

Question #:1

A Data Scientist needs to create a serverless ingestion and analytics solution for high-velocity, real-time streaming data.

The ingestion process must buffer and convert incoming records from JSON to a query-optimized, columnar format without data loss. The output datastore must be highly available, and Analysts must be able to run SQL queries against the data and connect to existing business intelligence dashboards.

Which solution should the Data Scientist build to satisfy the requirements?

- A. Create a schema in the AWS Glue Data Catalog of the incoming data format. Use an Amazon Kinesis Data Firehose delivery stream to stream the data and transform the data to Apache Parquet or ORC format using the AWS Glue Data Catalog before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.
- B. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and writes the data to a processed data location in Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena, and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.
- C. Write each JSON record to a staging location in Amazon S3. Use the S3 Put event to trigger an AWS Lambda function that transforms the data into Apache Parquet or ORC format and inserts it into an Amazon RDS PostgreSQL database. Have the Analysts query and run dashboards from the RDS database.
- D. Use Amazon Kinesis Data Analytics to ingest the streaming data and perform real-time SQL queries to convert the records to Apache Parquet before delivering to Amazon S3. Have the Analysts query the data directly from Amazon S3 using Amazon Athena and connect to BI tools using the Athena Java Database Connectivity (JDBC) connector.

Answer: A

Question #:2

A Machine Learning Specialist is using Amazon SageMaker to host a model for a highly available customer-facing application .

The Specialist has trained a new version of the model, validated it with historical data, and now wants to deploy it to production To limit any risk of a negative customer experience, the Specialist wants to be able to monitor the model and roll it back, if needed

What is the SIMPLEST approach with the LEAST risk to deploy the model and roll it back, if needed?

- A. Create a SageMaker endpoint and configuration for the new model version. Redirect production traffic to the new endpoint by updating the client configuration. Revert traffic to the last version if the model does not perform as expected.

- B. Create a SageMaker endpoint and configuration for the new model version. Redirect production traffic to the new endpoint by using a load balancer. Revert traffic to the last version if the model does not perform as expected.
- C. Update the existing SageMaker endpoint to use a new configuration that is weighted to send 5% of the traffic to the new variant. Revert traffic to the last version by resetting the weights if the model does not perform as expected.
- D. Update the existing SageMaker endpoint to use a new configuration that is weighted to send 100% of the traffic to the new variant. Revert traffic to the last version by resetting the weights if the model does not perform as expected.

Answer: A

Question #:3

A bank's Machine Learning team is developing an approach for credit card fraud detection. The company has a large dataset of historical data labeled as fraudulent. The goal is to build a model to take the information from new transactions and predict whether each transaction is fraudulent or not.

Which built-in Amazon SageMaker machine learning algorithm should be used for modeling this problem?

- A. Seq2seq
- B. XGBoost
- C. K-means
- D. Random Cut Forest (RCF)

Answer: C

Question #:4

A Machine Learning Specialist is configuring Amazon SageMaker so multiple Data Scientists can access notebooks, train models, and deploy endpoints. To ensure the best operational performance, the Specialist needs to be able to track how often the Scientists are deploying models, GPU and CPU utilization on the deployed SageMaker endpoints, and all errors that are generated when an endpoint is invoked.

Which services are integrated with Amazon SageMaker to track this information? (Select TWO.)

- A. AWS CloudTrail
- B. AWS Health
- C. AWS Trusted Advisor
- D. Amazon CloudWatch

E. AWS Config

Answer: A D

Question #:5

A Machine Learning Specialist works for a credit card processing company and needs to predict which transactions may be fraudulent in near-real time. Specifically, the Specialist must train a model that returns the probability that a given transaction may be fraudulent

How should the Specialist frame this business problem'?

- A. Streaming classification
- B. Binary classification
- C. Multi-category classification
- D. Regression classification

Answer: A

Question #:6

An interactive online dictionary wants to add a widget that displays words used in similar contexts. A Machine Learning Specialist is asked to provide word features for the downstream nearest neighbor model powering the widget.

What should the Specialist do to meet these requirements?

- A. Create one-hot word encoding vectors.
- B. Produce a set of synonyms for every word using Amazon Mechanical Turk.
- C. Create word embedding factors that store edit distance with every other word.
- D. Download word embedding's pre-trained on a large corpus.

Answer: A

Question #:7

A manufacturing company has a large set of labeled historical sales data The manufacturer would like to predict how many units of a particular part should be produced each quarter Which machine learning approach should be used to solve this problem?

- A. Logistic regression
- B. Random Cut Forest (RCF)
- C. Principal component analysis (PCA)

D. Linear regression

Answer: B

Question #:8

A Machine Learning Specialist needs to create a data repository to hold a large amount of time-based training data for a new model. In the source system, new files are added every hour. Throughout a single 24-hour period, the volume of hourly updates will change significantly. The Specialist always wants to train on the last 24 hours of the data.

Which type of data repository is the MOST cost-effective solution?

- A. An Amazon EBS-backed Amazon EC2 instance with hourly directories
- B. An Amazon RDS database with hourly table partitions
- C. An Amazon S3 data lake with hourly object prefixes
- D. An Amazon EMR cluster with hourly hive partitions on Amazon EBS volumes

Answer: C

Question #:9

A financial services company is building a robust serverless data lake on Amazon S3. The data lake should be flexible and meet the following requirements:

- * Support querying old and new data on Amazon S3 through Amazon Athena and Amazon Redshift Spectrum.
- * Support event-driven ETL pipelines.
- * Provide a quick and easy way to understand metadata.

Which approach meets these requirements?

- A. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Glue ETL job, and an AWS Glue Data catalog to search and discover metadata.
- B. Use an AWS Glue crawler to crawl S3 data, an AWS Lambda function to trigger an AWS Batch job, and an external Apache Hive metastore to search and discover metadata.
- C. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Batch job, and an AWS Glue Data Catalog to search and discover metadata.
- D. Use an AWS Glue crawler to crawl S3 data, an Amazon CloudWatch alarm to trigger an AWS Glue ETL job, and an external Apache Hive metastore to search and discover metadata.

Answer: B