

Data Engineering Associate with Databricks Practice Exam (Sample)

Study Guide



Everything you need from our exam experts!

Copyright © 2026 by Examzify - A Kaluba Technologies Inc. product.

ALL RIGHTS RESERVED.

No part of this book may be reproduced or transferred in any form or by any means, graphic, electronic, or mechanical, including photocopying, recording, web distribution, taping, or by any information storage retrieval system, without the written permission of the author.

Notice: Examzify makes every reasonable effort to obtain accurate, complete, and timely information about this product from reliable sources.

SAMPLE

Table of Contents

Copyright	1
Table of Contents	2
Introduction	3
How to Use This Guide	4
Questions	5
Answers	8
Explanations	10
Next Steps	16

Introduction

Preparing for a certification exam can feel overwhelming, but with the right tools, it becomes an opportunity to build confidence, sharpen your skills, and move one step closer to your goals. At Examzify, we believe that effective exam preparation isn't just about memorization, it's about understanding the material, identifying knowledge gaps, and building the test-taking strategies that lead to success.

This guide was designed to help you do exactly that.

Whether you're preparing for a licensing exam, professional certification, or entry-level qualification, this book offers structured practice to reinforce key concepts. You'll find a wide range of multiple-choice questions, each followed by clear explanations to help you understand not just the right answer, but why it's correct.

The content in this guide is based on real-world exam objectives and aligned with the types of questions and topics commonly found on official tests. It's ideal for learners who want to:

- Practice answering questions under realistic conditions,
- Improve accuracy and speed,
- Review explanations to strengthen weak areas, and
- Approach the exam with greater confidence.

We recommend using this book not as a stand-alone study tool, but alongside other resources like flashcards, textbooks, or hands-on training. For best results, we recommend working through each question, reflecting on the explanation provided, and revisiting the topics that challenge you most.

Remember: successful test preparation isn't about getting every question right the first time, it's about learning from your mistakes and improving over time. Stay focused, trust the process, and know that every page you turn brings you closer to success.

Let's begin.

How to Use This Guide

This guide is designed to help you study more effectively and approach your exam with confidence. Whether you're reviewing for the first time or doing a final refresh, here's how to get the most out of your Examzify study guide:

1. Start with a Diagnostic Review

Skim through the questions to get a sense of what you know and what you need to focus on. Your goal is to identify knowledge gaps early.

2. Study in Short, Focused Sessions

Break your study time into manageable blocks (e.g. 30 - 45 minutes). Review a handful of questions, reflect on the explanations.

3. Learn from the Explanations

After answering a question, always read the explanation, even if you got it right. It reinforces key points, corrects misunderstandings, and teaches subtle distinctions between similar answers.

4. Track Your Progress

Use bookmarks or notes (if reading digitally) to mark difficult questions. Revisit these regularly and track improvements over time.

5. Simulate the Real Exam

Once you're comfortable, try taking a full set of questions without pausing. Set a timer and simulate test-day conditions to build confidence and time management skills.

6. Repeat and Review

Don't just study once, repetition builds retention. Re-attempt questions after a few days and revisit explanations to reinforce learning. Pair this guide with other Examzify tools like flashcards, and digital practice tests to strengthen your preparation across formats.

There's no single right way to study, but consistent, thoughtful effort always wins. Use this guide flexibly, adapt the tips above to fit your pace and learning style. You've got this!

Questions

- 1. What is the primary purpose of Delta Sharing?**
 - A. To exchange data only within a single organization.**
 - B. To support data governance and security protocols while sharing data across platforms.**
 - C. To create data backups.**
 - D. To compile data from various sources into one database.**
- 2. Which JSON file management technique helps maintain version control?**
 - A. Editing in real-time through the interface.**
 - B. Creating backups before editing.**
 - C. Downloading the JSON description before changes.**
 - D. Using cloud storage to manage configurations.**
- 3. What SQL command is used to view the history of a Delta table?**
 - A. SELECT * FROM history**
 - B. DESCRIBE HISTORY**
 - C. SHOW TABLE HISTORY**
 - D. VIEW DATABASE HISTORY**
- 4. Which SQL function is utilized to generate the current timestamp?**
 - A. Input_file_name()**
 - B. Current_timestamp()**
 - C. Filter()**
 - D. Collect_set()**
- 5. What SQL syntax is used to define a Common Table Expression (CTE)?**
 - A. DECLARE cte AS SELECT ...;**
 - B. WITH cte AS SELECT ...;**
 - C. CACHE cte SELECT ...;**
 - D. CREATE CTE AS SELECT ...;**

- 6. What is one of the significant benefits of Delta Live Tables?**
- A. It requires extensive coding for scalability**
 - B. It simplifies both batch and streaming data processing**
 - C. It eliminates the need for data modeling**
 - D. It restricts data accessibility to few users**
- 7. What is the command to start a Delta Live Table?**
- A. CREATE STREAM TABLE**
 - B. CREATE INCREMENTAL LIVE TABLE**
 - C. START LIVE TABLE**
 - D. LAUNCH DELTA TABLE**
- 8. In Databricks, what is meant by a cluster?**
- A. A user account for data access**
 - B. A set of computational resources for running workloads**
 - C. A tool for data visualization**
 - D. A template for machine learning models**
- 9. What is the function of autoscaling in Databricks?**
- A. It modifies the data structure in the cluster**
 - B. It adjusts the number of nodes based on workload**
 - C. It locks the cluster to a fixed number of nodes**
 - D. It allows users to add data manually**
- 10. What SQL function is used to transform an array into a set of rows?**
- A. Collect_set**
 - B. Explode**
 - C. Flatten**
 - D. Array_distinct**

Answers

SAMPLE

1. B
2. C
3. B
4. B
5. B
6. B
7. B
8. B
9. B
10. B

SAMPLE

Explanations

SAMPLE

1. What is the primary purpose of Delta Sharing?

- A. To exchange data only within a single organization.
- B. To support data governance and security protocols while sharing data across platforms.**
- C. To create data backups.
- D. To compile data from various sources into one database.

The primary purpose of Delta Sharing is to support data governance and security protocols while sharing data across platforms. This initiative facilitates secure and efficient data sharing among organizations without duplication, ensuring that data can be accessed and managed appropriately while adhering to compliance and regulatory standards. Delta Sharing allows data providers to control permissions and access rights, allowing external users to retrieve data securely while maintaining the integrity and confidentiality of the shared information. This capability is essential in today's data-driven landscape, where organizations often collaborate but need to ensure that sensitive data is protected and shared responsibly. Other options do not align with the core function of Delta Sharing. For instance, the focus on data exchange within a single organization does not capture the essence of Delta Sharing, which is designed specifically for cross-organizational collaboration. Creating data backups is unrelated to sharing data for analytical or operational purposes, and compiling data from various sources into one database misses the emphasis on sharing existing datasets securely and effectively.

2. Which JSON file management technique helps maintain version control?

- A. Editing in real-time through the interface.
- B. Creating backups before editing.
- C. Downloading the JSON description before changes.**
- D. Using cloud storage to manage configurations.

The technique that effectively helps maintain version control is downloading the JSON description before changes. By saving the current state of the JSON file prior to making any modifications, you create a snapshot of that specific version. This approach enables you to revert to the previous version if needed, ensuring that you can track changes over time and manage configurations without losing important data. In the context of version control, this process is crucial because it provides a systematic way to maintain a history of changes. Users can compare different versions and understand the evolution of the file structure, which is essential for debugging and collaboration in data engineering projects. While the other techniques may offer some level of protection or convenience, they don't inherently establish a clear versioning system like downloading the JSON file does. Editing in real-time can introduce risks if changes are not tracked. Creating backups before editing offers safety but does not provide a streamlined process for version comparison. Using cloud storage can aid in accessibility and sharing but does not inherently provide version control without additional mechanisms in place.

3. What SQL command is used to view the history of a Delta table?

A. SELECT * FROM history

B. DESCRIBE HISTORY

C. SHOW TABLE HISTORY

D. VIEW DATABASE HISTORY

The command used to view the history of a Delta table is "DESCRIBE HISTORY." This SQL command provides detailed information about all changes made to the Delta table, including transaction details like the timestamp, user, operation type, and any relevant metadata. This allows users to track the evolution of data within the table, making it easier to audit changes, troubleshoot issues, or revert to previous states if necessary. In contrast, the other options do not serve the purpose of retrieving Delta table change history. For example, selecting all records from a table or using commands that concern other types of database objects would lack the specific context of tracking Delta table operations. Therefore, using "DESCRIBE HISTORY" is the correct approach to accessing the relevant history for a Delta table.

4. Which SQL function is utilized to generate the current timestamp?

A. Input_file_name()

B. Current_timestamp()

C. Filter()

D. Collect_set()

The function used to generate the current timestamp in SQL is the one that retrieves the system's current date and time. This function is essential for time-sensitive operations, whenever you need to log or track the timing of events within a database. The correct function, Current_timestamp(), returns the current date and time in the format of a timestamp, which is useful for inserting time data into tables, performing time-based comparisons, and more. It provides a precise measure of when a particular operation occurs, which can be crucial in applications such as logging, analytics, and transaction management. The other options serve different functions: Input_file_name() retrieves the name of the file being read in a data processing job, Filter() is used to filter records from a dataset based on given criteria, and Collect_set() aggregates unique values from a dataset into a set. Therefore, none of these options fulfill the requirement of generating the current timestamp as effectively as Current_timestamp() does.

5. What SQL syntax is used to define a Common Table Expression (CTE)?

- A. DECLARE cte AS SELECT ...;
- B. WITH cte AS SELECT ...;**
- C. CACHE cte SELECT ...;
- D. CREATE CTE AS SELECT ...;

The use of the "WITH" clause in SQL is the correct syntax for defining a Common Table Expression (CTE). This structure allows you to create a temporary result set that can be referenced within a SELECT, INSERT, UPDATE, or DELETE statement. When you use "WITH cte AS", you establish a CTE named "cte" that you can then utilize in subsequent queries. This enables organizing complex queries in a more readable and manageable way. For instance, a CTE can simplify a query that requires multiple steps, as you can define the CTE once and reference it multiple times as if it were a table. The other options do not conform to the established SQL syntax for defining CTEs. The terms "DECLARE", "CACHE", and "CREATE" do not form valid commands for this purpose, as they either pertain to variable declaration, caching mechanisms, or general object creation, none of which correctly set up a CTE. Thus, using "WITH" is essential for enabling structured, powerful queries that require modular subqueries.

6. What is one of the significant benefits of Delta Live Tables?

- A. It requires extensive coding for scalability
- B. It simplifies both batch and streaming data processing**
- C. It eliminates the need for data modeling
- D. It restricts data accessibility to few users

One of the significant benefits of Delta Live Tables is its ability to simplify both batch and streaming data processing. Delta Live Tables provides a unified framework that allows data engineers to define data pipelines with declarative syntax, effectively handling data in various forms - whether that be large volumes of historical batch data or real-time streaming data. This capability enables users to build and manage data pipelines more efficiently, promoting ease of use and reducing the complexity typically associated with managing separate batch and streaming processes. Moreover, this benefit enhances the overall workflow by allowing for seamless governance, reliability, and optimization, ensuring that data pipelines are both robust and easy to maintain. Users can focus on their data transformations and processing logic without having to worry about the underlying complexities of different data ingestion methods, as Delta Live Tables handles these aspects, streamlining the entire process.

7. What is the command to start a Delta Live Table?

- A. CREATE STREAM TABLE
- B. CREATE INCREMENTAL LIVE TABLE**
- C. START LIVE TABLE
- D. LAUNCH DELTA TABLE

The command to start a Delta Live Table is accurately identified as creating an incremental live table. This approach allows you to define a live table that processes data incrementally as new data arrives, providing real-time insights and updates. This command integrates seamlessly with the Delta Lake architecture, which underpins Delta Live Tables, and facilitates continuous data ingestion and processing. In the context of Delta Live Tables, this command is pivotal because it helps in managing the flow of data efficiently. Leveraging such live tables ensures that your data pipeline remains dynamic, automatically handling updates and changes to the underlying data sources. This feature is particularly beneficial for applications requiring up-to-date information without the need for a complete refresh of the data. The other options provided do not represent valid commands related to starting Delta Live Tables. For instance, a command suggesting the creation of a standard live table or any function that implies launching or starting without the incremental context does not align with the specific operational characteristics of Delta Live Tables. Understanding this command helps solidify knowledge about employing Delta Lake's capabilities for managing data efficiently in a streaming context.

8. In Databricks, what is meant by a cluster?

- A. A user account for data access
- B. A set of computational resources for running workloads**
- C. A tool for data visualization
- D. A template for machine learning models

In the context of Databricks, a cluster refers to a set of computational resources designated for executing tasks and workloads. Clusters can consist of various virtual machines (VMs) that provide the necessary processing power, memory, and storage capabilities to handle data processing, analytics, and machine learning tasks. When data engineers or data scientists run jobs, they utilize these clusters to leverage parallel processing and manage large datasets efficiently. The ability to configure clusters—such as adjusting the number of nodes, the type of instances, and the runtime versions—allows users to optimize their workload performance based on specific project requirements. Clusters are central to the operation of Databricks as they facilitate the distributed computing environment essential for handling big data and large-scale processing. This means that for any intensive data operation, the compute resources allocated through a cluster are what allow for quick, efficient, and scalable execution of numerous tasks concurrently.

9. What is the function of autoscaling in Databricks?

- A. It modifies the data structure in the cluster
- B. It adjusts the number of nodes based on workload**
- C. It locks the cluster to a fixed number of nodes
- D. It allows users to add data manually

Autoscaling in Databricks is designed to dynamically adjust the number of nodes in a cluster based on the current workload. This feature enhances performance and efficiency by automatically scaling up the number of resources during peak demand and scaling them down during periods of lower usage. By optimizing resource allocation, autoscaling helps ensure that jobs can be completed in a timely manner while minimizing costs associated with over-provisioning resources. The capability to flexibly manage the number of nodes allows organizations to run diverse workloads effectively, adapting to varying sizes and complexities of data processing tasks without manual intervention. This automated aspect of resource management is particularly advantageous in cloud environments, where resource costs can be optimized based on actual utilization. The other choices don't accurately represent the function of autoscaling within the Databricks environment, as they suggest modifications to the data structure, do not reflect the dynamic nature of resource management, or refer to manual data handling, which is not related to the scaling mechanism.

10. What SQL function is used to transform an array into a set of rows?

- A. Collect_set
- B. Explode**
- C. Flatten
- D. Array_distinct

The function that transforms an array into a set of rows is the Explode function. When applied to an array, Explode takes the elements within that array and expands them into separate rows in a result set. Each element of the array becomes a distinct entry in its own row, making it a valuable tool for working with array data types in SQL. This is particularly useful in scenarios where you need to perform further analysis or operations on each individual element of the array. For instance, if you have a dataset with a column containing arrays of values, using the Explode function allows you to normalize your data, facilitating easier aggregations and joins with other tables. In contrast, the other functions mentioned serve different purposes: Collect_set is used to aggregate values into a set, avoiding duplicates; Flatten refers to the process of merging nested data structures but does not necessarily output rows; Array_distinct is used to return distinct values from an array, not to expand it into rows. Understanding the specific function each serves is key to effectively manipulating and analyzing data within a database environment.

Next Steps

Congratulations on reaching the final section of this guide. You've taken a meaningful step toward passing your certification exam and advancing your career.

As you continue preparing, remember that consistent practice, review, and self-reflection are key to success. Make time to revisit difficult topics, simulate exam conditions, and track your progress along the way.

If you need help, have suggestions, or want to share feedback, we'd love to hear from you. Reach out to our team at hello@examzify.com.

Or visit your dedicated course page for more study tools and resources:

<https://dataengineeringassociate-databricks.examzify.com>

We wish you the very best on your exam journey. You've got this!