
INSIGHTFULSAGA

CERTIFICATION PREPARATION SERIES

DATABRICKS

CERTIFICATION PREPARATION

Complete Practice Workbook

Practice Tests Comprehensive Question Bank Verified Explanations

"Prepare Smarter. Practice With Purpose."

WORKBOOK SCOPE & COVERAGE

12 Complete Practice Tests covering Questions 1 to 300 (299 total questions)

Core architectural concepts, production scenarios, performance, and security questions

Verified answer keys with concise technical explanations and architectural rationales

Formatted as a high-density, printable technical reference workbook for exam preparation

TABLE OF CONTENTS

Practice Test 01: Practice Test 01	Questions 1 20 (20 Qs)
Practice Test 02: Practice Test 02	Questions 21 40 (20 Qs)
Practice Test 03: Practice Test 03	Questions 41 60 (20 Qs)
Practice Test 04: Practice Test 04	Questions 61 80 (20 Qs)
Practice Test 05: Practice Test 05	Questions 81 90 (10 Qs)
Practice Test 06: Practice Test 06	Questions 91 110 (20 Qs)
Practice Test 07: Practice Test 07	Questions 111 160 (50 Qs)
Practice Test 08: Practice Test 08	Questions 161 173 (13 Qs)
Practice Test 09: Practice Test 09	Questions 175 200 (26 Qs)
Practice Test 10: Practice Test 10	Questions 201 230 (30 Qs)
Practice Test 11: Practice Test 11	Questions 231 260 (30 Qs)
Practice Test 12: Practice Test 12	Questions 261 300 (40 Qs)

DATABRICKS CERTIFICATION PRACTICE TEST 01

Practice Test 01 Questions 1 20 (20 Questions)

QUESTION 1

Question 1

A data engineer wants to store data in a format that supports ACID transactions, schema enforcement, and time travel.

Which storage format should be used?

- A. CSV
- B. JSON
- C. Delta Lake
- D. Avro

ANSWER

C. Delta Lake

EXPLANATION

Delta Lake provides:

- ACID transactions
- Schema enforcement
- Schema evolution
- Time Travel
- Reliable batch and streaming support

CSV, JSON, and Avro do not provide all of these capabilities natively.

QUESTION 2

Question 2

Which component in a Databricks cluster coordinates Spark job execution?

- A. Executor
- B. Worker Node
- C. Driver Node
- D. Delta Engine

ANSWER**C. Driver Node****EXPLANATION**

The Driver Node:

- Accepts submitted code
- Creates execution plans
- Coordinates worker nodes
- Tracks job progress

Worker nodes execute the actual tasks.

QUESTION 3

Question 3

Which architecture layer should contain raw, unmodified source data?

- A. Gold
- B. Silver
- C. Bronze
- D. Semantic

ANSWER**C. Bronze****EXPLANATION**

Bronze Layer stores:

- Raw source data
- Historical records
- Minimal transformation

Silver is used for cleansing and Gold is used for business-ready reporting.

QUESTION 4

Question 4

A data engineer wants to remove duplicate customer records and standardize column formats.

Which layer should perform these operations?

- A. Bronze
- B. Silver
- C. Gold
- D. Dashboard

ANSWER**B. Silver****EXPLANATION**

Silver Layer is responsible for:

- Data cleansing
- Deduplication
- Standardization
- Data quality improvements

QUESTION 5

Question 5

Which SQL operation should be used to perform inserts and updates in a Delta table?

- A. UNION
- B. APPEND
- C. MERGE
- D. CACHE

ANSWER**C. MERGE****EXPLANATION**

MERGE supports:

- INSERT
- UPDATE
- DELETE

in a single statement and is commonly used for SCD processing.

QUESTION 6

Question 6

A Databricks user needs to review previous versions of a table.

Which feature should be used?

- A. VACUUM
- B. OPTIMIZE
- C. Time Travel
- D. CACHE

ANSWER**C. Time Travel****EXPLANATION**

Time Travel allows users to query historical versions of Delta tables using a version number or timestamp.

QUESTION 7

Question 7

Which Databricks feature improves query performance by compacting small files?

- A. CACHE
- B. OPTIMIZE
- C. MERGE
- D. HISTORY

ANSWER**B. OPTIMIZE****EXPLANATION**

OPTIMIZE:

- Compacts small files
- Improves read performance
- Reduces file fragmentation

QUESTION 8

Question 8

What is the purpose of the VACUUM command?

- A. Deduplicate data
- B. Improve joins
- C. Remove obsolete files
- D. Create partitions

ANSWER**C. Remove obsolete files**

EXPLANATION

VACUUM removes old data files that are no longer referenced by the Delta transaction log.

QUESTION 9

Question 9

A developer wants a query to automatically adapt when data distribution changes during execution. Which feature provides this capability?

- A. Photon
- B. AQE
- C. Delta Cache
- D. Unity Catalog

ANSWER

B. AQE

EXPLANATION

Adaptive Query Execution (AQE):

- Optimizes joins dynamically
- Handles skew automatically
- Improves runtime performance

QUESTION 10

Question 10

Which join strategy is most efficient when one table is very small?

- A. Sort Merge Join
- B. Shuffle Join
- C. Broadcast Join
- D. Cartesian Join

ANSWER

C. Broadcast Join

EXPLANATION

Broadcast Join sends the small table to all executors, avoiding large shuffle operations.

QUESTION 11

Question 11

Which command creates a temporary view from a DataFrame?

- A. saveAsTable()
- B. createTempView()
- C. write.format()
- D. cache()

ANSWER

B. createTempView()

EXPLANATION

Temporary views allow SQL queries to run against DataFrames during the current session.

QUESTION 12

Question 12

A company requires centralized access control for catalogs, schemas, and tables. Which Databricks service should be implemented?

- A. MLflow
- B. Workflows
- C. Unity Catalog
- D. Delta Cache

ANSWER

C. Unity Catalog**EXPLANATION**

Unity Catalog provides:

- Centralized governance
 - Fine-grained permissions
 - Data lineage
 - Auditing
-

QUESTION 13

Question 13

In Unity Catalog hierarchy, which object comes directly below Catalog?

- A. Workspace
- B. Table
- C. Schema
- D. View

ANSWER

C. Schema

EXPLANATION

Hierarchy:

Catalog Schema Table/View

QUESTION 14

Question 14

Which file format is used internally by Delta Lake for data storage?

- A. XML
- B. CSV
- C. TXT
- D. Parquet

ANSWER

D. Parquet

EXPLANATION

Delta Lake stores actual data using Parquet files while transaction information is stored in ' _delta_log '.

QUESTION 15

Question 15

Which statement best describes schema evolution?

- A. Deletes invalid columns
- B. Allows controlled schema changes
- C. Encrypts schema metadata
- D. Creates backup columns

ANSWER

B. Allows controlled schema changes

EXPLANATION

Schema evolution allows adding new columns and adjusting schema without recreating tables.

QUESTION 16

Question 16

What is the main purpose of checkpointing in Structured Streaming?

- A. Improve notebook rendering
- B. Store dashboard metrics
- C. Maintain fault tolerance
- D. Compress files

ANSWER**C. Maintain fault tolerance****EXPLANATION**

Checkpointing stores progress information so streaming jobs can recover from failures.

QUESTION 17

Question 17

A nightly ETL process runs only once per day.

What is the most cost-effective compute option?

- A. All-Purpose Cluster
- B. Job Cluster
- C. SQL Warehouse
- D. Interactive Cluster

ANSWER**B. Job Cluster****EXPLANATION**

Job Clusters:

- Start automatically
 - Run the job
 - Terminate automatically
- This reduces compute costs.

QUESTION 18

Question 18

Which DataFrame operation stores frequently used data in memory?

- A. optimize()
- B. count()
- C. cache()
- D. merge()

ANSWER**C. cache()****EXPLANATION**

Caching reduces repeated reads and recomputation of frequently accessed datasets.

QUESTION 19

Question 19

A Delta table contains billions of rows and queries frequently filter by region.

Which design choice can improve performance?

- A. Cache everything
- B. Use partitioning
- C. Convert to CSV
- D. Remove Delta format

ANSWER**B. Use partitioning****EXPLANATION**

Partitioning reduces the amount of data that must be scanned during query execution.

QUESTION 20

Question 20

Which Databricks feature provides a high-performance vectorized query engine?

- A. Photon
- B. MLflow
- C. Repos
- D. Workflow

ANSWER**A. Photon****EXPLANATION**

Photon is Databricks' optimized execution engine designed to improve SQL and DataFrame workload performance.

title: Databricks Certification Practice Test 02

description: Scenario-Based Certification Questions

DATABRICKS CERTIFICATION PRACTICE TEST 02

Practice Test 02 Questions 21 40 (20 Questions)

QUESTION 21

Question 21

A data engineering team receives customer data from multiple source systems. The incoming files often contain additional columns that were not present previously.

The team wants new columns to be added automatically without recreating the destination Delta table.

Which feature should be enabled?

- A. Partition Pruning
- B. Schema Evolution
- C. Z-Ordering
- D. Time Travel

ANSWER**B. Schema Evolution****EXPLANATION**

Schema Evolution allows Delta tables to automatically incorporate compatible schema changes, such as newly added columns.

QUESTION 22

Question 22

A data pipeline writes thousands of small files every hour.

Users report increasing query latency.

Which action should be performed FIRST?

- A. VACUUM the table
- B. OPTIMIZE the table
- C. DELETE old data
- D. Increase notebook size

ANSWER**B. OPTIMIZE the table****EXPLANATION**

Large numbers of small files negatively impact query performance. OPTIMIZE compacts these files into larger files.

QUESTION 23

Question 23

A company wants to restrict analysts from viewing salary information while allowing access to employee names and departments.

Which Unity Catalog capability should be used?

- A. Cluster Policies
- B. Column-Level Security
- C. Job Scheduling
- D. Delta Cache

ANSWER**B. Column-Level Security****EXPLANATION**

Column-level security allows sensitive columns to remain hidden while exposing permitted columns.

QUESTION 24

Question 24

A streaming job fails unexpectedly.

After restart, processing should continue from the last successfully processed record.

Which feature makes this possible?

- A. Time Travel
- B. Checkpointing
- C. Z-Ordering
- D. Broadcast Join

ANSWER**B. Checkpointing****EXPLANATION**

Checkpointing stores state and progress information that allows recovery after failures.

QUESTION 25

Question 25

A data engineer needs to investigate who modified a table and when modifications occurred.

Which feature should be queried?

- A. DESCRIBE HISTORY
- B. OPTIMIZE
- C. CACHE TABLE
- D. SHOW PARTITIONS

ANSWER**A. DESCRIBE HISTORY****EXPLANATION**

DESCRIBE HISTORY provides information regarding Delta table operations, users, timestamps and versions.

QUESTION 26

Question 26

A query joins a 5 TB fact table with a 5 MB dimension table.

Which optimization is most appropriate?

- A. Cartesian Join
- B. Broadcast Join
- C. Repartition Fact Table
- D. Cache Driver Node

ANSWER**B. Broadcast Join****EXPLANATION**

Broadcast joins eliminate expensive shuffles by sending the small dimension table to each executor.

QUESTION 27

Question 27

An analyst accidentally deletes records from a Delta table.

The team must restore the table to yesterday's state.

Which feature supports this requirement?

- A. Auto Loader
- B. Time Travel
- C. Photon
- D. MLflow

ANSWER**B. Time Travel****EXPLANATION**

Delta Time Travel provides access to previous table versions using timestamps or versions.

QUESTION 28

Question 28

Which architecture layer should contain business-level KPIs such as Monthly Revenue and Total Profit?

- A. Bronze
- B. Silver
- C. Gold
- D. Landing Zone

ANSWER**C. Gold****EXPLANATION**

Gold Layer contains aggregated and business-ready datasets.

QUESTION 29

Question 29

A company wants compute resources to terminate automatically after ETL completion.

Which compute type should be selected?

- A. All-Purpose Cluster
- B. Interactive Cluster
- C. Job Cluster
- D. SQL Editor

ANSWER**C. Job Cluster****EXPLANATION**

Job Clusters start automatically and terminate after workload completion.

QUESTION 30

Question 30

A table contains millions of records distributed unevenly across partitions.

One partition contains significantly more data than others.

What issue is occurring?

- A. Broadcasting
- B. Data Skew
- C. Vacuuming
- D. Caching

ANSWER**B. Data Skew**

EXPLANATION

Data skew occurs when workload distribution is highly imbalanced across partitions.

QUESTION 31

Question 31

Which feature automatically adjusts query plans during execution?

- A. Photon
- B. AQE
- C. Workspace
- D. Repos

ANSWER

B. AQE

EXPLANATION

Adaptive Query Execution dynamically optimizes query execution while jobs are running.

QUESTION 32

Question 32

A team requires centralized governance across multiple Databricks workspaces.

Which service should be implemented?

- A. Delta Cache
- B. Unity Catalog
- C. Photon
- D. Auto Loader

ANSWER

B. Unity Catalog

EXPLANATION

Unity Catalog centralizes data governance, permissions, lineage and auditing across workspaces.

QUESTION 33

Question 33

An engineer wants to ingest cloud files incrementally as they arrive.

Which Databricks feature is specifically designed for this use case?

- A. Auto Loader
- B. Dashboard
- C. Repo
- D. Notebook Job

ANSWER

A. Auto Loader

EXPLANATION

Auto Loader efficiently discovers and processes newly arriving files.

QUESTION 34

Question 34

A Delta table is partitioned by Region.

A query filters only Region = 'India'.

What optimization occurs?

- A. Broadcast Execution
- B. Partition Pruning
- C. Time Travel
- D. Checkpoint Recovery

ANSWER

B. Partition Pruning

EXPLANATION

Only relevant partitions are scanned, reducing read cost and improving performance.

QUESTION 35

Question 35

Which Unity Catalog object directly contains tables?

- A. Catalog
- B. Workspace
- C. Schema
- D. Cluster

ANSWER

C. Schema

EXPLANATION

Hierarchy:

Catalog Schema Table

QUESTION 36

Question 36

A pipeline requires historical tracking of customer address changes.

Which design pattern is most suitable?

- A. Append Only
- B. SCD Type 2
- C. Truncate and Load
- D. Snapshot Delete

ANSWER

B. SCD Type 2

EXPLANATION

SCD Type 2 stores historical versions and tracks record changes over time.

QUESTION 37

Question 37

Which command provides metadata about Delta table columns and data types?

- A. DESCRIBE TABLE
- B. VACUUM
- C. OPTIMIZE
- D. MERGE

ANSWER

A. DESCRIBE TABLE

EXPLANATION

DESCRIBE TABLE returns schema-related information.

QUESTION 38

Question 38

A dimension table is frequently reused in multiple transformations.

What should be considered to reduce repeated computation?

- A. DELETE
- B. CACHE
- C. MERGE
- D. VACUUM

ANSWER

B. CACHE

EXPLANATION

Caching stores data in memory for faster reuse.

QUESTION 39

Question 39

A company wants to understand upstream and downstream dependencies of a dataset. Which Unity Catalog capability provides this information?

- A. Lineage
- B. Photon
- C. Auto Loader
- D. Delta Live Tables

ANSWER

A. Lineage

EXPLANATION

Lineage provides visibility into data flow and dependencies.

QUESTION 40

Question 40

An engineer wants the highest query performance available for SQL workloads in Databricks. Which feature should be enabled?

- A. Photon
- B. Dashboard
- C. Repo
- D. Notebook Scheduling

ANSWER

A. Photon

EXPLANATION

Photon is Databricks' high-performance native execution engine optimized for analytics workloads.

title: Databricks Certification Practice Test 03

description: Real Company Scenario Based Questions

DATABRICKS CERTIFICATION PRACTICE TEST 03

Practice Test 03 Questions 41 60 (20 Questions)

QUESTION 41

Question 41

A global retail company receives daily sales files from 500 stores. Each store uploads its data separately throughout the day. The data engineering team notices that some stores occasionally resend the same file due to network failures.

The team wants to ensure duplicate records are not loaded into the reporting layer.

What is the MOST appropriate solution?

- A. Store all records in Gold layer
- B. Perform deduplication in Silver layer
- C. Use VACUUM daily
- D. Use OPTIMIZE after every load

ANSWER

B. Perform deduplication in Silver layer

EXPLANATION

The Silver layer is responsible for:

- Data cleansing

- Standardization
 - Deduplication
 - Quality improvements
- Bronze should preserve raw incoming data.
-

QUESTION 42

Question 42

An online banking company stores customer transaction history in a Delta table. A deployment issue accidentally updates thousands of account balances incorrectly.

Management asks the engineering team to restore the table exactly as it existed six hours ago.

Which solution requires the LEAST effort?

- A. Reprocess all source files
- B. Restore from CSV backup
- C. Use Delta Time Travel
- D. Create a new cluster

ANSWER

C. Use Delta Time Travel

EXPLANATION

Delta Time Travel allows access to previous versions of a table using timestamps or version numbers.

QUESTION 43

Question 43

A healthcare company runs a nightly pipeline that joins a 10 TB patient records table with a 50 MB hospital reference table.

The nightly job takes much longer than expected.

Which optimization would MOST likely improve performance?

- A. Broadcast the hospital reference table
- B. Convert Delta table to CSV
- C. Increase notebook timeout
- D. Use VACUUM

ANSWER

A. Broadcast the hospital reference table

EXPLANATION

The hospital table is small enough to broadcast, eliminating expensive data shuffles during joins.

QUESTION 44

Question 44

A logistics company stores shipment data partitioned by shipment_date.

Analysts usually query only the last 7 days of data.

What advantage does this partitioning strategy provide?

- A. Faster cluster startup
- B. Partition pruning
- C. Better notebook rendering
- D. Increased storage

ANSWER

B. Partition pruning

EXPLANATION

Databricks scans only the required partitions instead of the entire dataset.

QUESTION 45

Question 45

A media company loads video viewing events every few seconds.
Business users require dashboards to reflect data almost immediately.
Which processing approach should be used?

- A. Monthly Batch
- B. Weekly Batch
- C. Structured Streaming
- D. Manual Upload

ANSWER**C. Structured Streaming****EXPLANATION**

Structured Streaming enables continuous ingestion and near real-time reporting.

QUESTION 46

Question 46

A company has many Databricks workspaces across development, testing and production environments.
Security teams want a single location to manage permissions and governance.
Which service should be implemented?

- A. MLflow
- B. Unity Catalog
- C. Auto Loader
- D. Photon

ANSWER**B. Unity Catalog****EXPLANATION**

Unity Catalog centralizes:

- Permissions
- Governance
- Auditing
- Lineage

across workspaces.

QUESTION 47

Question 47

A manufacturing company receives IoT sensor files every minute.
Engineers do not want to manually identify new files landing in cloud storage.
Which Databricks feature is best suited?

- A. Auto Loader
- B. Repartition
- C. Vacuum
- D. Dashboard

ANSWER**A. Auto Loader****EXPLANATION**

Auto Loader automatically discovers and processes newly arriving files.

QUESTION 48

Question 48

A customer dimension table stores customer status changes.
Business users need to know both current and historical customer statuses.
Which design approach should be implemented?

- A. Overwrite table daily

- B. Append all records blindly
- C. SCD Type 2
- D. Truncate and Reload

ANSWER**C. SCD Type 2****EXPLANATION**

SCD Type 2 preserves historical versions while maintaining the current active record.

QUESTION 49**## Question 49**

A Delta table contains billions of records. Queries frequently filter by Country and City. Which optimization strategy would provide the greatest benefit?

- A. Convert to JSON
- B. Partition by Country
- C. Store as CSV
- D. Remove indexing

ANSWER**B. Partition by Country****EXPLANATION**

Partitioning allows Databricks to read fewer files during query execution.

QUESTION 50**## Question 50**

A data engineer notices that one Spark task processes significantly more data than all other tasks. The job completes slowly because a single executor takes much longer than the rest. What issue is occurring?

- A. Unity Governance
- B. Data Skew
- C. Broadcast Failure
- D. Metadata Corruption

ANSWER**B. Data Skew****EXPLANATION**

Data skew occurs when data distribution is uneven across partitions.

QUESTION 51**## Question 51**

A retail company writes transaction data continuously throughout the day. Several analysts run queries against the same table simultaneously. Which Databricks feature ensures readers always see consistent results during writes?

- A. ACID Transactions
- B. Cache Table
- C. Auto Loader
- D. Photon

ANSWER**A. ACID Transactions****EXPLANATION**

Delta Lake ACID guarantees consistency even when multiple readers and writers access data concurrently.

QUESTION 52

Question 52

A technology company has a nightly ETL workflow that runs once and then sits idle for the rest of the day.

Management wants to minimize cloud costs.

Which compute option is most appropriate?

- A. Always-on Interactive Cluster
- B. All Purpose Cluster
- C. Job Cluster
- D. SQL Editor

ANSWER**C. Job Cluster****EXPLANATION**

Job Clusters automatically start and terminate, reducing idle costs.

QUESTION 53

Question 53

A travel company stores booking information in a Delta table.

A new data source begins providing a new column called loyalty_tier.

The team wants the table to accept the new column automatically.

Which Delta capability supports this requirement?

- A. Checkpointing
- B. Schema Evolution
- C. Broadcast Join
- D. Delta Cache

ANSWER**B. Schema Evolution****EXPLANATION**

Schema Evolution allows compatible schema changes without recreating tables.

QUESTION 54

Question 54

A streaming pipeline suddenly fails because of a temporary infrastructure outage.

When restarted, the engineers want processing to continue from the last committed event.

Which feature is required?

- A. Photon
- B. Checkpointing
- C. ZORDER
- D. Delta Sharing

ANSWER**B. Checkpointing****EXPLANATION**

Checkpointing stores processing state and offsets for recovery.

QUESTION 55

Question 55

A pharmaceutical company must identify who modified a regulated dataset and when the changes occurred.

Which command provides this information?

- A. DESCRIBE HISTORY
- B. SHOW DATABASES
- C. OPTIMIZE
- D. CACHE TABLE

ANSWER**A. DESCRIBE HISTORY****EXPLANATION**

DESCRIBE HISTORY displays operation type, user, timestamp and version details.

QUESTION 56

Question 56

A data engineering team notices thousands of tiny files being created every hour in a Delta table.

Query performance gradually declines.

What should the team do?

- A. Increase notebook size
- B. Run OPTIMIZE
- C. Delete partitions
- D. Use CREATE VIEW

ANSWER**B. Run OPTIMIZE****EXPLANATION**

OPTIMIZE compacts numerous small files into larger files and improves read performance.

QUESTION 57

Question 57

An airline company has Bronze, Silver and Gold layers.

Business analysts use a Power BI dashboard for monthly KPI reporting.

Which layer should Power BI primarily consume?

- A. Bronze
- B. Silver
- C. Gold
- D. Landing

ANSWER**C. Gold****EXPLANATION**

Gold contains business-ready, aggregated and curated datasets.

QUESTION 58

Question 58

A global e-commerce platform frequently reruns the same expensive transformation multiple times within a notebook.

Engineers want to reduce repeated processing.

Which action should be considered?

- A. VACUUM
- B. CACHE
- C. DELETE
- D. MERGE

ANSWER**B. CACHE****EXPLANATION**

Caching stores frequently reused datasets in memory.

QUESTION 59

Question 59

A security team wants to hide customer Social Security Numbers while still allowing analysts to query other customer attributes.

Which capability should be implemented?

- A. Column-Level Security
- B. Photon
- C. Workflow Scheduler
- D. Broadcast Join

ANSWER**A. Column-Level Security****EXPLANATION**

Column-level security protects sensitive fields while exposing other attributes.

QUESTION 60

Question 60

A new data engineer joins a company and needs to understand how a Gold KPI table is generated from source systems.

Which Unity Catalog capability would help most?

- A. Lineage
- B. Cache
- C. Repartition
- D. Vacuum

ANSWER**A. Lineage****EXPLANATION**

Lineage shows:

- Source datasets
- Transformations
- Downstream dependencies
- End-to-end data flow

This helps engineers quickly understand data movement across the platform.

title: Databricks Certification Practice Test 04

description: Advanced Company Scenario Based Questions

DATABRICKS CERTIFICATION PRACTICE TEST 04

Practice Test 04 Questions 61 80 (20 Questions)

QUESTION 61

Question 61

A retail company loads product data from multiple vendors. Different vendors use different column names for the same business field. For example:

- vendor_product_id
- product_code
- item_id

The analytics team wants a standardized dataset for reporting.

In which Medallion layer should these column names typically be standardized?

- A. Landing
- B. Bronze
- C. Silver
- D. Gold

ANSWER**C. Silver****EXPLANATION**

Silver is responsible for data standardization and business rule implementation. Different source system structures are converted into a consistent format before moving to Gold.

QUESTION 62

Question 62

A financial services company stores customer transactions in a Delta table. Auditors require a complete record of every INSERT, UPDATE, and DELETE operation performed against the table.

Which Delta component provides this capability?

- A. Checkpoint Files
- B. Delta Transaction Log
- C. Broadcast Cache
- D. Cluster Events

ANSWER**B. Delta Transaction Log****EXPLANATION**

The Delta Transaction Log records all committed operations against the table and enables features such as ACID transactions and table history.

QUESTION 63

Question 63

An insurance company runs a report that always filters by:

- claim_year
- claim_month

What is the best partitioning strategy?

- A. Partition by customer_name
- B. Partition by claim_amount
- C. Partition by claim_year and claim_month
- D. No partitioning

ANSWER**C. Partition by claim_year and claim_month****EXPLANATION**

Partition columns should align with common query filters. This minimizes file scanning and improves performance.

QUESTION 64

Question 64

A newly hired Data Engineer notices that data is being copied multiple times between Bronze and Silver tables with no transformations being applied.

What is the biggest concern?

- A. Excessive storage and processing costs
- B. Missing dashboards
- C. Driver memory shortage
- D. Schema evolution failure

ANSWER**A. Excessive storage and processing costs****EXPLANATION**

Copying data without business value increases storage consumption and processing costs while providing little benefit.

QUESTION 65

Question 65

A company wants all development notebooks to be reviewed before being deployed to Production. Which Databricks feature best supports collaborative source control?

- A. Repos
- B. Dashboards
- C. SQL Warehouse
- D. Delta Cache

ANSWER**A. Repos****EXPLANATION**

Databricks Repos integrates with Git repositories and supports version control, code reviews, and collaborative development.

QUESTION 66

Question 66

An airline company receives flight information as nested JSON documents. Before analysts can query the data efficiently, the nested structures must be flattened. Which layer should typically perform this activity?

- A. Bronze
- B. Silver
- C. Gold
- D. Reporting Layer

ANSWER**B. Silver****EXPLANATION**

Flattening, cleaning, and transforming semi-structured data are common Silver-layer responsibilities.

QUESTION 67

Question 67

A company stores historical sales data going back ten years. Most users query only the last three months. Which approach will likely reduce query cost?

- A. Query the whole table
- B. Store data as text files
- C. Archive rarely accessed data
- D. Duplicate all records

ANSWER**C. Archive rarely accessed data****EXPLANATION**

Archiving cold data reduces storage and compute costs while keeping frequently accessed data readily available.

QUESTION 68

Question 68

A streaming pipeline receives duplicate events from an external system. The business wants each event processed only once. Which capability should the engineer investigate?

- A. Event Deduplication
- B. Dashboard Refresh
- C. Time Travel
- D. Photon

ANSWER

A. Event Deduplication**EXPLANATION**

Deduplication logic helps ensure repeated events do not impact downstream analytics.

QUESTION 69

Question 69

A healthcare analytics team wants to share curated datasets with another business unit without physically copying the data.

Which Databricks capability is designed for this use case?

- A. Auto Loader
- B. Delta Sharing
- C. Cache
- D. Checkpointing

ANSWER

B. Delta Sharing

EXPLANATION

Delta Sharing enables secure sharing of datasets without creating additional physical copies.

QUESTION 70

Question 70

An engineer notices that a job spends most of its time moving data between executors instead of processing records.

What is the likely bottleneck?

- A. Shuffle Operations
- B. Dashboard Rendering
- C. Notebook Formatting
- D. Delta Log Growth

ANSWER

A. Shuffle Operations

EXPLANATION

Large shuffle activity is often one of the most expensive portions of Spark execution.

QUESTION 71

Question 71

A company needs a process that automatically runs notebook A, then notebook B, then notebook C every night.

Which Databricks capability is most appropriate?

- A. Workflow
- B. Catalog
- C. Dashboard
- D. Photon

ANSWER

A. Workflow

EXPLANATION

Databricks Workflows orchestrate and schedule dependent tasks and notebooks.

QUESTION 72

Question 72

A retail organization needs to prevent developers from creating oversized clusters that increase cloud costs.

Which feature should be implemented?

- A. Cluster Policy
- B. Auto Loader
- C. Delta Sharing
- D. Catalog Explorer

ANSWER**A. Cluster Policy****EXPLANATION**

Cluster Policies enforce rules regarding cluster size, configuration, and governance.

QUESTION 73**## Question 73**

A notebook fails midway through processing because one source file contains unexpected null values. What should be added to improve pipeline reliability?

- A. Data Quality Validation
- B. Time Travel
- C. Repartition
- D. Dashboard Filters

ANSWER**A. Data Quality Validation****EXPLANATION**

Validation checks help detect and handle bad data before it impacts downstream systems.

QUESTION 74**## Question 74**

An engineering team wants a dataset refreshed every 15 minutes instead of waiting for a nightly batch. What business benefit is gained?

- A. Lower availability
- B. Improved data freshness
- C. Smaller schema
- D. Reduced governance

ANSWER**B. Improved data freshness****EXPLANATION**

More frequent processing provides more current information for decision-making.

QUESTION 75**## Question 75**

A company wants to track notebook execution success and failures across hundreds of pipelines. Which capability should be used?

- A. Workflow Monitoring
- B. Repartition
- C. Broadcast Join
- D. Delta Vacuum

ANSWER**A. Workflow Monitoring****EXPLANATION**

Workflow Monitoring provides visibility into job execution, failures, retries, and run history.

QUESTION 76**## Question 76**

An online marketplace receives customer reviews in multiple languages.

The data engineering team stores the original text exactly as received before performing any translation.

Why is this considered a good practice?

- A. Preserves original source data
- B. Reduces cluster size
- C. Eliminates partitions
- D. Replaces governance

ANSWER**A. Preserves original source data****EXPLANATION**

Raw data should be retained so transformations can be reapplied if business requirements change.

QUESTION 77**## Question 77**

A company wants to understand how a failed dashboard ultimately depends on a source ERP table.

Which capability provides end-to-end visibility?

- A. Lineage
- B. Broadcast Join
- C. Auto Loader
- D. Cache

ANSWER**A. Lineage****EXPLANATION**

Lineage traces dependencies from source systems through transformations to reporting assets.

QUESTION 78**## Question 78**

A large e-commerce company notices that engineers repeatedly create tables with inconsistent naming conventions.

What should be introduced to improve governance?

- A. Data Standards and Naming Conventions
- B. More Partitions
- C. More Clusters
- D. More Notebooks

ANSWER**A. Data Standards and Naming Conventions****EXPLANATION**

Standardized naming improves maintainability, discoverability, and governance.

QUESTION 79**## Question 79**

A data platform team wants every production pipeline failure to automatically notify support engineers.

Which approach is most appropriate?

- A. Alerting and Notifications
- B. Additional Partitions
- C. Delta Cache
- D. Time Travel

ANSWER**A. Alerting and Notifications**

EXPLANATION

Automated alerts reduce downtime by enabling faster response to failures.

QUESTION 80

Question 80

A customer analytics dataset is used by Marketing, Finance, and Operations teams.

The company wants a single trusted version of customer metrics.

What principle is being implemented?

- A. Single Source of Truth
- B. Broadcast Execution
- C. File Compaction
- D. Executor Isolation

ANSWER

A. Single Source of Truth

EXPLANATION

A Single Source of Truth ensures all teams use consistent business definitions and metrics.

title: Databricks Certification Practice Test 05

description: Fill in the Blank Certification Questions

DATABRICKS CERTIFICATION PRACTICE TEST 05

Practice Test 05 Questions 81 90 (10 Questions)

QUESTION 81

Question 81

A retail company stores raw files exactly as received from source systems. According to Medallion

Architecture, these files should typically be stored in the **** _____ **** layer.

<Select>

<option>Gold</option>

<option>Silver</option>

<option>Bronze</option>

<option>Semantic</option>

</Select>

ANSWER

Bronze

EXPLANATION

The Bronze layer stores raw, unmodified source data before cleansing or business transformations occur.

QUESTION 82

Question 82

A Delta table is receiving thousands of small files every day. To improve query performance, the

engineer should run **** _____ ****.

<Select>

<option>MERGE</option>

<option>OPTIMIZE</option>

<option>DESCRIBE HISTORY</option>

<option>CACHE</option>

</Select>

ANSWER

OPTIMIZE

EXPLANATION

OPTIMIZE compacts small files into larger files and improves read performance.

QUESTION 83

Question 83

A streaming pipeline fails unexpectedly. To resume processing from the last successful state, engineers should configure ** _____ **.

<Select>

<option>Photon</option>

<option>Checkpointing</option>

<option>Partitioning</option>

<option>ZORDER</option>

</Select>

ANSWER

Checkpointing

EXPLANATION

Checkpointing stores offsets and state information so processing can resume after failures.

QUESTION 84

Question 84

A healthcare company wants to hide patient SSN values while allowing access to other columns. Unity Catalog feature ** _____ ** should be implemented.

<Select>

<option>Column-Level Security</option>

<option>Broadcast Join</option>

<option>Delta Sharing</option>

<option>AQE</option>

</Select>

ANSWER

Column-Level Security

EXPLANATION

Column-Level Security restricts access to sensitive columns.

QUESTION 85

Question 85

A company needs historical tracking of customer address changes. The most appropriate design pattern is ** _____ **.

<Select>

<option>SCD Type 1</option>

<option>SCD Type 2</option>

<option>Overwrite Load</option>

<option>Snapshot Delete</option>

</Select>

ANSWER

SCD Type 2

EXPLANATION

SCD Type 2 preserves historical versions of records while maintaining current values.

QUESTION 86

Question 86

A Data Engineer wants to view all changes made to a Delta table over time. The command **** _____ **** provides this information.

<Select>

<option>SHOW DATABASES</option>

<option>DESCRIBE HISTORY</option>

<option>CACHE TABLE</option>

<option>CREATE VIEW</option>

</Select>

ANSWER**DESCRIBE HISTORY****EXPLANATION**

DESCRIBE HISTORY displays operation type, timestamp, user and Delta table versions.

QUESTION 87

Question 87

A small reference table is joined with a multi-terabyte fact table. The most efficient join strategy is usually **** _____ ****.

<Select>

<option>Broadcast Join</option>

<option>Cross Join</option>

<option>Cartesian Join</option>

<option>Nested Join</option>

</Select>

ANSWER**Broadcast Join****EXPLANATION**

Broadcasting the small table avoids expensive data shuffling.

QUESTION 88

Question 88

A company wants centralized governance across multiple Databricks workspaces. The recommended Databricks service is **** _____ ****.

<Select>

<option>MLflow</option>

<option>Unity Catalog</option>

<option>Photon</option>

<option>Notebook Workflow</option>

</Select>

ANSWER**Unity Catalog****EXPLANATION**

Unity Catalog centralizes permissions, lineage and governance.

QUESTION 89

Question 89

The hierarchy in Unity Catalog is:

Catalog _____ Table

<Select>

<option>Workspace</option>

<option>Schema</option>

<option>Cluster</option>
<option>Photon</option>
</Select>

ANSWER**Schema****EXPLANATION**

Objects are organized as Catalog Schema Table/View.

QUESTION 90**## Question 90**

A logistics company continuously receives new files in cloud storage. Databricks feature **_____** automatically discovers and ingests new files.

<Select>
<option>Auto Loader</option>
<option>Photon</option>
<option>Vacuum</option>
<option>Cluster Policy</option>
</Select>

ANSWER**Auto Loader****EXPLANATION**

Auto Loader automatically detects and processes newly arriving files.

title: Databricks Certification Practice Test 06

description: Intermediate and Advanced Fill in the Blank Questions

DATABRICKS CERTIFICATION PRACTICE TEST 06

Practice Test 06 Questions 91 110 (20 Questions)

QUESTION 91**## Question 91**

A Delta table already contains data. A Data Engineer attempts to write a DataFrame containing an additional column that does not exist in the destination table.

To allow the write operation without manually recreating the table, the engineer should enable

_____.

<Select>
<option>Schema Enforcement</option>
<option>Schema Evolution</option>
<option>Checkpointing</option>
<option>Z-Ordering</option>
</Select>

ANSWER**Schema Evolution****EXPLANATION**

Schema Evolution allows compatible schema changes such as adding new columns during write operations.

QUESTION 92

Question 92

A Spark application repeatedly performs a large aggregation that causes memory pressure on executors. Instead of storing the DataFrame only in memory, the engineer chooses to store some partitions on disk when memory becomes full.

This is achieved using **** _____ ****.

<Select>

<option>cache()</option>

<option>persist()</option>

<option>broadcast()</option>

<option>coalesce()</option>

</Select>

ANSWER**persist()****EXPLANATION**

persist() supports different storage levels and can spill data to disk when memory is insufficient.

QUESTION 93

Question 93

A data platform team wants queries filtering by customer_id to skip reading unnecessary files within the same partition.

The team should implement **** _____ ****.

<Select>

<option>Z-Ordering</option>

<option>Checkpointing</option>

<option>Cataloging</option>

<option>Versioning</option>

</Select>

ANSWER**Z-Ordering****EXPLANATION**

Z-Ordering colocates related values to reduce file scanning during queries.

QUESTION 94

Question 94

A Data Engineer wants a transformation to execute only when an action such as count() or collect() is triggered.

This Spark behavior is known as **** _____ ****.

<Select>

<option>Lazy Evaluation</option>

<option>Eager Execution</option>

<option>Partition Pruning</option>

<option>Broadcasting</option>

</Select>

ANSWER**Lazy Evaluation****EXPLANATION**

Spark builds the execution plan first and executes only when an action is requested.

QUESTION 95

Question 95

A streaming pipeline performs aggregations over incoming events.

To maintain aggregation state across micro-batches, Spark internally maintains **** _____ ****.

<Select>

<option>Workflow Metadata</option>

<option>State Store</option>

<option>Driver Cache</option>

<option>Cluster Policy</option>

</Select>

ANSWER**State Store****EXPLANATION**

The State Store tracks aggregation and streaming state between batches.

QUESTION 96

Question 96

A Data Engineer wants to reduce the number of partitions from 500 to 50 without triggering a full shuffle operation.

The preferred operation is **** _____ ****.

<Select>

<option>repartition()</option>

<option>coalesce()</option>

<option>explode()</option>

<option>collect()</option>

</Select>

ANSWER**coalesce()****EXPLANATION**

coalesce() is often more efficient when reducing partitions because it minimizes shuffling.

QUESTION 97

Question 97

A company wants to create reusable data transformation pipelines using declarative definitions instead of manually writing orchestration logic.

The Databricks solution designed for this is **** _____ ****.

<Select>

<option>Delta Live Tables</option>

<option>Photon</option>

<option>Repos</option>

<option>SQL Warehouse</option>

</Select>

ANSWER**Delta Live Tables****EXPLANATION**

DLT simplifies ETL development, quality enforcement, and pipeline management.

QUESTION 98

Question 98

A Delta table receives updates from multiple users simultaneously.

Consistency is maintained through **** _____ ****.

<Select>

<option>ACID Transactions</option>
<option>Executor Caching</option>
<option>Workflow Scheduling</option>
<option>Notebook Versioning</option>
</Select>

ANSWER**ACID Transactions****EXPLANATION**

ACID guarantees reliable concurrent reads and writes.

QUESTION 99**## Question 99**

An engineer wants Spark to automatically infer the best number of shuffle partitions during query execution.

This optimization is largely supported by **_____**.

<Select>
<option>Adaptive Query Execution</option>
<option>Auto Loader</option>
<option>Delta Sharing</option>
<option>MLflow</option>
</Select>

ANSWER**Adaptive Query Execution****EXPLANATION**

AQE dynamically adjusts execution strategies while the query runs.

QUESTION 100**## Question 100**

A notebook contains several transformations but no action.

As a result, Spark performs **_____**.

<Select>
<option>No Execution</option>
<option>Full Execution</option>
<option>Driver Restart</option>
<option>Auto Optimization</option>
</Select>

ANSWER**No Execution****EXPLANATION**

Transformations alone do not trigger execution in Spark.

QUESTION 101**## Question 101**

A Data Engineer wants to split an array column into multiple rows.

The most appropriate Spark function is **_____**.

<Select>
<option>flatten()</option>
<option>explode()</option>
<option>collect_set()</option>
<option>substring()</option>
</Select>

ANSWER**explode()****EXPLANATION**

explode() converts array elements into separate rows.

QUESTION 102

Question 102

A query requires ranking employees by salary within each department.

The engineer should use a ** ____ ** function.

<Select>

<option>Aggregate</option>

<option>Window</option>

<option>Scalar</option>

<option>String</option>

</Select>

ANSWER**Window****EXPLANATION**

Window functions perform calculations across related rows while retaining row-level detail.

QUESTION 103

Question 103

A team wants pipeline execution to stop automatically whenever mandatory columns contain null values.

This requirement is commonly implemented through ** ____ ** rules.

<Select>

<option>Data Quality</option>

<option>Cluster Policy</option>

<option>Autoscaling</option>

<option>Partitioning</option>

</Select>

ANSWER**Data Quality****EXPLANATION**

Data quality rules prevent bad data from reaching downstream consumers.

QUESTION 104

Question 104

A Spark job contains significantly more tasks than available CPU cores.

The extra tasks are placed into a ** ____ ** until resources become available.

<Select>

<option>Queue</option>

<option>Schema</option>

<option>Catalog</option>

<option>Checkpoint</option>

</Select>

ANSWER**Queue****EXPLANATION**

Spark schedules pending tasks until executor resources become free.

QUESTION 105

Question 105

A company wants machine learning experiments, parameters and metrics to be tracked centrally. Databricks provides ** _____ ** for this purpose.

<Select>

<option>MLflow</option>

<option>Photon</option>

<option>Delta Engine</option>

<option>State Store</option>

</Select>

ANSWER**MLflow****EXPLANATION**

MLflow manages experiment tracking and model lifecycle management.

QUESTION 106

Question 106

An engineer needs to transform a nested struct column into separate top-level columns. This process is commonly known as ** _____ ** the data.

<Select>

<option>Flattening</option>

<option>Vacuuming</option>

<option>Broadcasting</option>

<option>Versioning</option>

</Select>

ANSWER**Flattening****EXPLANATION**

Flattening converts complex nested structures into a relational format.

QUESTION 107

Question 107

A Spark application spends excessive time serializing objects between executors. This activity contributes to execution ** _____ **.

<Select>

<option>Overhead</option>

<option>Governance</option>

<option>Lineage</option>

<option>Partitioning</option>

</Select>

ANSWER**Overhead****EXPLANATION**

Serialization overhead can significantly impact large-scale Spark workloads.

QUESTION 108

Question 108

A data scientist needs a read-only SQL endpoint dedicated to BI dashboards. The most suitable Databricks compute option is a ** _____ **.

<Select>

<option>SQL Warehouse</option>

<option>Job Cluster</option>

<option>Interactive Notebook</option>
<option>Workflow Task</option>
</Select>

ANSWER

SQL Warehouse

EXPLANATION

SQL Warehouses are optimized for BI and analytics workloads.

QUESTION 109

Question 109

A company wants notebook changes to be stored in Git and reviewed through pull requests.

Databricks supports this workflow using **** _____ ****.

<Select>
<option>Repos</option>
<option>DLT</option>
<option>AQE</option>
<option>Photon</option>
</Select>

ANSWER

Repos

EXPLANATION

Repos integrates Databricks development with Git providers.

QUESTION 110

Question 110

A large table is being processed by Spark. The engineer wants records with the same key to be redistributed evenly across executors before a transformation.

The engineer should use **** _____ ****.

<Select>
<option>repartition()</option>
<option>display()</option>
<option>collect()</option>
<option>cache()</option>
</Select>

ANSWER

repartition()

EXPLANATION

repartition() performs a shuffle and redistributes data across partitions more evenly.

title: Databricks Certification Practice Test 07

description: Advanced Multi-Select Questions

DATABRICKS CERTIFICATION PRACTICE TEST 07

Practice Test 07 Questions 111 160 (50 Questions)

QUESTION 111

Question 111

A company wants to improve reliability of production ETL pipelines.
Which TWO practices should be implemented?

- A. Data Quality Checks
- B. Manual Execution
- C. Monitoring & Alerting
- D. Random Sampling

ANSWER**A, C****EXPLANATION**

Data quality checks prevent bad data.

Monitoring and alerting help detect failures quickly.

QUESTION 112

Question 112

Which TWO actions trigger Spark execution?

- A. select()
- B. filter()
- C. count()
- D. collect()

ANSWER**C, D****EXPLANATION**

count() and collect() are actions. Spark executes the DAG only when an action occurs.

QUESTION 113

Question 113

Which TWO operations are considered wide transformations?

- A. groupBy()
- B. join()
- C. select()
- D. withColumn()

ANSWER**A, B****EXPLANATION**

Wide transformations require shuffling data between partitions.

QUESTION 114

Question 114

Which TWO metrics should a Data Engineer monitor during workload tuning?

- A. Shuffle Read
- B. Shuffle Write
- C. Notebook Font
- D. Dashboard Color

ANSWER**A, B**

QUESTION 115

Question 115

A company wants code promotions from Dev QA Prod.
Which TWO capabilities support this process?

- A. Repos
- B. Git Branching
- C. Photon
- D. SQL Warehouse

ANSWER**A, B**

QUESTION 116

Question 116

Which TWO benefits are provided by Delta Live Tables?

- A. Data Quality Enforcement
- B. Pipeline Management
- C. Driver Scaling
- D. Dashboard Creation

ANSWER**A, B**

QUESTION 117

Question 117

A Data Engineer wants to reduce cloud costs.
Which TWO actions are recommended?

- A. Auto Termination
- B. Right Sizing Clusters
- C. Duplicate Tables
- D. Increase Executors Permanently

ANSWER**A, B**

QUESTION 118

Question 118

Which TWO datasets are usually suitable for caching?

- A. Frequently Reused Lookup Tables
- B. Intermediate Results Reused Multiple Times
- C. One-Time Reads
- D. Archived Data

ANSWER**A, B**

QUESTION 119

Question 119

A Spark job processes customer records.
Which TWO factors can cause performance degradation?

- A. Skewed Partitions
- B. Excessive Shuffle
- C. Proper Partitioning
- D. Predicate Pushdown

ANSWER**A, B**

QUESTION 120

Question 120

Which TWO activities belong to Data Governance?

- A. Access Control
- B. Data Classification
- C. Repartitioning
- D. Broadcasting

ANSWER**A, B**

QUESTION 121

Question 121

Which TWO features help improve discoverability of datasets?

- A. Business Metadata
- B. Data Catalog
- C. VACUUM
- D. Photon

ANSWER**A, B**

QUESTION 122

Question 122

Which TWO attributes make a dataset trusted?

- A. High Data Quality
- B. Clear Ownership
- C. Large File Size
- D. More Partitions

ANSWER**A, B**

QUESTION 123

Question 123

A company wants to audit data access.

Which TWO details are important?

- A. Who accessed data
- B. When access occurred
- C. Notebook Theme
- D. Cluster Color

ANSWER**A, B**

QUESTION 124

Question 124

Which TWO tasks are commonly automated in ETL platforms?

- A. Scheduling
- B. Retry Logic
- C. Manual Approvals
- D. Documentation Writing

ANSWER

A, B

QUESTION 125

Question 125

Which TWO indicators suggest a pipeline requires optimization?

- A. Growing Runtime
- B. Increasing Resource Usage
- C. Smaller Input Volumes
- D. Better SLA Compliance

ANSWER

A, B

QUESTION 126

Question 126

Which TWO benefits come from partition pruning?

- A. Reduced Data Scan
- B. Faster Query Execution
- C. Increased Storage
- D. More Metadata

ANSWER

A, B

QUESTION 127

Question 127

A production deployment failed.

Which TWO capabilities help rollback changes?

- A. Git Version Control
- B. Release Management
- C. Dashboard Filters
- D. Photon

ANSWER

A, B

QUESTION 128

Question 128

Which TWO characteristics are associated with streaming systems?

- A. Continuous Processing
- B. Low Latency
- C. Annual Refreshes
- D. Manual Loads

ANSWER

A, B

QUESTION 129

Question 129

Which TWO business benefits come from data lineage?

- A. Faster Impact Analysis
- B. Easier Troubleshooting
- C. Larger Clusters
- D. More Workers

ANSWER**A, B**

QUESTION 130

Question 130

An organization follows medallion architecture.

Which TWO responsibilities belong in Gold?

- A. Business KPIs
- B. Aggregated Metrics
- C. Raw Files
- D. Source Backups

ANSWER**A, B**

QUESTION 131

Question 131

Which TWO techniques help reduce data duplication?

- A. Master Data Management
- B. Standardized Data Sources
- C. More Copies
- D. More Tables

ANSWER**A, B**

QUESTION 132

Question 132

Which TWO things should a support engineer check first when a workflow fails?

- A. Error Logs
- B. Failed Task Details
- C. Dashboard Layout
- D. Notebook Colors

ANSWER**A, B**

QUESTION 133

Question 133

A company wants SLA compliance.

Which TWO metrics are most important?

- A. Pipeline Runtime
- B. Pipeline Success Rate
- C. Screen Resolution
- D. Notebook Size

ANSWER**A, B**

QUESTION 134

Question 134

Which TWO practices improve maintainability?

- A. Modular Code
- B. Reusable Functions

- C. Duplicate Logic
- D. Hardcoded Values

ANSWER

A, B

QUESTION 135

Question 135

Which TWO characteristics define a production-ready pipeline?

- A. Monitoring
- B. Error Handling
- C. Random Processing
- D. No Logging

ANSWER

A, B

QUESTION 136

Question 136

Which TWO activities are part of root cause analysis?

- A. Log Investigation
- B. Dependency Review
- C. Ignoring Errors
- D. Restart Everything

ANSWER

A, B

QUESTION 137

Question 137

Which TWO factors influence cluster sizing?

- A. Data Volume
- B. Concurrency
- C. Notebook Font
- D. Dashboard Theme

ANSWER

A, B

QUESTION 138

Question 138

A data product is consumed by multiple teams.

Which TWO things should exist?

- A. Documentation
- B. Data Owner
- C. Random Naming
- D. Local Spreadsheets

ANSWER

A, B

QUESTION 139

Question 139

Which TWO goals are achieved through standard schemas?

- A. Easier Integration
- B. Consistency
- C. More Duplicates
- D. More Failures

ANSWER**A, B**

QUESTION 140

Question 140

Which TWO areas should be included in production monitoring?

- A. Resource Utilization
- B. Job Failures
- C. Keyboard Shortcuts
- D. Browser Settings

ANSWER**A, B**

QUESTION 141

Question 141

Which TWO symptoms suggest under-partitioning?

- A. Large Tasks
- B. Long Processing Time
- C. Empty Tables
- D. Missing Catalogs

ANSWER**A, B**

QUESTION 142

Question 142

Which TWO activities improve platform security?

- A. Least Privilege Access
- B. Regular Access Reviews
- C. Shared Credentials
- D. Open Permissions

ANSWER**A, B**

QUESTION 143

Question 143

Which TWO practices improve data trustworthiness?

- A. Validation Rules
- B. Data Ownership
- C. Duplicate Definitions
- D. Untracked Changes

ANSWER**A, B**

QUESTION 144

Question 144

Which TWO reasons justify workflow retries?

- A. Temporary Network Issues
- B. Transient Cloud Errors
- C. Bad Business Logic
- D. Missing Requirements

ANSWER**A, B**

QUESTION 145

Question 145

Which TWO items should be documented for datasets?

- A. Refresh Frequency
- B. Business Definition
- C. Cluster Wallpaper
- D. Notebook Background

ANSWER**A, B**

QUESTION 146

Question 146

Which TWO principles support Data Mesh initiatives?

- A. Domain Ownership
- B. Self-Service Data Platforms
- C. Centralized Spreadsheets
- D. Manual Data Requests

ANSWER**A, B**

QUESTION 147

Question 147

Which TWO indicators show healthy operations?

- A. Stable Runtime
- B. Low Failure Rate
- C. Rising Incidents
- D. Frequent Outages

ANSWER**A, B**

QUESTION 148

Question 148

Which TWO components should be considered before deleting production data?

- A. Downstream Dependencies
- B. Retention Policies
- C. Notebook Theme
- D. Dashboard Style

ANSWER**A, B**

QUESTION 149

Question 149

Which TWO benefits come from reusable ETL frameworks?

- A. Reduced Development Time
- B. Consistent Standards
- C. More Manual Work
- D. Duplicate Logic

ANSWER**A, B**

QUESTION 150

Question 150

Which TWO skills are most important for a Senior Data Engineer?

- A. System Design
- B. Troubleshooting
- C. Dashboard Coloring
- D. PowerPoint Animations

ANSWER**A, B**

QUESTION 151

Question 151

A global retail company processes customer transactions from stores located in 50 countries.

The company currently experiences the following issues:

- Duplicate customer records
- Inconsistent product naming
- Missing customer identifiers
- Frequent downstream reporting failures

The data engineering team is redesigning the Silver layer to improve overall data quality and reliability.

Which actions should be implemented?

- Remove duplicate records
- Standardize product names
- Validate mandatory fields
- Build executive dashboards
- Apply business quality rules
- Store raw files unchanged

ANSWER**Remove duplicate records****Standardize product names****Validate mandatory fields****Apply business quality rules****EXPLANATION**

Silver layer responsibilities include:

- Cleansing
- Deduplication
- Standardization
- Validation
- Data quality enforcement

Executive dashboards belong to Gold. Raw files belong to Bronze.

QUESTION 152

Question 152

A large airline company runs a nightly Databricks workflow that occasionally misses its SLA because processing times fluctuate significantly between executions.

Investigation reveals:

Uneven data distribution

Large shuffle operations

Rapid growth in data volume

Which actions may help improve performance?

Review partition strategy

Analyze shuffle metrics

Investigate skewed keys

Increase dashboard refresh rate

Evaluate Adaptive Query Execution

Rename notebook files

ANSWER

Review partition strategy

Analyze shuffle metrics

Investigate skewed keys

Evaluate Adaptive Query Execution

EXPLANATION

All four actions directly address Spark execution bottlenecks.

QUESTION 153

Question 153

A healthcare company stores highly sensitive patient information in Databricks.

Security auditors require stronger governance controls.

Which controls should be considered?

Least privilege access

Data classification

Access auditing

Open permissions for all users

Periodic access reviews

Shared administrator accounts

ANSWER

Least privilege access

Data classification

Access auditing

Periodic access reviews

EXPLANATION

These are standard governance and security controls for regulated environments.

QUESTION 154

Question 154

A newly hired Data Engineer reviews an ETL framework that is used by dozens of pipelines across the organization.

The engineering manager wants future pipelines to be easier to maintain.

Which practices support maintainability?

Modular code

Reusable components

Consistent naming conventions

Copy-paste transformations

Documentation standards

Hardcoded parameters

ANSWER

Modular code

Reusable components
Consistent naming conventions
Documentation standards

QUESTION 155

Question 155

An e-commerce company plans to create a trusted enterprise data platform.

Business users often complain that reports from different departments show different revenue numbers.

Which actions help establish a trusted data foundation?

Define common business metrics
Assign dataset ownership
Implement data quality validation
Allow independent metric definitions
Document data definitions
Eliminate governance reviews

ANSWER

Define common business metrics
Assign dataset ownership
Implement data quality validation
Document data definitions

QUESTION 156

Question 156

A support engineer is troubleshooting a failed production workflow.

Which information should be reviewed first?

Error messages
Task execution history
Pipeline logs
Cluster resource utilization
User profile pictures
Dashboard color themes

ANSWER

Error messages
Task execution history
Pipeline logs
Cluster resource utilization

QUESTION 157

Question 157

A manufacturing company expects IoT data volume to triple during the next year.

The architecture team wants to ensure scalability.

Which factors should be evaluated?

Growth projections
Storage strategy
Processing capacity
Cost management
Employee parking spaces
Office seating plans

ANSWER

Growth projections
Storage strategy
Processing capacity
Cost management

QUESTION 158

Question 158

A company wants better operational visibility across all production pipelines.
Which capabilities should be implemented?

- Alerting
- Monitoring dashboards
- SLA tracking
- Failure notifications
- Random notebook execution
- Manual health checks only

ANSWER

- Alerting**
 - Monitoring dashboards**
 - SLA tracking**
 - Failure notifications**
-

QUESTION 159

Question 159

A global bank is preparing for a regulatory audit.
Which metadata information would auditors likely consider valuable?

- Data owners
- Data lineage
- Access history
- Dataset classifications
- Wallpaper configuration
- Browser themes

ANSWER

- Data owners**
 - Data lineage**
 - Access history**
 - Dataset classifications**
-

QUESTION 160

Question 160

A company wants to build data products that can be safely consumed by multiple business units.
Which characteristics should those products have?

- Clear ownership
- Documentation
- Quality controls
- Defined SLAs
- Unknown refresh schedule
- Inconsistent schema

ANSWER

- Clear ownership**
- Documentation**
- Quality controls**
- Defined SLAs**

title: Databricks Certification Practice Test 08**description: Select All That Apply - Advanced Scenario Questions**

DATABRICKS CERTIFICATION PRACTICE TEST 08Practice Test 08 Questions 161 173 (13 Questions)

QUESTION 161

Question 161

A global retail company has hundreds of Databricks jobs running daily. Recently, several jobs failed but nobody was notified until business users reported missing reports.

Which actions should the Data Engineering team implement?

- Configure failure notifications
- Create operational dashboards
- Implement workflow monitoring
- Disable job logging
- Define SLA tracking
- Increase notebook font size

ANSWER

- Configure failure notifications**
- Create operational dashboards**
- Implement workflow monitoring**
- Define SLA tracking**

EXPLANATION

Production systems should proactively detect, monitor and alert on failures.

QUESTION 162

Question 162

A company has 50 engineers writing pipelines independently.

Over time they created inconsistent table names such as:

- cust
- customer_tbl
- cst_master

Which actions would improve platform governance?

- Establish naming standards
- Create documentation guidelines
- Define data ownership
- Allow teams to create any naming pattern they prefer
- Create governance reviews

ANSWER

- Establish naming standards**
- Create documentation guidelines**
- Define data ownership**
- Create governance reviews**

QUESTION 163

Question 163

A banking company wants developers to submit code changes through Pull Requests before deployment.

Which practices support this requirement?

- Git Integration
- Branch Strategy
- Code Review Process
- Store all code in notebooks only
- CI/CD Validation

ANSWER

- Git Integration**
- Branch Strategy**
- Code Review Process**

CI/CD Validation

QUESTION 164

Question 164

A company notices rising cloud costs every month.

The FinOps team has been asked to investigate.

Which areas should they analyze?

- Idle clusters
- Cluster utilization
- Job frequency
- Orphaned datasets
- Employee attendance records

ANSWER

- Idle clusters**
- Cluster utilization**
- Job frequency**
- Orphaned datasets**

QUESTION 165

Question 165

A newly joined Data Engineer wants to understand how a dashboard metric is created.

Which information would be most useful?

- Dataset lineage
- Data owners
- Business definitions
- Dependency mapping
- Desktop wallpaper

ANSWER

- Dataset lineage**
- Data owners**
- Business definitions**
- Dependency mapping**

QUESTION 166

Question 166

An insurance company wants to ensure that all production datasets are trustworthy.

Which controls should be implemented?

- Data quality validation
- Ownership assignment
- Certification process
- Monitoring
- Random manual edits

ANSWER

- Data quality validation**
- Ownership assignment**
- Certification process**
- Monitoring**

QUESTION 167

Question 167

A platform team is designing reusable ETL frameworks.

Which benefits would they expect?

- Reduced development effort
- Consistent standards
- Faster onboarding
- Simplified maintenance
- More duplicated code

ANSWER

- Reduced development effort**
- Consistent standards**
- Faster onboarding**
- Simplified maintenance**

QUESTION 168

Question 168

A company has strict compliance requirements.

Auditors want visibility into who accessed datasets.

Which capabilities should exist?

- Access Logs
- Audit Trails
- User Activity Tracking
- Data Access Reporting
- Hidden Access History

ANSWER

- Access Logs**
- Audit Trails**
- User Activity Tracking**
- Data Access Reporting**

QUESTION 169

Question 169

A telecommunications company processes billions of records daily.

The architecture team expects volumes to double next year.

Which factors should be considered during architecture planning?

- Scalability
- Storage growth
- Processing growth
- Cost forecasting
- Office paint color

ANSWER

- Scalability**
- Storage growth**
- Processing growth**
- Cost forecasting**

QUESTION 170

Question 170

A company wants production incidents resolved faster.

Which actions will help?

- Alerting
- Root Cause Analysis
- Runbook Documentation

Monitoring Dashboards
Ignore warning messages

ANSWER

Alerting
Root Cause Analysis
Runbook Documentation
Monitoring Dashboards

QUESTION 171

Question 171

A Lakehouse platform serves Finance, Marketing and Sales departments.
Which practices help maintain a Single Source of Truth?

Standard metric definitions
Centralized datasets
Governance process
Documented business rules
Department-specific revenue calculations

ANSWER

Standard metric definitions
Centralized datasets
Governance process
Documented business rules

QUESTION 172

Question 172

A Data Product is consumed by multiple teams.
What characteristics should the product have?

Defined SLA
Data owner
Documentation
Quality checks
Unknown refresh schedule

ANSWER

Defined SLA
Data owner
Documentation
Quality checks

QUESTION 173

Question 173

A company wants to improve developer productivity.
Which engineering practices most commonly help?

Reusable libraries
Shared frameworks
Coding standards
Automated deployments
Manual environment setup

ANSWER

Reusable libraries
Shared frameworks
Coding standards
Automated deployments

title: Databricks Certification Practice Test 09

description: Statement Analysis Questions

DATABRICKS CERTIFICATION PRACTICE TEST 09Practice Test 09 Questions 175 200 (26 Questions)

QUESTION 175

Question 175

Statement 1:

A Data Contract helps data producers and consumers agree on schema expectations.

Statement 2:

Data Contracts eliminate the need for data quality validation.

- A. Only Statement 1 is correct
- B. Only Statement 2 is correct
- C. Both Statements are correct
- D. Neither Statement is correct

ANSWER**A****EXPLANATION**

Data Contracts define expectations, but data quality validation is still required.

QUESTION 176

Question 176

Statement 1:

A CI/CD pipeline can automate promotion of code between environments.

Statement 2:

Production deployments should always be performed manually.

ANSWER**A****EXPLANATION**

CI/CD enables automated deployments. Manual deployment is not always required.

QUESTION 177

Question 177

Statement 1:

Metadata helps users discover datasets.

Statement 2:

Metadata has no value for governance initiatives.

ANSWER**A**

QUESTION 178

Question 178

Statement 1:

Data Products should have clearly defined owners.

Statement 2:

Data Products should have defined SLAs.

ANSWER**C**

EXPLANATION

Ownership and SLA definitions are both core data product principles.

QUESTION 179

Question 179

Statement 1:

A business glossary helps standardize business terminology.

Statement 2:

Every department should define revenue independently.

ANSWER

A

QUESTION 180

Question 180

Statement 1:

Observability helps engineers understand platform health.

Statement 2:

Observability is useful only after failures occur.

ANSWER

A

EXPLANATION

Observability is useful before, during and after incidents.

QUESTION 181

Question 181

Statement 1:

Data lineage helps measure downstream impact of changes.

Statement 2:

Lineage information is useful during incident analysis.

ANSWER

C

QUESTION 182

Question 182

Statement 1:

Data ownership improves accountability.

Statement 2:

Datasets without owners are easier to govern.

ANSWER

A

QUESTION 183

Question 183

Statement 1:

Release approvals can reduce deployment risk.

Statement 2:

All code changes should go directly to production.

ANSWER

A

QUESTION 184

Question 184

Statement 1:

Operational dashboards help monitor pipeline health.

Statement 2:

Operational dashboards replace logging.

ANSWER

A

QUESTION 185

Question 185

Statement 1:

Domain teams are responsible for their data in Data Mesh.

Statement 2:

Data Mesh removes all governance requirements.

ANSWER

A

QUESTION 186

Question 186

Statement 1:

A runbook helps responders handle production incidents.

Statement 2:

Runbooks can reduce incident resolution time.

ANSWER

C

QUESTION 187

Question 187

Statement 1:

Monitoring can identify trends before failures occur.

Statement 2:

Monitoring should only focus on failed jobs.

ANSWER

A

QUESTION 188

Question 188

Statement 1:

Dataset certification increases consumer confidence.

Statement 2:

Certification guarantees data can never contain errors.

ANSWER

A

QUESTION 189

Question 189

Statement 1:

A semantic layer helps standardize business metrics.

Statement 2:

Semantic layers have no value for reporting consistency.

ANSWER

A

QUESTION 190

Question 190

Statement 1:

Documentation supports onboarding new engineers.

Statement 2:

Documentation becomes unnecessary after deployment.

ANSWER

A

QUESTION 191

Question 191

Statement 1:

Root Cause Analysis aims to identify underlying problems.

Statement 2:

Root Cause Analysis should focus only on symptoms.

ANSWER

A

QUESTION 192

Question 192

Statement 1:

Alert fatigue can occur when excessive notifications are generated.

Statement 2:

Every informational event should trigger an alert.

ANSWER

A

QUESTION 193

Question 193

Statement 1:

A service level objective helps define reliability expectations.

Statement 2:

Reliability targets should be measurable.

ANSWER

C

QUESTION 194

Question 194

Statement 1:

Data stewardship supports governance goals.

Statement 2:

Data stewardship is unrelated to data quality.

ANSWER

A

QUESTION 195

Question 195

Statement 1:

A platform team often provides shared capabilities for engineering teams.

Statement 2:

Each team should build its own governance framework independently.

ANSWER

A

QUESTION 196

Question 196

Statement 1:

Impact analysis should occur before major schema changes.

Statement 2:

Consumer dependencies should be reviewed before deployment.

ANSWER

C

QUESTION 197

Question 197

Statement 1:

Incident postmortems help organizations learn from failures.

Statement 2:

Postmortems should focus on assigning blame.

ANSWER

A

QUESTION 198

Question 198

Statement 1:

Dataset discoverability improves self-service analytics.

Statement 2:

Searchable catalogs can improve discoverability.

ANSWER

C

QUESTION 199

Question 199

Statement 1:

Data freshness can influence business decision-making.

Statement 2:

Refresh schedules should be documented.

ANSWER**C**

QUESTION 200

Question 200

Statement 1:

Standardized engineering practices improve maintainability.

Statement 2:

Reusable design patterns can reduce development effort.

ANSWER**C****EXPLANATION**

Both practices improve scalability, consistency and long-term maintainability across enterprise data platforms.

title: Databricks Certification Practice Test 10

description: Machine Learning and Lakehouse AI Scenario Questions

DATABRICKS CERTIFICATION PRACTICE TEST 10Practice Test 10 Questions 201-230 (30 Questions)

QUESTION 201

Question 201

A retail company has trained a demand forecasting model.

The model performs well in development but business users cannot reproduce previous prediction results after a retraining cycle.

What is the BEST action?

- A. Create larger clusters
- B. Track model versions
- C. Increase partitions
- D. Run OPTIMIZE

ANSWER**B. Track model versions****EXPLANATION**

Model versioning ensures engineers can identify, reproduce and compare model outputs over time.

QUESTION 202

Question 202

A bank has multiple data scientists training fraud detection models.

Management wants all experiments tracked in a central location.

What is the BEST solution?

- A. Store results in Excel
- B. Email experiment metrics
- C. Use MLflow Experiment Tracking

D. Create more notebooks

ANSWER

C. Use MLflow Experiment Tracking

QUESTION 203

Question 203

A telecom company has 500 machine learning features stored in different notebooks.

Data scientists frequently calculate the same features multiple times.

What is the BEST action?

- A. Build a Feature Store
- B. Add more clusters
- C. Create additional workspaces
- D. Store features in CSV files

ANSWER

A. Build a Feature Store

QUESTION 204

Question 204

A recommendation model suddenly becomes less accurate.

Customer behavior changed significantly during the last month.

What is the MOST likely issue?

- A. Schema Drift
- B. Data Drift
- C. Photon Failure
- D. Notebook Timeout

ANSWER

B. Data Drift

QUESTION 205

Question 205

A healthcare company must explain why an AI model approved one patient while rejecting another.

What is the BEST approach?

- A. Remove model documentation
- B. Improve model explainability
- C. Delete predictions
- D. Hide training logic

ANSWER

B. Improve model explainability

QUESTION 206

Question 206

A data science team repeatedly trains models using different hyperparameter values.

What is the BEST capability to track?

- A. Cluster Names
- B. Experiment Parameters
- C. Dashboard Colors
- D. Catalog Structure

ANSWER

B. Experiment Parameters

QUESTION 207

Question 207

A manufacturing company wants machine learning models automatically retrained when performance drops below a threshold.

What is the BEST approach?

- A. MLOps Pipeline
- B. Manual Emails
- C. CSV Exports
- D. More SQL Warehouses

ANSWER**A. MLOps Pipeline**

QUESTION 208

Question 208

A credit scoring model predicts significantly better in training than production.

What should engineers investigate first?

- A. Overfitting
- B. Notebook Font
- C. Cluster Policy
- D. Unity Catalog

ANSWER**A. Overfitting**

QUESTION 209

Question 209

A company wants consistent customer segmentation features shared across dozens of ML models.

What is the BEST strategy?

- A. Duplicate Logic
- B. Feature Store
- C. Manual Feature Creation
- D. Separate Definitions

ANSWER**B. Feature Store**

QUESTION 210

Question 210

A machine learning pipeline fails because expected columns are missing from incoming datasets.

What should be implemented?

- A. Feature Validation
- B. Dashboard Alerts Only
- C. Additional Catalogs
- D. More Workers

ANSWER**A. Feature Validation**

QUESTION 211

Question 211

A logistics company wants to compare Model V1 and Model V2 before deployment.

What is the BEST method?

- A. A/B Testing
- B. Create More Tables
- C. Rename Models

D. Delete Old Version

ANSWER

A. A/B Testing

QUESTION 212

Question 212

A fraud model is retrained every night.

Management wants automatic approval only if accuracy improves.

What is the BEST design?

- A. Model Validation Step
- B. Manual SQL Query
- C. Repartition Data
- D. Change Clusters

ANSWER

A. Model Validation Step

QUESTION 213

Question 213

A company wants a complete history of model deployments.

What should be maintained?

- A. Deployment Registry
- B. Temporary Views
- C. Query Cache
- D. Dashboard History

ANSWER

A. Deployment Registry

QUESTION 214

Question 214

A marketing model is deployed to production.

Engineers need a way to quickly rollback after deployment issues.

What is the BEST strategy?

- A. Versioned Model Registry
- B. Delete Previous Models
- C. Disable Tracking
- D. Remove Metadata

ANSWER

A. Versioned Model Registry

QUESTION 215

Question 215

A recommendation engine uses stale customer activity data.

What business risk exists?

- A. Reduced Prediction Quality
- B. Better Governance
- C. Faster Clusters
- D. Improved Recall

ANSWER

A. Reduced Prediction Quality

QUESTION 216

Question 216

A company wants training and inference features generated using identical business logic. What should be prioritized?

- A. Feature Consistency
- B. More Partitions
- C. Manual Processing
- D. Larger Notebooks

ANSWER**A. Feature Consistency**

QUESTION 217

Question 217

A support team wants alerts whenever model accuracy drops below 90%. What capability should be implemented?

- A. Model Monitoring
- B. Dashboard Themes
- C. Auto Loader
- D. Repartitioning

ANSWER**A. Model Monitoring**

QUESTION 218

Question 218

An ML Engineer wants to understand which dataset version was used during training. What should be tracked?

- A. Training Data Lineage
- B. Worker Count
- C. Notebook Colors
- D. Browser Extensions

ANSWER**A. Training Data Lineage**

QUESTION 219

Question 219

A company trains hundreds of models every month. Management wants standardized deployment procedures. What should be created?

- A. MLOps Framework
- B. Manual Checklist
- C. Shared Spreadsheet
- D. Email Process

ANSWER**A. MLOps Framework**

QUESTION 220

Question 220

A model succeeds technically but produces decisions that violate business policies. What should be added?

- A. Business Rule Validation
- B. More Training Data
- C. Smaller Clusters

D. More Partitions

ANSWER

A. Business Rule Validation

QUESTION 221

Question 221

A customer churn model heavily favors one customer group and disadvantages another. What should be investigated?

- A. Bias Detection
- B. Auto Scaling
- C. Data Compression
- D. Driver Memory

ANSWER

A. Bias Detection

QUESTION 222

Question 222

A retailer wants near real-time product recommendations. What architecture is most appropriate?

- A. Batch Recommendation Weekly
- B. Real-Time Inference Pipeline
- C. Monthly Export
- D. CSV Prediction Files

ANSWER

B. Real-Time Inference Pipeline

QUESTION 223

Question 223

A company stores prediction results but not input features. Months later they cannot explain model decisions. What key practice was missing?

- A. Prediction Traceability
- B. Partition Pruning
- C. AQE
- D. Checkpointing

ANSWER

A. Prediction Traceability

QUESTION 224

Question 224

An ML team wants reproducible model training. What should be versioned?

- A. Code
- B. Data
- C. Parameters
- D. All of the Above

ANSWER

D. All of the Above

QUESTION 225**## Question 225**

A fraud model generates thousands of predictions daily.

Engineers need confidence that production behavior matches training expectations.

What should be implemented?

- A. Continuous Model Monitoring
- B. Less Logging
- C. Manual Reviews Only
- D. More Catalogs

ANSWER

A. Continuous Model Monitoring

QUESTION 226**## Question 226**

A company wants to prevent deployment of unapproved models.

What is the BEST control?

- A. Deployment Approval Workflow
- B. Manual Naming Standards
- C. More Clusters
- D. Additional Views

ANSWER

A. Deployment Approval Workflow

QUESTION 227**## Question 227**

A machine learning team discovers duplicate features across multiple projects.

What should be created?

- A. Central Feature Repository
- B. Additional Workspaces
- C. More Notebooks
- D. More Schemas

ANSWER

A. Central Feature Repository

QUESTION 228**## Question 228**

A model requires customer age, income and transaction history.

These are examples of:

- A. Features
- B. Experiments
- C. Workflows
- D. Dashboards

ANSWER

A. Features

QUESTION 229**## Question 229**

A prediction service experiences latency spikes.

What should be reviewed first?

- A. Endpoint Performance Metrics
- B. Notebook Theme
- C. Workspace Name

D. Catalog Naming

ANSWER

A. Endpoint Performance Metrics

QUESTION 230

Question 230

A company wants to compare thousands of experiment runs quickly.
What capability is most valuable?

- A. Experiment Tracking Dashboard
- B. SQL Formatting
- C. Table Comments
- D. Repartitioning

ANSWER

A. Experiment Tracking Dashboard

title: Databricks Certification Practice Test 11

description: AI, GenAI and Advanced ML Scenario Questions

DATABRICKS CERTIFICATION PRACTICE TEST 11

Practice Test 11 Questions 231 260 (30 Questions)

QUESTION 231

Question 231

A retail company has deployed a Large Language Model (LLM) chatbot for customer support.
Customers report that the chatbot occasionally provides incorrect return policies.
What is the MOST appropriate recommendation?

- A. Increase cluster size
- B. Add Retrieval-Augmented Generation (RAG)
- C. Create more workspaces
- D. Reduce notebook count

ANSWER

B. Add Retrieval-Augmented Generation (RAG)

EXPLANATION

RAG allows the chatbot to retrieve current company policies before generating answers.

QUESTION 232

Question 232

A banking organization wants an AI assistant to answer questions using internal documents without retraining the model every week.
What is the BEST approach?

- A. Retrain daily
- B. Store documents in Excel
- C. Implement Vector Search
- D. Create more catalogs

ANSWER

C. Implement Vector Search

QUESTION 233**## Question 233**

A healthcare provider needs to identify whether an LLM-generated answer originated from approved company knowledge sources.

What capability should be prioritized?

- A. Response Traceability
- B. More GPUs
- C. More Partitions
- D. Dashboard Scheduling

ANSWER

A. Response Traceability

QUESTION 234**## Question 234**

A company wants AI-generated responses to cite supporting documents.

What is the BEST recommendation?

- A. Use manual screenshots
- B. Enable source attribution
- C. Increase worker nodes
- D. Disable monitoring

ANSWER

B. Enable source attribution

QUESTION 235**## Question 235**

A customer service chatbot sometimes generates fabricated information.

This problem is commonly known as:

- A. Serialization
- B. Lineage
- C. Hallucination
- D. Partition Drift

ANSWER

C. Hallucination

QUESTION 236**## Question 236**

An enterprise wants to prevent employees from sending confidential information to external AI systems.

What should be implemented?

- A. AI Governance Controls
- B. More Storage
- C. Additional Dashboards
- D. More Pipelines

ANSWER

A. AI Governance Controls

QUESTION 237**## Question 237**

A legal department requires every model deployment to pass compliance reviews before production release.

What is the BEST approach?

- A. Manual Emails
- B. Governance Approval Workflow

- C. Larger Clusters
- D. More Repositories

ANSWER**B. Governance Approval Workflow**

QUESTION 238

Question 238

A logistics company wants to measure whether generated AI responses are helpful.

What should be collected?

- A. User Feedback Metrics
- B. Cluster Names
- C. Table Comments
- D. CPU Count

ANSWER**A. User Feedback Metrics**

QUESTION 239

Question 239

A retail recommendation engine stores product descriptions as embeddings.

What is the PRIMARY purpose?

- A. Dashboard Creation
- B. Semantic Similarity Search
- C. Workflow Scheduling
- D. Cluster Monitoring

ANSWER**B. Semantic Similarity Search**

QUESTION 240

Question 240

An LLM solution must answer only from approved company documentation.

What architecture is most appropriate?

- A. Open Internet Search Only
- B. RAG Architecture
- C. Manual Text Entry
- D. Weekly Retraining

ANSWER**B. RAG Architecture**

QUESTION 241

Question 241

A company discovers that a recently deployed AI model responds differently to the same question after multiple releases.

What should be investigated?

- A. Model Version Changes
- B. Column Aliases
- C. Dashboard Layout
- D. Storage Accounts

ANSWER**A. Model Version Changes**

QUESTION 242**## Question 242**

A financial institution wants complete visibility into prompts sent to AI systems.
What should be captured?

- A. Prompt Logs
- B. Cluster Policies
- C. Job Names
- D. Partition Metadata

ANSWER**A. Prompt Logs**

QUESTION 243**## Question 243**

A company wants faster AI response retrieval from millions of documents.
What should be introduced?

- A. Vector Index
- B. CSV Files
- C. Static Dashboards
- D. More Schemas

ANSWER**A. Vector Index**

QUESTION 244**## Question 244**

A support chatbot needs access to recently updated product documentation every hour.
What is the BEST recommendation?

- A. Monthly Retraining
- B. Continuous Knowledge Refresh
- C. Delete Old Documents
- D. Disable Indexing

ANSWER**B. Continuous Knowledge Refresh**

QUESTION 245**## Question 245**

An enterprise AI team wants to compare answer quality across different LLMs.
What should be established?

- A. Evaluation Framework
- B. Additional Catalogs
- C. New Visual Theme
- D. More Worker Nodes

ANSWER**A. Evaluation Framework**

QUESTION 246**## Question 246**

A company wants to identify questions where an AI assistant performs poorly.
What should engineers analyze?

- A. Failed Interaction Patterns
- B. Notebook Colors
- C. Cluster Ownership
- D. Driver Memory Only

ANSWER**A. Failed Interaction Patterns**

QUESTION 247

Question 247

An insurance provider requires AI responses to remain consistent with regulatory guidelines.
What should be introduced?

- A. Guardrails
- B. More Clusters
- C. More Dashboards
- D. More Catalogs

ANSWER**A. Guardrails**

QUESTION 248

Question 248

A retail company wants to understand which documents contributed to a generated answer.
What capability provides this?

- A. Retrieval Metadata
- B. Checkpointing
- C. Repartitioning
- D. Compaction

ANSWER**A. Retrieval Metadata**

QUESTION 249

Question 249

An AI application serves thousands of users simultaneously.
Response latency increases significantly during peak hours.
What should be reviewed first?

- A. Model Serving Metrics
- B. Schema Definitions
- C. Catalog Structure
- D. Dataset Names

ANSWER**A. Model Serving Metrics**

QUESTION 250

Question 250

A company has multiple AI assistants serving different departments.
Management wants consistent governance, monitoring and deployment standards across all assistants.
What is the MOST appropriate strategy?

- A. Centralized AI Platform Governance
- B. Independent Department Standards
- C. Separate Monitoring Per User
- D. Manual Tracking Documents

ANSWER**A. Centralized AI Platform Governance****EXPLANATION**

Centralized governance improves consistency, compliance, security, monitoring and operational control across enterprise AI solutions.

QUESTION 251

Question 251

An airline company plans to build an AI assistant that helps maintenance engineers troubleshoot aircraft issues using maintenance manuals, repair histories, and engineering bulletins.

What should be created first?

- A. Vectorized Knowledge Base
- B. Additional Dashboards
- C. New Workspaces
- D. SQL Warehouse

ANSWER**A. Vectorized Knowledge Base**

QUESTION 252

Question 252

A company wants an AI assistant to answer only questions related to HR policies and refuse unrelated topics.

What is the BEST recommendation?

- A. Domain Restriction Rules
- B. More GPUs
- C. More Clusters
- D. Additional Repositories

ANSWER**A. Domain Restriction Rules**

QUESTION 253

Question 253

An ML team wants to understand why users are abandoning an AI application.

What should be analyzed?

- A. Conversation Analytics
- B. Cluster Tags
- C. Catalog Comments
- D. Data Compression Ratios

ANSWER**A. Conversation Analytics**

QUESTION 254

Question 254

A company deploys a document-search AI solution.

Users complain that answers are outdated.

What is the MOST likely recommendation?

- A. Refresh Source Indexes More Frequently
- B. Add More Dashboards
- C. Increase Notebook Size
- D. Add More Schemas

ANSWER**A. Refresh Source Indexes More Frequently**

QUESTION 255

Question 255

A GenAI application serves five countries.

Management wants to compare response quality by region.

What should be implemented?

- A. Regional Evaluation Metrics
- B. More Storage Accounts
- C. More Catalogs
- D. Cluster Renaming

ANSWER**A. Regional Evaluation Metrics**

QUESTION 256

Question 256

A company wants AI-generated content reviewed before being sent to customers.

What process should be introduced?

- A. Human-in-the-Loop Validation
- B. Daily Cluster Restarts
- C. Dashboard Archiving
- D. Workspace Duplication

ANSWER**A. Human-in-the-Loop Validation**

QUESTION 257

Question 257

A support organization wants AI systems to automatically escalate uncertain responses to human agents.

What capability should be implemented?

- A. Confidence Threshold Routing
- B. Additional Partitions
- C. Bulk Retraining
- D. Dashboard Monitoring

ANSWER**A. Confidence Threshold Routing**

QUESTION 258

Question 258

A company wants to prevent prompt injection attacks against an enterprise AI assistant.

What should be prioritized?

- A. Input Validation and Prompt Security Controls
- B. More Compute Nodes
- C. Additional Databases
- D. More BI Reports

ANSWER**A. Input Validation and Prompt Security Controls**

QUESTION 259

Question 259

An enterprise wants to compare the cost and quality of multiple foundation models before selecting one.

What should be performed?

- A. Benchmark Evaluation
- B. Notebook Cleanup
- C. Schema Optimization

D. Workspace Migration

ANSWER

A. Benchmark Evaluation

QUESTION 260

Question 260

A newly hired Data Engineer joins a company building a Lakehouse AI platform.

Management wants a scalable solution where Data Engineering, Analytics, Machine Learning, and Generative AI share the same trusted data foundation.

What architectural principle should guide the implementation?

- A. Unified Lakehouse Architecture
- B. Separate Data Platforms per Team
- C. Department-Owned CSV Files
- D. Independent AI Data Sources

ANSWER

A. Unified Lakehouse Architecture

EXPLANATION

A Unified Lakehouse Architecture reduces silos, improves governance, promotes reuse, and supports analytics, ML, and GenAI workloads from a single trusted platform.

title: Databricks Certification Practice Test 12

description: Advanced AI & ML Engineering Scenarios

DATABRICKS CERTIFICATION PRACTICE TEST 12

Practice Test 12 Questions 261 300 (40 Questions)

QUESTION 261

Question 261

A retail company has developed a fraud detection model.

Training accuracy remains above 95%, but production accuracy falls every month even though no code changes were deployed.

What is the MOST likely explanation?

- A. Concept Drift
- B. Cluster Drift
- C. Catalog Drift
- D. Workspace Drift

ANSWER

A. Concept Drift

Explanation

Concept drift occurs when the relationship between features and target outcomes changes over time.

QUESTION 262

Question 262

A healthcare company retrains models every week.

Management wants assurance that newly trained models outperform currently deployed models before promotion.

What should be implemented?

- A. Champion-Challenger Framework
- B. Additional Partitions
- C. More Clusters
- D. Additional Workflows

ANSWER**A. Champion-Challenger Framework**

QUESTION 263

Question 263

A recommendation system requires feature values that are less than 30 seconds old.

What architecture is most suitable?

- A. Batch Feature Pipeline
- B. Real-Time Feature Serving
- C. Monthly Feature Refresh
- D. Offline Feature Store Only

ANSWER**B. Real-Time Feature Serving**

QUESTION 264

Question 264

An ML team discovers that training data contains significantly fewer examples of rare fraud events than normal transactions.

What ML challenge exists?

- A. Class Imbalance
- B. Schema Evolution
- C. Partition Pruning
- D. Feature Lineage

ANSWER**A. Class Imbalance**

QUESTION 265

Question 265

A company wants to understand which input features most influenced a model prediction.

What capability should be implemented?

- A. Feature Importance Analysis
- B. VACUUM
- C. Partition Compaction
- D. Cluster Policy

ANSWER**A. Feature Importance Analysis**

QUESTION 266

Question 266

A retailer trains a model using historical sales data.

Engineers later discover future sales information accidentally existed in training data.

What issue occurred?

- A. Data Leakage
- B. Time Travel
- C. Data Skew
- D. Caching

ANSWER**A. Data Leakage**

QUESTION 267

Question 267

A data scientist wants to automatically search thousands of parameter combinations.
What capability is most appropriate?

- A. Hyperparameter Optimization
- B. Table Optimization
- C. Auto Scaling
- D. Metadata Indexing

ANSWER**A. Hyperparameter Optimization**

QUESTION 268

Question 268

An enterprise AI platform wants standardized model packaging regardless of programming language.
What should be prioritized?

- A. Portable Model Artifacts
- B. Cluster Standardization
- C. Workspace Naming
- D. Dashboard Consistency

ANSWER**A. Portable Model Artifacts**

QUESTION 269

Question 269

A manufacturing company wants to predict machine failures before they occur.
What ML category best fits this use case?

- A. Predictive Maintenance
- B. Sentiment Analysis
- C. Recommendation Engine
- D. Semantic Search

ANSWER**A. Predictive Maintenance**

QUESTION 270

Question 270

A bank requires every prediction generated by a model to be reproducible years later.
What is most critical?

- A. Model and Data Versioning
- B. Additional Storage Accounts
- C. Extra Clusters
- D. More Dashboards

ANSWER**A. Model and Data Versioning**

QUESTION 271

Question 271

A company wants to detect previously unseen fraudulent behavior patterns.
Which approach is most appropriate?

- A. Anomaly Detection
- B. Batch Aggregation
- C. Data Cataloging
- D. SQL Warehousing

ANSWER**A. Anomaly Detection**

QUESTION 272

Question 272

A model performs well in North America but poorly in Asia.

What should engineers investigate first?

- A. Regional Feature Differences
- B. Notebook Language
- C. Storage Tier
- D. Driver Configuration

ANSWER**A. Regional Feature Differences**

QUESTION 273

Question 273

A company wants an AI model deployment strategy where only 10% of traffic initially reaches the new model.

What should be implemented?

- A. Canary Deployment
- B. Full Cutover
- C. Batch Retraining
- D. Reindexing

ANSWER**A. Canary Deployment**

QUESTION 274

Question 274

An ML platform team wants to compare production predictions with actual outcomes over time.

What should be measured?

- A. Model Performance Decay
- B. Table Compression
- C. Catalog Growth
- D. Cluster Tags

ANSWER**A. Model Performance Decay**

QUESTION 275

Question 275

A company has hundreds of trained models.

Management wants visibility into which models are actively used.

What should be tracked?

- A. Model Usage Analytics
- B. Dashboard Themes
- C. Schema Comments
- D. Table Vacuums

ANSWER**A. Model Usage Analytics**

QUESTION 276

Question 276

A recommendation model requires user clickstream behavior within seconds.
Which data pattern is required?

- A. Event-Driven Feature Updates
- B. Quarterly Refresh
- C. Weekly Snapshot
- D. Annual Archive

ANSWER**A. Event-Driven Feature Updates**

QUESTION 277

Question 277

A support organization wants AI systems to explain confidence for each answer generated.
What should be exposed?

- A. Confidence Scores
- B. Driver Memory
- C. Storage Metrics
- D. Cluster IDs

ANSWER**A. Confidence Scores**

QUESTION 278

Question 278

A team trains multiple models using the same customer features.
Management wants one approved version of features.
What principle should be applied?

- A. Feature Governance
- B. Workspace Isolation
- C. Manual Duplication
- D. Independent Feature Definitions

ANSWER**A. Feature Governance**

QUESTION 279

Question 279

A logistics company predicts shipment delays.
Business users want predictions before delays occur, not after.
What type of analytics is being requested?

- A. Predictive Analytics
- B. Descriptive Analytics
- C. Historical Analytics
- D. Static Reporting

ANSWER**A. Predictive Analytics**

QUESTION 280

Question 280

A model generates excellent results but requires 30 seconds per prediction.
The business requires sub-second responses.
What should be optimized?

- A. Inference Latency

- B. Metadata Size
- C. Catalog Structure
- D. Workspace Count

ANSWER**A. Inference Latency**

QUESTION 281

Question 281

A company notices predictions change after retraining despite identical parameters.

What should be investigated?

- A. Training Dataset Changes
- B. Dashboard Updates
- C. Workspace Names
- D. Table Comments

ANSWER**A. Training Dataset Changes**

QUESTION 282

Question 282

An AI assistant must answer customer questions in 25 different languages.

What capability becomes critical?

- A. Multilingual Model Support
- B. Cluster Auto Scaling
- C. Dataset Partitioning
- D. Notebook Repositories

ANSWER**A. Multilingual Model Support**

QUESTION 283

Question 283

A company wants AI applications to learn from customer feedback over time.

What process should exist?

- A. Feedback Loop Integration
- B. Quarterly Shutdown
- C. Manual Retraining Only
- D. Schema Enforcement

ANSWER**A. Feedback Loop Integration**

QUESTION 284

Question 284

A fraud detection model consumes 10,000 features.

Only 50 significantly affect outcomes.

What engineering task may improve performance?

- A. Feature Selection
- B. Larger Dashboards
- C. Additional Catalogs
- D. Repartitioning Views

ANSWER**A. Feature Selection**

QUESTION 285

Question 285

A company wants every feature used in production predictions to be traceable back to source systems. What should be established?

- A. Feature Lineage
- B. Dashboard Ownership
- C. Workspace Hierarchy
- D. Cluster Pools

ANSWER**A. Feature Lineage**

QUESTION 286

Question 286

A model serving endpoint experiences sudden traffic spikes. What architecture principle helps absorb demand?

- A. Horizontal Scalability
- B. Manual Scaling
- C. Fixed Capacity
- D. Static Resources

ANSWER**A. Horizontal Scalability**

QUESTION 287

Question 287

An AI application produces responses that vary significantly for identical prompts. What should engineers investigate?

- A. Model Determinism Settings
- B. File Compaction
- C. Catalog Structure
- D. Auto Loader Configuration

ANSWER**A. Model Determinism Settings**

QUESTION 288

Question 288

A company wants to identify whether model performance differs across age groups. What discipline is being applied?

- A. Fairness Evaluation
- B. Cluster Governance
- C. Data Cataloging
- D. Query Optimization

ANSWER**A. Fairness Evaluation**

QUESTION 289

Question 289

A financial institution wants to evaluate risk before deploying a new model. What practice should be performed?

- A. Model Risk Assessment
- B. Repartitioning
- C. Data Archiving
- D. Workflow Scheduling

ANSWER**A. Model Risk Assessment**

QUESTION 290

Question 290

A chatbot must access company policies, legal agreements, and procedure manuals from multiple repositories.

What architecture is most suitable?

- A. Enterprise Knowledge Retrieval Architecture
- B. Static FAQ Page
- C. Manual Search Process
- D. Spreadsheet Repository

ANSWER**A. Enterprise Knowledge Retrieval Architecture**

QUESTION 291

Question 291

A model predicts customer churn.

Business leaders want an explanation for the top reasons contributing to churn.

What should be implemented?

- A. Explainable AI Techniques
- B. More Clusters
- C. More Catalogs
- D. Additional Dashboards

ANSWER**A. Explainable AI Techniques**

QUESTION 292

Question 292

A company wants to compare model quality across multiple business units using a common evaluation standard.

What should be established?

- A. Standard Evaluation Metrics
- B. Independent Scoring Methods
- C. Department-Specific KPIs Only
- D. Manual Reviews

ANSWER**A. Standard Evaluation Metrics**

QUESTION 293

Question 293

A model retraining workflow consumes significant compute resources.

Management wants to train only when statistically necessary.

What strategy should be adopted?

- A. Trigger-Based Retraining
- B. Daily Forced Retraining
- C. Weekly Manual Retraining
- D. Continuous Full Retraining

ANSWER**A. Trigger-Based Retraining**

QUESTION 294

Question 294

An enterprise wants to identify duplicate AI models solving the same business problem.
What capability would help?

- A. Central Model Catalog
- B. Additional Workspaces
- C. Larger Clusters
- D. More SQL Warehouses

ANSWER**A. Central Model Catalog**

QUESTION 295

Question 295

A recommendation system serves millions of customers globally.
The business requires uninterrupted availability during deployments.
What approach should be used?

- A. Blue-Green Deployment
- B. Hard Cutover
- C. Manual Replacement
- D. Service Shutdown

ANSWER**A. Blue-Green Deployment**

QUESTION 296

Question 296

A company wants to ensure AI-generated content follows corporate brand guidelines.
What should be implemented?

- A. Response Policy Validation
- B. Additional Partitions
- C. More Compute Nodes
- D. New Workspaces

ANSWER**A. Response Policy Validation**

QUESTION 297

Question 297

A data scientist wants to quickly identify which training runs produced the highest business value.
What should be tracked alongside technical metrics?

- A. Business Outcome Metrics
- B. Notebook Size
- C. Workspace Count
- D. Table Counts

ANSWER**A. Business Outcome Metrics**

QUESTION 298

Question 298

A GenAI platform serves thousands of users.
Management wants visibility into token consumption and operational costs.
What should be monitored?

- A. AI Usage Cost Metrics
- B. Dashboard Layout

- C. Schema Definitions
- D. Cluster Names

ANSWER**A. AI Usage Cost Metrics**

QUESTION 299

Question 299

An organization wants a single framework governing ML models, GenAI applications, feature definitions, evaluations, and deployments.

What should be established?

- A. Enterprise AI Operating Model
- B. Independent Team Processes
- C. Local Tracking Spreadsheets
- D. Department-Specific Governance

ANSWER**A. Enterprise AI Operating Model**

QUESTION 300

Question 300

A global enterprise is building a Lakehouse AI platform expected to support analytics, machine learning, GenAI, AI agents, and future AI workloads.

What architectural characteristic is MOST important?

- A. Extensibility
- B. Fixed Design
- C. Department Isolation
- D. Manual Scaling

ANSWER**A. Extensibility****EXPLANATION**

An extensible architecture allows AI platforms to evolve as new models, frameworks, and business requirements emerge.
