TCS Azure Data Engineer Interview Guide – Experienced 3+

Technical Round 1:

1. **Explain Delta Live Tables and their features, such as declarative pipeline definition and automatic data validation.**

   Follow-ups:

   - How do Delta Live Tables ensure data quality during transformations?
   - Can you give a use case where Delta Live Tables would be ideal?

2. **What is Databricks Auto Loader, and how does it handle new files?**

   Follow-ups:

   - How does Auto Loader avoid reloading files with the same name?
   - What are the performance considerations when using Auto Loader?

3. **What is the importance of the checkpoint location in Databricks?**

   Follow-ups:

   - How does checkpointing ensure fault tolerance in streaming workflows?
   - What happens if the checkpoint location is accidentally deleted?

4. **How do you install a Python library that is not in the Databricks runtime?**

   Follow-ups:

   - What are the differences between %pip and %conda commands in Databricks?
   - How do you handle version conflicts for libraries?

5. **How do you implement row and column-level security in Databricks?**

   Follow-ups:

   - Can you describe the role of user groups in setting up these policies?
   - How do these policies affect query performance?

6. **Explain data encryption in Databricks, both at rest and in transit.**

   Follow-ups:

   - How is Azure Key Vault used to manage encryption keys in Databricks?
   - What are the implications of enabling encryption at rest on storage performance?

7. **How do you move a Databricks notebook to higher environments?**

   Follow-ups:

   - What role do workspace APIs play in this process?
   - How do you ensure version control when migrating notebooks?

Technical Round 2:

1. **Explain the architecture of Databricks, including the control plane and data plane.**

   Follow-ups:

   - How does Databricks integrate with external storage systems?
   - What are the security considerations for the control plane?

2. **When would you use flatten, explode, or collect_list in Spark?**

   Follow-ups:

   - Can you give an example of processing nested JSON data using these functions?
   - How do these transformations impact memory usage?

3. **How would you read a large file (e.g., 15GB) efficiently in Spark by increasing parallelism?**

   Follow-ups:

   - What factors determine the optimal number of partitions for a large file?
   - How do you monitor and debug skewed partitions?

4. **What is dynamic partition pruning, and how does it optimize query execution?**

   Follow-ups:

   - Can you provide an example where dynamic partition pruning improved performance?
   - How does it differ from static partition pruning?

5. **How do you call one Databricks notebook from another?**

   Follow-ups:

   - What are the differences between %run and dbutils.notebook.run?
   - How do you handle passing parameters between notebooks?

6. **What are the steps to debug a failed workflow in Databricks?**

   Follow-ups:

   - How do you identify resource bottlenecks in cluster logs?
   - What strategies do you use to retry failed steps in workflows?

7. **What determines the maximum parallelism achievable in Databricks?**

   Follow-ups:

   - How does cluster size impact parallelism limits?
   - What role does executor memory and CPU configuration play in maximizing parallelism?

Glassdoor TCS Review –

https://www.glassdoor.co.in/Reviews/Tata-Consultancy-Services-Reviews-E13461.htm

TCS Careers –

https://www.tcs.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