Presidio Data Engineer Interview Guide – Experienced 3+

Technical round one and two combined.

1. How to Create Clustered and Non-Clustered Index – Syntax and Examples

- **Clustered Index** sorts the data rows in the table based on the key columns.
  Syntax:

```
CREATE CLUSTERED INDEX idx_name ON table_name(column_name);
```

- **Non-Clustered Index** creates a separate structure with pointers to the data rows.
  Syntax:

```
CREATE NONCLUSTERED INDEX idx_name ON table_name(column_name);
```

2. Difference Between Truncate/Delete and Union/Union All – Performance and Usage

- **Truncate**: Deletes all rows without logging each row, faster, cannot use WHERE.
- **Delete**: Deletes rows with logging and supports WHERE clause.
- **Union**: Removes duplicates, slower.
- **Union All**: Keeps duplicates, faster.

3. Database vs Data Warehouse vs Data Mart vs Data Lake

Feature	Database	Data Warehouse	Data Mart	Data Lake
Purpose	Transactional processing	Analytical reporting	Subset of Data Warehouse	Storage for raw, unprocessed data
Schema	Structured	Structured	Structured	Schema-on-read
Usage	OLTP	OLAP	Department-level analysis	Big data processing

4. Copy Large Files from On-Premises to Azure in ADF

- Use **Azure Data Box** or **AzCopy** for high-throughput transfer.
- In ADF, use **self-hosted IR** with **parallel copy activity** and **block blob storage** for large files.

5. Stored Procedure vs Function

Feature	Stored Procedure	Function
Return Value	Can return multiple values	Returns only one value
Usage	Used for complex logic	Used for computations
Syntax	CREATE PROCEDURE	CREATE FUNCTION

6. ACID Properties

- **Atomicity**: Ensures all or none of the transactions are executed.
- **Consistency**: Maintains database integrity before and after the transaction.
- **Isolation**: Transactions occur independently.
- **Durability**: Data persists after transaction completion.

Example: A bank transfer involving two accounts.

7. How to Connect to Salesforce Without Typing Credentials Manually

Use OAuth for secure authentication:

1. Create a connected app in Salesforce.
2. Use OAuth tokens to access APIs instead of passwords.

Code Interview Questions

1. SQL Query Analysis

```sql
SELECT first_name, last_name
FROM customers
WHERE (first_name LIKE 'R%' AND first_name LIKE '%t')
   OR (last_name LIKE 'R%' AND last_name LIKE '%t')
ORDER BY first_name DESC;
```

Explanation:

- Finds names starting with **'R'** and ending with **'t'**.
- Sorts the result by first_name in descending order.

2. PySpark Code Examples

- **Read CSV**:

```
df = spark.read.csv("path", header=True, delimiter=',', inferSchema=True)
```

- **Concatenate Columns**:

```
df = df.withColumn('Name', concat(col('first_name'), col('last_name')))
```

- **Partition and Save as Parquet**:

```
df.write.format('parquet').partitionBy('Age')
```

- **Add Row Numbers**:

from pyspark.sql.window import Window

df = df.withColumn('row_id',
row_number().over(Window.partitionBy('Dept').orderBy(col('sal').desc())))

df_filtered = df.filter(col('row_id') == 2)

3. Executor vs Driver

- **Driver**: Coordinates tasks, maintains metadata, and collects results.
- **Executor**: Executes tasks and performs computations.

4. Spark Architecture

Components include **Driver**, **Executors**, **Cluster Manager**, and **Tasks**.

5. Types of Integration Runtime in ADF

- **Auto-resolve**: Automatically managed by Azure.
- **Self-hosted**: For on-premises data.
- **Azure**: Native Azure data transfer.

6. On-Premises to Cloud Integration Runtime

- Use **self-hosted IR** for secure connectivity.

7. What is Integration Runtime?

It provides compute resources for data movement and transformation in ADF.

8. Broadcast Join in PySpark

Used when one table is small to avoid shuffling:

df_result = df_large.join(broadcast(df_small), "key")

9. Coalesce vs Repartition

- **Coalesce** reduces partitions (without shuffle).
- **Repartition** increases partitions (with shuffle).

10. Managed Table vs External Table

Managed Table	External Table
Data stored in Hive	Data stored externally
Dropping table deletes data	Data remains intact

11. Python Code for Prime Numbers

```python
def is_prime(num):
    for i in range(2, int(num**0.5) + 1):
        if num % i == 0:
            return False
    return True

print([x for x in range(2, 100) if is_prime(x)])
```

12. What is Parquet Format? Advantages Over CSV

Parquet is a **columnar storage format** designed for storing and processing large datasets efficiently. It is commonly used in big data frameworks like Apache Spark and Hive. Unlike row-based formats like CSV, Parquet optimizes for both storage efficiency and query performance.

Advantages of Parquet over CSV

1. **Columnar Storage**: Parquet stores data column-wise, making it more efficient for queries that access only a subset of columns.
2. **Compression**: It supports built-in compression algorithms (e.g., Snappy, Gzip), reducing storage costs significantly.
3. **Schema Support**: Parquet includes schema information with metadata, enabling easier data management and reducing errors.
4. **Efficient I/O**: It reads only the relevant columns needed for a query, minimizing disk I/O and improving performance.
5. **Data Type Support**: Parquet supports complex nested data structures and multiple data types better than CSV.

Example Usage in PySpark

```
df = spark.read.csv("data.csv", header=True, inferSchema=True)
df.write.parquet("data_parquet")
```

By converting a CSV to Parquet, you reduce storage requirements and gain performance benefits for large-scale analytics.

13. Activities in Azure Data Factory (ADF)

- **Copy Activity**: Transfers data from a source to a destination.
- **ForEach Activity**: Iterates over a collection of items and executes child activities.
- **Data Flow Activity**: Provides data transformation capabilities with a visual interface.

14. How to Run a Notebook in ADF

Steps to integrate Databricks notebooks in ADF:

1. Create a **Databricks Linked Service** in ADF.
2. Add a **Notebook Activity** in a pipeline.
3. Configure the activity to point to the linked service and specify the notebook path and parameters.

Example:

```
{
  "name": "RunNotebook",
  "type": "DatabricksNotebook",
  "linkedServiceName": "AzureDatabricksLinkedService",
  "notebookPath": "/path/to/notebook",
  "parameters": {
    "param1": "value"
  }
}
```

Glassdoor Presidio Review –

https://www.glassdoor.co.in/Reviews/Presidio-Reviews-E3340.htm

Presidio Careers –

https://www.presidio.com/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