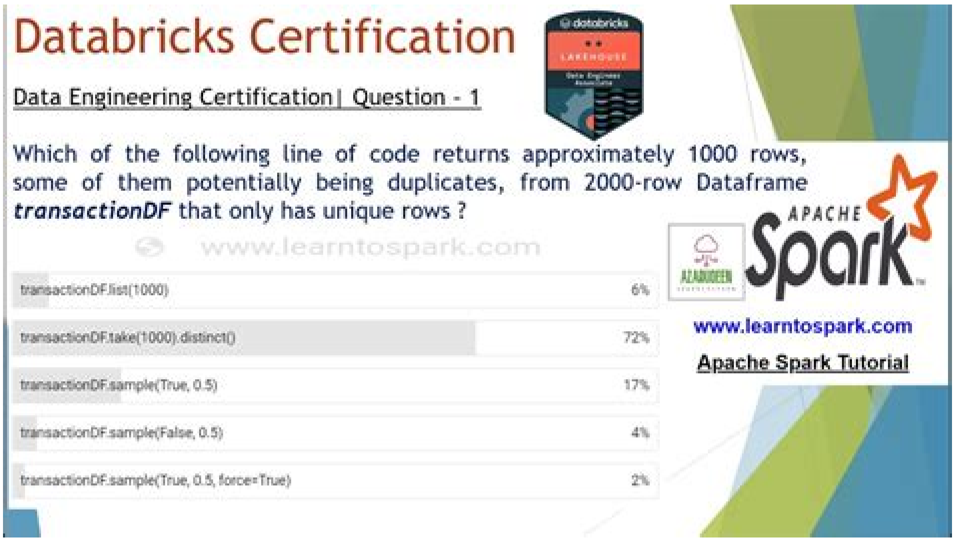
Databricks Data Engineer Associate Exam Questions



Databricks Certification

Data Engineering Certification | Question - 1

Which of the following line of code returns approximately 1000 rows, some of them potentially being duplicates, from 2000-row Dataframe *transactionDF* that only has unique rows ?

www.learntospark.com

transactionDF.list(1000)	6%
transactionDF.take(1000).distinct()	72%
transactionDF.sample(True, 0.5)	17%
transactionDF.sample(False, 0.5)	4%
transactionDF.sample(True, 0.5, force=True)	2%

www.learntospark.com

Apache Spark Tutorial

Databricks Data Engineer Associate Exam Questions are essential for candidates preparing for one of the most sought-after certifications in the data engineering domain. As organizations increasingly leverage big data technologies, the need for skilled data engineers has surged. The Databricks Data Engineer Associate certification validates a candidate's proficiency in data engineering concepts and tools, focusing on the Databricks platform and its integration with Apache Spark. In this article, we will explore common exam questions, essential topics to study, and tips for success, providing you with a comprehensive guide to mastering the Databricks Data Engineer Associate exam.

Understanding the Databricks Data Engineer Associate Exam

The Databricks Data Engineer Associate exam is designed to assess candidates on their ability to perform data engineering tasks using the Databricks platform. This certification is recommended for individuals who have experience in data engineering and are familiar with the functionalities of Databricks and Apache Spark.

Exam Format and Structure

Before diving into specific questions, it's crucial to understand the exam format:

- Number of Questions: The exam typically consists of 45-60 questions.
- Question Type: The questions can be multiple-choice or multiple-select.
- Duration: Candidates usually have 120 minutes to complete the exam.
- Passing Score: The passing score varies but generally hovers around 70%.

Topics Covered in the Exam

The Databricks Data Engineer Associate exam covers various domains. Understanding these topics will help you focus your preparation effectively. Here are the primary areas of focus:

1. Data Ingestion
- Methods for ingesting data into Databricks
- Integration with data sources like Amazon S3, Azure Blob Storage, and more
- Handling streaming data

2. Data Transformation
- Using Apache Spark for data processing
- Writing efficient Spark jobs
- Understanding DataFrames and Datasets

3. Data Storage and Management
- Managing Delta Lake
- Understanding the Delta Lake architecture
- Benefits of using Delta Lake over traditional data lakes

4. Data Governance and Security
- Implementing access controls
- Managing data privacy and compliance
- Monitoring and auditing data access

5. Performance Tuning
- Optimizing Spark jobs
- Analyzing execution plans
- Utilizing caching effectively

6. Collaborative Development
- Working with notebooks in Databricks
- Version control and collaboration practices
- CI/CD for data pipelines

Sample Questions for Databricks Data Engineer Associate Exam

While the actual exam questions are proprietary, understanding the types of questions can significantly aid your preparation. Here are some sample questions that reflect the style and content of the exam:

1. Data Ingestion Questions

- What is the purpose of the `spark.read` method in Databricks?
- a) To write data to external sources
- b) To read data into a DataFrame
- c) To create a new Spark session
- d) To cache data in memory

- Which of the following is NOT a method to ingest data into Databricks?
- a) Using Apache Kafka
- b) Writing a custom Python script
- c) Directly uploading files through the UI
- d) Using Databricks SQL Analytics

2. Data Transformation Questions

- How can you improve the performance of a Spark job?
- a) Increase the number of partitions
- b) Use DataFrames instead of RDDs
- c) Cache intermediate data
- d) All of the above

- Which function would you use to filter rows in a DataFrame?
- a) select()
- b) filter()
- c) groupBy()
- d) join()

3. Data Storage and Management Questions

- What is Delta Lake primarily used for?
- a) Storing unstructured data
- b) Providing ACID transactions on big data
- c) Serving as a data visualization tool
- d) None of the above

- Which command would you use to convert a DataFrame to a Delta table?
- a) write.format("delta").save("path")
- b) saveAsTable("delta_table")
- c) createOrReplaceTempView("delta_table")
- d) write.format("parquet").save("path")

4. Data Governance and Security Questions

- Which feature provides secure access to data in Databricks?
- a) Delta Lake
- b) Token-based authentication
- c) Cluster policies
- d) All of the above

- What is the role of Unity Catalog in Databricks?
- a) To manage compute resources
- b) To facilitate data discovery and governance
- c) To provide real-time analytics
- d) To optimize query performance

Preparation Strategies for the Exam

To effectively prepare for the Databricks Data Engineer Associate exam, consider the following strategies:

1. Hands-On Practice

- Use Databricks Notebooks: Engage with Databricks notebooks to write and run Spark jobs. Familiarize yourself with data ingestion, transformation, and storage techniques.
- Work on Real Projects: If possible, work on real-world data projects that involve data engineering tasks, particularly those using Databricks.

2. Study Resources

- Official Databricks Documentation: Make sure to review the latest documentation and resources provided by Databricks.
- Online Courses: Enroll in online courses focused on Apache Spark and Databricks, which often include hands-on labs.

3. Join Study Groups

- Networking: Connect with other data engineers preparing for the exam. Join forums or communities to discuss concepts and share resources.

4. Mock Exams

- Practice Tests: Take mock exams to familiarize yourself with the exam format and timing. This will help you identify areas where you need improvement.

Conclusion

Preparing for the **Databricks Data Engineer Associate Exam Questions** requires a thorough understanding of data engineering principles, hands-on experience with the Databricks platform, and familiarity with Apache Spark. By focusing on the key topics and utilizing effective study strategies, candidates can enhance their chances of passing the exam and advancing their careers in data engineering. Whether you are new to the field or looking to validate your skills, this certification can open new doors in the ever-evolving world of big data.

Frequently Asked Questions

What is the main focus of the Databricks Data Engineer Associate exam?

The exam primarily focuses on testing the candidate's knowledge and skills in data engineering concepts, including data ingestion, transformation, and storage using Databricks and Apache Spark.

What types of data formats are commonly used in Databricks for data engineering tasks?

Common data formats include Parquet, Delta Lake, CSV, JSON, and Avro, which are used for efficient data storage and processing in Databricks.

How does Delta Lake enhance data reliability in Databricks?

Delta Lake provides ACID transactions, scalable metadata handling, and data versioning, which enhance data reliability and enable efficient data processing and querying.

What are the key components of a data pipeline in Databricks?

Key components include data ingestion (using Apache Spark), data transformation (using Spark SQL and DataFrames), and data storage (using

Delta Lake or external storage solutions).

What is the importance of using notebooks in Databricks for data engineering?

Notebooks in Databricks allow for interactive data exploration, visualization, and collaboration among team members, making it easier to develop, test, and document data engineering workflows.

Databricks Data Engineer Associate Exam Questions

Printing secret value in Databricks - Stack Overflow
Nov 11, 2021 · First, install the Databricks Python SDK and configure authentication per the docs here. pip install databricks-sdk Then you can use the approach below to print out secret values. Because the code doesn't run in Databricks, the secret values aren't redacted. For my particular use case, I wanted to print values for all secrets in a given scope.

Databricks shows REDACTED on a hardcoded value
Mar 16, 2023 · It's not possible, Databricks just scans entire output for occurences of secret values and replaces them with " [REDACTED]". It is helpless if you transform the value. For example, like you tried already, you could insert spaces between characters and that would reveal the value. You can use a trick with an invisible character - for example Unicode invisible …

Databricks: How do I get path of current notebook?
Nov 29, 2018 · Databricks is smart and all, but how do you identify the path of your current notebook? The guide on the website does not help. It suggests: %scala dbutils.notebook.getContext.notebookPath res1: ...

Is there a way to use parameters in Databricks in SQL with …
Sep 29, 2024 · There is a lot of confusion wrt the use of parameters in SQL, but I see Databricks has started harmonizing heavily (for example, 3 months back, IDENTIFIER () didn't work with catalog, now it does). Check my answer for a working solution.

Databricks - Download a dbfs:/FileStore file to my Local Machine
Method3: Using third-party tool named DBFS Explorer DBFS Explorer was created as a quick way to upload and download files to the Databricks filesystem (DBFS). This will work with both AWS and Azure instances of Databricks. You will need to create a bearer token in the web interface in order to connect.

Databricks: managed tables vs. external tables - Stack Overflow
Jun 21, 2024 · The decision to use managed table or external table depends on your use case and also the existing setup of your delta lake, framework code and workflows. Your understanding of the

Managed tables is partially correct based on the explanation that you have given. For managed tables, databricks handles the storage and metadata of the tables, including the …

databricks: writing spark dataframe directly to excel
Nov 29, 2019 · Are there any method to write spark dataframe directly to xls/xlsx format ???? Most of the example in the web showing there is example for panda dataframes. but I would like to use spark datafr...

How to read xlsx or xls files as spark dataframe - Stack Overflow
Jun 3, 2019 · Can anyone let me know without converting xlsx or xls files how can we read them as a spark dataframe I have already tried to read with pandas and then tried to convert to spark dataframe but got...

Connecting C# Application to Azure Databricks - Stack Overflow
The Datalake is hooked to Azure Databricks. The requirement asks that the Azure Databricks is to be connected to a C# application to be able to run queries and get the result all from the C# application. The way we are currently tackling the problem is that we have created a workspace on Databricks with a number of queries that need to be executed.

Installing multiple libraries 'permanently' on Databricks' cluster ...
Feb 28, 2024 · Installing multiple libraries 'permanently' on Databricks' cluster Asked 1 year, 4 months ago Modified 1 year, 4 months ago Viewed 4k times

Printing secret value in Databricks - Stack Overflow
Nov 11, 2021 · First, install the Databricks Python SDK and configure authentication per the docs here. pip install databricks-sdk Then you can use the approach below to print out secret values. Because the code doesn't run in Databricks, the secret values aren't redacted. For my particular use case, I wanted to print values for all secrets in a given scope.

Databricks shows REDACTED on a hardcoded value
Mar 16, 2023 · It's not possible, Databricks just scans entire output for occurences of secret values and replaces them with " [REDACTED]". It is helpless if you transform the value. For example, like you tried already, you could insert spaces between characters and that would reveal the value. You can use a trick with an invisible character - for example Unicode invisible …

Databricks: How do I get path of current notebook?
Nov 29, 2018 · Databricks is smart and all, but how do you identify the path of your current notebook? The guide on the website does not help. It suggests: %scala dbutils.notebook.getContext.notebookPath res1: ...

Is there a way to use parameters in Databricks in SQL with ...
Sep 29, 2024 · There is a lot of confusion wrt the use of parameters in SQL, but I see Databricks has started harmonizing heavily (for example, 3 months back, IDENTIFIER () didn't work with catalog, now it does). Check my answer for a working solution.

Databricks - Download a dbfs:/FileStore file to my Local Machine
Method3: Using third-party tool named DBFS Explorer DBFS Explorer was created as a quick way to upload and download files to the Databricks filesystem (DBFS). This will work with both AWS and Azure instances of Databricks. You will need to create a bearer token in the web interface in order to connect.

Databricks: managed tables vs. external tables - Stack Overflow
Jun 21, 2024 · The decision to use managed table or external table depends on your use case and also the existing setup of your delta lake, framework code and workflows. Your understanding of the Managed tables is partially correct based on the explanation that you have given. For managed tables, databricks handles the storage and metadata of the tables, including the …

databricks: writing spark dataframe directly to excel
Nov 29, 2019 · Are there any method to write spark dataframe directly to xls/xlsx format ???? Most of the example in the web showing there is example for panda dataframes. but I would like to use spark datafr...

How to read xlsx or xls files as spark dataframe - Stack Overflow
Jun 3, 2019 · Can anyone let me know without converting xlsx or xls files how can we read them as a spark dataframe I have already tried to read with pandas and then tried to convert to spark dataframe but got...

Connecting C# Application to Azure Databricks - Stack Overflow
The Datalake is hooked to Azure Databricks. The requirement asks that the Azure Databricks is to be connected to a C# application to be able to run queries and get the result all from the C# application. The way we are currently tackling the problem is that we have created a workspace on Databricks with a number of queries that need to be executed.

Installing multiple libraries 'permanently' on Databricks' cluster ...
Feb 28, 2024 · Installing multiple libraries 'permanently' on Databricks' cluster Asked 1 year, 4 months ago Modified 1 year, 4 months ago Viewed 4k times

Prepare for the Databricks Data Engineer Associate Exam with our comprehensive guide on exam questions. Discover how to ace your certification today!

Back to Home