Amazon

BDS-C00
AWS Certified Big Data - Specialty

Questions And Answers PDF Format:

For More Information – Visit link below:
https://www.certsgrade.com/

Version = **Product**

Latest Version: 10.0

Question: 1

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 need to run a clustering algorithm on the dataset and wants to use Elastic Map Reduce service in the closes 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. Establishing 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 in 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: 2

A company needs a churn prevention model to predict which customers will NOT review their yearly subscription to the company's service. The company plans to provide these customers with a promotional offer. A binary classification model that uses Amazon Machine Learning is required.
On which basis should this binary classification model be built?

A. User profiles (age, gender, income, occupation)
B. Last user session
C. Each user time series events in the past 3 months
D. Quarterly results

Answer: C

Question: 3

A company that provides economics data dashboards needs to be able to develop software to display rich, interactive, data-driven graphics that run in web browsers and leverages the full stack of web standards (HTML, SVG and CSS).
Which technology provides the most appropriate for this requirement?

A. D3.js
B. Python/Jupyter
C. R Studio
D. Hue

Answer: A

Refer : https://d3js.org/. D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

Question: 4

A customer needs to determine the optimal distribution strategy for the ORDERS fact table in its Redshift schem

a. The ORDERS table has foreign key relationships with multiple dimension tables in this schema. How should the company determine the most appropriate distribution key for the ORDRES table?
A. Identity the largest and most frequently joined dimension table and ensure that it and the ORDERS table both have EVEN distribution
B. Identify the target dimension table and designate the key of this dimension table as the distribution key of the ORDERS table
C. Identity the smallest dimension table and designate the key of this dimension table as the distribution key of ORDERS table
D. Identify the largest and most frequently joined dimension table and designate the key of this dimension table as the distribution key for the orders table

Answer: D

https://aws.amazon.com/blogs/big-data/optimizing-for-star-schemas-and-interleaved-sorting-on-amazon-redshift/

Question: 5

A company has several teams of analytics. Each team of analysts has their own cluster. The teams need to run SQL queries using Hive, Spark-SQL and Presto with Amazon EMR. The company needs to enable a centralized metadata layer to expose the Amazon S3 objects as tables to the analysts.
Which approach meets the requirement for a centralized metadata layer?

A. EMRFS consistent view with a common Amazon DynamoDB table
B. Bootstrap action to change the Hive Metastore to an Amazon RDS database
C. s3distcp with the outputManifest option to generate RDS DDL
D. naming scheme support with automatic partition discovery from Amazon S3

Answer: B

https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-metastore-external-hive.html

Question: 6

The department of transportation for a major metropolitan area has placed sensors on roads at key locations around the city. The goal is to analyze the flow of traffic and notifications from emergency services to identity potential issues and to help planners correct trouble spots.
A data engineer needs a scalable and fault-tolerant solution that allows planners to respond to issues within 30 seconds of their occurrence.
Which solution should the data engineer choose?

A. Collect the sensor data with Amazon Kinesis Firehose and store it in Amazon Redshift for analysis. Collect emergency services events with Amazon SQS and store in Amazon DynamoDB for analysis
B. Collect the sensor data with Amazon SQS and store in Amazon DynamoDB for analysis.
C. Collect both sensor data and emergency services events with Amazon Kinesis Streams and use Amazon DynamoDB for analysis
D. Collect both sensor data and emergency services events with Amazon Kinesis Firehouse and use Amazon Redshift for Analysis

Answer: C

https://sookocheff.com/post/aws/comparing-kinesis-and-sqs/

Question: 7

An online photo album app has a key design feature to support multiple screens (e.g. desktop, mobile phone, and tablet) with high quality displays. Multiple versions of the image must be saved in different resolutions and layouts.
The image processing Java program takes an average of five seconds per upload, depending on the image size and format. Each image upload captures the following image metadata: user, album, photo label, upload timestamp
The app should support the following requirements:
• Hundreds of user image uploads per second
• Maximum image metadata size of 10 MB
• Maximum image metadata size of 1 KB
•Image displayed in optimized resolution in all supported screens no later than one minute after image upload
Which strategy should be used to meet these requirements?

A. Write images and metadata to Amazon Kinesis, Use a Kinesis Client Library (KCL) application to run the image processing and save the image output to Amazon S3 and metadata to the app repository DB
B. Write image and metadata RDS with BLOB data type. Use AWS Data Pipeline to run the image processing and save the image output to Amazon S3 and metadata to the app repository DB
C. Upload image with metadata to Amazon S3 use Lambda function to run the image processing and save the image output to Amazon S3 and metadata to the app repository DB
D. Write image and metadata to Amazon kinesis. Use Amazon Elastic MapReduce (EMR) with Spark Streaming to run image processing and save image output to Amazon

Answer: D

Question: 8

A data engineer is running a DWH on a 25-node Redshift cluster of a SaaS service. The data engineer needs to build a dashboard that will be used by customers. Five big customers represent 80% of usage, and there is a long tail of dozens of smaller customers. The data engineer has selected the dashboarding tool.
How should the data engineer make sure that the larger customer workloads do NOT interfere with the smaller customer workloads?

A. Apply query filters based on customer-id that can NOT be changed by the user and apply distribution keys on customer id
B. Place the largest customers into a single user group with a dedicated query queue and place the rest of the customer into a different query queue

C. Push aggregations into an RDS for Aurora instance. Connect the dashboard application to Aurora rather than Redshift for faster queries
D. Route the largest customers to a dedicated Redshift cluster, Raise the concurrency of the multi-tenant Redshift cluster to accommodate the remaining customers

Answer: B

https://docs.aws.amazon.com/redshift/latest/dg/c_workload_mngmt_classification.html

Question: 9

A solutions architect works for a company that has a data lake based on a central Amazon S3 bucket. The data contains sensitive information. The architect must be able to specify exactly which files each user can access. Users access the platform through SAML federation Single Sign On platform.
The architect needs to build a solution that allows fine grained access control, traceability of access to the objects, and usage of the standard tools (AWS Console, AWS CLI) to access the data.
Which solution should the architect build?

A. Use Amazon S3 Server-Side Encryption with AWS KMS-Managed Keys for strong data.
B. Use Amazon S3 Server-Side Encryption with Amazon S3 Managed Keys. Set Amazon S3
C. Use Amazon S3 Client-Side Encryption with Client-Side Master Key. Set Amazon S3 ACI to allow access to specific elements of the platform. Use Amazon S3 access logs for auditing
D. Use Amazon S3 Client-Side Encryption with AWS KMS-Managed keys for storing data.

Answer: D

Question: 10

An Amazon Kinesis stream needs to be encrypted.
Which approach should be used to accomplish this task?

A. Perform a client-side encryption of the data before it enters the Amazon Kinesis stream on the producer
B. Use a partition key to segment the data by MD5 hash functions which makes indecipherable while in transit
C. Perform a client-side encryption of the data before it enters the Amazon Kinesis stream on the consumer
D. Use a shard to segment the data which has built-in functionality to make it indecipherable while in transit

Answer: A

https://aws.amazon.com/blogs/big-data/encrypt-and-decrypt-amazon-kinesis-records-using-aws-kms/

Question: 11

A telecommunications company needs to predict customer churn (i.e. customers eho decide to switch a computer). The company has historic records of each customer, including monthly consumption patterns, calls to customer service, and whether the customer ultimately quit th eservice. All of this data is stored in Amazon S3. The company needs to know which customers are likely going to churn soon so that they can win back their loyalty.
What is the optimal approach to meet these requirements?

A. Use the Amazon Machine Learning service to build the binary classification model based on the dataset stored in Amazon S3. The model will be used regularly to predict churn attribute for existing customers
B. Use AWS QuickSight to connect it to data stored in Amazon S3 to obtain the necessary
C. Use EMR to run the Hive queries to build a profile of a churning customer. Apply the profile to existing customers to determine the likelihood of churn
D. Use a Redshift cluster to COPY the data from Amazon S3. Create a user Define Function in Redshift that computers the likelihood of churn

Answer: A

https://aws.amazon.com/blogs/machine-learning/predicting-customer-churn-with-amazon-machine-learning/

Question: 12

An administrator is deploying Spark on Amazon EMR for two distinct use cases: machine learning algorithms and ad hoc querying. All data will be stored in Amazon S3. Two separate clusters for each use case will be deployed. The data volumes on Amazon S3 are less than 10 GB.
How should the administrator align instance types with the cluster's purpose?

A. Machine Learning on C instance types and ad-hoc queries on R instance types
B. Machine Learning on R instance types and ad-hoc queries on G2 instance types
C. Machine Learning on T instance types and ad-hoc queries on M instance types
D. Machine Learning on D instance types and ad-hoc queries on I instance types

Answer: A

Question: 13

A company hosts a portfolio of e-commerce websites across the Oregon, N.Virginia, Ireland and Sydney AWS regions. Each site keeps log files that captures user behavior. The company has built an application that generates batches of product recommendations with collaborative filtering in Oregon. Oregon was selected because the flagship site is hosted there and provides the largest collection of data to train machine learning models against. The other regions do NOT have enough historic data to train accurate machine learning models.
Which set of data processing steps improves recommendations for each region?

A. Use the e-commerce application in Oregon to write replica log files in each other region
B. Use Amazon S3 bucket replication to consolidate log entries and builds a single model in
C. Use Kinesis as a butler for web logs and replicate logs to the Kinesis streams of a neighboring region
D. Use the CloudWatch Logs agent to consolidate logs into a single CloudWatch logs group

Answer: D

https://cloudacademy.com/blog/centralized-log-management-with-aws-cloudwatch-part-1-of-3/

Question: 14

A media advertising company handles a large number of real-time messages sourced from over 200 websites. The company's data engineer needs to collect and process records in real time for analysis using Spark Streaming on Amazon Elastic MapReduce (EMR). The data engineer needs to fulfill a corporate mandate to keep ALL raw messages as they are received as a top priority.
Which Amazon Kinesis configuration meets these requirements?

A. Publish messages to Amazon Kinesis Firehose backed by Amazon Simple Storage Service (S3). Pull messages off Firehose with Spark Streaming in parallel to persistence to Amazon S3
B. Publish messages to Amazon Kinesis Streams. Pull messages off Stream with Spark Streaming in parallel to AWS messages from Streams to Firehose backed by Amazon Simple Storage Service (S3)
C. Publish messages to Amazon Kinesis Firehose backed by Amazon Simple Storage (S3).
D. Publish messages to Amazon Kinesis Streams, pull messages off with Spark Streaming and write data new data to Amazon Simple Storage Service (S3) before and after processing

Answer: C

Question: 15

A company receives data sets coming from external providers on Amazon S3. Data sets from different providers are dependent on one another. Data sets will drive at different and is no particular order.
A data architect needs to design a solution that enables the company to do the following:
• Rapidly perform cross data set analysis as soon as the data becomes available
• Manage dependencies between data sets that arrives at different times
Which architecture strategy offers a scalable and cost-effective solution that meets these requirements?

A. Maintain data dependency information in Amazon RDS for MySQL. Use an AWS Pipeline job to load an Amazon EMR Hive Table based on task dependencies and event notification triggers in Amazon S3
B. Maintain data dependency information in an Amazon DynamoDB table. Use Amazon SNS and event notification to publish data to a fleet of Amazon EC2 workers. Once the task dependencies have been resolved process the data with Amazon EMR
C. Maintain data dependency information in an Amazon ElasticCache Redis cluster. Use Amazon S3 event notifications to trigger an AWS Lambda function that maps the S3 object to Redis. Once the dependencies have been resolved process the data with Amazon EMR
D. Maintain data dependency information in an Amazon DynamoDB table. Use Amazon S3 event notifications to trigger an AWS Lambda function that maps the S3 object to the task associated with it in DynamoDB. Once all task dependencies have been resolved process the data with Amazon EMR

Answer: C

For More Information – Visit link below:

http://www.certsgrade.com/

PRODUCT FEATURES

- 100% Money Back Guarantee
- 90 Days Free updates
- Special Discounts on Bulk Orders
- Guaranteed Success
- 50,000 Satisfied Customers
- 100% Secure Shopping
- Privacy Policy
- Refund Policy

Discount Coupon Code: CERTSGRADE10