Top 25 ETL Testing Interview Questions & Answers

Following are frequently asked questions in interviews for freshers as well experienced ETL tester and developer.

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 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 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 analyzed. 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. Portioning enables Informatica 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 portioning 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 Data Reader Destination Adapter?

The advantage of using the Data Reader 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 data Reader 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 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 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 Power Connect tool
- Import the source into the Source Analyzer. Between Informatica and SAP Power connect act as a gateway. 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 dimension 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