Impetus Data Engineer Interview Guide – Experienced 2+

Interview Process Breakdown

Data engineering interviews are typically structured across four distinct rounds, each testing specific competencies:

1. **Round 1: Technical (Core Technical Skills)**
    - Focus: Spark (or PySpark), SQL, and Python fundamentals.
    - Goal: Assess technical depth, coding proficiency, and problem-solving abilities in a distributed computing environment and relational databases.
2. **Round 2: Advanced Technical Concepts & Projects**
    - Focus: AWS cloud services, data architecture, and real-world project experience.
    - Goal: Test your understanding of cloud-native tools and your ability to explain real-life project implementations.
3. **Round 3: Managerial**
    - Focus: Past experiences, soft skills, and decision-making abilities in various scenarios.
    - Goal: Assess how you handle team dynamics, stakeholder management, and project execution.
4. **Round 4: HR**
    - Focus: Career motivations, alignment with the company, and salary negotiations.
    - Goal: Determine cultural fit and finalize logistics.

Detailed Insights on Each Round

Round 1: Technical - Core Technical Skills

Spark (PySpark)

- **Word Count Problem**:
    - **Task**: Modify a word count script to output results in descending frequency order.
    - **Key Learning**: Reduce computational overhead by using reduceByKey instead of groupByKey because reduceByKey combines values locally before shuffling, minimizing data movement across nodes.
    - **Bonus Question**: Explain why lineage in Spark is crucial for fault tolerance. (Tip: Describe how Spark DAGs track transformations to rebuild data if a partition fails.)

- **Common Spark Concepts**:
    - **Cache vs. Persist**: Highlight scenarios where persist's storage levels (e.g., MEMORY_AND_DISK) offer flexibility compared to cache.
    - **Fault Tolerance**: Compare Spark's lineage recovery with Hadoop's block replication mechanism.

SQL

- **Query Execution Order**: Understanding this concept is vital for optimizing query performance. Explain the standard sequence: FROM → WHERE → GROUP BY → HAVING → SELECT → ORDER BY.
- **Joins**: Detail examples of inner, outer, left, and right joins to demonstrate practical knowledge.
- **Rank vs. Dense_Rank**: Emphasize the absence of rank gaps in DENSE_RANK and scenarios where it is more appropriate.
- **Advanced Concepts**: Define cursors and stored procedures, including their use cases in iterative data operations and modular query design.

Python

- **Foundational Questions**:
    - Docstrings are essential for code documentation. Use examples to highlight their role in creating self-explanatory code.
    - The pass statement acts as a placeholder and is often used during function stubbing or maintaining syntactic structure.
    - Which data structure occupies more memory: list or tuple? Why?
- **Coding Problems**:
    - **Character Frequency in a Text File**: Use collections.Counter for an efficient solution.
    - **Palindrome Generation**: Write concise code to mirror strings, ensuring edge cases (e.g., single characters) are handled.

Round 2: Advanced Technical Concepts & Projects

AWS Concepts

- **AWS Glue Data Catalog**: Explain how it organizes metadata for structured and unstructured data across the cloud.
- **Athena vs. Aurora**: Emphasize that Athena is a serverless query engine for data lakes, while Aurora is a relational database service.
- **Versioning in S3**: Discuss its role in data recovery and audit tracking.
- **Redshift Data Distribution**: Explain the significance of EVEN, KEY, and ALL distribution styles in optimizing query performance.

Project Discussions

- Describe your project's problem statement and technical implementation clearly. Use diagrams and workflows to outline tools, architecture, and challenges faced.

Round 3: Managerial Round

- **Past Experiences**: Prepare STAR (Situation, Task, Action, Result) responses to demonstrate impact.
- **Scenario-Based Questions**: Expect inquiries like, *"How would you handle a deadline conflict between two high-priority projects?"* Offer solutions that balance stakeholder expectations and resource management.

Round 4: HR Round

- **Why Change Roles?** Frame your answer around growth opportunities, alignment with the company's mission, or interest in new challenges.
- **Salary Negotiations**: Research market trends and articulate your value while remaining flexible.

Glassdoor Impetus Technologies Review –

https://www.glassdoor.co.in/Reviews/Impetus-Technologies-Reviews-E259493.htm

Impetus Technologies Careers –

https://www.impetus.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