Tredence Data Engineer Interview Guide – Experienced 3+

Round 1: Database Concepts and Azure Fundamentals

This round assessed foundational database knowledge and the ability to work with Azure's data ecosystem. The questions were scenario-based and focused on evaluating a candidate's understanding of data models, indexing, metadata management, and file formats.

Topics Covered

1. **Data Modeling**

   Snowflake vs Star Schema vs 3NF: Compare their advantages and appropriate use cases.

2. **Indexing and Query Optimization**

   Clustered vs Non-Clustered Index: Explain the differences and use cases.

   Indexing in Databases: Discuss how indexing improves query performance.

3. **Metadata and Temporary Data Structures**

   Where is Metadata Stored for Internal Tables? Explain the storage mechanisms.

   Common Table Expressions (CTE): Clarify whether CTEs persist data or act as temporary structures.

4. **Azure File Formats**

   Delta Format vs Parquet Format: Compare in terms of performance and use cases.

   Storage in Delta Lake: Explain the file format used in Delta Lake.

Questions Asked

1. How would you decide between a Snowflake schema and a Star schema for a reporting database?
2. Explain the difference between clustered and non-clustered indexes. Provide a scenario where each would be beneficial.
3. What is the purpose of Delta format, and how does it differ from Parquet in terms of storage and querying?
4. Describe how metadata is stored and accessed for internal tables in a relational database.
5. Does a Common Table Expression store data? If not, how does it function in SQL?

Round 2: PySpark and Advanced SQL

This round focused on practical implementation and coding skills in PySpark and SQL. Candidates were required to solve data transformation problems and optimize queries.

Topics Covered

1. **SQL Query Writing**

   Finding the 2nd Highest Salary: Multiple solutions to solve the same query, demonstrating a thorough understanding of SQL techniques.

2. **PySpark Data Transformation**

   Transforming input data to add derived columns such as Age, FirstName, and LastName.

3. **Performance Optimization**

   Caching in PySpark: Discuss caching techniques and their impact on distributed computing performance.

4. **Scenario-Based Questions**

   Using PySpark to handle real-world data engineering challenges.

Questions Asked

1. **SQL Challenge:**

   Find the second-highest salary in the employees table using three different methods:

Solution 1:

```sql
SELECT id, name, sal
FROM (SELECT *, DENSE_RANK() OVER (ORDER BY sal DESC) AS rank_sal
      FROM employees)
WHERE rank_sal = 2;
```

Solution 2:

```sql
SELECT *
FROM employees
WHERE sal NOT IN (SELECT e.sal
                  FROM employees e, employees e1
                  WHERE e.sal > e1.sal)
ORDER BY sal DESC
LIMIT 1;
```

Solution 3:

```sql
SELECT MAX(sal)
FROM employees
WHERE sal < (SELECT MAX(sal) FROM employees);
```

2. **PySpark Coding Challenge:** Transform the following input dataset:

   Input Columns: id, dob, name
   Output Columns: id, dob, age, name, firstname, lastname
   Input Example:
   id: 1, dob: 1/Jan/1987, name: Nitheesh Pranesh

   Solution Code:

```
from pyspark.sql.functions import year, current_date, split, col

df = df.withColumn('Age', (year(current_date()) - year(col('dob'))).cast('int')) \
       .withColumn('FirstName', split(col('name'), ' ')[0]) \
       .withColumn('LastName', split(col('name'), ' ')[1])
```

3. What is the advantage of caching in PySpark? When and why would you use it?
4. How would you optimize a SQL query for better performance when working with large datasets?
5. Write a PySpark script to process data stored in Delta format and transform it into Parquet.

Glassdoor Tredence Review –

https://www.glassdoor.co.in/Reviews/Tredence-Reviews-E1141078.htm

Tredence Careers –

https://www.tredence.com/company-careers.

Subscribe to my YouTube Channel for Free Data Engineering Content –

https://www.youtube.com/@shubhamwadekar27

Connect with me here –

https://bento.me/shubhamwadekar

Checkout more Interview Preparation Material on –

https://topmate.io/shubham_wadekar