Bristol Myers Squibb Data Engineer Interview Guide – Experienced 3+

Interview Process Breakdown

Round 1: Technical (Live Coding)

Objective: Assess hands-on expertise in SQL, Python, PySpark, Spark optimization, and AWS technologies through live problem-solving tasks.

Round 2: Techno-Managerial

Objective: Evaluate the ability to optimize data workflows, analyze past projects, and handle situational questions.

Round 3: HR Discussion

Objective: Negotiate salary, benefits, and confirm cultural fit.

Detailed Insights on Each Round

Round 1: Technical (Live Coding)

The technical round tested both breadth and depth of technical skills. Here's what it covered:

- **SQL Proficiency:**
  - Focused on window functions (e.g., ROW_NUMBER, RANK, and PARTITION BY) and their applications in scenarios like ranking, deduplication, and aggregation.
  - Output-based questions required writing SQL queries to produce given results. Example: Transforming a raw dataset into a formatted summary using GROUP BY and CASE.
  - Handling null values during joins with SQL functions like COALESCE to replace nulls or ensure correct join conditions.
  - Counting records with nulls in joins and numerical comparisons required solid understanding of LEFT JOIN and NULL behavior.
- **Python Coding:**
  - Questions revolved around basic string manipulations (e.g., reversing strings, case transformations) and array operations (e.g., finding subarrays, sorting).
  - Medium-difficulty Python problems focused on logical flow and performance.
- **PySpark Skills:**
  - Writing PySpark code to:
    - Pass a schema explicitly when reading data using StructType and StructField.
    - Add a new column based on conditions using withColumn() and when() functions.

    - Discussed Spark optimization techniques like using broadcast joins, caching intermediate results, and partitioning strategies to minimize shuffling.

- **Spark Architecture:**

  Questions on Spark's internal working, including DAG creation, lazy evaluation, shuffling, and job execution stages.

- **AWS Questions:**

  Covered services like Glue (ETL workflows), Lambda (event-driven execution), S3 (data storage), and cross-questioning about real-world project implementations using these tools.

Example Question:
Write a PySpark script to filter out invalid records from a dataset and calculate the average for a specific column, ensuring the schema is strictly defined at runtime.

Round 2: Techno-Managerial

This round combined technical depth with situational and managerial insights.

- **SQL Optimization:**

  Questions explored strategies like using indexes, avoiding nested subqueries, and choosing appropriate joins for faster performance.

- **Spark Optimization:**

  Discussion on reducing data shuffling, repartitioning large datasets, and managing task parallelism for better Spark job efficiency.

- **Pandas and Numpy:**

  Focused on transformation functions such as apply(), groupby(), and array manipulation using Numpy for data preprocessing.

- **Project Analysis:**
    - In-depth discussion of prior data engineering projects, with questions about architectural decisions, tools used, and challenges faced.
    - Example: "Why did you choose Spark for processing this dataset instead of a traditional database?"

- **Situation-Based Questions:**
    - Non-technical scenarios like resolving team conflicts, managing project delays, and prioritizing tasks.
    - Example: If a critical pipeline breaks an hour before delivery, how would you handle it?

- **Company-Specific Question:**

  Why are you interested in joining BMS? They emphasized unique, thoughtful responses that reflected research about the company and alignment with its goals.

Round 3: HR Discussion

- Focused on salary negotiation, explaining the variable component, and discussing long-term growth opportunities within the company.

Key Tip: Be prepared to justify your salary expectations with a mix of industry research and examples of the value you bring to the role.

Example Questions for Each Round

Round 1 (Technical):

1. Write a query using window functions to find the top 3 employees by salary in each department.
2. How would you handle nulls in a SQL join? Provide examples using COALESCE.
3. Write PySpark code to filter records based on specific conditions and add a calculated column.

Round 2 (Techno-Managerial):

1. What is the most common performance bottleneck in Spark jobs, and how would you resolve it?
2. Explain a time when you optimized an SQL query for a large dataset. What was the impact?
3. If your team disagrees on the approach to solving a problem, how do you manage the situation?

Round 3 (HR Discussion):

1. What are your expectations for the role beyond the salary?
2. How do you see your career evolving in the next 3-5 years?

Glassdoor Bristol Myers Squibb Review –

https://www.glassdoor.co.in/Reviews/Bristol-Myers-Squibb-Reviews-E107.htm

Bristol Myers Squibb Careers –

https://careers.bms.com/

Subscribe to my YouTube Channel for Free Data Engineering Content –

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

Connect with me here –

https://bento.me/shubhamwadekar