

# Quiz Amazon - Data-Engineer-Associate - Perfect Latest AWS Certified Data Engineer - Associate (DEA-C01) Braindumps Pdf



2026 Latest TestPassed Data-Engineer-Associate PDF Dumps and Data-Engineer-Associate Exam Engine Free Share:  
[https://drive.google.com/open?id=1Uysu74sdffLckCu\\_HCXyLSKykevJlhFmM](https://drive.google.com/open?id=1Uysu74sdffLckCu_HCXyLSKykevJlhFmM)

The TestPassed is a leading and reliable platform that has been offering real, valid, and updated AWS Certified Data Engineer - Associate (DEA-C01) (Data-Engineer-Associate) exam practice test questions for many years. Over this long time period thousands of candidates have passed their dream AWS Certified Data Engineer - Associate (DEA-C01) (Data-Engineer-Associate) certification exam. And the one thing has come in their success that was the usage of top-notch Data-Engineer-Associate Exam Practice test questions. So you can also get help from TestPassed practice test questions and make the Amazon Data-Engineer-Associate exam preparation simple, smart and quick.

We are determined to be the best vendor in this career to help more and more candidates to accomplish their dream and get their desired Data-Engineer-Associate certification. No only that we provide the most effective Data-Engineer-Associate Study Materials, but also we offer the first-class after-sale service to all our customers. Our professional online service are pleased to give guide in 24 hours.

>> Latest Data-Engineer-Associate Braindumps Pdf <<

## Premium Data-Engineer-Associate Exam - Data-Engineer-Associate Valid Exam Discount

Our Data-Engineer-Associate training braindumps are famous for its wonderful advantages. The content is carefully designed for the Data-Engineer-Associate exam, rich question bank and answer to enable you to master all the test knowledge in a short period of time. Our Data-Engineer-Associate Exam Questions have helped a large number of candidates pass the Data-Engineer-Associate exam yet. Hope you can join us, and we work together to create a miracle.

## Amazon AWS Certified Data Engineer - Associate (DEA-C01) Sample Questions (Q80-Q85):

### NEW QUESTION # 80

A data engineer is configuring Amazon SageMaker Studio to use AWS Glue interactive sessions to prepare data for machine learning (ML) models.

The data engineer receives an access denied error when the data engineer tries to prepare the data by using SageMaker Studio. Which change should the engineer make to gain access to SageMaker Studio?

- **A. Add a policy to the data engineer's IAM user that includes the sts:AssumeRole action for the AWS Glue and SageMaker service principals in the trust policy.**

- B. Add the AmazonSageMakerFullAccess managed policy to the data engineer's IAM user.
- C. Add a policy to the data engineer's IAM user that allows the sts:AddAssociation action for the AWS Glue and SageMaker service principals in the trust policy.
- D. Add the AWSGlueServiceRole managed policy to the data engineer's IAM user.

**Answer: A**

Explanation:

This solution meets the requirement of gaining access to SageMaker Studio to use AWS Glue interactive sessions. AWS Glue interactive sessions are a way to use AWS Glue DataBrew and AWS Glue Data Catalog from within SageMaker Studio. To use AWS Glue interactive sessions, the data engineer's IAM user needs to have permissions to assume the AWS Glue service role and the SageMaker execution role. By adding a policy to the data engineer's IAM user that includes the sts:AssumeRole action for the AWS Glue and SageMaker service principals in the trust policy, the data engineer can grant these permissions and avoid the access denied error. The other options are not sufficient or necessary to resolve the error. Reference:

Get started with data integration from Amazon S3 to Amazon Redshift using AWS Glue interactive sessions Troubleshoot Errors - Amazon SageMaker AccessDeniedException on sagemaker:CreateDomain in AWS SageMaker Studio, despite having SageMakerFullAccess

### NEW QUESTION # 81

A company is building an analytics solution. The solution uses Amazon S3 for data lake storage and Amazon Redshift for a data warehouse. The company wants to use Amazon Redshift Spectrum to query the data that is in Amazon S3.

Which actions will provide the FASTEST queries? (Choose two.)

- A. Use gzip compression to compress individual files to sizes that are between 1 GB and 5 GB.
- B. Split the data into files that are less than 10 KB.
- C. Use a columnar storage file format.
- D. Partition the data based on the most common query predicates.
- E. Use file formats that are not

**Answer: C,D**

Explanation:

Amazon Redshift Spectrum is a feature that allows you to run SQL queries directly against data in Amazon S3, without loading or transforming the data. Redshift Spectrum can query various data formats, such as CSV, JSON, ORC, Avro, and Parquet.

However, not all data formats are equally efficient for querying. Some data formats, such as CSV and JSON, are row-oriented, meaning that they store data as a sequence of records, each with the same fields. Row-oriented formats are suitable for loading and exporting data, but they are not optimal for analytical queries that often access only a subset of columns. Row-oriented formats also do not support compression or encoding techniques that can reduce the data size and improve the query performance.

On the other hand, some data formats, such as ORC and Parquet, are column-oriented, meaning that they store data as a collection of columns, each with a specific data type. Column-oriented formats are ideal for analytical queries that often filter, aggregate, or join data by columns. Column-oriented formats also support compression and encoding techniques that can reduce the data size and improve the query performance. For example, Parquet supports dictionary encoding, which replaces repeated values with numeric codes, and run-length encoding, which replaces consecutive identical values with a single value and a count. Parquet also supports various compression algorithms, such as Snappy, GZIP, and ZSTD, that can further reduce the data size and improve the query performance.

Therefore, using a columnar storage file format, such as Parquet, will provide faster queries, as it allows Redshift Spectrum to scan only the relevant columns and skip the rest, reducing the amount of data read from S3. Additionally, partitioning the data based on the most common query predicates, such as date, time, region, etc., will provide faster queries, as it allows Redshift Spectrum to prune the partitions that do not match the query criteria, reducing the amount of data scanned from S3. Partitioning also improves the performance of joins and aggregations, as it reduces data skew and shuffling.

The other options are not as effective as using a columnar storage file format and partitioning the data. Using gzip compression to compress individual files to sizes that are between 1 GB and 5 GB will reduce the data size, but it will not improve the query performance significantly, as gzip is not a splittable compression algorithm and requires decompression before reading. Splitting the data into files that are less than 10 KB will increase the number of files and the metadata overhead, which will degrade the query performance. Using file formats that are not supported by Redshift Spectrum, such as XML, will not work, as Redshift Spectrum will not be able to read or parse the data. Reference:

Amazon Redshift Spectrum

Choosing the Right Data Format

AWS Certified Data Engineer - Associate DEA-C01 Complete Study Guide, Chapter 4: Data Lakes and Data Warehouses, Section 4.3: Amazon Redshift Spectrum

## NEW QUESTION # 82

A security company stores IoT data that is in JSON format in an Amazon S3 bucket. The data structure can change when the company upgrades the IoT devices. The company wants to create a data catalog that includes the IoT data. The company's analytics department will use the data catalog to index the data.

Which solution will meet these requirements MOST cost-effectively?

- A. Create an AWS Glue Data Catalog. Configure an AWS Glue Schema Registry. Create a new AWS Glue workload to orchestrate the ingestion of the data that the analytics department will use into Amazon Redshift Serverless.
- B. Create an Amazon Redshift provisioned cluster. Create an Amazon Redshift Spectrum database for the analytics department to explore the data that is in Amazon S3. Create Redshift stored procedures to load the data into Amazon Redshift.
- C. Create an Amazon Athena workgroup. Explore the data that is in Amazon S3 by using Apache Spark through Athena. Provide the Athena workgroup schema and tables to the analytics department.
- D. Create an AWS Glue Data Catalog. Configure an AWS Glue Schema Registry. Create AWS Lambda user defined functions (UDFs) by using the Amazon Redshift Data API. Create an AWS Step Functions job to orchestrate the ingestion of the data that the analytics department will use into Amazon Redshift Serverless.

**Answer: C**

Explanation:

The best solution to meet the requirements of creating a data catalog that includes the IoT data, and allowing the analytics department to index the data, most cost-effectively, is to create an Amazon Athena workgroup, explore the data that is in Amazon S3 by using Apache Spark through Athena, and provide the Athena workgroup schema and tables to the analytics department. Amazon Athena is a serverless, interactive query service that makes it easy to analyze data directly in Amazon S3 using standard SQL or Python<sup>1</sup>. Amazon Athena also supports Apache Spark, an open-source distributed processing framework that can run large-scale data analytics applications across clusters of servers<sup>2</sup>. You can use Athena to run Spark code on data in Amazon S3 without having to set up, manage, or scale any infrastructure. You can also use Athena to create and manage external tables that point to your data in Amazon S3, and store them in an external data catalog, such as AWS Glue Data Catalog, Amazon Athena Data Catalog, or your own Apache Hive metastore<sup>3</sup>. You can create Athena workgroups to separate query execution and resource allocation based on different criteria, such as users, teams, or applications<sup>4</sup>. You can share the schemas and tables in your Athena workgroup with other users or applications, such as Amazon QuickSight, for data visualization and analysis<sup>5</sup>.

Using Athena and Spark to create a data catalog and explore the IoT data in Amazon S3 is the most cost-effective solution, as you pay only for the queries you run or the compute you use, and you pay nothing when the service is idle<sup>1</sup>. You also save on the operational overhead and complexity of managing data warehouse infrastructure, as Athena and Spark are serverless and scalable. You can also benefit from the flexibility and performance of Athena and Spark, as they support various data formats, including JSON, and can handle schema changes and complex queries efficiently.

Option A is not the best solution, as creating an AWS Glue Data Catalog, configuring an AWS Glue Schema Registry, creating a new AWS Glue workload to orchestrate the ingestion of the data that the analytics department will use into Amazon Redshift Serverless, would incur more costs and complexity than using Athena and Spark. AWS Glue Data Catalog is a persistent metadata store that contains table definitions, job definitions, and other control information to help you manage your AWS Glue components<sup>6</sup>. AWS Glue Schema Registry is a service that allows you to centrally store and manage the schemas of your streaming data in AWS Glue Data Catalog<sup>7</sup>. AWS Glue is a serverless data integration service that makes it easy to prepare, clean, enrich, and move data between data stores<sup>8</sup>. Amazon Redshift Serverless is a feature of Amazon Redshift, a fully managed data warehouse service, that allows you to run and scale analytics without having to manage data warehouse infrastructure<sup>9</sup>. While these services are powerful and useful for many data engineering scenarios, they are not necessary or cost-effective for creating a data catalog and indexing the IoT data in Amazon S3. AWS Glue Data Catalog and Schema Registry charge you based on the number of objects stored and the number of requests made<sup>6,7</sup>. AWS Glue charges you based on the compute time and the data processed by your ETL jobs<sup>8</sup>. Amazon Redshift Serverless charges you based on the amount of data scanned by your queries and the compute time used by your workloads<sup>9</sup>. These costs can add up quickly, especially if you have large volumes of IoT data and frequent schema changes. Moreover, using AWS Glue and Amazon Redshift Serverless would introduce additional latency and complexity, as you would have to ingest the data from Amazon S3 to Amazon Redshift Serverless, and then query it from there, instead of querying it directly from Amazon S3 using Athena and Spark.

Option B is not the best solution, as creating an Amazon Redshift provisioned cluster, creating an Amazon Redshift Spectrum database for the analytics department to explore the data that is in Amazon S3, and creating Redshift stored procedures to load the data into Amazon Redshift, would incur more costs and complexity than using Athena and Spark. Amazon Redshift provisioned clusters are clusters that you create and manage by specifying the number and type of nodes, and the amount of storage and compute capacity<sup>10</sup>. Amazon Redshift Spectrum is a feature of Amazon Redshift that allows you to query and join data across your data warehouse and your data lake using standard SQL<sup>11</sup>. Redshift stored procedures are SQL statements that you can define and store in Amazon Redshift, and then call them by using the CALL command<sup>12</sup>. While these features are powerful and useful for many data warehousing scenarios, they are not necessary or cost-effective for creating a data catalog and indexing the IoT data in Amazon

S3. Amazon Redshift provisioned clusters charge you based on the node type, the number of nodes, and the duration of the cluster<sup>10</sup>. Amazon Redshift Spectrum charges you based on the amount of data scanned by your queries<sup>11</sup>.

These costs can add up quickly, especially if you have large volumes of IoT data and frequent schema changes. Moreover, using Amazon Redshift provisioned clusters and Spectrum would introduce additional latency and complexity, as you would have to provision and manage the cluster, create an external schema and database for the data in Amazon S3, and load the data into the cluster using stored procedures, instead of querying it directly from Amazon S3 using Athena and Spark.

Option D is not the best solution, as creating an AWS Glue Data Catalog, configuring an AWS Glue Schema Registry, creating AWS Lambda user defined functions (UDFs) by using the Amazon Redshift Data API, and creating an AWS Step Functions job to orchestrate the ingestion of the data that the analytics department will use into Amazon Redshift Serverless, would incur more costs and complexity than using Athena and Spark. AWS Lambda is a serverless compute service that lets you run code without provisioning or managing servers<sup>13</sup>. AWS Lambda UDFs are Lambda functions that you can invoke from within an Amazon Redshift query. Amazon Redshift Data API is a service that allows you to run SQL statements on Amazon Redshift clusters using HTTP requests, without needing a persistent connection. AWS Step Functions is a service that lets you coordinate multiple AWS services into serverless workflows. While these services are powerful and useful for many data engineering scenarios, they are not necessary or cost-effective for creating a data catalog and indexing the IoT data in Amazon S3. AWS Glue Data Catalog and Schema Registry charge you based on the number of objects stored and the number of requests made<sup>67</sup>. AWS Lambda charges you based on the number of requests and the duration of your functions<sup>13</sup>. Amazon Redshift Serverless charges you based on the amount of data scanned by your queries and the compute time used by your workloads<sup>9</sup>. AWS Step Functions charges you based on the number of state transitions in your workflows. These costs can add up quickly, especially if you have large volumes of IoT data and frequent schema changes. Moreover, using AWS Glue, AWS Lambda, Amazon Redshift Data API, and AWS Step Functions would introduce additional latency and complexity, as you would have to create and invoke Lambda functions to ingest the data from Amazon S3 to Amazon Redshift Serverless using the Data API, and coordinate the ingestion process using Step Functions, instead of querying it directly from Amazon S3 using Athena and Spark. References:

What is Amazon Athena?

Apache Spark on Amazon Athena

Creating tables, updating the schema, and adding new partitions in the Data Catalog from AWS Glue ETL jobs Managing Athena workgroups Using Amazon QuickSight to visualize data in Amazon Athena AWS Glue Data Catalog AWS Glue Schema Registry

What is AWS Glue?

Amazon Redshift Serverless

Amazon Redshift provisioned clusters

Querying external data using Amazon Redshift Spectrum

Using stored procedures in Amazon Redshift

What is AWS Lambda?

[Creating and using AWS Lambda UDFs]

[Using the Amazon Redshift Data API]

[What is AWS Step Functions?]

AWS Certified Data Engineer - Associate DEA-C01 Complete Study Guide

### NEW QUESTION # 83

A company has three subsidiaries. Each subsidiary uses a different data warehousing solution. The first subsidiary hosts its data warehouse in Amazon Redshift. The second subsidiary uses Teradata Vantage on AWS. The third subsidiary uses Google BigQuery.

The company wants to aggregate all the data into a central Amazon S3 data lake. The company wants to use Apache Iceberg as the table format.

A data engineer needs to build a new pipeline to connect to all the data sources, run transformations by using each source engine, join the data, and write the data to Iceberg.

Which solution will meet these requirements with the LEAST operational effort?

- A. Use the Amazon Athena federated query connectors for Amazon Redshift, Teradata, and BigQuery to build the pipeline in Athena. Write a SQL query to read from all the data sources, join the data, and run a Merge operation on the data lake Iceberg table.
- B. Use native Amazon Redshift, Teradata, and BigQuery connectors to build the pipeline in AWS Glue. Use native AWS Glue transforms to join the data. Run a Merge operation on the data lake Iceberg table.
- C. Use the native Amazon Redshift, Teradata, and BigQuery connectors in Amazon Appflow to write data to Amazon S3 and AWS Glue Data Catalog. Use Amazon Athena to join the data. Run a Merge operation on the data lake Iceberg table.
- D. Use the native Amazon Redshift connector, the Java Database Connectivity (JDBC) connector for Teradata, and the open source Apache Spark BigQuery connector to build the pipeline in Amazon EMR. Write code in PySpark to join the data. Run a Merge operation on the data lake Iceberg table.

**Answer: A**

Explanation:

Amazon Athena provides federated query connectors that allow querying multiple data sources, such as Amazon Redshift, Teradata, and Google BigQuery, without needing to extract the data from the original source. This solution is optimal because it offers the least operational effort by avoiding complex data movement and transformation processes.

Amazon Athena Federated Queries:

Athena's federated queries allow direct querying of data stored across multiple sources, including Amazon Redshift, Teradata, and BigQuery. With Athena's support for Apache Iceberg, the company can easily run a Merge operation on the Iceberg table.

The solution reduces complexity by centralizing the query execution and transformation process in Athena using SQL queries.

Reference:

Alternatives Considered:

A (AWS Glue pipeline): This would work but requires more operational effort to manage and transform the data in AWS Glue.

C (Amazon EMR): Using EMR and writing PySpark code introduces more operational overhead and complexity compared to a SQL-based solution in Athena.

D (Amazon AppFlow): AppFlow is more suitable for transferring data between services but is not as efficient for transformations and joins as Athena federated queries.

Amazon Athena Documentation

Federated Queries in Amazon Athena

**NEW QUESTION # 84**

A company receives a data file from a partner each day in an Amazon S3 bucket. The company uses a daily AWS Glue extract, transform, and load (ETL) pipeline to clean and transform each data file. The output of the ETL pipeline is written to a CSV file named Dairy.csv in a second S3 bucket.

Occasionally, the daily data file is empty or is missing values for required fields. When the file is missing data, the company can use the previous day's CSV file.

A data engineer needs to ensure that the previous day's data file is overwritten only if the new daily file is complete and valid.

Which solution will meet these requirements with the LEAST effort?

- A. Run a SQL query in Amazon Athena to read the CSV file and drop missing rows. Copy the corrected CSV file to the second S3 bucket.
- **B. Configure the AWS Glue ETL pipeline to use AWS Glue Data Quality rules. Develop rules in Data Quality Definition Language (DQDL) to check for missing values in required files and empty files.**
- C. Invoke an AWS Lambda function to check the file for missing data and to fill in missing values in required fields.
- D. Use AWS Glue Studio to change the code in the ETL pipeline to fill in any missing values in the required fields with the most common values for each field.

**Answer: B**

Explanation:

Problem Analysis:

The company runs a daily AWS Glue ETL pipeline to clean and transform files received in an S3 bucket.

If a file is incomplete or empty, the previous day's file should be retained.

Need a solution to validate files before overwriting the existing file.

Key Considerations:

Automate data validation with minimal human intervention.

Use built-in AWS Glue capabilities for ease of integration.

Ensure robust validation for missing or incomplete data.

Solution Analysis:

Option A: Lambda Function for Validation

Lambda can validate files, but it would require custom code.

Does not leverage AWS Glue's built-in features, adding operational complexity.

Option B: AWS Glue Data Quality Rules

AWS Glue Data Quality allows defining Data Quality Definition Language (DQDL) rules.

Rules can validate if required fields are missing or if the file is empty.

Automatically integrates into the existing ETL pipeline.

If validation fails, retain the previous day's file.

Option C: AWS Glue Studio with Filling Missing Values

Modifying ETL code to fill missing values with most common values risks introducing inaccuracies.

Does not handle empty files effectively.

Option D: Athena Query for Validation

Athena can drop rows with missing values, but this is a post-hoc solution.

Requires manual intervention to copy the corrected file to S3, increasing complexity.

Final Recommendation:

Use AWS Glue Data Quality to define validation rules in DQDL for identifying missing or incomplete data.

This solution integrates seamlessly with the ETL pipeline and minimizes manual effort.

Implementation Steps:

Enable AWS Glue Data Quality in the existing ETL pipeline.

Define DQDL Rules, such as:

Check if a file is empty.

Verify required fields are present and non-null.

Configure the pipeline to proceed with overwriting only if the file passes validation.

In case of failure, retain the previous day's file.

AWS Glue Data Quality Overview

Defining DQDL Rules

AWS Glue Studio Documentation

## NEW QUESTION # 85

.....

The Data-Engineer-Associate practice test pdf contains the most updated and verified questions & answers, which cover all the exam topics and course outline completely. The Data-Engineer-Associate vce dumps can simulate the actual test environment, which can help you to be more familiar about the Data-Engineer-Associate Real Exam. Now, you can free download Amazon Data-Engineer-Associate updated demo and have a try. If you have any questions about Data-Engineer-Associate pass-guaranteed dumps, contact us at any time.

**Premium Data-Engineer-Associate Exam:** <https://www.testpassed.com/Data-Engineer-Associate-still-valid-exam.html>

In addition, these experts and professors from our company are responsible for constantly updating the Data-Engineer-Associate guide questions, Yes, we do, Amazon Latest Data-Engineer-Associate Braindumps Pdf You will have a great advantage over the other people, Amazon Latest Data-Engineer-Associate Braindumps Pdf We treasure time as all customers do, Amazon Latest Data-Engineer-Associate Braindumps Pdf Please don't worry about the purchase process because it's really simple for you.

Securing the Borderless Network reviews the Data-Engineer-Associate latest Cisco technology solutions for managing identity and securing networks, content, endpoints, and applications, Use Latest Data-Engineer-Associate Braindumps Pdf the security-focused Tails distribution as a quick path to a hardened workstation.

## Excellent Latest Data-Engineer-Associate Braindumps Pdf to Obtain Amazon Certification

In addition, these experts and professors from our company are responsible for constantly updating the Data-Engineer-Associate Guide questions, Yes, we do, You will have a great advantage over the other people.

We treasure time as all customers do, Please Premium Data-Engineer-Associate Exam don't worry about the purchase process because it's really simple for you.

- Quiz Data-Engineer-Associate - AWS Certified Data Engineer - Associate (DEA-C01) Accurate Latest Braindumps Pdf   Search for  Data-Engineer-Associate  and obtain a free download on  [www.verifieddumps.com](http://www.verifieddumps.com)   Valid Data-Engineer-Associate Cram Materials
- Data-Engineer-Associate Testking Learning Materials  Question Data-Engineer-Associate Explanations  Exam Data-Engineer-Associate Preparation  Easily obtain free download of  Data-Engineer-Associate  by searching on  [www.pdfvce.com](http://www.pdfvce.com)   Data-Engineer-Associate Reliable Real Exam
- Quiz Data-Engineer-Associate - AWS Certified Data Engineer - Associate (DEA-C01) Accurate Latest Braindumps Pdf   Open  [www.verifieddumps.com](http://www.verifieddumps.com)  and search for [  Data-Engineer-Associate  ] to download exam materials for free  Exam Data-Engineer-Associate Introduction
- Pdfvce Amazon Data-Engineer-Associate Exam Questions are Available in Three Different Formats  Search for  Data-Engineer-Associate  and obtain a free download on  [www.pdfvce.com](http://www.pdfvce.com)    Data-Engineer-Associate Reliable Real Exam
- Amazon - Professional Latest Data-Engineer-Associate Braindumps Pdf  Download  Data-Engineer-Associate  for free by simply searching on  [www.prepawayexam.com](http://www.prepawayexam.com)    Valid Data-Engineer-Associate Cram Materials

