Testpassport Q&A

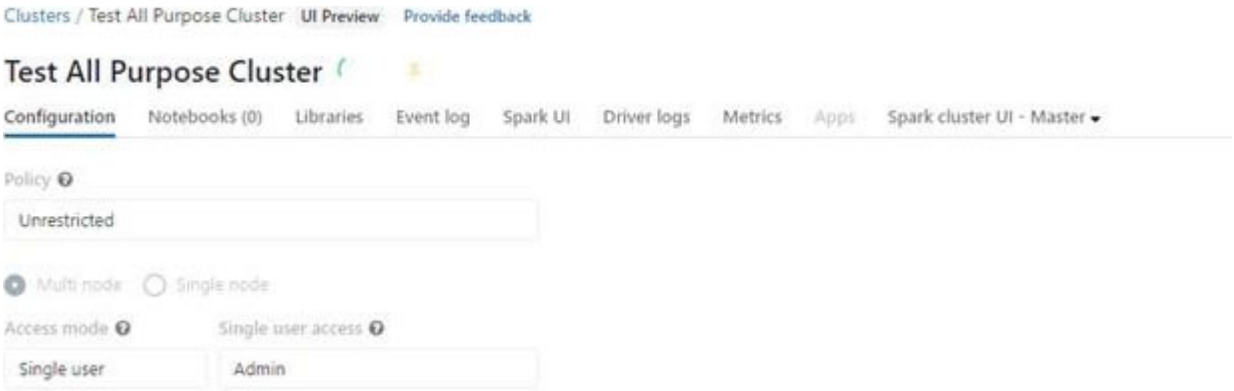
La meilleure qualité le meilleur service

Exam : **Databricks Certified Professional Data Engineer**

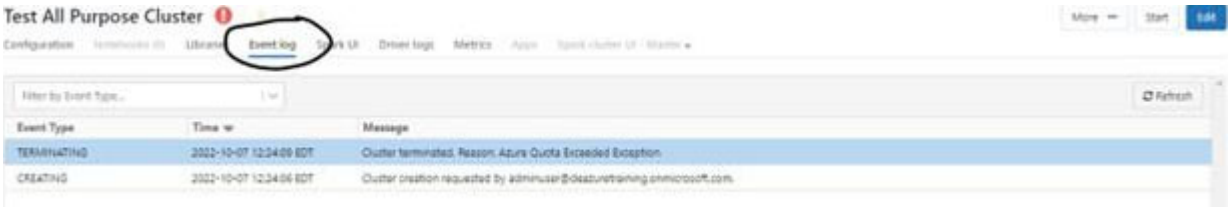
Title : Databricks Certified Data Engineer Professional Exam

Version : DEMO

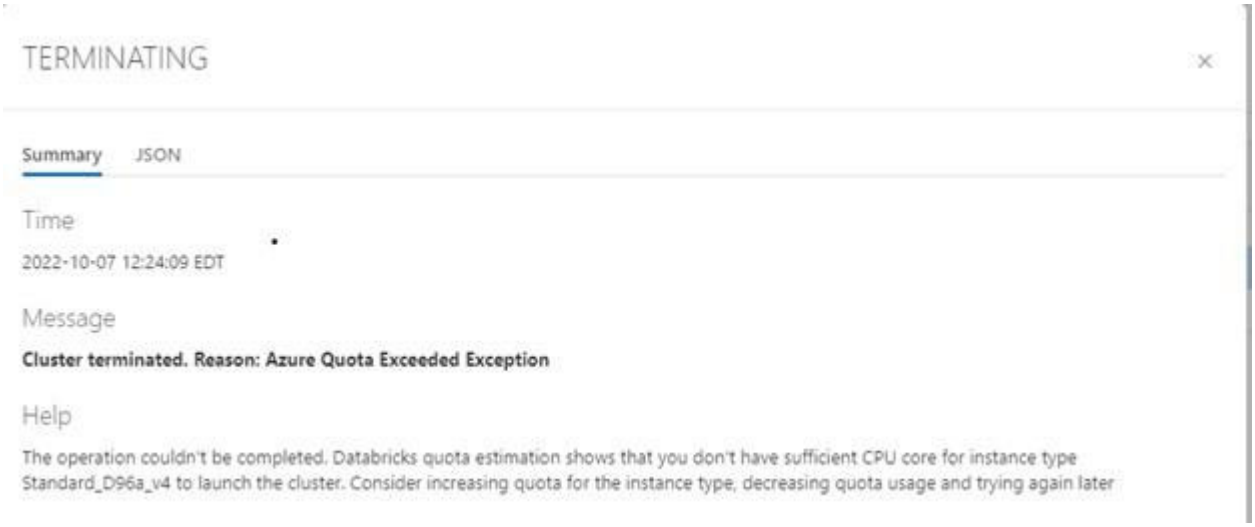
1.You were asked to setup a new all-purpose cluster, but the cluster is unable to start which of the following steps do you need to take to identify the root cause of the issue and the reason why the cluster was unable to start?
A. Check the cluster driver logs
B. Check the cluster event logs
(Correct)
C. Workspace logs
D. Storage account
E. Data plane
Answer: B
Explanation:
Cluster event logs are very useful, to identify issues pertaining to cluster availability. Cluster may not start due to resource limitations or issues with the cloud providers.
Some of the common issues include a subnet for compute VM reaching its limits or exceeding the subscription or cloud account CPU quota limit.
Here is an example where the cluster did not start due to subscription reaching the quota limit on a certain type of cpu cores for a VM type.



Graphical user
interface, text, application, email
Description automatically generated
Click on event logs

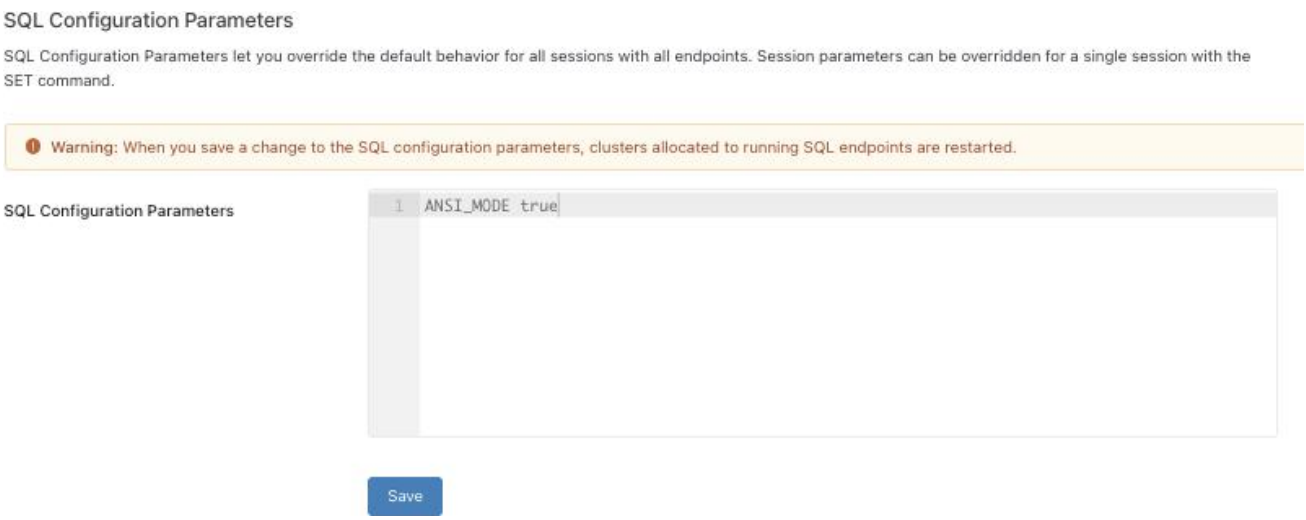


Graphical user
interface, text, application, email
Description automatically generated
Click on the message to see the detailed error message on why the cluster did not start.



Graphical user
interface, text, application, email
Description automatically generated

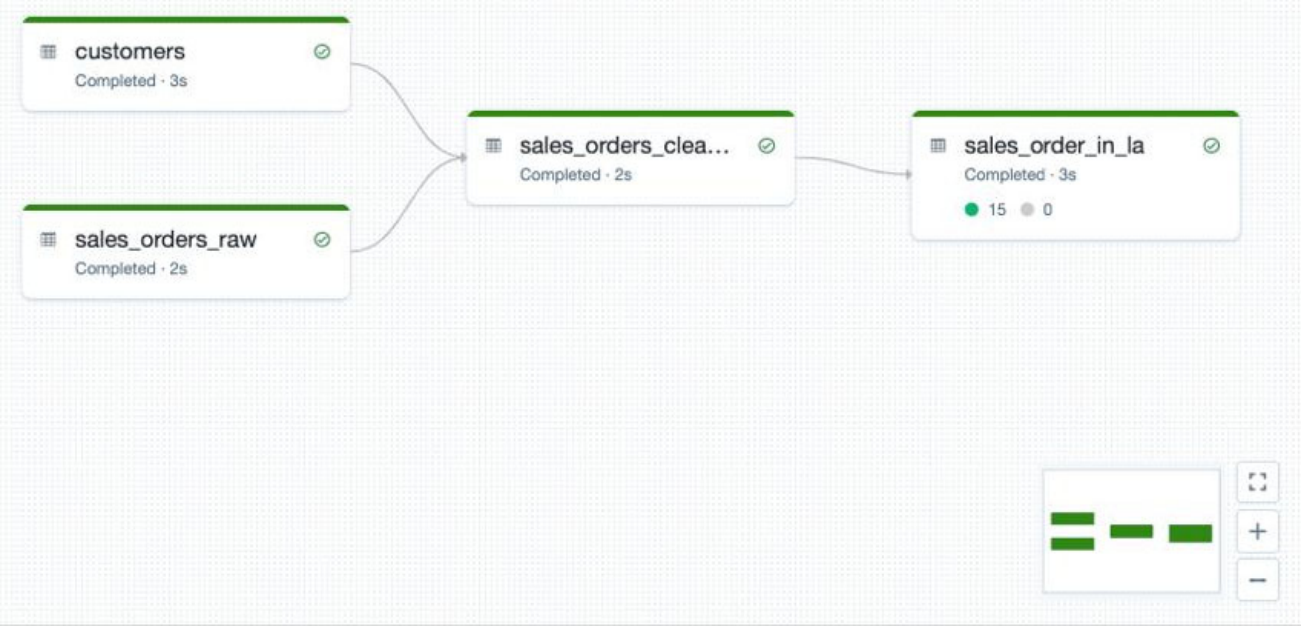
2.A SQL Dashboard was built for the supply chain team to monitor the inventory and product orders, but all of the timestamps displayed on the dashboards are showing in UTC format, so they requested to change the time zone to the location of New York.
How would you approach resolving this issue?
A. Move the workspace from Central US zone to East US Zone
B. Change the timestamp on the delta tables to America/New_York format
C. Change the spark configuration of SQL endpoint to format the timestamp to Ameri-ca/New_York
D. Under SQL Admin Console, set the SQL configuration parameter time zone to Ameri-ca/New_York
E. Add SET Timezone = America/New_York on every of the SQL queries in the dashboard.
Answer: D
Explanation:
The answer is, Under SQL Admin Console, set the SQL configuration parameter time zone to America/New_York
Here are steps you can take this to configure, so the entire dashboard is changed without changing individual queries
Configure SQL parameters
To configure all warehouses with SQL parameters:
1.Click Settings at the bottom of the sidebar and select SQL Admin Console.
2.Click the SQL Warehouse Settings tab.
3.In the SQL Configuration Parameters textbox, specify one key-value pair per line. Sepa-rate the name of the parameter from its value using a space. For example, to ena-ble ANSI_MODE:



Graphical user
interface, text, application
Description automatically generated
Similarly, we can add a line in the SQL Configuration parameters timezone America/New_York
SQL configuration parameters | Databricks on AWS

3.You are currently asked to work on building a data pipeline, you have noticed that you are currently working on a very large scale ETL many data dependencies, which of the following tools can be used to address this problem?
A. AUTO LOADER
B. JOBS and TASKS
C. SQL Endpoints
D. DELTA LIVE TABLES
E. STRUCTURED STREAMING with MULTI HOP
Answer: D
Explanation:
The answer is, DELTA LIVE TABLES
DLT simplifies data dependencies by building DAG-based joins between live tables. Here is a view of how the dag looks with data dependencies without additional meta data,
1.create or replace live view customers
2.select * from customers;
3.
4.create or replace live view sales_orders_raw
5.select * from sales_orders;
6.
7.create or replace live view sales_orders_cleaned
8.as
9.select sales.* from
10.live.sales_orders_raw s
11. join live.customers c
12.on c.customer_id = s.customer_id

13.where c.city = 'LA';
14.
15.create or replace live table sales_orders_in_la
16.selects from sales_orders_cleaned;
Above code creates below dag



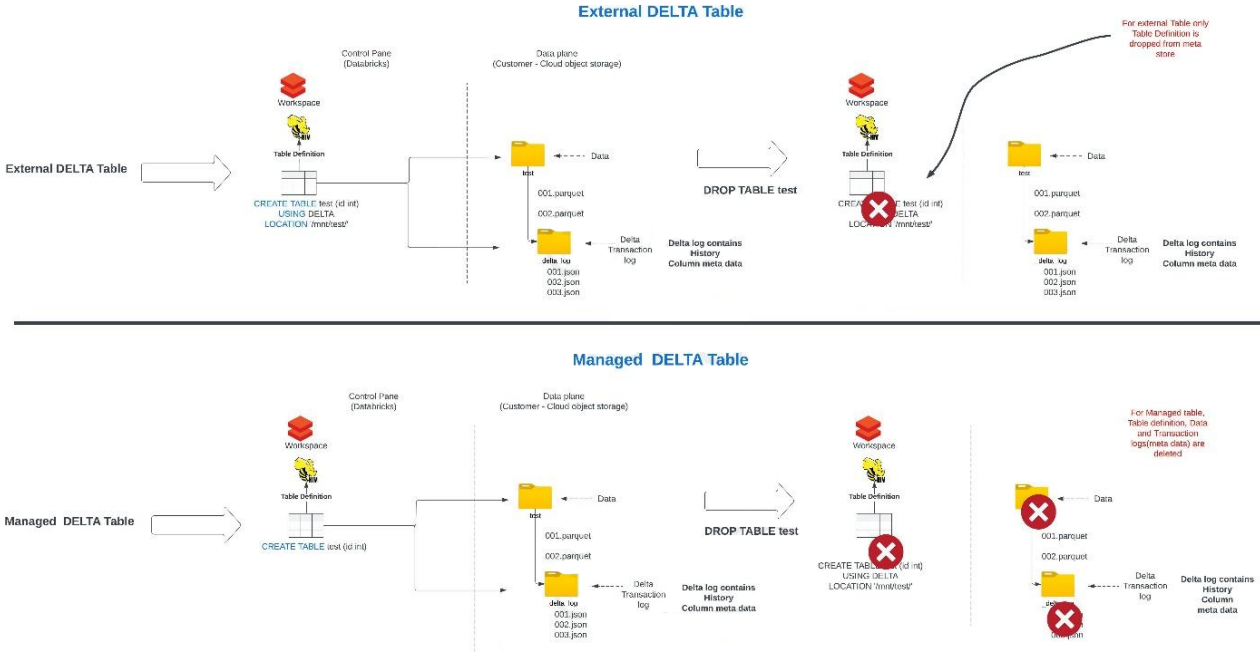
Documentation on DELTA LIVE TABLES,
https://databricks.com/product/delta-live-tables
https://databricks.com/blog/2022/04/05/announcing-generally-availability-of-databricks-delta-live-tables-dlt.html
DELTA LIVE TABLES, addresses below challenges when building ETL processes
1. Complexities of large scale ETL
a. Hard to build and maintain dependencies
b. Difficult to switch between batch and stream
2. Data quality and governance
a. Difficult to monitor and enforce data quality
b. Impossible to trace data lineage
3. Difficult pipeline operations
a. Poor observability at granular data level
b. Error handling and recovery is laborious

4.When you drop a managed table using SQL syntax DROP TABLE table_name how does it impact metadata, history, and data stored in the table?
A. Drops table from meta store, drops metadata, history, and data in storage.
B. Drops table from meta store and data from storage but keeps metadata and history in storage
C. Drops table from meta store, meta data and history but keeps the data in storage
D. Drops table but keeps meta data, history and data in storage
E. Drops table and history but keeps meta data and data in storage
Answer: A

Explanation:

For a managed table, a drop command will drop everything from metastore and storage.

See the below image to understand the differences between dropping an external table.



Diagram

Description automatically generated

5.Which of the following approaches can the data engineer use to obtain a version-controllable configuration of the Job's schedule and configuration?

A. They can link the Job to notebooks that are a part of a Databricks Repo.

B. They can submit the Job once on a Job cluster.

C. They can download the JSON equivalent of the job from the Job's page.

D. They can submit the Job once on an all-purpose cluster.

E. They can download the XML description of the Job from the Job's page

Answer: D