ETL Testing Interview Questions

1. What is ETL testing?

Answer: ETL testing is validating data extraction, transformation and loading to ensure data integrity, accuracy and performance in a data warehouse environment. It checks if data is extracted correctly from source systems, transformed as per business rules and loaded into target system without data loss or corruption.

2. Explain ETL process.

Answer: ETL process consists of three steps:

- Extraction: Data is extracted from multiple source systems.
- Transformation: Extracted data is transformed into analysis ready format. This can include data cleaning, aggregation and enrichment.
- Loading: Transformed data is loaded into target data warehouse or database.

3. What are the types of ETL testing?

Answer: Types of ETL testing:

- Data Completeness Testing: All expected data is loaded into target system.
- Data Accuracy Testing: Data transformation rules are applied correctly.
- Data Quality Testing: Data accuracy, consistency and reliability.
- Performance Testing: ETL process performance and scalability.
- Regression Testing: New changes do not break existing functionality.

4. What are the challenges in ETL testing?

Answer:

- Data Volume: Handling big data is tough.
- Complex Transformations: Complex transformation logic is hard to test.
- Data Quality Issues: Data quality across multiple sources is tough.
- Performance Bottlenecks: Finding and fixing ETL performance issues.
- Tool Limitations: ETL tool limitations.

5. What is data validation in ETL testing?

Answer: Data validation in ETL testing is checking if data loaded into target system matches expected data. This includes data completeness, accuracy and consistency. Validation can be done at various stages of ETL process to ensure data integrity.

6. What is data mapping in ETL?

Answer: Data mapping is how source fields map to target fields. It involves specifying the transformation rules and logic to convert source data into target format. Data mapping is a key step in ETL process to ensure data is transformed correctly.

7. What is a Source List in ETL testing?

Answer: A Source List in ETL testing is a list of all the source systems and data sources from which data will be extracted. It includes details about the source systems, data formats and extraction methods. The Source List is used to plan and manage the data extraction process.

8. What tools are used for ETL testing?

Answer: ETL testing tools:

- Informatica: ETL tool
- Talend: Open source ETL
- Microsoft SSIS: SQL Server based ETL
- DataStage: ETL tool from IBM
- Pentaho: Open source ETL

9. How to handle performance issues in ETL testing?

Answer: To handle performance issues in ETL testing:

- Optimize Queries: Optimize SQL queries.
- Indexing: Indexing
- Parallel Processing: Process in parallel for large data.
- Batch Processing: Batch processing for large data.
- Resource Allocation: Ensure enough system resources (CPU, memory etc) for ETL process.

10. Why data quality in ETL testing?

Answer: Data quality is important in ETL testing as it ensures the data loaded into the target system is accurate, complete and consistent. Poor data quality can lead to incorrect business decisions, operational inefficiencies and loss of trust in the data warehouse. Data quality checks help to identify and resolve data issues early in the ETL process.

11. What is data transformation?

Answer: Data transformation is the process of converting data from its original format to a format suitable for analysis or reporting. This includes data cleaning, aggregation and enrichment. Data transformation ensures the data is consistent, accurate and ready to use in the target system.

12. What is a data warehouse?

Answer: A data warehouse is a central repository that stores data from multiple sources in a structured format, optimized for analysis and reporting. It allows organizations to consolidate data from various sources, ensures data consistency and supports business intelligence activities.

13. What is data lineage in ETL?

Answer: Data lineage is the tracking of data as it flows from source to ETL to target in the data warehouse. It provides a detailed view of the data transformation journey, helps to understand data origin, movement and transformation which is important for data governance and auditing.

14. Why staging areas in ETL?

Answer: Staging areas are intermediate storage locations where raw data is stored temporarily before it is transformed and loaded into target system. Staging areas allow data cleaning, transformation and validation without impacting source or target systems, ensures data accuracy and consistency.

15. How do you do data validation in ETL?

Answer: Data validation in ETL is:

- Source-to-Target Validation: Source data matches target data.
- Transformation Validation: Transformation rules are applied correctly.
- Data Quality Checks: Data is accurate, complete and consistent.
- Performance Testing: ETL performance and scalability.
- Error Handling: Errors in ETL.

16. What are the ETL testing errors?

Answer: ETL testing errors are:

- Data Loss: Data is missing in ETL.
- Data Duplication: Duplicate data in target system.
- Transformation Errors: Transformation rules not applied correctly.
- Performance Issues: ETL performance or scalability issues.
- Integration Issues: Issues with data from different sources.

17. How do you ensure data completeness in ETL testing?

Answer: To ensure data completeness in ETL testing:

- Source-to-Target Comparison: Compare source and target data to ensure all data is loaded.
- Row Count Verification: Verify source and target row count.
- Data Profiling: Profile data to identify and resolve data completeness issues.

18. What is data profiling?

Answer: Data profiling is analyzing data from multiple sources to understand the data structure, content and quality. It helps to identify data quality issues like missing

values, duplicates and inconsistencies and is a part of ETL process to ensure data accuracy and completeness.

19. What is a surrogate key in ETL?

Answer: A surrogate key is a unique identifier for each record in a data warehouse, usually a sequential number, used for data integrity and query performance. Not derived from application data and is the primary key in dimension tables.

20. How do you handle data quality issues in ETL testing?

Answer: To handle data quality issues in ETL testing:

- Data Cleansing: Fix the errors in the data.
- Validation Rules: Apply the validation rules.
- Monitoring: Monitor data quality throughout the ETL process.
- Error Logging: Log the data quality issues and resolutions.

21. What is incremental load in ETL?

Answer: Incremental load is updating the target data warehouse with only the new or changed data since the last load. This is more efficient and reduces processing time and resource utilization compared to full load.

22. What is SCD (Slowly Changing Dimensions)?

Answer: Slowly Changing Dimensions (SCD) are dimensions that change slowly over time, not on a schedule. There are three types:

- SCD Type 1: Overwrite old data with new data.
- SCD Type 2: Create a new record with a new surrogate key for each change.
- SCD Type 3: Track changes with separate columns for old and new data.

23. What is the role of ETL scheduler?

Answer: ETL scheduler automates the execution of ETL processes at specific times or intervals. It ensures data is up to date, reduces manual intervention and improves ETL efficiency and reliability.

24. How do you manage metadata in ETL?

Answer: Metadata management in ETL is capturing and maintaining information about the data, such as source and target data definitions, transformation rules, data lineage and data quality metrics. Good metadata management provides transparency, traceability and consistency in ETL.

25.What are Cubes and OLAP Cubes?

Cubes are data processing units made up of fact tables and dimension tables from the data warehouse. They provide multi-dimensional analysis.

OLAP stands for ‘Online Analytics Processing’ and OLAP Cubes store large amount of data in multi-dimensional form for reporting. They have facts called ‘measures’ grouped by dimensions.