DUMPSBOSS.COM

AWS Certified Data Analytics - Specialty

Amazon AWS DAS-C01

Version Demo

Total Demo Questions: 10

Total Premium Questions: 164

Buy Premium PDF

https://dumpsboss.com

support@dumpsboss.com

QUESTION NO: 1

A company is building a data lake and needs to ingest data from a relational database that has time-series data. The company wants to use managed services to accomplish this. The process needs to be scheduled daily and bring incremental data only from the source into Amazon S3.

What is the MOST cost-effective approach to meet these requirements?

A. Use AWS Glue to connect to the data source using JDBC Drivers. Ingest incremental records only using job bookmarks.

B. Use AWS Glue to connect to the data source using JDBC Drivers. Store the last updated key in an Amazon DynamoDB table and ingest the data using the updated key as a filter.

C. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the entire dataset. Use appropriate Apache Spark libraries to compare the dataset, and find the delta.

D. Use AWS Glue to connect to the data source using JDBC Drivers and ingest the full data. Use AWS DataSync to ensure the delta only is written into Amazon S3.

ANSWER: B

QUESTION NO: 2

A company provides an incentive to users who are physically active. The company wants to determine how active the users are by using an application on their mobile devices to track the number of steps they take each day. The company needs to ingest and perform near-real-time analytics on live data. The processed data must be stored and must remain available for 1 year for analytics purposes.

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

A. Use Amazon Cognito to write the data from the application to Amazon DynamoDB. Use an AWS Step Functions workflow to create a transient Amazon EMR cluster every hour and process the new data from DynamoDB. Output the processed data to Amazon Redshift for analytics. Archive the data from Amazon Redshift after 1 year.

B. Ingest the data into Amazon DynamoDB by using an Amazon API Gateway API as a DynamoDB proxy. Use an AWS Step Functions workflow to create a transient Amazon EMR cluster every hour and process the new data from DynamoDOutput the processed data to Amazon Redshift to run analytics calculations. Archive the data from Amazon Redshift after 1 year.

C. Ingest the data into Amazon Kinesis Data Streams by using an Amazon API Gateway API as a Kinesis proxy. Run Amazon Kinesis Data Analytics on the stream data. Output the processed data into Amazon S3 by using Amazon Kinesis Data Firehose. Use Amazon Athena to run analytics calculations. Use S3 Lifecycle rules to transition objects to S3 Glacier after 1 year.

D. Write the data from the application into Amazon S3 by using Amazon Kinesis Data Firehose. Use Amazon Athena to run the analytics on the data in Amazon S3. Use S3 Lifecycle rules to transition objects to S3 Glacier after 1 year.

ANSWER: C

QUESTION NO: 3

A company has collected more than 100 TB of log files in the last 24 months. The files are stored as raw text in a dedicated Amazon S3 bucket. Each object has a key of the form year-month-

day_log_HHmmss.txt where HHmmss represents the time the log file was initially created. A table was created in Amazon Athena that points to the S3 bucket. One-time queries are run against a subset of columns in the table several times an hour.

A data analyst must make changes to reduce the cost of running these queries. Management wants a solution with minimal maintenance overhead. Which combination of steps should the data analyst take to meet these requirements? (Choose three.)

A. Convert the log files to Apace Avro format.

B. Add a key prefix of the form date=year-month-day/ to the S3 objects to partition the data.

C. Convert the log files to Apache Parquet format.

D. Add a key prefix of the form year-month-day/ to the S3 objects to partition the data.

E. Drop and recreate the table with the PARTITIONED BY clause. Run the ALTER TABLE ADD PARTITION statement.

F. Drop and recreate the table with the PARTITIONED BY clause. Run the MSCK REPAIR TABLE statement.

ANSWER: B C F

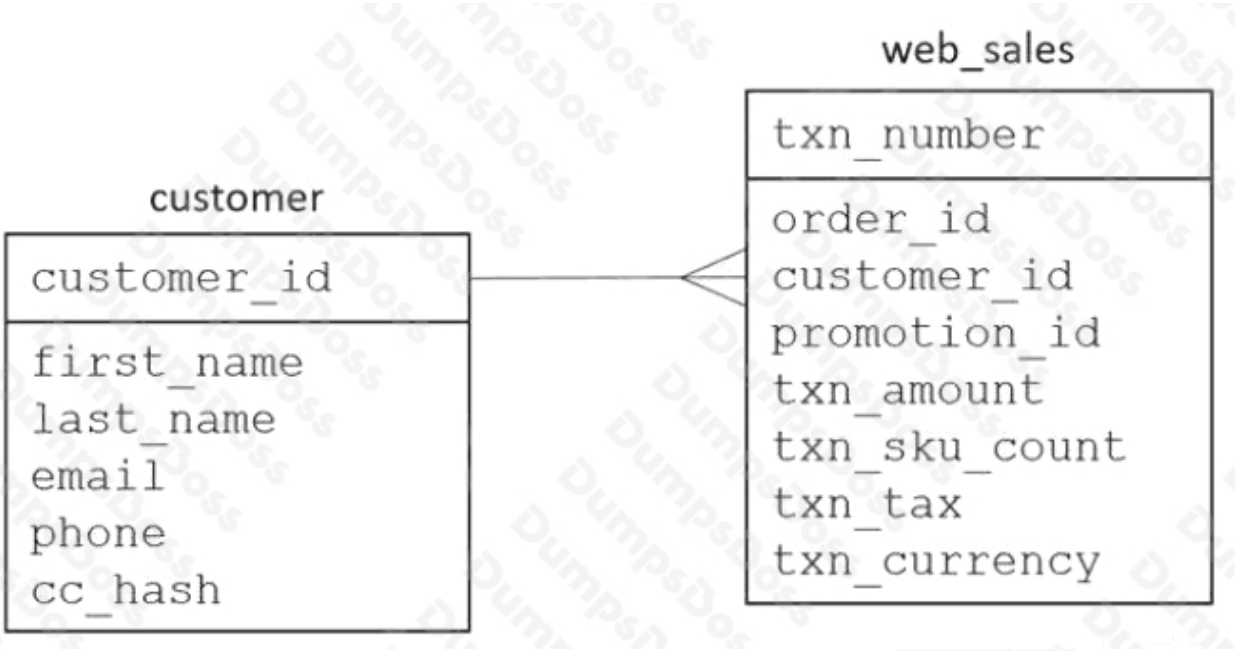
Explanation:

Reference: https://docs.aws.amazon.com/athena/latest/ug/msck-repair-table.html

QUESTION NO: 4

A retail company is using an Amazon S3 bucket to host an ecommerce data lake. The company is using AWS Lake Formation to manage the data lake.

A data analytics specialist must provide access to a new business analyst team. The team will use Amazon Athena from the AWS Management Console to query data from existing web_sales and customer tables in the ecommerce database. The team needs read-only access and the ability to uniquely identify customers by using first and last names. However, the team must not be able to see any other personally identifiable data. The table structure is as follows:



Which combination of steps should the data analytics specialist take to provide the required permission by using the principle of least privilege? (Choose three.)

A. In AWS Lake Formation, grant the business_analyst group SELECT and ALTER permissions for the web_sales table.

B. In AWS Lake Formation, grant the business_analyst group the SELECT permission for the web_sales table.

C. In AWS Lake Formation, grant the business_analyst group the SELECT permission for the customer table. Under columns, choose filter type “Include columns” with columns fisrt_name, last_name, and customer_id.

D. In AWS Lake Formation, grant the business_analyst group SELECT and ALTER permissions for the customer table. Under columns, choose filter type “Include columns” with columns fisrt_name and last_name.

E. Create users under a business_analyst IAM group. Create a policy that allows the lakeformation:GetDataAccess action, the athena:* action, and the glue:Get* action.

F. Create users under a business_analyst IAM group. Create a policy that allows the lakeformation:GetDataAccess action, the athena:* action, and the glue:Get* action. In addition, allow the s3:GetObject action, the s3:PutObject action, and the s3:GetBucketLocation action for the Athena query results S3 bucket.

ANSWER: B D F

QUESTION NO: 5

A data analytics specialist is setting up workload management in manual mode for an Amazon Redshift environment. The data analytics specialist is defining query monitoring rules to manage system performance and user experience of an Amazon Redshift cluster.

Which elements must each query monitoring rule include?

A. A unique rule name, a query runtime condition, and an AWS Lambda function to resubmit any failed queries in off hours

B. A queue name, a unique rule name, and a predicate-based stop condition

C. A unique rule name, one to three predicates, and an action

D. A workload name, a unique rule name, and a query runtime-based condition

ANSWER: C

Explanation:

Reference: https://docs.aws.amazon.com/redshift/latest/dg/cm-c-wlm-query-monitoring-rules.html

QUESTION NO: 6

A gaming company is building a serverless data lake. The company is ingesting streaming data into Amazon Kinesis Data Streams and is writing the data to Amazon S3 through Amazon Kinesis Data Firehose. The company is using 10 MB as the S3 buffer size and is using 90 seconds as the buffer interval. The company runs an AWS Glue ETL job to merge and transform the data to a different format before writing the data back to Amazon S3.

Recently, the company has experienced substantial growth in its data volume. The AWS Glue ETL jobs are frequently showing an OutOfMemoryError error.

Which solutions will resolve this issue without incurring additional costs? (Choose two.)

A. Place the small files into one S3 folder. Define one single table for the small S3 files in AWS Glue Data Catalog. Rerun the AWS Glue ETL jobs against this AWS Glue table.

B. Create an AWS Lambda function to merge small S3 files and invoke them periodically. Run the AWS Glue ETL jobs after successful completion of the Lambda function.

C. Run the S3DistCp utility in Amazon EMR to merge a large number of small S3 files before running the AWS Glue ETL jobs.

D. Use the groupFiles setting in the AWS Glue ETL job to merge small S3 files and rerun AWS Glue ETL jobs.

E. Update the Kinesis Data Firehose S3 buffer size to 128 MB. Update the buffer interval to 900 seconds.

ANSWER: A D

Explanation:

Reference: https://docs.aws.amazon.com/glue/latest/dg/grouping-input-files.html
https://docs.aws.amazon.com/glue/latest/dg/grouping-input-files.html

QUESTION NO: 7

A company is providing analytics services to its sales and marketing departments. The departments can access the data only through their business intelligence (BI) tools, which run queries on Amazon Redshift using an Amazon Redshift internal user to connect. Each department is assigned a user in the Amazon Redshift database with the permissions needed for that department. The marketing data analysts must be granted direct access to the advertising table, which is stored in Apache

Parquet format in the marketing S3 bucket of the company data lake. The company data lake is managed by AWS Lake Formation. Finally, access must be limited to the three promotion columns in the table. Which combination of steps will meet these requirements? (Choose three.)

A. Grant permissions in Amazon Redshift to allow the marketing Amazon Redshift user to access the three promotion columns of the advertising external table.

B. Create an Amazon Redshift Spectrum IAM role with permissions for Lake Formation. Attach it to the Amazon Redshift cluster.

C. Create an Amazon Redshift Spectrum IAM role with permissions for the marketing S3 bucket. Attach it to the Amazon Redshift cluster.

D. Create an external schema in Amazon Redshift by using the Amazon Redshift Spectrum IAM role. Grant usage to the marketing Amazon Redshift user.

E. Grant permissions in Lake Formation to allow the Amazon Redshift Spectrum role to access the three promotion columns of the advertising table.

F. Grant permissions in Lake Formation to allow the marketing IAM group to access the three promotion columns of the advertising table.

ANSWER: B D E

QUESTION NO: 8

A company is streaming its high-volume billing data (100 MBps) to Amazon Kinesis Data Streams. A data analyst partitioned the data on account_id to ensure that all records belonging to an account go to the same Kinesis shard and order is maintained. While building a custom consumer using the Kinesis Java SDK, the data analyst notices that, sometimes, the messages arrive out of order for account_id.

Upon further investigation, the data analyst discovers the messages that are out of order seem to be arriving from different shards for the same account_id and are seen when a stream resize runs.

What is an explanation for this behavior and what is the solution?

A. There are multiple shards in a stream and order needs to be maintained in the shard. The data analyst needs to make sure there is only a single shard in the stream and no stream resize runs.

B. The hash key generation process for the records is not working correctly. The data analyst should generate an explicit hash key on the producer side so the records are directed to the appropriate shard accurately.

C. The records are not being received by Kinesis Data Streams in order. The producer should use the PutRecords API call instead of the PutRecord API call with the SequenceNumberForOrdering parameter.

D. The consumer is not processing the parent shard completely before processing the child shards after a stream resize. The data analyst should process the parent shard completely first before processing the child shards.

ANSWER: A

QUESTION NO: 9

A social media company is using business intelligence tools to analyze its data for forecasting. The company is using Apache Kafka to ingest the low-velocity data in near-real time. The company wants to build dynamic dashboards with machine learning (ML) insights to forecast key business trends. The dashboards must provide hourly updates from data in Amazon S3. Various teams at the company want to view the dashboards by using Amazon QuickSight with ML insights. The solution also must correct the scalability problems that the company experiences when it uses its current architecture to ingest data.

Which solution will MOST cost-effectively meet these requirements?

A. Replace Kafka with Amazon Managed Streaming for Apache Kafka. Ingest the data by using AWS Lambda, and store the data in Amazon S3. Use QuickSight Standard edition to refresh the data in SPICE from Amazon S3 hourly and create a dynamic dashboard with forecasting and ML insights.

B. Replace Kafka with an Amazon Kinesis data stream. Use an Amazon Kinesis Data Firehose delivery stream to consume the data and store the data in Amazon S3. Use QuickSight Enterprise edition to refresh the data in SPICE from Amazon S3 hourly and create a dynamic dashboard with forecasting and ML insights.

C. Configure the Kafka-Kinesis-Connector to publish the data to an Amazon Kinesis Data Firehose delivery stream that is configured to store the data in Amazon S3. Use QuickSight Enterprise edition to refresh the data in SPICE from Amazon S3 hourly and create a dynamic dashboard with forecasting and ML insights.

D. Configure the Kafka-Kinesis-Connector to publish the data to an Amazon Kinesis Data Firehose delivery stream that is configured to store the data in Amazon S3. Configure an AWS Glue crawler to crawl the data. Use an Amazon Athena data source with QuickSight Standard edition to refresh the data in SPICE hourly and create a dynamic dashboard with forecasting and ML insights.

ANSWER: B

Explanation:

Reference: https://noise.getoto.net/tag/amazon-kinesis-data-firehose/

QUESTION NO: 10

An airline has been collecting metrics on flight activities for analytics. A recently completed proof of concept demonstrates how the company provides insights to data analysts to improve on-time departures. The proof of concept used objects in Amazon S3, which contained the metrics in .csv format, and used Amazon Athena for querying the data. As the amount of data increases, the data analyst wants to optimize the storage solution to improve query performance.

Which options should the data analyst use to improve performance as the data lake grows? (Choose three.)

A. Add a randomized string to the beginning of the keys in S3 to get more throughput across partitions.

B. Use an S3 bucket in the same account as Athena.

C. Compress the objects to reduce the data transfer I/O.

D. Use an S3 bucket in the same Region as Athena.

E. Preprocess the .csv data to JSON to reduce I/O by fetching only the document keys needed by the query.

F. Preprocess the .csv data to Apache Parquet to reduce I/O by fetching only the data blocks needed for predicates.

ANSWER: A C E