

## Deloitte Data Engineer Interview Guide – Experienced 5+

### **Role - Databricks/PySpark Developer**

#### Round 1: Technical Interview

##### **SQL Query: Managers with at least 7 employees reporting to them**

###### **Question:**

Find the names of managers who have at least 7 employees directly reporting to them.

###### **Answer:**

```
SELECT emp_name
FROM employees
WHERE emp_id IN (
    SELECT manager_id
    FROM employees
    GROUP BY manager_id
    HAVING COUNT(emp_id) >= 7
);
```

##### **SQL Query: Rows with the highest scores for each student in a year**

###### **Question:**

Fetch the rows with the highest scores for each student in a year.

###### **Answer:**

```
WITH RankedScores AS (
    SELECT name, year, scores,
           ROW_NUMBER() OVER (PARTITION BY name ORDER BY scores DESC) AS rank
    FROM scores_table
)
SELECT name, year, scores
FROM RankedScores
WHERE rank = 1;
```

## SQL Query: Using LAG Function to Find Previous Year's Scores

### Question:

Find each student's previous year's scores using the LAG function.

### Answer:

```
SELECT name, year, scores,
       LAG(scores) OVER (PARTITION BY name ORDER BY year) AS prev_year_scores
  FROM scores_table;
```

## SQL Query: Aggregate Surface Areas and Calculate Cumulative Surface Area

### Question:

Aggregate surface areas and calculate cumulative surface area using the LAG function.

### Answer:

```
SELECT continent, surface_area,
       surface_area + COALESCE(
           LAG(surface_area) OVER (PARTITION BY continent ORDER BY surface_area), 0
       ) AS cumulative_surface_area
  FROM surface_areas_table;
```

## Azure Data Factory Scenarios

### Key Concepts:

1. **GetMetadata**: Used to fetch information about files, folders, and data sets (e.g., structure, size, last modified time).
2. **ForEach**: Iterates over collections (e.g., lists of files or records) and processes them individually.
3. **Copy Data**: Facilitates data movement from source to destination, supporting a wide range of formats and protocols.

## Round 2: Managerial Round

### **1. Self-Introduction and Career Background**

Focus on a concise summary of your expertise, key projects, and tools you've mastered.

Example:

*"I'm a data engineer with 5 years of experience specializing in building scalable pipelines using Spark and Azure services. My recent project involved migrating a 2TB Oracle database to BigQuery while reducing query latency by 30%."*

### **2. Discussion on Previous Projects (Spark Optimization)**

- Highlight specific optimizations like:
  - Partitioning and bucketing for large datasets.
  - Caching frequently accessed datasets.
  - Avoiding shuffles by reusing RDDs or DataFrames.
- Example: *"I optimized a Spark job by partitioning data based on query patterns, reducing shuffle write time by 50%."*

### **3. Specific to Azure Services**

- Talk about tools like Azure Data Factory, Synapse Analytics, and Databricks.
- Example: *"In Azure Data Factory, I used dynamic content and parameterization to automate pipeline creation, reducing manual intervention."*

## Round 3: Director Interview (Data & AI Unit)

### **1. In-Depth Questions on Data Engineering Projects**

Discuss impactful projects with measurable outcomes. Example:

*"In a real-time analytics project, I integrated Kafka with Spark Streaming to process 500k events per second and stored results in Snowflake for dashboarding."*

### **2. SQL Queries Focused on GROUP BY Scenarios and Aggregate Functions**

- Example Question: *"Find the total and average revenue for each product category in 2024."*

**Answer:**

```
SELECT category, SUM(revenue) AS total_revenue, AVG(revenue) AS avg_revenue
FROM sales
WHERE year = 2024
GROUP BY category;
```

### **3. Scenario-Based Questions on Databricks Optimization**

- Example Question: *How would you optimize a slow-running notebook in Databricks?*

**Answer:**

1. Use cluster with appropriate instance types and autoscaling.
2. Partition and cache frequently accessed data.
3. Use Delta Lake for ACID transactions and efficient queries.
4. Enable adaptive query execution to optimize join strategies.

**Glassdoor Deloitte Review –**

<https://www.glassdoor.co.in/Reviews/Deloitte-Reviews-E2763.htm>

**Deloitte Careers –**

<https://jobsindia.deloitte.com/>

**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](https://topmate.io/shubham_wadekar)