Microsoft

DP-750

Implementing Data Engineering Solutions Using Azure Databricks

For More Information – Visit link below:

https://www.examsempire.com/

Product Version

1. **Up to Date products, reliable and verified.**
2. **Questions and Answers in PDF Format.**

Latest Version: 6.0

Topic 1, Contoso

Overview

Contoso has a single Azure Databricks workspace named Workspace1 in the West US Azure region. Workspace1 is enabled for Unity Catalog.

Workspace1 contains all-purpose clusters for both development and production workloads.

The company's Azure environment contains:

- In the West US, Central US, and East US Azure regions, Azure event hubs that stream telemetry data and an Azure Data Lake Storage Gen2 account in each region for each hub
- A single Azure SQL database in the West US region that hosts enterprise resource planning (ERP) data
- An Azure Database for PostgreSQL server in the West US region that stores operational maintenance data

Company information

Contoso, Inc. is a renewable energy provider that operates solar and wind farms across North America.

Data Environment

Contoso ingests the following operational and business data:

- Telemetry data: More than 40,000 loT sensors across 28 sites emit JSON telemetry events every few seconds. Each site sends the events to the nearest event hub, which writes the data into the corresponding Data Lake Storage Gen2 account. These files frequently experience schema drift.
- Maintenance logs: Maintenance systems generate historical repair logs, daily incremental updates, technician notes, and unstructured attachments that are stored in the Data Lake Storage Gen2 accounts.
- Operational maintenance data: Structured operational maintenance data is stored on the Azure Database for
PostgreSQL server.
- External weather data: Hourly weather forecasts are retrieved from a REST API and written to the Data Lake Storage Gen2 accounts.
- ERP data: Daily CSV extracts of 50 to 100 GB contain equipment metadata, work orders, and purchase order information.

Problem Statements

The company's existing analytics environment has several issues:

Ingestion

- Telemetry pipelines fall behind during peak loads.
- Telemetry ingestion fails when schema drift occurs.
- Streaming pipelines reprocess events after a pipeline restarts.

Compute

- Production and development workloads run on the same all-purpose clusters.

- Production and development workloads do NOT support autoscaling or workload isolation.

Governance

- The ERP data is duplicated across systems and development teams.
- Naming conventions are inconsistent across development teams, regions, and products.
- Ownership of the loT sensors changes over time, and analysts must track the full history of the ownership.
- Occasionally, equipment manufacturers must correct data-entry mistakes in equipment names. Historical values are NOT required.

Pipeline operations

- Pipelines lack resiliency, alerting, and centralized scheduling.

Planned Changes

Contoso plans to implement the following changes:

- Implement scalable data pipeline orchestration.
- Create a managed analytics catalog in Unity Catalog.
- Implement a consistent approach to creating curated datasets.
- Establish a centralized governance model across ingestion, cleansed, and curated layers.
- Grant data engineers access to the ERP tables by using minimal development effort.
- Adopt a compute strategy that isolates production workloads and supports autoscaling.
- Adopt a slowly changing dimension (SCD) approach to address current data modeling issues.

Technical Requirements

Contoso identifies the following environment and compute requirements:

- Ensure that production ingestion workloads run on compute clusters that can scale automatically during telemetry spikes.
- Provide fast and consistent performance for business intelligence (Bl) workloads.
- Prevent development activity from affecting production pipelines.
- Production ingestion workloads must run as scheduled, non-interactive pipelines rather than on shared interactive development clusters.

Contoso identifies the following data ingestion and processing requirements:

- Auto-scale ingestion pipelines to handle bursty workloads.
- Handle schema drift for the maintenance and telemetry data.
- Ingest file-based telemetry data by using minimal operational effort.
- Store all the ingested data in a format that supports incremental processing.
- Support the continuous ingestion of telemetry data from the event hubs by using exactly-once semantics.
- Support the ingestion of the structured maintenance data from the Azure Database for PostgreSQL server.
- Build a new telemetry pipeline that ingests raw events from the event hubs, cleanses the data, and publishes curated tables to Unity Catalog.
- Ensure that the Apache Spark Structured Streaming pipelines reading from the event hubs write the data into a managed Delta table named telemetry.raw_events. The pipelines

must support schema drift and resume processing after failures without reprocessing the data.
Contoso identifies the following data modeling and optimization requirements:
- Build curated tables that standardize business logic.
- Overwrite equipment metadata attributes, such as name, manufacturer, model, and commissioning date, when the attributes change. Historical values are NOT required.

Contoso identifies the following pipeline deployment and operation requirements: |^ •
Orchestrate multi-step ingestion and transformation workflows.
- Define a clear execution order and dependencies.
- Automatically retry failed steps and notify operators.
- Schedule ingestion and transformation workloads consistently.

Governance Requirements
Contoso identifies the following governance requirements:
- Centralize the metadata catalog.
- Provide isolated development areas that follow standard naming conventions.
- Establish a consistent structure for organizing raw, cleansed, and curated data.
- Provide a read-only mechanism to reference the ERP data through a foreign catalog.

Business Requirements
Contoso identifies the following business requirements:
- Improve ingestion reliability and reduce operational effort.
- Standardize data definitions across development teams.

Question: 1

You need to develop the task logic for a new job in Lakeflow Jobs that processes telemetry data. Each task must contain only the appropriate logic for its step in the pipeline. The solution must support the planned changes and meet the data ingestion and processing requirements.
What should you do?

A. Use a single Databricks notebook task that performs ingestion, cleansing, and curation in one script.
B. Create three tasks that each contains the identical logic and use task retries.
C. Use a single SQL task that performs ingestion, cleansing, and curation by running merge commands.
D. Create separate tasks for ingestion, cleansing, and curation.

Answer: D

Explanation:
CORRECT ANSWE R: D - Create separate tasks for ingestion, cleansing, and curation.
According to Microsoft Learn, Lakeflow Jobs (formerly Databricks Workflows) supports multi-task pipelines where each task encapsulates a single, well-defined step. The official documentation states that best practice is to decompose complex pipelines into discrete tasks

— ingestion, cleansing, and curation — so that each task contains only the logic appropriate for that stage. This approach aligns with the Contoso planned change to 'implement scalable data pipeline orchestration' and the requirement to 'define a clear execution order and dependencies.' Option A is incorrect because combining all logic in one notebook violates the single-responsibility principle and makes retry/recovery difficult. Option B is incorrect because duplicating identical logic across tasks wastes resources and defeats the purpose of a modular pipeline. Option C is incorrect because a single SQL MERGE task cannot cleanly separate the ingestion, cleansing, and curation concerns required.

Question: 2

You need to configure compute for the ingestion of telemetry data. The solution must meet the data ingestion and processing requirements.
What should you do?

A. Enable Photon acceleration for a job compute cluster.
B. Move the ingestion pipelines to shared compute.
C. Increase an all-purpose cluster to a larger fixed node type.
D. Disable autoscaling for a job compute cluster.

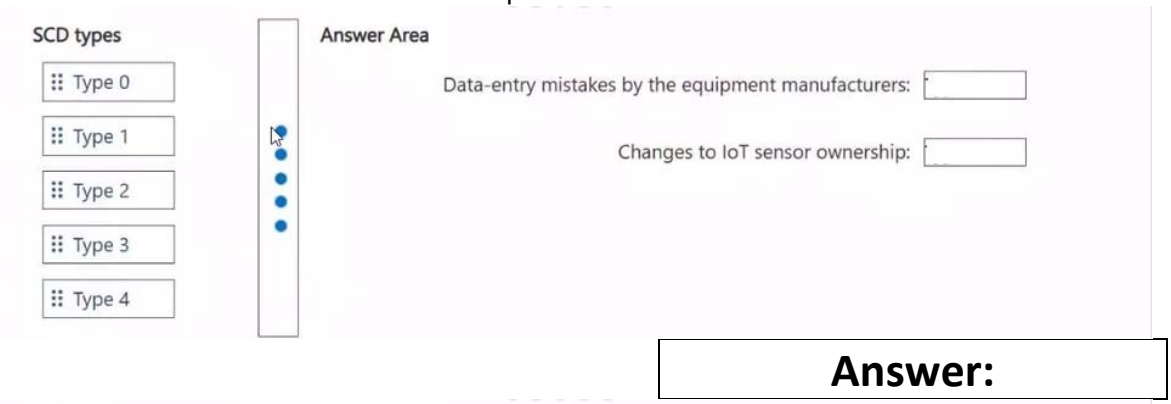
Answer: A

Explanation:
CORRECT ANSWE R: A - Enable Photon acceleration for a job compute cluster.
According to Microsoft Learn and the Azure Databricks documentation, Photon is a high-performance vectorized query engine written in C++ that accelerates Apache Spark workloads, especially ingestion and SQL operations. The Contoso technical requirement states: 'Ensure that production ingestion workloads run on compute clusters that can scale automatically during telemetry spikes' and 'Provide fast and consistent performance for BI workloads.' Photon on a job compute cluster directly addresses both speed and consistency for ingestion pipelines.
Option B is incorrect because moving ingestion to shared compute would violate the requirement to isolate production from development. Option C is incorrect because increasing a fixed-node all-purpose cluster does not provide autoscaling. Option D is incorrect because disabling autoscaling would prevent the cluster from handling bursty telemetry workloads, directly contradicting the stated requirements.

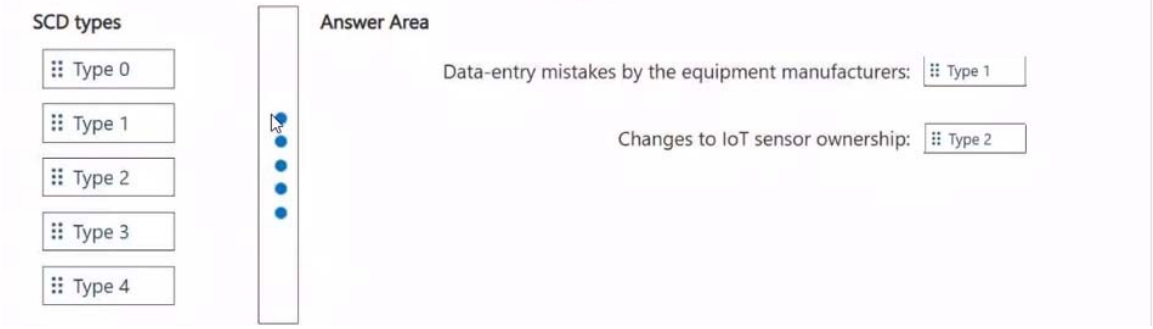
Question: 3

DRAGDROP
Which SCD type should you use to support the planned data modeling changes? To answer, drag the appropriate types to the correct issues. Each type may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.



Answer:

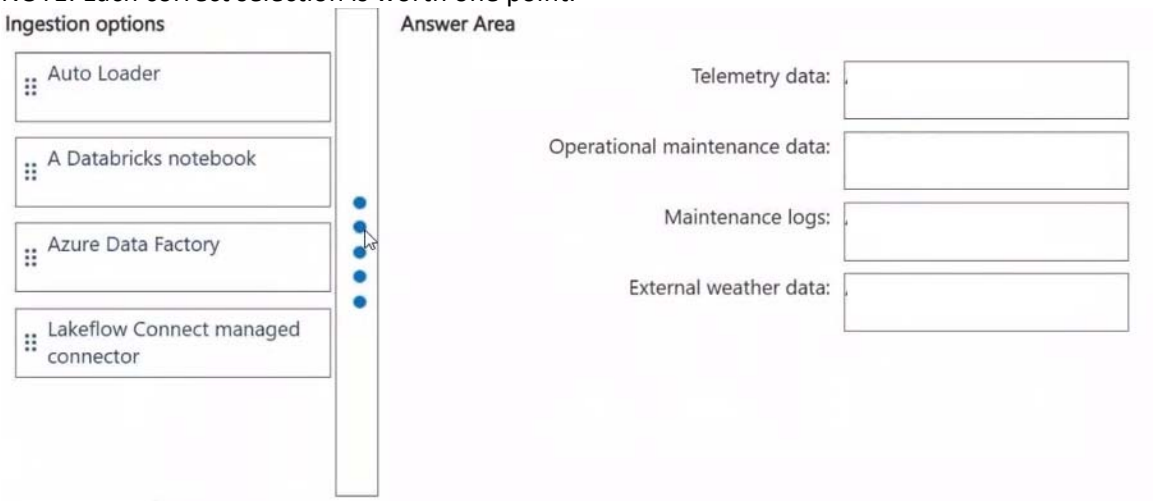


Question: 4

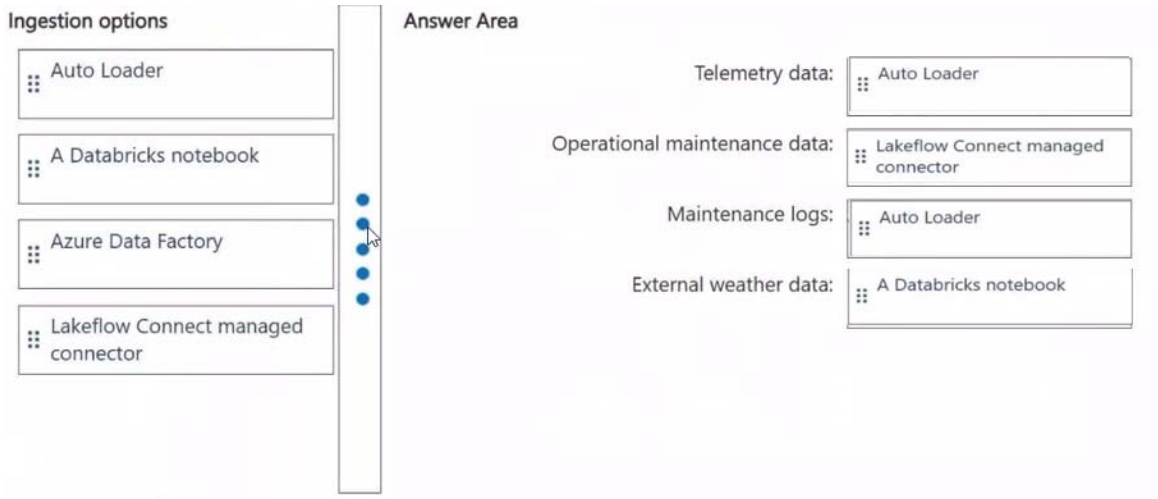
DRAGDROP

Which ingestion option should you recommend for each data source? To answer, drag the appropriate options to the correct data sources. Each option may be used once, more than once, or not at all. You may need to drag the split bar between panes or scroll to view content.

NOTE: Each correct selection is worth one point.



Answer:



Question: 5

HOTSPOT
You need to complete the PySpark code for the Spark Structured Streaming pipelines. The solution must meet the data ingestion and processing requirements.
How should you complete the code segment? To answer, select the appropriate options in the answer area.
NOTE: Each correct selection is worth one point.

Answer Area

```
telemetry_path = "abfss://telemetry@contosodata.dfs.core.windows.net/"

checkpoint_path = ""

raw_events = "analytics.telemetry.raw_events"


df = (spark.readStream

          .format("json")

          .load(telemetry_path))


df.writeStream

      .format
```

("csv")
("delta")
("json")
("parquet")

```
      .option
```

("checkPointLocation", checkpoint_path)
("failOnDataLoss", "false")
("path", checkpoint_path)
("startingOffsets", "earliest")

```
      .option("m
      .trigger(a
      .table(raw
```

Answer:

Answer Area

```
telemetry_path = "abfss://telemetry@contosodata.dfs.core.windows.net/"

checkpoint_path = ""

raw_events = "analytics.telemetry.raw_events"


df = (spark.readStream

          .format("json")

          .load(telemetry_path))


df.writeStream

      .format
```

("csv")
("delta")
("json")
("parquet")

```
      .option
```

("checkPointLocation", checkpoint_path)
("failOnDataLoss", "false")
("path", checkpoint_path)
("startingOffsets", "earliest")

```
      .option("m
      .trigger(a
      .table(raw
```

Thank You for Trying Our Product

Special 16 USD Discount Coupon: NSZUBG3X

Email: support@examsempire.com

Check our Customer Testimonials and ratings available on every product page.

Visit our website.

https://examsempire.com/