

Data-Engineer-Associate Exam Cram & Data-Engineer-Associate VCE Dumps & Data-Engineer-Associate Latest Dumps



BTW, DOWNLOAD part of ITExamDownload Data-Engineer-Associate dumps from Cloud Storage:
<https://drive.google.com/open?id=1skPLwEzrzqK-pxRnztldwvfl7V2bSSy->

Our users can prove to you that the hit rate of our Data-Engineer-Associate exam questions is very high. And you can just see the data how many customers are visiting our Data-Engineer-Associate study materials everyday. And the pass rate is also high as 98% to 100%. You can walk into the examination room with peace of mind, after which you will experience a very calm examination. As for the result, please come home and wait. Our Data-Engineer-Associate training prep will not disappoint you.

If you're still studying hard to pass the Amazon Data-Engineer-Associate exam, ITExamDownload help you to achieve your dream. We provide you with the best Amazon Data-Engineer-Associate exam materials. It passed the test of practice, and with the best quality. It is better than Amazon Data-Engineer-Associate tutorials and any other related materials. It can help you to pass the Amazon Data-Engineer-Associate exam, and help you to become a strong IT expert.

>> Exam Data-Engineer-Associate Objectives Pdf <<

Top Exam Data-Engineer-Associate Objectives Pdf | Reliable Latest Data-Engineer-Associate Cram Materials: AWS Certified Data Engineer - Associate (DEA-C01)

We provide the Amazon Data-Engineer-Associate exam questions in a variety of formats, including a web-based practice test, desktop practice exam software, and downloadable PDF files. ITExamDownload provides proprietary preparation guides for the certification exam offered by the Data-Engineer-Associate Exam Dumps. In addition to containing numerous questions similar to the Data-Engineer-Associate exam, the AWS Certified Data Engineer - Associate (DEA-C01) (Data-Engineer-Associate) exam questions are a great way to prepare for the Amazon Data-Engineer-Associate exam dumps.

Amazon AWS Certified Data Engineer - Associate (DEA-C01) Sample

Questions (Q80-Q85):

NEW QUESTION # 80

A company uses a variety of AWS and third-party data stores. The company wants to consolidate all the data into a central data warehouse to perform analytics. Users need fast response times for analytics queries.

The company uses Amazon QuickSight in direct query mode to visualize the data. Users normally run queries during a few hours each day with unpredictable spikes.

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

- A. Use Amazon Redshift provisioned clusters to load all the data into Amazon Redshift managed storage (RMS).
- **B. Use Amazon Redshift Serverless to load all the data into Amazon Redshift managed storage (RMS).**
- C. Use Amazon Aurora PostgreSQL to load all the data into Aurora.
- D. Use Amazon Athena to load all the data into Amazon S3 in Apache Parquet format.

Answer: B

Explanation:

Problem Analysis:

The company requires a centralized data warehouse for consolidating data from various sources.

They use Amazon QuickSight in direct query mode, necessitating fast response times for analytical queries.

Users query the data intermittently, with unpredictable spikes during the day.

Operational overhead should be minimal.

Key Considerations:

The solution must support fast, SQL-based analytics.

It must handle unpredictable spikes efficiently.

Must integrate seamlessly with QuickSight for direct querying.

Minimize operational complexity and scaling concerns.

Solution Analysis:

Option A: Amazon Redshift Serverless

Redshift Serverless eliminates the need for provisioning and managing clusters.

Automatically scales compute capacity up or down based on query demand.

Reduces operational overhead by handling performance optimization.

Fully integrates with Amazon QuickSight, ensuring low-latency analytics.

Reduces costs as it charges only for usage, making it ideal for workloads with intermittent spikes.

Option B: Amazon Athena with S3 (Apache Parquet)

Athena supports querying data directly from S3 in Parquet format.

While it's cost-effective, performance depends on the size and complexity of the data.

It is not optimized for high-speed analytics needed by QuickSight in direct query mode.

Option C: Amazon Redshift Provisioned Clusters

Requires manual cluster provisioning, scaling, and maintenance.

Higher operational overhead compared to Redshift Serverless.

Option D: Amazon Aurora PostgreSQL

Aurora is optimized for transactional databases, not data warehousing or analytics.

Does not meet the requirement for fast analytics queries.

Final Recommendation:

Amazon Redshift Serverless is the best choice for this use case because it provides fast analytics, integrates natively with QuickSight, and minimizes operational complexity while efficiently handling unpredictable spikes.

Reference:

Amazon Redshift Serverless Overview

Amazon QuickSight and Redshift Integration

NEW QUESTION # 81

A manufacturing company collects sensor data from its factory floor to monitor and enhance operational efficiency. The company uses Amazon Kinesis Data Streams to publish the data that the sensors collect to a data stream. Then Amazon Kinesis Data Firehose writes the data to an Amazon S3 bucket.

The company needs to display a real-time view of operational efficiency on a large screen in the manufacturing facility.

Which solution will meet these requirements with the LOWEST latency?

- A. Configure the S3 bucket to send a notification to an AWS Lambda function when any new object is created. Use the Lambda function to publish the data to Amazon Aurora. Use Aurora as a source to create an Amazon QuickSight dashboard.

- B. Use AWS Glue bookmarks to read sensor data from the S3 bucket in real time. Publish the data to an Amazon Timestream database. Use the Timestream database as a source to create a Grafana dashboard.
- C. Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to process the sensor data. Create a new Data Firehose delivery stream to publish data directly to an Amazon Timestream database. Use the Timestream database as a source to create an Amazon QuickSight dashboard.
- D. Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to process the sensor data. Use a connector for Apache Flink to write data to an Amazon Timestream database. Use the Timestream database as a source to create a Grafana dashboard.

Answer: C

Explanation:

This solution will meet the requirements with the lowest latency because it uses Amazon Managed Service for Apache Flink to process the sensor data in real time and write it to Amazon Timestream, a fast, scalable, and serverless time series database. Amazon Timestream is optimized for storing and analyzing time series data, such as sensor data, and can handle trillions of events per day with millisecond latency. By using Amazon Timestream as a source, you can create an Amazon QuickSight dashboard that displays a real-time view of operational efficiency on a large screen in the manufacturing facility. Amazon QuickSight is a fully managed business intelligence service that can connect to various data sources, including Amazon Timestream, and provide interactive visualizations and insights¹²³.

The other options are not optimal for the following reasons:

* A. Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to process the sensor data. Use a connector for Apache Flink to write data to an Amazon Timestream database. Use the Timestream database as a source to create a Grafana dashboard. This option is similar to option C, but it uses Grafana instead of Amazon QuickSight to create the dashboard.

Grafana is an open source visualization tool that can also connect to Amazon Timestream, but it requires additional steps to set up and configure, such as deploying a Grafana server on Amazon EC2, installing the Amazon Timestream plugin, and creating an IAM role for Grafana to access Timestream.

These steps can increase the latency and complexity of the solution.

* B. Configure the S3 bucket to send a notification to an AWS Lambda function when any new object is created. Use the Lambda function to publish the data to Amazon Aurora. Use Aurora as a source to create an Amazon QuickSight dashboard. This option is not suitable for displaying a real-time view of operational efficiency, as it introduces unnecessary delays and costs in the data pipeline. First, the sensor data is written to an S3 bucket by Amazon Kinesis Data Firehose, which can have a buffering interval of up to 900 seconds. Then, the S3 bucket sends a notification to a Lambda function, which can incur additional invocation and execution time. Finally, the Lambda function publishes the data to Amazon Aurora, a relational database that is not optimized for time series data and can have higher storage and performance costs than Amazon Timestream.

* D. Use AWS Glue bookmarks to read sensor data from the S3 bucket in real time. Publish the data to an Amazon Timestream database. Use the Timestream database as a source to create a Grafana dashboard.

This option is also not suitable for displaying a real-time view of operational efficiency, as it uses AWS Glue bookmarks to read sensor data from the S3 bucket. AWS Glue bookmarks are a feature that helps AWS Glue jobs and crawlers keep track of the data that has already been processed, so that they can resume from where they left off. However, AWS Glue jobs and crawlers are not designed for real-time data processing, as they can have a minimum frequency of 5 minutes and a variable start-up time.

Moreover, this option also uses Grafana instead of Amazon QuickSight to create the dashboard, which can increase the latency and complexity of the solution.

References:

- * 1: Amazon Managed Streaming for Apache Flink
- * 2: Amazon Timestream
- * 3: Amazon QuickSight
- * : Analyze data in Amazon Timestream using Grafana
- * : Amazon Kinesis Data Firehose
- * : Amazon Aurora
- * : AWS Glue Bookmarks
- * : AWS Glue Job and Crawler Scheduling

NEW QUESTION # 82

A company uses AWS Step Functions to orchestrate a data pipeline. The pipeline consists of Amazon EMR jobs that ingest data from data sources and store the data in an Amazon S3 bucket. The pipeline also includes EMR jobs that load the data to Amazon Redshift.

The company's cloud infrastructure team manually built a Step Functions state machine. The cloud infrastructure team launched an EMR cluster into a VPC to support the EMR jobs. However, the deployed Step Functions state machine is not able to run the EMR jobs.

Which combination of steps should the company take to identify the reason the Step Functions state machine is not able to run the EMR jobs? (Choose two.)

- A. Check the retry scenarios that the company configured for the EMR jobs. Increase the number of seconds in the interval between each EMR task. Validate that each fallback state has the appropriate catch for each decision state. Configure an Amazon Simple Notification Service (Amazon SNS) topic to store the error messages.
- **B. Verify that the Step Functions state machine code has all IAM permissions that are necessary to create and run the EMR jobs. Verify that the Step Functions state machine code also includes IAM permissions to access the Amazon S3 buckets that the EMR jobs use. Use Access Analyzer for S3 to check the S3 access properties.**
- C. Use AWS CloudFormation to automate the Step Functions state machine deployment. Create a step to pause the state machine during the EMR jobs that fail. Configure the step to wait for a human user to send approval through an email message. Include details of the EMR task in the email message for further analysis.
- D. Check for entries in Amazon CloudWatch for the newly created EMR cluster. Change the AWS Step Functions state machine code to use Amazon EMR on EKS. Change the IAM access policies and the security group configuration for the Step Functions state machine code to reflect inclusion of Amazon Elastic Kubernetes Service (Amazon EKS).
- **E. Query the flow logs for the VPC. Determine whether the traffic that originates from the EMR cluster can successfully reach the data providers. Determine whether any security group that might be attached to the Amazon EMR cluster allows connections to the data source servers on the informed ports.**

Answer: B,E

Explanation:

To identify the reason why the Step Functions state machine is not able to run the EMR jobs, the company should take the following steps:

* Verify that the Step Functions state machine code has all IAM permissions that are necessary to create and run the EMR jobs.

The state machine code should have an IAM role that allows it to invoke the EMR APIs, such as RunJobFlow, AddJobFlowSteps, and DescribeStep. The state machine code should also have IAM permissions to access the Amazon S3 buckets that the EMR jobs use as input and output locations. The company can use Access Analyzer for S3 to check the access policies and permissions of the S3 buckets¹². Therefore, option B is correct.

* Query the flow logs for the VPC. The flow logs can provide information about the network traffic to and from the EMR cluster that is launched in the VPC. The company can use the flow logs to determine whether the traffic that originates from the EMR cluster can successfully reach the data providers, such as Amazon RDS, Amazon Redshift, or other external sources. The company can also determine whether any security group that might be attached to the EMR cluster allows connections to the data source servers on the informed ports. The company can use Amazon VPC Flow Logs or Amazon CloudWatch Logs Insights to query the flow logs³. Therefore, option D is correct.

Option A is incorrect because it suggests using AWS CloudFormation to automate the Step Functions state machine deployment. While this is a good practice to ensure consistency and repeatability of the deployment, it does not help to identify the reason why the state machine is not able to run the EMR jobs. Moreover, creating a step to pause the state machine during the EMR jobs that fail and wait for a human user to send approval through an email message is not a reliable way to troubleshoot the issue. The company should use the Step Functions console or API to monitor the execution history and status of the state machine, and use Amazon CloudWatch to view the logs and metrics of the EMR jobs.

Option C is incorrect because it suggests changing the AWS Step Functions state machine code to use Amazon EMR on EKS. Amazon EMR on EKS is a service that allows you to run EMR jobs on Amazon Elastic Kubernetes Service (Amazon EKS) clusters. While this service has some benefits, such as lower cost and faster execution time, it does not support all the features and integrations that EMR on EC2 does, such as EMR Notebooks, EMR Studio, and EMRFS. Therefore, changing the state machine code to use EMR on EKS may not be compatible with the existing data pipeline and may introduce new issues.

Option E is incorrect because it suggests checking the retry scenarios that the company configured for the EMR jobs. While this is a good practice to handle transient failures and errors, it does not help to identify the root cause of why the state machine is not able to run the EMR jobs. Moreover, increasing the number of seconds in the interval between each EMR task may not improve the success rate of the jobs, and may increase the execution time and cost of the state machine. Configuring an Amazon SNS topic to store the error messages may help to notify the company of any failures, but it does not provide enough information to troubleshoot the issue.

:

1: Manage an Amazon EMR Job - AWS Step Functions

2: Access Analyzer for S3 - Amazon Simple Storage Service

3: Working with Amazon EMR and VPC Flow Logs - Amazon EMR

[4]: Analyzing VPC Flow Logs with Amazon CloudWatch Logs Insights - Amazon Virtual Private Cloud

[5]: Monitor AWS Step Functions - AWS Step Functions

[6]: Monitor Amazon EMR clusters - Amazon EMR

[7]: Amazon EMR on Amazon EKS - Amazon EMR

NEW QUESTION # 83

A company has a data processing pipeline that runs multiple SQL queries in sequence against an Amazon Redshift cluster. The company merges with a second company. The original company modifies a query that aggregates sales revenue data to join sales tables from both companies.

The sales table for the first company is named Table S1 and contains 10 billion records. The sales table for the second company is named Table S2 and contains 900 million records. The query becomes slow after the modification.

A data engineer must improve the query performance.

Which solutions will meet these requirements? (Select TWO)

- A. Use the EVEN distribution style for Table S1. Use the ALL distribution style for Table S2.
- B. Use the KEY distribution style for both sales tables. Select a low-cardinality column to use for the join.
- C. Use the Amazon Redshift query optimizer to review and select optimizations to implement.
- **D. Use Amazon Redshift Advisor to review and select optimizations to implement.**
- **E. Use the KEY distribution style for both sales tables. Select a high-cardinality column to use for the join.**

Answer: D,E

Explanation:

Amazon Redshift query performance for large joins depends heavily on data distribution and query optimization. Using the KEY distribution style on both tables ensures that rows with the same join key are co-located on the same compute nodes, minimizing costly data redistribution during joins.

Selecting a high-cardinality column as the distribution key is critical to evenly distribute data across nodes and prevent data skew. Low-cardinality keys can cause hotspots where too much data resides on a small number of nodes, degrading performance.

Using EVEN distribution on the 10-billion-row table and ALL distribution on the 900-million-row table is inefficient because ALL distribution replicates the entire table to every node, which is not appropriate for large datasets. The Amazon Redshift query optimizer operates automatically and does not require manual selection.

Amazon Redshift Advisor analyzes workload patterns and provides actionable recommendations for distribution styles, sort keys, and query optimizations. Using Redshift Advisor allows the data engineer to apply proven optimizations based on actual cluster usage.

Therefore, combining KEY distribution with a high-cardinality join column and leveraging Amazon Redshift Advisor recommendations is the most effective solution to improve query performance.

NEW QUESTION # 84

A manufacturing company wants to collect data from sensors. A data engineer needs to implement a solution that ingests sensor data in near real time.

The solution must store the data to a persistent data store. The solution must store the data in nested JSON format. The company must have the ability to query from the data store with a latency of less than 10 milliseconds.

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

- A. Use AWS Lambda to process the sensor data. Store the data in Amazon S3 for querying.
- B. Use a self-hosted Apache Kafka cluster to capture the sensor data. Store the data in Amazon S3 for querying.
- **C. Use Amazon Kinesis Data Streams to capture the sensor data. Store the data in Amazon DynamoDB for querying.**
- D. Use Amazon Simple Queue Service (Amazon SQS) to buffer incoming sensor data. Use AWS Glue to store the data in Amazon RDS for querying.

Answer: C

Explanation:

Amazon Kinesis Data Streams is a service that enables you to collect, process, and analyze streaming data in real time. You can use Kinesis Data Streams to capture sensor data from various sources, such as IoT devices, web applications, or mobile apps. You can create data streams that can scale up to handle any amount of data from thousands of producers. You can also use the Kinesis Client Library (KCL) or the Kinesis Data Streams API to write applications that process and analyze the data in the streams¹.

Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. You can use DynamoDB to store the sensor data in nested JSON format, as DynamoDB supports document data types, such as lists and maps. You can also use DynamoDB to query the data with a latency of less than 10 milliseconds, as DynamoDB offers single-digit millisecond performance for any scale of data. You can use the DynamoDB API or the AWS SDKs to perform queries on the data, such as using key-value lookups, scans, or queries².

The solution that meets the requirements with the least operational overhead is to use Amazon Kinesis Data Streams to capture the sensor data and store the data in Amazon DynamoDB for querying. This solution has the following advantages:

It does not require you to provision, manage, or scale any servers, clusters, or queues, as Kinesis Data Streams and DynamoDB are fully managed services that handle all the infrastructure for you. This reduces the operational complexity and cost of running your solution.

It allows you to ingest sensor data in near real time, as Kinesis Data Streams can capture data records as they are produced and deliver them to your applications within seconds. You can also use Kinesis Data Firehose to load the data from the streams to DynamoDB automatically and continuously.

It allows you to store the data in nested JSON format, as DynamoDB supports document data types, such as lists and maps. You can also use DynamoDB Streams to capture changes in the data and trigger actions, such as sending notifications or updating other databases.

It allows you to query the data with a latency of less than 10 milliseconds, as DynamoDB offers single-digit millisecond performance for any scale of data. You can also use DynamoDB Accelerator (DAX) to improve the read performance by caching frequently accessed data.

Option A is incorrect because it suggests using a self-hosted Apache Kafka cluster to capture the sensor data and store the data in Amazon S3 for querying. This solution has the following disadvantages:

It requires you to provision, manage, and scale your own Kafka cluster, either on EC2 instances or on-premises servers. This increases the operational complexity and cost of running your solution.

It does not allow you to query the data with a latency of less than 10 milliseconds, as Amazon S3 is an object storage service that is not optimized for low-latency queries. You need to use another service, such as Amazon Athena or Amazon Redshift Spectrum, to query the data in S3, which may incur additional costs and latency.

Option B is incorrect because it suggests using AWS Lambda to process the sensor data and store the data in Amazon S3 for querying. This solution has the following disadvantages:

It does not allow you to ingest sensor data in near real time, as Lambda is a serverless compute service that runs code in response to events. You need to use another service, such as API Gateway or Kinesis Data Streams, to trigger Lambda functions with sensor data, which may add extra latency and complexity to your solution.

It does not allow you to query the data with a latency of less than 10 milliseconds, as Amazon S3 is an object storage service that is not optimized for low-latency queries. You need to use another service, such as Amazon Athena or Amazon Redshift Spectrum, to query the data in S3, which may incur additional costs and latency.

Option D is incorrect because it suggests using Amazon Simple Queue Service (Amazon SQS) to buffer incoming sensor data and use AWS Glue to store the data in Amazon RDS for querying. This solution has the following disadvantages:

It does not allow you to ingest sensor data in near real time, as Amazon SQS is a message queue service that delivers messages in a best-effort manner. You need to use another service, such as Lambda or EC2, to poll the messages from the queue and process them, which may add extra latency and complexity to your solution.

It does not allow you to store the data in nested JSON format, as Amazon RDS is a relational database service that supports structured data types, such as tables and columns. You need to use another service, such as AWS Glue, to transform the data from JSON to relational format, which may add extra cost and overhead to your solution.

[1]: Amazon Kinesis Data Streams - Features

[2]: Amazon DynamoDB - Features

[3]: Loading Streaming Data into Amazon DynamoDB - Amazon Kinesis Data Firehose

[4]: Capturing Table Activity with DynamoDB Streams - Amazon DynamoDB

[5]: Amazon DynamoDB Accelerator (DAX) - Features

[6]: Amazon S3 - Features

[7]: AWS Lambda - Features

[8]: Amazon Simple Queue Service - Features

[9]: Amazon Relational Database Service - Features

[10]: Working with JSON in Amazon RDS - Amazon Relational Database Service

[11]: AWS Glue - Features

NEW QUESTION # 85

.....

It is browser-based; therefore no need to install it, and you can start practicing for the AWS Certified Data Engineer - Associate (DEA-C01) (Data-Engineer-Associate) exam by creating the Amazon Data-Engineer-Associate practice test. You don't need to install any separate software or plugin to use it on your system to practice for your actual AWS Certified Data Engineer - Associate (DEA-C01) (Data-Engineer-Associate) exam. ITExamDownload AWS Certified Data Engineer - Associate (DEA-C01) (Data-Engineer-Associate) web-based practice software is supported by all well-known browsers like Chrome, Firefox, Opera, Internet Explorer, etc.

Latest Data-Engineer-Associate Cram Materials: <https://www.itexamdownload.com/Data-Engineer-Associate-valid-questions.html>

Disposable vapes

What's more, part of that ITEXamDownload Data-Engineer-Associate dumps now are free: <https://drive.google.com/open?id=1skPLwEzrrqK-pxRnztwvf7V2bSSy->