PWC Data Engineer Interview Guide – Experienced 3+

Technical Round 1

1. **Explain the architecture of Spark, including the roles of driver, executors, DAGs, and SparkContext.**

   Follow-ups:

   - How does the driver program handle task scheduling?
   - What happens when an executor fails during a task execution?

2. **What are the advantages and disadvantages of Delta Tables?**

   Follow-ups:

   - How do Delta Tables handle large-scale data updates efficiently?
   - What limitations do you face when using Delta Tables in a multi-cloud environment?

3. **Explain Delta Time Travel and the purpose of the vacuum command.**

   Follow-ups:

   - What happens if the vacuum command is not run periodically?
   - How do you configure retention periods for Delta tables?

4. **Differentiate between Schema Enforcement and Schema Evolution.**

   Follow-ups:

   - Can Schema Evolution lead to data inconsistencies? If so, how do you manage them?
   - What are the implications of enabling schema auto-detection?

5. **What is Secret Scope, and how is it used in Databricks?**

   Follow-ups:

   - How do you handle expired secrets in a production environment?
   - What are the differences between Azure Key Vault-backed and Databricks-backed Secret Scopes?

6. **How do you use Spark UI to debug stages, tasks, and performance issues?**

   Follow-ups:

   - How would you identify and resolve a shuffle spill in Spark UI?
   - What insights can you gather from the DAG visualization in Spark UI?

7. **How do you handle bad data in Databricks?**

   Follow-ups:

   - How do quarantine tables ensure data quality in downstream pipelines?
   - What are the best practices for logging and monitoring bad data?

Technical Round 2:

1. **Explain how Adaptive Query Execution (AQE) works in Databricks.**

   Follow-ups:

   - How does AQE optimize join operations dynamically?
   - What configuration parameters are critical for enabling AQE effectively?

2. **Describe the role of Dynamic Resource Allocation in Databricks.**

   Follow-ups:

   - How does resource allocation adjust when a job experiences a sudden load increase?
   - What are the potential downsides of enabling dynamic resource allocation?

3. **What is the usage of Optimize and REORG commands in Databricks?**

   Follow-ups:

   - How does Optimize command improve query latency in Delta tables?
   - What are the limitations of the REORG command with respect to large datasets?

4. **How is Git version control implemented in Databricks?**

   Follow-ups:

   - What challenges do you face when managing multiple notebooks in Git?
   - How do you resolve merge conflicts in Databricks notebooks?

5. **What causes data skewness in Spark, and how can it be resolved?**

   Follow-ups:

   - How do you identify skewed partitions in a dataset?
   - What are the performance trade-offs of using salting to mitigate data skewness?

6. **How do you decide the number of partitions for repartitioning data in Spark?**

   Follow-ups:

   - What metrics would you analyze to determine if your partitioning strategy is effective?
   - How does improper partitioning affect Spark job performance?

7. **What causes Out of Memory (OOM) issues in Databricks, and how do you resolve them?**

   Follow-ups:

   - How do caching strategies impact memory management in Databricks?
   - What role does the executor heap size play in preventing OOM errors?

Round 3 – HR/ Managerial Round.

I will be attaching one separate PDF for commonly asked Managerial and HR questions.

Glassdoor PWC Review –

https://www.glassdoor.co.in/Interview/PwC-Data-Engineer-Interview-Questions-EI_IE8450.0,3_KO4,17.htm

PWC Careers –

https://www.pwc.in/careers.html

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