii. All attributes/columns for each entity/table are specified.

iii. The primary key for each entity is specified.

iv. Foreign keys (keys identifying the relationship between different entities) are specified.

v. Constraints are defined. (Unique, Not null, Check, default etc..)

1. **Physical Data model**

Actual implementation of logical model into Database is called Physical Data Model

1. **Dimensional Model**

**Q. What is Fact (Measures) ?**--- It is counted or measured event.

**Q. What is Dimension?**---It contains referential information about fact.

**Q. What is Fact Table ?**

* Fact table consist of measurements or facts of a business process.
* It is central table in dimension model surrounded by dimension tables.
* A fact table typically has two types of columns:

i. Those that contain facts.

ii. Those that are a foreign key to dimension tables.

**Q. What is Dimension Table?**

--Dimension tables are used to describe dimensions.

**\*\*\*\*Type of Dimension model**

1) Star Schema

2) Snowflake Schema

3) Galaxy or fact Constellation schema

**1) Star Schema**

1) It is simplest form of dimensional model

2) In Star schema design, central table is called fact table and

Radially connect other tables are called as dimension tables.

3) It is known as star schema because the entity-relationship

Diagram of this schemas look like a star.

4) Dimension tables in star schema are in De-Normalized form.

5) Star Schema is good for data marts with simple relationships.

**2) Snowflake Schema**

1) The process of normalizing dimension tables is called snow flaking.

2) In Snowflake schema, Dimension Tables are in Normalized form.

3) Snowflake schema is a extension of star schema.

4) It’s ER diagram look like a snowflake shape that’s why is called as snowflake schema.

**3) Galaxy Schema**

1) Galaxy Schema contains two and more fact tables that share Same dimension tables between them.

2) It is also called Fact Constellation Schema.

3) The schema is viewed as a collection of stars hence the name Galaxy Schema

**Type Of Facts**

1. Additive Fact

2. Non-Additive Fact

3. Semi-Additive

**1. Addictive Fact**

-- Additive facts are facts that can be summed up through all of the dimensions in the fact table.

**2. Non-Addictive Fact**

---- Non-additive facts are facts that cannot be summed up for any of the dimensions present in the fact table.

**3. Semi-Addictive Fact**

---- Semi-additive facts are facts that can be summed up for some of the dimensions in the fact table, but not

**Types of Dimensions**

1. Slowly Changing Dimensions

2. Conformed Dimensions

3. Degenerated Dimensions

4. Junk Dimensions

**Slowly Changing Dimensions :**

* --Dimensions that changes slowly over a period of time, rather than changing on regular schedule.
* A Slowly Changing Dimension (SCD) is a dimension that stores and manages both current and historical data over time in a data warehouse
* It is considered and implemented as one of the most critical ETL tasks in tracking the history of dimension records.

There are many approaches how to deal with SCD. The most popular are:

Type 0 - The passive method

Type 1 - Overwriting the old value

Type 2 - Creating a new additional record

Type 3 - Adding a new column

Type 4 - Using historical table

Type 6 - Combine approaches of types 1,2,3 (1+2+3=6)

**Type 0 - The passive method.**

In type 0, no special action is performed upon dimensional changes. Dimension data that remains same as it was first time inserted.

**Type 1 - Overwriting the old value**

* old value is simply overwritten by new value. Only new value is maintained.
* History of dimension changes is not kept in the database.
* This type is easy to maintain and is often use for data which changes are caused by processing corrections. (e.g. removal special characters, correcting spelling Errors).

**Type 2 - New row is created for new data**.

* Old value and new value is present is same table.
* New row is created for new value in dimension table.
* In this method all history of dimension changes is kept in the database.

**Type 3 - Adding a new column.**

* old and new value is kept in same table and same row.
* The new value is loaded into 'new' column and the old one into 'previous' Column.
* History is limited to the number of columns which are created for storing Historical data.
* This is the least commonly needed technique.

**Type 4 -Using historical table**.

In Type 4, separate table are there for old value and new value Separate historical table is used to track all historical changes for each of the dimension.

The 'main' table keeps only the New data (current data ) . e.g. customer and customer history tables.

**Type 6 - Combine approaches of types 1,2,3 (1+2+3=6).**

In this type we have additional columns in dimension table such as Current\_Address, Current\_Year : for keeping current value of the attribute.

Previous\_Address, Previous\_Year : for keeping historical value of the attribute.

Current\_Flag : for keeping information about the most recent record.

**2.Conformed Dimensions**

Conformed dimensions are those dimensions which have been designing such way that the dimensions can be used across many fact tables (data marts) in different subject areas of data warehouse.

**3.Degenerated Dimensions**

Degenerated Dimensions are the dimension which are directly present in fact table not in separate dimension table.

**4. Junk Dimensions**

When group of independent dimensions store in one dimension table that dimension table is called as Junk dimension.

**Surrogate key**

1. Some times in database table we cannot make primary key from real data.

2. In this situation, we have to add one artificial column in table which is unique and not null, and make this column as primary in table.

3. This primary which is generated from artificial column as surrogate key

**Data Mapping Document**

1. Data mapping document defines relationship between source data fields to their related target data fields which are involved in ETL process.

2. In simple terms, data mapping document is nothing but the map between source data and target data in ETL process.

**Why ETL Testers need mapping document?**

ETL tester needs mapping document because, while testing on data of target tables we have to refer data in source tables and mapping document having detail information of each and every table that is part of ETL process from source to target along with transformation logic.

**Types of ETL TESTING:**

1. Metadata Testing

2. Data Completeness Testing

3. Data Transformation Testing

4. Data Quality Testing

**1. Metadata Testing:**

In meta data testing, we have to validate physical model against it’s logical model. Metadata testing involves verification of :

* --Table Name
* --Column Name
* --Column data Type
* --Column data length
* --Constraints

**2. Data Completeness Testing:**

In Data Completeness testing, we are going to ensures that all the expected data is loaded in target from the source system.

Data Completeness testing involves :

* Checking and validating the record counts and null counts between the source and target for columns.
* Also, checking and validating the count of records with aggregate functions, filter, against incremental and Historical data loads.

**3. Data Transformation testing**:

In Data Transformation Testing, we are going to ensures source data which is converted (or transformed) as per business rules/requirement is loaded correctly in Target (Data warehouse).

Data Transformations can be :

* + Concatenation
  + Joins, splits
  + Conditional
  + Aggregation
  + Data Conversion

**4. Data Quality Testing**

In Data Quality testing, we are going to ensures accuracy of the data in target system.

Data Quality testing involves checking and validating :

* --Duplicate data
* --Rejected Records
* --Data Validation Rules
* --Data Integrity

**ETL Tester - Roles and Responsibilities**

* Understand logical flow of the application/ project.
* --Understand and review database design documents. (Logical model vs Physical model)
* --Understand and review source to target data mapping document.
* Analysis on business rules/data transformations/validation rules provided by Client.
* Create and maintain test plan document for ETL testing. (> 3 year)
* Identify the Test scenarios from mapping document for ETL testing.
* --After that we have to develop test cases from test scenarios for ETL testing.
* Create test data and write Sql queries for all test cases.
* Review the test cases.
* --After getting build from Development Team we are going to start executing the test cases.
* Run ETL Jobs to load source data in target system.
* Document test result and log the defects for failed test cases.
* --Retesting for failed test cases.
* Regression testing for code changes.(Defect fixes, Enhancements, New feature)
* Test closure.
* Sign Off from testing team to Project Manager and Development team.

**ETL Testing Challenges :**

• Data loss during the ETL process.

• Large volume of data.

• Invalid and duplicate data at target system.

• Many number of Source data stores.

• Source to target mapping information may not be provided to ETL Tester.

• Does not have permission to execute ETL code/Job.

• Unstable Testing environment.