ETL Developer Interview Questions and Answers

Introduction

Are you getting ready for a job interview in data engineering or ETL? With this collection of the most common **ETL interview questions and answers** on a range of ETL tools, you may be ready for your upcoming ETL job interview. Excel in data engineering with our ETL Developer course in Chennai.

List of ETL Developer Interview Questions for Freshers

1. What is an ETL process?
2. How many steps are involved in an ETL process?
3. What do initial and full loads mean?
4. What features do snapshots have, and what does their significance entail?

5. Which applications is PowerCenter compatible with?
6. What distinguishes PowerMart from PowerCenter?
7. Define views.
8. What does incremental load mean?
9. What does partitioning in ETL mean?
10. Does PowerMart offer links to sources of ERP data?

Explore our ETL Testing course syllabus to get started.

ETL Developer Interview Questions and Answers for Freshers

1. What is an ETL process?

ETL is abbreviated as extraction, transformation, and loading.

2. How many steps are involved in an ETL process?

In its most basic form, data extraction, transformation, and loading comprise the ETL process. Even so, the acronym represents a clear and sequential three-step process: extract, transform, and load.

3. What do initial and full loads mean?

The process of filling every data warehousing table for the first time is known as the first load in ETL. Depending on the volume of the data, all set records are loaded at once during full load, when the data is loaded for the first time. It would reload the most recent data and remove everything from the table.

4. What features do snapshots have, and what does their significance entail?

The read-only data that is kept in the master table is duplicated in snapshots.

To record changes to the master table, snapshots are stored on remote nodes and are frequently refreshed. They are duplicates of tables as well.

5. Which applications are PowerCenter compatible with?

SAP, Oracle Apps, PeopleSoft, and other ERP sources can be integrated with PowerCenter.

6. What distinguishes PowerMart from PowerCenter?

While Power Mart handles modest volumes of data, PowerCenter processes huge volumes of data.

7. Define views.

One or more tables' properties are used to build views. It is possible to update a view with a single table, but not one with many tables.

8. What does incremental load mean?

Applying dynamic modifications as needed within predetermined timeframes is known as incremental load.

9. What does partitioning in ETL mean?

In ETL, partitioning is the splitting of the transactions into smaller groups to boost efficiency.

10. Does PowerMart offer links to sources of ERP data?

Not at all! PowerMart does not provide ERP source connectors.

Explore our ETL Tutorial for Beginners and get started with your data engineering career.

List of ETL Developer Interview Questions for Experienced

1. What does a materialized view log entail, and what exactly is a materialized view, in your opinion?
2. Describe the idea of data skewness in ETL procedures.
3. In terms of performance, which is preferable: joining data first and then filtering it, or filtering it first and then joining it with other sources?
4. In ETL pipelines, how would you manage incremental data loading?
5. In ETL workflows, how would you manage exceptions and errors?
6. How would you create an ETL architecture that is both fault-tolerant and scalable?
7. When source schemas alter over time, how would you manage schema evolution in ETL pipelines?
8. In ETL, how are the tables analyzed?
9. How would you set up the ETL process's logging?
10. Could you define OLAP cubes and cubes?

ETL Developer Technical Interview Questions and Answers for Experienced

1. What does a materialized view log entail, and what exactly is a materialized view, in your opinion?

A table that records modifications to the basis tables used in a materialized view is called a materialized view log. A pre-computed aggregate table with combined or condensed data from fact and dimension tables is called a materialized view.

2. Describe the idea of data skewness in ETL procedures.

Data skewness is the term for an imbalance in the distribution of data among processing nodes or partitions, which can lead to issues with the performance of ETL processes. An imbalance occurs when certain keys or values occur far more frequently than others.

As a result, some nodes get overloaded with data while others remain unused. Techniques like data splitting, data shuffling, and the use of sophisticated processing frameworks like Apache Spark can all be used to address data skewness.

3. In terms of performance, which is preferable: joining data first and then filtering it, or filtering it first and then joining it with other sources?

- Before joining data with other sources, it is preferable to filter the data.
- Eliminating unneeded data as early in the process as possible is an excellent strategy to boost the performance of the ETL process. It cuts down on the amount of time needed for memory processing, I/O, and/or data transfer.
- As a general rule, fewer rows should be processed and data that never reaches the objective should not be transformed.

4. In ETL pipelines, how would you manage incremental data loading?

Incremental data loading refreshes the new or modified data since the last ETL run, as opposed to reloading the entire dataset.

This can be done by noting the last successful ETL run and utilizing timestamps or metadata to determine the delta changes that have happened subsequently.

Strategies including database triggers, timestamp comparisons, and CDC (Change Data Capture) can be used to detect incremental changes and load them into the target system.

5. In ETL workflows, how would you manage exceptions and errors?

ETL operations need to handle errors and exceptions to ensure data dependability and integrity.

To do this, robust error-handling strategies can be applied

- Adding alerting mechanisms to notify stakeholders about severe failures,
- Logging error details for troubleshooting,

- Setting up checkpoints to resume from the point of failure,
- Retrying methods for transient issues.

Moreover, fault-tolerant processing frameworks like Apache Airflow or Apache NiFi can help with ETL workflow management.

6. How would you create an ETL architecture that is both fault-tolerant and scalable?

When creating a scalable and fault-tolerant ETL architecture, many considerations need to be made, including:

- Implementing distributed processing
- Utilizing fault-tolerant storage systems
- Arranging for horizontal scalability
- Incorporating redundancy and failover techniques.

Using microservices design, cloud-based ETL services, containerization technologies like Docker and Kubernetes, data replication, and backup plans are some of the strategies that can help with building a strong ETL architecture.

7. When source schemas alter over time, how would you manage schema evolution in ETL pipelines?

The practice of managing changes to the structure or schema of the source data over time is known as schema evolution. Among the techniques for controlling schema evolution in ETL pipelines are schema inference, schema mapping, and schema evolution policies.

Furthermore, flexible data models like schema-on-read, data serialization formats like Avro or Parquet, and schema versioning approaches can be used to adapt changes to source schemas without affecting downstream processes.

8. In ETL, how are the tables analyzed?

A cost-based optimizer uses the statistics produced by the '*ANALYZE*' statement to determine the most economical strategy for data retrieval. The ANALYZE statement can be used to support the system's object structure validation and space management.

COMPUTER, ESTIMATE, and DELETE are examples of operations.

Example Oracle 7 Code:

select OWNER,

sum(decode(nvl(NUM_ROWS,9999), 9999,0,1)) analyzed,

sum(decode(nvl(NUM_ROWS,9999), 9999,1,0)) not_analyzed,

count(TABLE_NAME) total

from dba_tables

where OWNER not in ('SYS', 'SYSTEM')

group by OWNER

To obtain table information, this software runs a SQL query against the database's `dba_tables` view. The goal is to count the number of analyzed and unanalyzed tables for each owner by analyzing the data in the tables.

9. How would you set up the ETL process's logging?

To monitor all changes and failures during a load, logging is crucial. Using flat files or logging tables is the most popular method for getting ready for logging. In other words, counts, timestamps, and metadata related to the source and target are added during the operation and subsequently deposited into a table or flat file.

To accomplish that in an ETL process, a developer would utilize variables (such as the Transact-SQL system variable @@ROWCOUNT) or event handlers (SSIS) to track inserted, updated, and deleted entries.

The SSISDB database allows us to monitor each processed package while utilizing SSIS:

```
SELECT TOP 10000   o.Object_Name            [Project Name]

    ,REPLACE(e.package_name    , '.dtsx', '') [Package Name]

    ,o.start_time                  [Start Time]

    ,o.end_time                    [End Time]

    ,e.message_source_name              [Message Source Name]

    ,e.event_name                  [Event Name]

    ,e.subcomponent_name               [Subcomponent Name]

    ,e.message_code                [Message Code]

    ,m.message_time                [Event Time]

    ,m.message                     [Error Message]

    ,m.message_time                [Error Date]

    ,o.caller_name                 [Caller Name]

    ,o.Stopped_By_Name                 [Stopped By]

    ,ROW_NUMBER() OVER (PARTITION BY m.operation_id
ORDER BY m.message_source_type DESC [Source Type Order]

FROM   SSISDB.internal.operations o
```

JOIN SSISDB.internal.operation_messages m

ON o.operation_id = m.operation_id

JOIN SSISDB.internal.event_messages e

ON m.operation_id = e.operation_id

AND m.operation_message_id = e.event_message_id

WHERE o.Object_Name LIKE '%[object name]%' — database

AND event_name LIKE '%Error%'

ORDER BY o.end_time DESC

,o.Object_Name

,[Source Type Order] ASC

ETL tools provide native alerting and logging features, such as a dashboard showing the current load status, in addition to flat files and the database itself.

10. Could you define OLAP cubes and cubes?

The cube is an essential component and plays a major part in data processing. Cubes are essentially the Data Warehouse's data processing units; they include dimensions and fact tables. They help clients by providing a multifaceted view of the data in addition to searching and analyzing it.

However, software called Online Analytical Processing (OLAP) makes it possible to analyze data from several databases at once. An OLAP cube can be used to store data in a multidimensional format for reporting purposes.

Cubes simplify the process of creating and reading reports, which improves and streamlines the reporting procedure. End users are in charge of maintaining and managing these cubes, which means they have to manually update the data within.

Conclusion

We hope that our list of ETL interview questions and answers will be helpful to you as you get ready for data science jobs. Enroll in our **software training institute in Chennai** and earn IBM certification that helps you launch your career with in-demand IT skills.