Amazon

AWS-CERTIFIED-BIG-DATA-SPECIALTY Exam

Amazon AWS Certified Big Data - Specialty Exam

Questions & Answers Demo

Version: 12.0

Question: 1

A data engineer in a manufacturing company is designing a data processing platform that receives a large volume of unstructured data. The data engineer must populate a well-structured star schema in Amazon Redshift.
What is the most efficient architecture strategy for this purpose?

A. Transform the unstructured data using Amazon EMR and generate CSV data. COPY the CSV data into the analysis schema within Redshift.
B. Load the unstructured data into Redshift, and use string parsing functions to extract structured data for inserting into the analysis schema.
C. When the data is saved to Amazon S3, use S3 Event Notifications and AWS Lambda to transform the file contents. Insert the data into the analysis schema on Redshift.
D. Normalize the data using an AWS Marketplace ETL tool, persist the results to Amazon S3, and use AWS Lambda to INSERT the data into Redshift.

Answer: A

Question: 2

A new algorithm has been written in Python to identify SPAM e-mails. The algorithm analyzes the free text contained within a sample set of 1 million e-mails stored on Amazon S3. The algorithm must be scaled across a production dataset of 5 PB, which also resides in Amazon S3 storage.
Which AWS service strategy is best for this use case?

A. Copy the data into Amazon ElastiCache to perform text analysis on the in-memory data and export the results of the model into Amazon Machine Learning.
B. Use Amazon EMR to parallelize the text analysis tasks across the cluster using a streaming program step.
C. Use Amazon Elasticsearch Service to store the text and then use the Python Elasticsearch Client to run analysis against the text index.
D. Initiate a Python job from AWS Data Pipeline to run directly against the Amazon S3 text files.

Answer: C

Explanation:
Reference:
https://aws.amazon.com/blogs/database/indexing-metadata-in-amazon-elasticsearch-service-using-aws-lambda-and-python/

Question: 3

A data engineer chooses Amazon DynamoDB as a data store for a regulated application. This application must be submitted to regulators for review. The data engineer needs to provide a control framework that lists the security controls from the process to follow to add new users down to the physical controls of the data center, including items like security guards and cameras.
How should this control mapping be achieved using AWS?

A. Request AWS third-party audit reports and/or the AWS quality addendum and map the AWS responsibilities to the controls that must be provided.
B. Request data center Temporary Auditor access to an AWS data center to verify the control mapping.
C. Request relevant SLAs and security guidelines for Amazon DynamoDB and define these guidelines within the application's architecture to map to the control framework.
D. Request Amazon DynamoDB system architecture designs to determine how to map the AWS responsibilities to the control that must be provided.

Answer: A

Question: 4

An administrator needs to design a distribution strategy for a star schema in a Redshift cluster. The administrator needs to determine the optimal distribution style for the tables in the Redshift schema. In which three circumstances would choosing Key-based distribution be most appropriate? (Select three.)
A. When the administrator needs to optimize a large, slowly changing dimension table.
B. When the administrator needs to reduce cross-node traffic.
C. When the administrator needs to optimize the fact table for parity with the number of slices.
D. When the administrator needs to balance data distribution and collocation data.
E. When the administrator needs to take advantage of data locality on a local node for joins and aggregates.

Answer: A,C,D

Question: 5

Company A operates in Country X. Company A maintains a large dataset of historical purchase orders that contains personal data of their customers in the form of full names and telephone numbers. The dataset consists of 5 text files, 1TB each. Currently the dataset resides on-premises due to legal requirements of storing personal data in-country. The research and development department needs to run a clustering algorithm on the dataset and wants to use Elastic Map Reduce service in the closest AWS region. Due to geographic distance, the minimum latency between the on-premises system and the closet AWS region is 200 ms.
Which option allows Company A to do clustering in the AWS Cloud and meet the legal requirement of maintaining personal data in-country?

A. Anonymize the personal data portions of the dataset and transfer the data files into Amazon S3 in the AWS region. Have the EMR cluster read the dataset using EMRFS.

B. Establish a Direct Connect link between the on-premises system and the AWS region to reduce latency. Have the EMR cluster read the data directly from the on-premises storage system over Direct Connect.
C. Encrypt the data files according to encryption standards of Country X and store them on AWS region in Amazon S3. Have the EMR cluster read the dataset using EMRFS.
D. Use AWS Import/Export Snowball device to securely transfer the data to the AWS region and copy the files onto an EBS volume. Have the EMR cluster read the dataset using EMRFS.

Answer: B

Question: 6

An administrator needs to design a strategy for the schema in a Redshift cluster. The administrator needs to determine the optimal distribution style for the tables in the Redshift schema.
In which two circumstances would choosing EVEN distribution be most appropriate? (Choose two.)
A. When the tables are highly denormalized and do NOT participate in frequent joins.
B. When data must be grouped based on a specific key on a defined slice.
C. When data transfer between nodes must be eliminated.
D. When a new table has been loaded and it is unclear how it will be joined to dimension.

Answer: B,D