

1. How can you read parameters using `dbutils.widgets.text` and retrieve their values?
2. How do you provide read access for a production notebook to a new data engineer for review?
3. When attaching a notebook to a cluster, which permission allows you to run the notebook?
4. Should production DLT pipelines be run on a job cluster or an all-purpose cluster?
5. Does a CTAS (CREATE TABLE AS SELECT) operation execute the load every time or only during table creation?
6. How can you control access to read production secrets using scope access control?
7. Where does the `%sh` command run in Databricks?
8. If a query contains a filter, how does Databricks use file statistics in the transaction log?
9. What happens when you run a `VACUUM` command on a shallow clone table?
10. Which type of join (left, inner, right) is not possible when performing a join between a static `DataFrame` and a streaming `DataFrame`?
11. When the source is a CDC (Change Data Capture), should you use `MERGE INTO` or leverage the Change Data Feed (CDF) feature?
12. How can you find the difference between the previous and present commit in a Delta table?
13. What is the best approach for nightly jobs that overwrite a table for the business team with the least latency?
14. What does the `OPTIMIZE TABLE` command do, and what is the target file size?
15. In a streaming scenario, what does the `.withWatermark` function do with a delay of 10 minutes?
16. How does aggregating on the source and then overwriting/appending to the target impact data load?
17. Why did you receive three email notifications when a job was set to trigger an email if `mean(temp) > 120`?
18. Why should the checkpoint directory be unique for each stream in a streaming job?
19. How would you set up an Autoloader scenario to load data into a bronze table with history and update the target table?
20. How can you handle streaming deduplication based on a given code scenario?
21. For batch loading, what happens if the load is set to overwrite or append?
22. In a Change Data Feed (CDF) scenario, if `readChangeFeed` starts at version 0 and append is used, will there be deduplication?
23. How can you identify whether a table is SCD Type 1 or 2 based on an upsert operation?
24. To avoid performance issues, should you decrease the trigger time or not?
25. How does Delta Lake decide on file skipping based on columns in a query, and what are the implications for nested columns?
26. What does granting "Usage" and "Select" permissions on a Delta table allow a user to do?
27. How do you create an unmanaged table in Databricks?
28. What makes a date column a good candidate for partitioning in a Delta table?
29. What happens in the transaction log when you rename a Delta table using `ALTER TABLE xx RENAME xx`?
30. How would you handle an error with a check constraint and what would you recommend?
31. When using `DESCRIBE` commands, how can you retrieve table properties, comments, and partition details?
32. How are file statistics used in Delta Lake, and why are they important?
33. In the Ganglia UI, how can you detect a spill during query execution?
34. If a repo branch is missing locally, how can you retrieve that branch with the latest code changes?
35. After deleting records with a query like `DELETE FROM A WHERE id IN (SELECT id FROM B)`, can you time travel to see the deleted records, and how can you prevent their permanent deletion?
36. What are the differences between DBFS and mounts in Databricks?
37. If the `API 2.0/jobs/create` is executed three times with the same JSON, what will happen? Will it execute or create three jobs?
38. What is DBFS in Databricks?
39. How do you install a Python library using `%pip` in a Databricks notebook?
40. If Task 1 has downstream Task 2 and Task 3 running in parallel, and Task 1 and Task 2 succeed while Task 3 fails, what will be the final job status?
41. How do you handle streaming job retries in production, specifically with job clusters, unlimited retries, and a maximum of one concurrent run?
42. How can you clone an existing job and version it using the Databricks CLI?
43. When converting a large JSON file (1TB) to Parquet with a partition size of 512 MB, what is the correct order of steps?

44. What happens in the target table when duplicates are dropped during a batch read and append operation?
45. If a column was missed during profiling from Kafka, how can you ensure that the data is fully replayable in the future?
46. How do you handle access control for users in Databricks?
47. What is the use of the `pyspark.sql.functions.broadcast` function in a Spark job?
48. What happens when performing a join on `orders_id` with a condition "when not matched, insert *"?
49. How would you write a silver load function to transform and update downstream tables from bronze?
50. What will be the output of `CASE WHEN is_member("group") THEN email ELSE 'redacted' END AS email`?
51. How can you use the Ganglia UI to view logs and troubleshoot a Databricks job?
52. How do you list or get the tasks using the API `2.0/jobs/list` or `2.0/jobs/run/list`?
53. What is unit testing, and how is it applied in a Databricks environment?
54. What happens when multiple `display()` commands are executed repeatedly in development, and what is the impact in production?
55. Will the option `"readChangeFeed"` work on a source Delta table with no CDC enabled?
56. How can you identify whether a tumbling or sliding window is being used based on the code provided?
57. What performance tuning considerations are involved with `spark.sql.files.maxPartitionBytes` and `spark.sql.shuffle.partitions`?