ETL Testing – Interview Questions

Previous Page Next Page

What do you understand by an ETL?

ETL stands for Extract, Transform, and Load. It is an important concept in Data Warehousing systems. **Extraction** stands for extracting data from different data sources such as transactional systems or applications. **Transformation** stands for applying the conversion rules on data so that it becomes suitable for analytical reporting. The **loading** process involves moving the data into the target system, normally a data warehouse.

Explain the 3-layer architecture of an ETL cycle.

The three layers involved in an ETL cycle are –

Staging Layer – The staging layer is used to store the data extracted from different source data systems.

Data Integration Layer – The integration layer transforms the data from the staging layer and moves the data to a database, where the data is arranged into hierarchical groups, often called **dimensions**, and into facts and aggregate facts. The combination of facts and dimensions tables in a DW system is called a **schema**.

Access Layer – The access layer is used by end-users to retrieve the data for analytical reporting.

What is the difference between and ETL and BI tools?

An ETL tool is used to extract data from different data sources, transform the data, and load it into a DW system. In contrast, a BI tool is used to generate interactive and adhoc reports for end-users, dashboard for senior management, data visualizations for monthly, quarterly, and annual board meetings.

Most common ETL tools include – SAP BO Data Services (BODS), Informatica, Microsoft – SSIS, Oracle Data Integrator ODI, Talend Open Studio, Clover ETL Open source, etc.

Most common BI tools include – SAP Business Objects, SAP Lumira, IBM Cognos, JasperSoft, Microsoft BI Platform, Tableau, Oracle Business Intelligence Enterprise Edition, etc.

What are the popular ETL tools available in the market?

The popular ETL tools available in the market are –

- Informatica – Power Center
- IBM – Websphere DataStage (Formerly known as Ascential DataStage)
- SAP – Business Objects Data Services BODS
- IBM – Cognos Data Manager (Formerly known as Cognos Decision Stream)
- Microsoft – SQL Server Integration Services SSIS
- Oracle – Data Integrator ODI (Formerly known as Sunopsis Data Conductor)
- SAS – Data Integration Studio
- Oracle – Warehouse Builder
- ABInitio
- Open source Clover ETL

Why do we need a staging area in an ETL process?

Staging area is an intermediate area that sits between data sources and data warehouse/data marts systems. Staging areas can be designed to provide many benefits, but the primary motivations for their use are to increase efficiency of ETL processes, ensure data integrity, and support data quality operations.

What is the difference between data warehousing and data mining?

Data warehousing is a broader concept as compared to data mining. Data mining involves extracting hidden information from data and interpret it for future predictions. In contrast data warehousing includes operations such as analytical reporting to generate detailed reports and ad-hoc reports, information processing to generate interactive dashboards and charts.

What are the structural differences between an OLTP and OLAP system?

OLTP stands for Online Transactional Processing system which is commonly a relational database and is used to manage day-to-day transactions.

OLAP stands for Online Analytical Processing system which is commonly a multidimensional system and is also called data warehouse.

What is a Dimension table and how is it different from a Fact table?

Suppose a company sells its products to customers. Every sale is a fact that takes place within the company and the fact table is used to record these facts. Each fact table stores the primary keys to join the fact table to dimension tables and measures/facts.

Example – Fact_Units

Cust_ID	Prod_Id	Time_Id	No. of units sold
101	24	1	25
102	25	2	15
103	26	3	30

A dimension table stores attributes or dimensions that describe the objects in a fact table. It is a set of companion tables to a fact table.

Example – Dim_Customer

Cust_id	Cust_Name	Gender
101	Jason	M
102	Anna	F

What is a Data Mart?

A data mart is a simple form of data warehouse and it is focused on a single functional area. It usually gets data only from a few sources.

Example – In an organization, data marts may exists for Finance, Marketing, Human Resource, and other individual departments which store data related to their specific functions.

What is an Aggregate function? Name a few common aggregate functions.

Aggregate functions are used to group multiple rows of a single column to form a more significant measurement. They are also used for performance optimization when we save aggregated tables in data warehouse.

Common Aggregate functions are –

MIN	returns the smallest value in a given column
MAX	returns the largest value in a given column
SUM	returns the sum of the numeric values in a given column
AVG	returns the average value of a given column
COUNT	returns the total number of values in a given column
COUNT(*)	returns the number of rows in a table

Example

```
SELECT AVG(salary)
FROM employee
WHERE title = 'developer';
```

> Explain the difference between DDL, DML, and DCL statements.

Data Definition Language (DDL) statements are used to define the database structure or schema.

Examples –

CREATE – to create objects in a database

ALTER – alters the structure of a database

Data Manipulation Language (DML) statements are used for manipulate data within database.

Examples –

SELECT – retrieves data from the a database

INSERT – inserts data into a table

UPDATE – updates existing data within a table

DELETE – deletes all records from a table, the space for the records remain

Data Control Language (DCL) statements are used to control access on database objects.

Examples –

GRANT – gives user's access privileges to database

REVOKE – withdraws access privileges given with the GRANT command

> What is an Operator in SQL? Explain common operator types.

Operators are used to specify conditions in an SQL statement and to serve as conjunctions for multiple conditions in a statement. The common operator types are –

- Arithmetic Operators
- Comparison/Relational Operators
- Logical Operators
- Set Operators
- Operators used to negate conditions

What are the common set operators in SQL?

The common set operators in SQL are –

- UNION
- UNION ALL
- INTERSECT
- MINUS

What is the difference between Minus and Intersect? What is their use in ETL testing?

Intersect operation is used to combine two SELECT statements, but it only returns the records which are common from both SELECT statements. In case of Intersect, the number of columns and datatype must be same. MySQL does not support INTERSECT operator. An Intersect query looks as follows –

```
select * from First
INTERSECT
select * from second
```

Minus operation combines result of two Select statements and return only those result which belongs to first set of result. A Minus query looks as follows –

```
select * from First
MINUS
select * from second
```

If you perform source minus target and target minus source, and if the minus query returns a value, then it should be considered as a case of mismatching rows.

If the minus query returns a value and the count intersect is less than the source count or the target table, then the source and target tables contain duplicate rows.

Explain 'Group-by' and 'Having' clause with an example.

Group-by clause is used with **select** statement to collect similar type of data. **HAVING** is very similar to **WHERE** except the statements within it are of an aggregate nature.

Syntax –

```
SELECT dept_no, count ( 1 ) FROM employee GROUP BY dept_no;
SELECT dept_no, count ( 1 ) FROM employee GROUP BY dept_no HAVING COUNT( 1 ) > 1;
```

Example – Employee table

Country	Salary
India	3000
US	2500
India	500
US	1500

Group by Country

Country	Salary
India	3000
India	500
US	2500
US	1500

What do you understand by ETL Testing?

ETL Testing is done before data is moved into a production Data Warehouse system. It is sometimes also called as Table Balancing or production reconciliation.

The main objective of ETL testing is to identify and mitigate data defects and general errors that occur prior to processing of data for analytical reporting.

How ETL Testing is different from database testing?

The following table captures the key features of Database and ETL testing and their comparison –

Function	Database Testing	ETL Testing
Primary Goal	Data validation and Integration	Data Extraction, Transform and Loading for BI Reporting
Applicable System	Transactional system where	System containing historical data

	business flow occurs	and not in business flow environment
Common Tools in market	QTP, Selenium, etc.	QuerySurge, Informatica, etc.
Business Need	It is used to integrate data from multiple applications, Severe impact.	It is used for Analytical Reporting, information and forecasting.
Modeling	ER method	Multidimensional
Database Type	It is normally used in OLTP systems	It is applied to OLAP systems
Data Type	Normalized data with more joins	De-normalized data with less joins, more indexes and Aggregations.

What are the different ETL Testing categories as per their function?

ETL testing can be divided into the following categories based on their function –

Source to Target Count Testing – It involves matching of count of records in source and target system.

Source to Target Data Testing – It involves data validation between source and target system. It also involves data integration and threshold value check and Duplicate data check in target system.

Data Mapping or Transformation Testing – It confirms the mapping of objects in source and target system. It also involves checking functionality of data in target system.

End-User Testing – It involves generating reports for end users to verify if data in reports are as per expectation. It involves finding deviation in reports and cross check the data in target system for report validation.

Retesting – It involves fixing the bugs and defects in data in target system and running the reports again for data validation.

System Integration Testing – It involves testing all the individual systems, and later combine the result to find if there is any deviation.

Explain the key challenges that you face while performing ETL Testing.

Data loss during the ETL process.

Incorrect, incomplete or duplicate data.

DW system contains historical data so data volume is too large and really complex to perform ETL testing in target system.

ETL testers are normally not provided with access to see job schedules in ETL tool. They hardly have access on BI Reporting tools to see final layout of reports and data inside the reports.

Tough to generate and build test cases as data volume is too high and complex.

ETL testers normally doesn't have an idea of end user report requirements and business flow of the information.

ETL testing involves various complex SQL concepts for data validation in target system.

Sometimes testers are not provided with source to target mapping information.

Unstable testing environment results delay in development and testing the process.

What are your responsibilities as an ETL Tester?

The key responsibilities of an ETL tester include –

Verifying the tables in the source system – Count check, Data type check, keys are not missing, duplicate data.

Applying the transformation logic before loading the data: Data threshold validation, surrogate ky check, etc.

Data Loading from the Staging area to the target system: Aggregate values and calculated measures, key fields are not missing, Count Check in target table, BI report validation, etc.

Testing of ETL tool and its components, Test cases – Create, design and execute test plans, test cases, Test ETL tool and its function, Test DW system, etc.

What do you understand by the term 'transformation'?

A transformation is a set of rules which generates, modifies, or passes data. Transformation can be of two types – Active and Passive.

What do you understand by Active and Passive Transformations?

In an active transformation, the number of rows that is created as output can be changed once a transformation has occurred. This does not happen during a passive transformation.

The information passes through the same number given to it as input.

What is Partitioning? Explain different types of partitioning.

Partitioning is when you divide the area of data store in parts. It is normally done to improve the performance of transactions.

If your DW system is huge in size, it will take time to locate the data. Partitioning of storage space allows you to find and analyze the data easier and faster.

Parting can be of two types – round-robin partitioning and Hash partitioning.

What is the difference between round-robin partitioning and Hash partitioning?

In round-robin partitioning, data is evenly distributed among all the partitions so the number of rows in each partition is relatively same. Hash partitioning is when the server uses a hash function in order to create partition keys to group the data.

Explain the terms – mapplet, session, mapping, workflow – in an ETL process?

A Mapplet defines the Transformation rules.

Sessions are defined to instruct the data when it is moved from source to target system.

A Workflow is a set of instructions that instructs the server on task execution.

Mapping is the movement of data from the source to the destination.

What is lookup transformation and when is it used?

Lookup transformation allows you to access data from relational tables which are not defined in mapping documents. It allows you to update slowly changing dimension tables to determine whether the records already exist in the target or not.

What is a surrogate key in a database?

A Surrogate key is something having sequence-generated numbers with no meaning, and just to identify the row uniquely. It is not visible to users or application. It is also called as Candidate key.

What is the difference between surrogate key and primary key?

A Surrogate key has sequence-generated numbers with no meaning. It is meant to identify the rows uniquely.

A Primary key is used to identify the rows uniquely. It is visible to users and can be changed as per requirement.

If there are thousands of records in the source system, how do you ensure that all the records are loaded to the target in a timely manner?

In such cases, you can apply the checksum method. You can start by checking the number of records in the source and the target systems. Select the sums and compare the information.

What do you understand by Threshold value validation Testing? Explain with an example.

In this testing, a tester validates the range of data. All the threshold values in the target system are to be checked to ensure they are as per the expected result.

Example – Age attribute shouldn't have a value greater than 100. In Date column DD/MM/YY, month field shouldn't have a value greater than 12.

Write an SQL statement to perform Duplicate Data check Testing.

```
Select Cust_Id, Cust_NAME, Quantity, COUNT (*)
FROM Customer GROUP BY Cust_Id, Cust_NAME, Quantity HAVING COUNT (*) >1;
```

How does duplicate data appear in a target system?

When no primary key is defined, then duplicate values may appear.

Data duplication may also arise due to incorrect mapping, and manual errors while transferring data from source to target system.

What is Regression testing?

Regression testing is when we make changes to data transformation and aggregation rules to add a new functionality and help the tester to find new errors. The bugs that appear in data which comes in Regression testing are called Regression.

Name the three approaches that can be followed for system integration.

The three approaches are – top-down, bottom-up, and hybrid.

What are the common ETL Testing scenarios?

The most common ETL testing scenarios are –

Structure validation

Validating Mapping document

Validate Constraints

Data Consistency check

Data Completeness Validation

Data Correctness Validation

Data Transform validation

Data Quality Validation

Null Validation

Duplicate Validation

Date Validation check

Full Data Validation using minus query

Other Test Scenarios

Data Cleaning

What is data purging?

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

What do you understand by a cosmetic bug in ETL testing?

Cosmetic bug is related to the GUI of an application. It can be related to font style, font size, colors, alignment, spelling mistakes, navigation, etc.

What do you call the testing bug that comes while performing threshold validation testing?

It is called Boundary Value Analysis related bug.

I have 50 records in my source system but I want to load only 5 records to the target for each run. How can I achieve this?

You can do it by creating a mapping variable and a filtered transformation. You might need to generate a sequence in order to have the specifically sorted record you require.

Name a few checks that can be performed to achieve ETL Testing Data accuracy.

Value comparison – It involves comparing the data in the source and the target systems with minimum or no transformation. It can be done using various ETL Testing tools such as

Source Qualifier Transformation in Informatica.

Critical data columns can be checked by comparing distinct values in source and target systems.

Which SQL statements can be used to perform Data completeness validation?

You can use Minus and Intersect statements to perform data completeness validation. When you perform source minus target and target minus source and the minus query returns a value, then it is a sign of mismatching rows.

If the minus query returns a value and the count intersect is less than the source count or the target table, then duplicate rows exist.

What is the difference between shortcut and reusable transformation?

Shortcut Transformation is a reference to an object that is available in a shared folder. These references are commonly used for various sources and targets which are to be shared between different projects or environments.

In the Repository Manager, a shortcut is created by assigning 'Shared' status. Later, objects can be dragged from this folder to another folder. This process allows a single point of control for the object and multiple projects do not have all import sources and targets into their local folders.

Reusable Transformation is local to a folder. **Example** – Reusable sequence generator for allocating warehouse Customer ids. It is useful to load customer details from multiple source systems and allocating unique ids to each new source-key.

What is Self-Join?

When you join a single table to itself, it is called Self-Join.

What do you understand by Normalization?

Database normalization is the process of organizing the attributes and tables of a relational database to minimize data redundancy.

Normalization involves decomposing a table into less redundant (and smaller) tables but without losing information.

What do you understand by fact-less fact table?

A fact-less fact table is a fact table that does not have any measures. It is essentially an intersection of dimensions. There are two types of fact-less tables: One is for capturing an event, and the other is for describing conditions.

What is a slowly changing dimension and what are its types?

Slowly Changing Dimensions refer to the changing value of an attribute over time. SCDs are of three types – Type 1, Type 2, and Type 3.

Previous Page Next Page

Enter email for newsletter go