Nihilent Azure Data Engineer Interview Guide – Experienced 3+

Round 1 & Round 2 – Technical

Key Questions and Topics:

1. Introduction and Experience

- Share your journey as a Data Engineer.
- Highlight the tools and technologies you've used in your current project.
- Discuss the challenges you've faced in your projects and the lessons learned.
- Explain your roles and responsibilities in your current project.

2. Azure Data Factory (ADF)

- **Types of triggers in ADF:**
  Discuss the types of triggers, such as tumbling window triggers, schedule triggers, and event-based triggers, along with their practical use cases.
- **Fault tolerance in ADF pipelines:**
  Explain strategies to achieve fault tolerance, including retry policies, activity dependencies, and robust logging mechanisms.
- **Difference between pipelines and data flows:**
  Pipelines are workflows defining the sequence of activities, while data flows allow for data transformations within ADF.
- **Types of Integration Runtimes (IR):**
  Explain self-hosted IR, Azure IR, and SSIS IR with their use cases.

3. Spark Architecture

- Explain the components of Spark Architecture:
  Driver, Executors, Cluster Manager, and how they interact during job execution.

4. PySpark and Spark

- **Reading data:**
  - Explain how to read a Parquet or CSV file into a DataFrame using spark.read.parquet() or spark.read.csv().
- **Catalyst Optimizer:**
  Discuss its role in query optimization and how it transforms logical plans into optimized physical plans.
- **Accumulators:**
  Describe their use as a shared variable for write-only operations and how they help in monitoring or debugging tasks.

- **Optimization techniques:**
  Explain techniques like partitioning, caching, broadcast joins, and bucketing for improving performance in Spark.
- **Repartition vs. Coalesce:**

  Repartition: Increases the number of partitions and shuffles data across nodes.

  Coalesce: Reduces the number of partitions without a shuffle.
- **Broadcast join:**
  Explain how it optimizes joins by broadcasting a smaller dataset to all nodes, and its use cases in scenarios with unevenly sized datasets.

5. SQL

- **Second-highest salary:**
  Write a query using LIMIT, OFFSET, or ROW_NUMBER().

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

- **Removing duplicates:**
  Use ROW_NUMBER() or DISTINCT to handle duplicate entries.

```sql
DELETE FROM table_name
WHERE id NOT IN (
  SELECT MIN(id)
  FROM table_name
  GROUP BY column_name);
```

- **Joins and window functions:**
  Explain types of joins (INNER, LEFT, RIGHT, FULL OUTER) and window functions like ROW_NUMBER(), RANK(), and DENSE_RANK().

6. Python

- **Schema evolution:**
  Discuss techniques for handling schema changes in PySpark using options like mergeSchema or schema inference during read operations.
- **Python libraries:**
  Mention libraries like Pandas, NumPy, Matplotlib, and how you've used them in data processing and analysis tasks.
- **Writing Excel sheets to Delta tables:**
  Use Pandas to read multiple Excel sheets and convert them into Delta tables in Databricks.

7. Azure Databricks

- **Running multiple notebooks:**
  Discuss methods like using dbutils.notebook.run() to execute multiple notebooks in sequence or concurrently.
- **Databricks notebooks vs. Fabric notebooks:**
  Highlight differences in performance, integration, and features.
- **Unity Catalog:**
  Explain its role in managing and securing data across Databricks workspaces.
- **Spark engine:**
  Describe the distributed computing framework of Spark and its advantages in processing big data.

8. Data Warehouse and Fabric

- **Serverless vs. Dedicated SQL pools:**
  Serverless pools are on-demand, while dedicated pools are provisioned for specific workloads.
- **Fabric pipelines vs. ADF pipelines:**
  Discuss the differences in capabilities, flexibility, and integration.
- **Fabric dataflows vs. ADF dataflows:**
  Compare their features and preferred use cases.
- **Lakehouse vs. Warehouse:**
  Lakehouses unify the storage of structured and unstructured data, whereas warehouses are optimized for structured data.

9. Agile Methodologies

- Explain your experience working in Agile environments, focusing on sprint planning, standups, and retrospectives.

10. Azure DevOps and CI/CD

- **CI/CD process:**
  Discuss how you've used Azure DevOps for version control, automated deployments, and monitoring.

Round 3 – Managerial

Key Questions:

1. Professional Background

- Provide a detailed walkthrough of your career journey.
- Describe your ownership and collaboration roles in your current project.

2. Project Collaboration

- Share examples of successful stakeholder communication and how you resolved conflicts or ensured alignment.

3. Analytical Thinking

- Discuss designing a data pipeline for a specific use case, addressing challenges and optimizations.
- Explain techniques for ensuring data quality in cross-functional team scenarios.

4. Team Management

- Highlight your experience in mentoring junior team members and leading initiatives within the team.

Round 4 – HR Discussion

Key Topics:

- Discussed salary expectations based on market standards and project value.
- Negotiated notice period and relocation preferences to ensure smooth onboarding.

Glassdoor Nihilent Review –

https://www.glassdoor.co.in/Reviews/Nihilent-Reviews-E30452.htm

Nihilent Careers –

https://www.virtusa.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