AWS Certified Big Data - Specialty.VCEplus.premium.exam.65q

Number: AWSBigData
Passing Score: 800
Time Limit: 120 min
File Version: 1.1

AWS Certified Big Data - Specialty

Version 1.1

Exam A

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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

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.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

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.

Correct Answer: ACD
Section: (none)
Explanation

Explanation/Reference:

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.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

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.

Correct Answer: BD
Section: (none)
Explanation

Explanation/Reference:

QUESTION 7
A large grocery distributor receives daily depletion reports from the field in the form of gzip archives od CSV files uploaded to Amazon S3. The files range from 500MB to 5GB. These files are processed daily by an EMR job.

Recently it has been observed that the file sizes vary, and the EMR jobs take too long. The distributor needs to tune and optimize the data processing workflow with this limited information to improve the performance of the EMR job.

Which recommendation should an administrator provide?

A. Reduce the HDFS block size to increase the number of task processors.
B. Use bzip2 or Snappy rather than gzip for the archives.
C. Decompress the gzip archives and store the data as CSV files.
D. Use Avro rather than gzip for the archives.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 8
A web-hosting company is building a web analytics tool to capture clickstream data from all of the websites hosted within its platform and to provide near-real-time business intelligence. This entire system is built on AWS services. The webhosting company is interested in using Amazon Kinesis to collect this data and perform sliding window analytics.

What is the most reliable and fault-tolerant technique to get each website to send data to Amazon Kinesis with every click?

A. After receiving a request, each web server sends it to Amazon Kinesis using the Amazon Kinesis PutRecord API. Use the sessionID as a partition key and set up a loop to retry until a success response is received.
B. After receiving a request, each web server sends it to Amazon Kinesis using the Amazon Kinesis Producer Library .addRecords method.
C. Each web server buffers the requests until the count reaches 500 and sends them to Amazon Kinesis using the Amazon Kinesis PutRecord API.
D. After receiving a request, each web server sends it to Amazon Kinesis using the Amazon Kinesis PutRecord API. Use the exponential back-off algorithm for retries until a successful response is received.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 9
A customer has an Amazon S3 bucket. Objects are uploaded simultaneously by a cluster of servers from multiple streams of data. The customer maintains a catalog of objects uploaded in Amazon S3 using an Amazon DynamoDB table. This catalog has the following fileds: StreamName, TimeStamp, and ServerName, from which ObjectName can be obtained.

The customer needs to define the catalog to support querying for a given stream or server within a defined time range.

Which DynamoDB table scheme is most efficient to support these queries?

A. Define a Primary Key with ServerName as Partition Key and TimeStamp as Sort Key. Do NOT define a Local Secondary Index or Global Secondary Index.
B. Define a Primary Key with StreamName as Partition Key and TimeStamp followed by ServerName as Sort Key. Define a Global Secondary Index with ServerName as partition key and TimeStamp followed by StreamName.
C. Define a Primary Key with ServerName as Partition Key. Define a Local Secondary Index with StreamName as Partition Key. Define a Global Secondary Index with TimeStamp as Partition Key.
D. Define a Primary Key with ServerName as Partition Key. Define a Local Secondary Index with TimeStamp as Partition Key. Define a Global Secondary Index with StreamName as Partition Key and TimeStamp as Sort Key.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 10
A company has several teams of analysts. 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

Correct Answer: A
Section: (none)

Explanation
Explanation/Reference:

QUESTION 11
An administrator needs to manage a large catalog of items from various external sellers. The administrator needs to determine if the items should be identified as minimally dangerous, dangerous, or highly dangerous based on their textual descriptions. The administrator already has some items with the danger attribute, but receives hundreds of new item descriptions every day without such classification.

The administrator has a system that captures dangerous goods reports from customer support team of from user feedback.

What is a cost-effective architecture to solve this issue?

A. Build a set of regular expression rules that are based on the existing examples, and run them on the DynamoDB Streams as every new item description is added to the system.
B. Build a Kinesis Streams process that captures and marks the relevant items in the dangerous goods reports using a Lambda function once more than two reports have been filed.
C. Build a machine learning model to properly classify dangerous goods and run it on the DynamoDB Streams as every new item description is added to the system.
D. Build a machine learning model with binary classification for dangerous goods and run it on the DynamoDB Streams as every new item description is added to the system.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 12
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 arrive at different times and in 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 arrive 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 Data 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 notifications to publish data to fleet of Amazon EC2 workers. Once the task dependencies have been resolved, process the data withAmazon EMR.
C. Maintain data dependency information in an Amazon ElastiCache Redis cluster. Use Amazon S3 event notifications to trigger an AWS Lambda function that maps the S3 object to Redis. Once the task 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 taskdependencies have been resolved, process the data with Amazon EMR.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 13
A media advertising company handles a large number of real-time messages sourced from over 200 websites in real time. Processing latency must be kept low. Based on calculations, a 60-shard Amazon Kinesis stream is more than sufficient to handle the maximum data throughput, even with traffic spikes. The company also uses an Amazon Kinesis Client Library (KCL) application running on Amazon Elastic Compute Cloud (EC2) managed by an Auto Scaling group. Amazon CloudWatch indicates an average of 25% CPU and a modest level of network traffic across all running servers.

The company reports a 150% to 200% increase in latency of processing messages from Amazon Kinesis during peak times. There are NO reports of delay from the sites publishing to Amazon Kinesis.

What is the appropriate solution to address the latency?

A. Increase the number of shards in the Amazon Kinesis stream to 80 for greater concurrency.
B. Increase the size of the Amazon EC2 instances to increase network throughput.

C. Increase the minimum number of instances in the Auto Scaling group.
D. Increase Amazon DynamoDB throughput on the checkpoint table.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 14
A Redshift data warehouse has different user teams that need to query the same table with very different query types. These user teams are experiencing poor performance.

Which action improves performance for the user teams in this situation?

A. Create custom table views.
B. Add interleaved sort keys per team.
C. Maintain team-specific copies of the table.
D. Add support for workload management queue hopping.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/redshift/latest/dg/cm-c-implementing-workload-management.html

QUESTION 15
A company operates an international business served from a single AWS region. The company wants to expand into a new country. The regulator for that country requires the Data Architect to maintain a log of financial transactions in the country within 24 hours of the product transaction. The production application is latency insensitive. The new country contains another AWS region.

What is the most cost-effective way to meet this requirement?

A. Use CloudFormation to replicate the production application to the new region.
B. Use Amazon CloudFront to serve application content locally in the country; Amazon CloudFront logs will satisfy the requirement.
C. Continue to serve customers from the existing region while using Amazon Kinesis to stream transaction data to the regulator.
D. Use Amazon S3 cross-region replication to copy and persist production transaction logs to a bucket in the new country's region.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 16 An administrator needs to design the event log storage architecture for events from mobile devices. The event data will be processed by an Amazon EMR cluster daily for aggregated reporting and analytics before being archived.

How should the administrator recommend storing the log data?

A. Create an Amazon S3 bucket and write log data into folders by device. Execute the EMR job on the device folders.
B. Create an Amazon DynamoDB table partitioned on the device and sorted on date, write log data to table. Execute the EMR job on the Amazon DynamoDB table.
C. Create an Amazon S3 bucket and write data into folders by day. Execute the EMR job on the daily folder.
D. Create an Amazon DynamoDB table partitioned on EventID, write log data to table. Execute the EMR job on the table.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 17

A data engineer wants to use an Amazon Elastic Map Reduce for an application. The data engineer needs to make sure it complies with regulatory requirements. The auditor must be able to confirm at any point which servers are running and which network access controls are deployed.

Which action should the data engineer take to meet this requirement?

A. Provide the auditor IAM accounts with the SecurityAudit policy attached to their group.
B. Provide the auditor with SSH keys for access to the Amazon EMR cluster.
C. Provide the auditor with CloudFormation templates.
D. Provide the auditor with access to AWS DirectConnect to use their existing tools.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 18

A social media customer has data from different data sources including RDS running MySQL, Redshift, and Hive on EMR. To support better analysis, the customer needs to be able to analyze data from different data sources and to combine the results.

What is the most cost-effective solution to meet these requirements?

A. Load all data from a different database/warehouse to S3. Use Redshift COPY command to copy data to Redshift for analysis.
B. Install Presto on the EMR cluster where Hive sits. Configure MySQL and PostgreSQL connector to select from different data sources in a single query.
C. Spin up an Elasticsearch cluster. Load data from all three data sources and use Kibana to analyze.
D. Write a program running on a separate EC2 instance to run queries to three different systems. Aggregate the results after getting the responses from all three systems.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 19

An Amazon EMR cluster using EMRFS has access to petabytes of data on Amazon S3, originating from multiple unique data sources. The customer needs to query common fields across some of the data sets to be able to perform interactive joins and then display results quickly.

Which technology is most appropriate to enable this capability?

A. Presto
B. MicroStrategy
C. Pig
D. R Studio

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 20

A game company needs to properly scale its game application, which is backed by DynamoDB. Amazon Redshift has the past two years of historical data. Game traffic varies throughout the year based on various factors such as season, movie release, and holiday season. An administrator needs to calculate how much read and write throughput should be provisioned for DynamoDB table for each week in advance.

How should the administrator accomplish this task?

A. Feed the data into Amazon Machine Learning and build a regression model.
B. Feed the data into Spark Mlib and build a random forest modest.
C. Feed the data into Apache Mahout and build a multi-classification model.
D. Feed the data into Amazon Machine Learning and build a binary classification model.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 21
A data engineer is about to perform a major upgrade to the DDL contained within an Amazon Redshift cluster to support a new data warehouse application. The upgrade scripts will include user permission updates, view and table structure changes as well as additional loading and data manipulation tasks.

The data engineer must be able to restore the database to its existing state in the event of issues.

Which action should be taken prior to performing this upgrade task?

A. Run an UNLOAD command for all data in the warehouse and save it to S3.
B. Create a manual snapshot of the Amazon Redshift cluster.
C. Make a copy of the automated snapshot on the Amazon Redshift cluster.
D. Call the waitForSnapshotAvailable command from either the AWS CLI or an AWS SDK.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-snapshots.html#working-with-snapshot-restore-table-from-snapshot

QUESTION 22
A large oil and gas company needs to provide near real-time alerts when peak thresholds are exceeded in its pipeline system. The company has developed a system to capture pipeline metrics such as flow rate, pressure, and temperature using millions of sensors. The sensors deliver to AWS IoT.

What is a cost-effective way to provide near real-time alerts on the pipeline metrics?

A. Create an AWS IoT rule to generate an Amazon SNS notification.
B. Store the data points in an Amazon DynamoDB table and poll if for peak metrics data from an Amazon EC2 application.
C. Create an Amazon Machine Learning model and invoke it with AWS Lambda.
D. Use Amazon Kinesis Streams and a KCL-based application deployed on AWS Elastic Beanstalk.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 23
A company is using Amazon Machine Learning as part of a medical software application. The application will predict the most likely blood type for a patient based on a variety of other clinical tests that are available when blood type knowledge is unavailable.

What is the appropriate model choice and target attribute combination for this problem?
A. Multi-class classification model with a categorical target attribute.

B. Regression model with a numeric target attribute.
C. Binary Classification with a categorical target attribute.
D. K-Nearest Neighbors model with a multi-class target attribute.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 24
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 customers 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.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 25 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 function, which makes it undecipherable 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.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/firehose/latest/dev/encryption.html

QUESTION 26
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 upload 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 images 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 the images output to Amazon S3 and metadata to app repository DB.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 27
A customer needs to determine the optimal distribution strategy for the ORDERS fact table in its Redshift schema. 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 ORDERS table?

A. Identify the largest and most frequently joined dimension table and ensure that it and the ORDERS table both have EVEN distribution.
B. Identify the largest dimension table and designate the key of this dimension table as the distribution key of the ORDERS table.
C. Identify the smallest dimension table and designate the key of this dimension table as the distribution key of the ORDERS table.
D. Identify the largest and the most frequently joined dimension table and designate the key of this dimension table as the distribution key of the ORDERS table.

Correct Answer: D
Section: (none)
Explanation

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

QUESTION 28
A customer is collecting clickstream data using Amazon Kinesis and is grouping the events by IP address into 5-minute chunks stored in Amazon S3.

Many analysts in the company use Hive on Amazon EMR to analyze this data. Their queries always reference a single IP address. Data must be optimized for querying based on IP address using Hive running on Amazon EMR.

What is the most efficient method to query the data with Hive?

A. Store an index of the files by IP address in the Amazon DynamoDB metadata store for EMRFS.
B. Store the Amazon S3 objects with the following naming scheme: bucket_name/source=ip_address/year=yy/month=mm/day=dd/hour=hh/filename.
C. Store the data in an HBase table with the IP address as the row key.
D. Store the events for an IP address as a single file in Amazon S3 and add metadata with keys: Hive_Partitioned_IPAddress.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 29
An online retailer is using Amazon DynamoDB to store data related to customer transactions. The items in the table contains several string attributes describing the transaction as well as a JSON attribute containing the shopping cart and other details corresponding to the transaction. Average item size is – 250KB, most of which is associated with the JSON attribute. The average customer generates – 3GB of data per month.

Customers access the table to display their transaction history and review transaction details as needed. Ninety percent of the queries against the table are executed when building the transaction history view, with the other 10% retrieving transaction details. The table is partitioned on CustomerID and sorted on transaction date.

The client has very high read capacity provisioned for the table and experiences very even utilization, but complains about the cost of Amazon DynamoDB compared to other NoSQL solutions.

Which strategy will reduce the cost associated with the client's read queries while not degrading quality?
A. Modify all database calls to use eventually consistent reads and advise customers that transaction history may be one second out-of-date.
B. Change the primary table to partition on TransactionID, create a GSI partitioned on customer and sorted on date, project small attributes into GSI, and then query GSI for summary data and the primary table for JSON details.

C. Vertically partition the table, store base attributes on the primary table, and create a foreign key reference to a secondary table containing the JSON data. Query the primary table for summary data and the secondary table for JSONdetails.
D. Create an LSI sorted on date, project the JSON attribute into the index, and then query the primary table for summary data and the LSI for JSON details.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 30
A company that manufactures and sells smart air conditioning units also offers add-on services so that customers can see real-time dashboards in a mobile application or a web browser. Each unit sends its sensor information in JSON format every two seconds for processing and analysis. The company also needs to consume this data to predict possible equipment problems before they occur. A few thousand pre-purchased units will be delivered in the next couple of months. The company expects high market growth in the next year and needs to handle a massive amount of data and scale without interruption.

Which ingestion solution should the company use?

A. Write sensor data records to Amazon Kinesis Streams. Process the data using KCL applications for the end-consumer dashboard and anomaly detection workflows.
B. Batch sensor data to Amazon Simple Storage Service (S3) every 15 minutes. Flow the data downstream to the end-consumer dashboard and to the anomaly detection application.
C. Write sensor data records to Amazon Kinesis Firehose with Amazon Simple Storage Service (S3) as the destination. Consume the data with a KCL application for the end-consumer dashboard and anomaly detection.
D. Write sensor data records to Amazon Relational Database Service (RDS). Build both the end-consumer dashboard and anomaly detection application on top of Amazon RDS.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 31
An organization needs a data store to handle the following data types and access patterns: ▪
Faceting
▪ Search
▪ Flexible schema (JSON) and fixed schema ▪
Noise word elimination

Which data store should the organization choose?

A. Amazon Relational Database Service (RDS)
B. Amazon Redshift
C. Amazon DynamoDB
D. Amazon Elasticsearch Service

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 32
A travel website needs to present a graphical quantitative summary of its daily bookings to website visitors for marketing purposes. The website has millions of visitors per day, but wants to control costs by implementing the least-expensive solution for this visualization.

What is the most cost-effective solution?
A. Generate a static graph with a transient EMR cluster daily, and store it an Amazon S3.
B. Generate a graph using MicroStrategy backed by a transient EMR cluster.
C. Implement a Jupyter front-end provided by a continuously running EMR cluster leveraging spot instances for task nodes.

D. Implement a Zeppelin application that runs on a long-running EMR cluster.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 33
A system engineer for a company proposes digitalization and backup of large archives for customers. The systems engineer needs to provide users with a secure storage that makes sure that data will never be tampered with once it has been uploaded.

How should this be accomplished?

A. Create an Amazon Glacier Vault. Specify a "Deny" Vault Lock policy on this Vault to block "glacier:DeleteArchive".
B. Create an Amazon S3 bucket. Specify a "Deny" bucket policy on this bucket to block "s3:DeleteObject".
C. Create an Amazon Glacier Vault. Specify a "Deny" vault access policy on this Vault to block "glacier:DeleteArchive".
D. Create secondary AWS Account containing an Amazon S3 bucket. Grant "s3:PutObject" to the primary account.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/amazonglacier/latest/dev/vault-lock-policy.html

QUESTION 34
An organization needs to design and deploy a large-scale data storage solution that will be highly durable and highly flexible with respect to the type and structure of data being stored. The data to be stored will be sent or generated from a variety of sources and must be persistently available for access and processing by multiple applications.

What is the most cost-effective technique to meet these requirements?

A. Use Amazon Simple Storage Service (S3) as the actual data storage system, coupled with appropriate tools for ingestion/acquisition of data and for subsequent processing and querying.
B. Deploy a long-running Amazon Elastic MapReduce (EMR) cluster with Amazon Elastic Block Store (EBS) volumes for persistent HDFS storage and appropriate Hadoop ecosystem tools for processing and querying.
C. Use Amazon Redshift with data replication to Amazon Simple Storage Service (S3) for comprehensive durable data storage, processing, and querying.
D. Launch an Amazon Relational Database Service (RDS), and use the enterprise grade and capacity of the Amazon Aurora engine for storage, processing, and querying.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 35
A customer has a machine learning workflow that consists of multiple quick cycles of reads-writes-reads on Amazon S3. The customer needs to run the workflow on EMR but is concerned that the reads in subsequent cycles will miss new data critical to the machine learning from the prior cycles.

How should the customer accomplish this?

A. Turn on EMRFS consistent view when configuring the EMR cluster.
B. Use AWS Data Pipeline to orchestrate the data processing cycles.
C. Set `hadoop.data.consistency = true` in the `core-site.xml` file.

D. Set `hadoop.s3.consistency = true` in the `core-site.xml` file.

Correct Answer: A

Section: (none)
Explanation

Explanation/Reference:

QUESTION 36
An Amazon Redshift Database is encrypted using KMS. A data engineer needs to use the AWS CLI to create a KMS encrypted snapshot of the database in another AWS region. Which

three steps should the data engineer take to accomplish this task? (Choose three.)

A. Create a new KMS key in the destination region.
B. Copy the existing KMS key to the destination region.
C. Use CreateSnapshotCopyGrant to allow Amazon Redshift to use the KMS key from the source region.
D. In the source region, enable cross-region replication and specify the name of the copy grant created.
E. In the destination region, enable cross-region replication and specify the name of the copy grant created.

Correct Answer: ABD
Section: (none)
Explanation

Explanation/Reference:
Reference: https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-db-encryption.html#working-with-aws-kms

QUESTION 37
Managers in a company need access to the human resources database that runs on Amazon Redshift, to run reports about their employees. Managers must only see information about their direct reports.

Which technique should be used to address this requirement with Amazon Redshift?

A. Define an IAM group for each manager with each employee as an IAM user in that group, and use that to limit the access.
B. Use Amazon Redshift snapshot to create one cluster per manager. Allow the manager to access only their designated clusters.
C. Define a key for each manager in AWS KMS and encrypt the data for their employees with their private keys.
D. Define a view that uses the employee's manager name to filter the records based on current user names.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 38
A company is building a new application in AWS. The architect needs to design a system to collect application log events. The design should be a repeatable pattern that minimizes data loss if an application instance fails, and keeps a durable copy of a log data for at least 30 days.

What is the simplest architecture that will allow the architect to analyze the logs?

A. Write them directly to a Kinesis Firehose. Configure Kinesis Firehose to load the events into an Amazon Redshift cluster for analysis.
B. Write them to a file on Amazon Simple Storage Service (S3). Write an AWS Lambda function that runs in response to the S3 event to load the events into Amazon Elasticsearch Service for analysis.
C. Write them to the local disk and configure the Amazon CloudWatch Logs agent to load the data into CloudWatch Logs and subsequently into Amazon Elasticsearch Service.
D. Write them to CloudWatch Logs and use an AWS Lambda function to load them into HDFS on an Amazon Elastic MapReduce (EMR) cluster for analysis.

Correct Answer: B
Section: (none)
Explanation
Explanation/Reference:

QUESTION 39
An organization uses a custom map reduce application to build monthly reports based on many small data files in an Amazon S3 bucket. The data is submitted from various business units on a frequent but unpredictable schedule. As the dataset continues to grow, it becomes increasingly difficult to process all of the data in one day. The organization has scaled up its Amazon EMR cluster, but other optimizations could improve performance.

The organization needs to improve performance with minimal changes to existing processes and applications.

What action should the organization take?

A. Use Amazon S3 Event Notifications and AWS Lambda to create a quick search file index in DynamoDB.
B. Add Spark to the Amazon EMR cluster and utilize Resilient Distributed Datasets in-memory.
C. Use Amazon S3 Event Notifications and AWS Lambda to index each file into an Amazon Elasticsearch Service cluster.
D. Schedule a daily AWS Data Pipeline process that aggregates content into larger files using S3DistCp.
E. Have business units submit data via Amazon Kinesis Firehose to aggregate data hourly into Amazon S3.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 40 An administrator is processing events in near real-time using Kinesis streams and Lambda. Lambda intermittently fails to process batches from one of the shards due to a 5-munite time limit.

What is a possible solution for this problem?

A. Add more Lambda functions to improve concurrent batch processing.
B. Reduce the batch size that Lambda is reading from the stream.
C. Ignore and skip events that are older than 5 minutes and put them to Dead Letter Queue (DLQ).
D. Configure Lambda to read from fewer shards in parallel.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 41 An organization uses Amazon Elastic MapReduce(EMR) to process a series of extract-transform-load (ETL) steps that run in sequence. The output of each step must be fully processed in subsequent steps but will not be retained.

Which of the following techniques will meet this requirement most efficiently?

A. Use the EMR File System (EMRFS) to store the outputs from each step as objects in Amazon Simple Storage Service (S3).
B. Use the s3n URI to store the data to be processed as objects in Amazon S3.
C. Define the ETL steps as separate AWS Data Pipeline activities.
D. Load the data to be processed into HDFS, and then write the final output to Amazon S3.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 42
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 identify 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 DynampDB for analysis.
B. Collect the sensor data with Amazon SQS and store in Amazon DynamoDB for analysis. Collect emergency services events with Amazon Kinesis Firehose and store in Amazon Redshift for analysis.
C. Collect both sensor data and emergency services events with Amazon Kinesis Streams and use DynamoDB for analysis.
D. Collect both sensor data and emergency services events with Amazon Kinesis Firehose and use Amazon Redshift for analysis.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 43
A telecommunications company needs to predict customer churn (i.e., customers who decide to switch to a competitor). The company has historic records of each customer, including monthly consumption patterns, calls to customer service, and whether the customer ultimately quit the service. 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 business insight. Plot the churn trend graph to extrapolate churn likelihood for existing customers.
C. Use EMR to run the Hive queries to build a profile of a churning customer. Apply a 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 Defined Function in Redshift that computes the likelihood of churn.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 44
A system needs to collect on-premises application spool files into a persistent storage layer in AWS. Each spool file is 2 KB. The application generates 1 M files per hour. Each source file is automatically deleted from the local server after an hour.

What is the most cost-efficient option to meet these requirements?

A. Write file contents to an Amazon DynamoDB table.
B. Copy files to Amazon S3 Standard Storage.
C. Write file contents to Amazon ElastiCache.
D. Copy files to Amazon S3 infrequent Access Storage.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 45
An administrator receives about 100 files per hour into Amazon S3 and will be loading the files into Amazon Redshift. Customers who analyze the data within Redshift gain significant value when they receive data as quickly as possible. The customers have agreed to a maximum loading interval of 5 minutes.
Which loading approach should the administrator use to meet this objective?

A. Load each file as it arrives because getting data into the cluster as quickly as possibly is the priority.

B. Load the cluster as soon as the administrator has the same number of files as nodes in the cluster.
C. Load the cluster when the administrator has an event multiple of files relative to Cluster Slice Count, or 5 minutes, whichever comes first.
D. Load the cluster when the number of files is less than the Cluster Slice Count.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 46
An enterprise customer is migrating to Redshift and is considering using dense storage nodes in its Redshift cluster. The customer wants to migrate 50 TB of data. The customer's query patterns involve performing many joins with thousands of rows.

The customer needs to know how many nodes are needed in its target Redshift cluster. The customer has a limited budget and needs to avoid performing tests unless absolutely needed.

Which approach should this customer use?

A. Start with many small nodes.
B. Start with fewer large nodes.
C. Have two separate clusters with a mix of a small and large nodes.
D. Insist on performing multiple tests to determine the optimal configuration.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 47 A company is centralizing a large number of unencrypted small files from multiple Amazon S3 buckets. The company needs to verify that the files contain the same data after centralization.

Which method meets the requirements?

A. Compare the S3 Etags from the source and destination objects.
B. Call the S3 CompareObjects API for the source and destination objects.
C. Place a HEAD request against the source and destination objects comparing SIG v4.
D. Compare the size of the source and destination objects.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 48
An online gaming company uses DynamoDB to store user activity logs and is experiencing throttled writes on the company's DynamoDB table. The company is **NOT** consuming close to the provisioned capacity. The table contains a large number of items and is partitioned on user and sorted by date. The table is 200GB and is currently provisioned at 10K WCU and 20K RCU. Which two additional pieces of information are required to determine the cause of the throttling? (Choose two.)

A. The structure of any GSIs that have been defined on the table
B. CloudWatch data showing consumed and provisioned write capacity when writes are being throttled
C. Application-level metrics showing the average item size and peak update rates for each attribute
D. The structure of any LSIs that have been defined on the table

E. The maximum historical WCU and RCU for the table

Correct Answer: AD
Section: (none)
Explanation

Explanation/Reference:

QUESTION 49
A city has been collecting data on its public bicycle share program for the past three years. The 5PB dataset currently resides on Amazon S3. The data contains the following datapoints: ▪
Bicycle origination points
▪ Bicycle destination points
▪ Mileage between the points
▪ Number of bicycle slots available at the station (which is variable based on the station location) ▪
Number of slots available and taken at a given time

The program has received additional funds to increase the number of bicycle stations available. All data is regularly archived to Amazon Glacier.

The new bicycle stations must be located to provide the most riders access to bicycles.

How should this task be performed?

A. Move the data from Amazon S3 into Amazon EBS-backed volumes and use an EC-2 based Hadoop cluster with spot instances to run a Spark job that performs a stochastic gradient descent optimization.
B. Use the Amazon Redshift COPY command to move the data from Amazon S3 into Redshift and perform a SQL query that outputs the most popular bicycle stations.
C. Persist the data on Amazon S3 and use a transient EMR cluster with spot instances to run a Spark streaming job that will move the data into Amazon Kinesis.
D. Keep the data on Amazon S3 and use an Amazon EMR-based Hadoop cluster with spot instances to run a Spark job that performs a stochastic gradient descent optimization over EMRFS.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 50
An administrator tries to use the Amazon Machine Learning service to classify social media posts that mention the administrator's company into posts that require a response and posts that do not. The training dataset of 10,000 posts contains the details of each post including the timestamp, author, and full text of the post. The administrator is missing the target labels that are required for training.

Which Amazon Machine Learning model is the most appropriate for the task?

A. Binary classification model, where the target class is the require-response post
B. Binary classification model, where the two classes are the require-response post and does-not-require-response
C. Multi-class prediction model, with two classes: require-response post and does-not-require-responseD. Regression model where the predicted value is the probability that the post requires a response

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 51
A medical record filing system for a government medical fund is using an Amazon S3 bucket to archive documents related to patients. Every patient visit to a physician creates a new file, which can add up millions of files each month. Collection of these files from each physician is handled via a batch process that runs every night using AWS Data Pipeline. This is sensitive data, so the data and any associated metadata must be encrypted at rest.

Auditors review some files on a quarterly basis to see whether the records are maintained according to regulations. Auditors must be able to locate any physical file in the S3 bucket for a given date, patient, or physician. Auditors spend a significant amount of time location such files.

What is the most cost- and time-efficient collection methodology in this situation?

A. Use Amazon Kinesis to get the data feeds directly from physicians, batch them using a Spark application on Amazon Elastic MapReduce (EMR), and then store them in Amazon S3 with folders separated per physician.
B. Use Amazon API Gateway to get the data feeds directly from physicians, batch them using a Spark application on Amazon Elastic MapReduce (EMR), and then store them in Amazon S3 with folders separated per physician.
C. Use Amazon S3 event notification to populate an Amazon DynamoDB table with metadata about every file loaded to Amazon S3, and partition them based on the month and year of the file.
D. Use Amazon S3 event notification to populate an Amazon Redshift table with metadata about every file loaded to Amazon S3, and partition them based on the month and year of the file.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 52 A clinical trial will rely on medical sensors to remotely assess patient health. Each physician who participates in the trial requires visual reports each morning. The reports are built from aggregations of all the sensor data taken each minute.

What is the most cost-effective solution for creating this visualization each day?

A. Use Kinesis Aggregators Library to generate reports for reviewing the patient sensor data and generate a QuickSight visualization on the new data each morning for the physician to review.
B. Use a transient EMR cluster that shuts down after use to aggregate the patient sensor data each night and generate a QuickSight visualization on the new data each morning for the physician to review.
C. Use Spark streaming on EMR to aggregate the patient sensor data in every 15 minutes and generate a QuickSight visualization on the new data each morning for the physician to review.
D. Use an EMR cluster to aggregate the patient sensor data each night and provide Zeppelin notebooks that look at the new data residing on the cluster each morning for the physician to review.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 53
A company uses Amazon Redshift for its enterprise data warehouse. A new on-premises PostgreSQL OLTP DB must be integrated into the data warehouse. Each table in the PostgreSQL DB has an indexed *last_modified* timestamp column. The data warehouse has a staging layer to load source data into the data warehouse environment for further processing.

The data lag between the source PostgreSQL DB and the Amazon Redshift staging layer should NOT exceed four hours.

What is the most efficient technique to meet these requirements?

A. Create a DBLINK on the source DB to connect to Amazon Redshift. Use a PostgreSQL trigger on the source table to capture the new insert/update/delete event and execute the event on the Amazon Redshift staging table.
B. Use a PostgreSQL trigger on the source table to capture the new insert/update/delete event and write it to Amazon Kinesis Streams. Use a KCL application to execute the event on the Amazon Redshift staging table.
C. Extract the incremental changes periodically using a SQL query. Upload the changes to multiple Amazon Simple Storage Service (S3) objects, and run the COPY command to load to the Amazon Redshift staging layer.
D. Extract the incremental changes periodically using a SQL query. Upload the changes to a single Amazon Simple Storage Service (S3) object, and run the COPY command to load to the Amazon Redshift staging layer.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 54
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 typesD. Machine Learning on D instance types and ad-hoc queries on I instance types

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 55
An organization is designing an application architecture. The application will have over 100 TB of data and will support transactions that arrive at rates from hundreds per second to tens of thousands per second, depending on the day of the week and time of the day. All transaction data, must be durably and reliably stored. Certain read operations must be performed with strong consistency.

Which solution meets these requirements?

A. Use Amazon DynamoDB as the data store and use strongly consistent reads when necessary.
B. Use an Amazon Relational Database Service (RDS) instance sized to meet the maximum anticipated transaction rate and with the High Availability option enabled.
C. Deploy a NoSQL data store on top of an Amazon Elastic MapReduce (EMR) cluster, and select the HDFS High Durability option.
D. Use Amazon Redshift with synchronous replication to Amazon Simple Storage Service (S3) and row-level locking for strong consistency.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 56
A company generates a large number of files each month and needs to use AWS import/export to move these files into Amazon S3 storage. To satisfy the auditors, the company needs to keep a record of which files were imported into Amazon S3.

What is a low-cost way to create a unique log for each import job?

A. Use the same log file prefix in the import/export manifest files to create a versioned log file in Amazon S3 for all imports.
B. Use the log file prefix in the import/export manifest files to create a unique log file in Amazon S3 for each import.
C. Use the log file checksum in the import/export manifest files to create a unique log file in Amazon S3 for each import.
D. Use a script to iterate over files in Amazon S3 to generate a log after each import/export job.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 57
A company needs a churn prevention model to predict which customers will **NOT** renew 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 monthsD. Quarterly results

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 58
A company with a support organization needs support engineers to be able to search historic cases to provide fast responses on new issues raised. The company has forwarded all support messages into an Amazon Kinesis Stream. This meets a company objective of using only managed services to reduce operational overhead.

The company needs an appropriate architecture that allows support engineers to search on historic cases and find similar issues and their associated responses.

Which AWS Lambda action is most appropriate?

A. Ingest and index the content into an Amazon Elasticsearch domain.
B. Stem and tokenize the input and store the results into Amazon ElastiCache.
C. Write data as JSON into Amazon DynamoDB with primary and secondary indexes.
D. Aggregate feedback in Amazon S3 using a columnar format with partitioning.

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:

QUESTION 59
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 a 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 storing data. Use AWS KMS Grants to allow access to specific elements of the platform. Use AWS CloudTrail for auditing.
B. Use Amazon S3 Server-Side Encryption with Amazon S3-Managed Keys. Set Amazon S3 ACLs to allow access to specific elements of the platform. Use Amazon S3 to access logs for auditing.
C. Use Amazon S3 Client-Side Encryption with Client-Side Master Key. Set Amazon S3 ACLs to allow access to specific elements of the platform. Use Amazon S3 to access logs for auditing.
D. Use Amazon S3 Client-Side Encryption with AWS KMS-Managed Keys for storing data. Use AWS KMS Grants to allow access to specific elements of the platform. Use AWS CloudTrail for auditing.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 60 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 support for this requirements?

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

Correct Answer: A
Section: (none)
Explanation

Explanation/Reference:
Reference: https://sa.udacity.com/course/data-visualization-and-d3js--ud507

QUESTION 61
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 capture 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 build a single model in Oregon.
C. Use Kinesis as a buffer for web logs and replicate logs to the Kinesis stream of a neighboring region.
D. Use the CloudWatch Logs agent to consolidate logs into a single CloudWatch Logs group.

Correct Answer: D
Section: (none)
Explanation

Explanation/Reference:

QUESTION 62
There are thousands of text files on Amazon S3. The total size of the files is 1 PB. The files contain retail order information for the past 2 years. A data engineer needs to run multiple interactive queries to manipulate the data. The Data Engineer has AWS access to spin up an Amazon EMR cluster. The data engineer needs to use an application on the cluster to process this data and return the results in interactive time frame.

Which application on the cluster should the data engineer use?

A. Oozie
B. Apache Pig with Tachyon
C. Apache Hive
D. Presto

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:

QUESTION 63
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 Streams with Spark Streaming in parallel to AWS Lambda pushing messages from Streams to Firehose backed by Amazon Simple Storage Service (S3).
C. Publish messages to Amazon Kinesis Firehose backed by Amazon Simple Storage Service (S3). Use AWS Lambda to pull messages from Firehose to Streams for processing with Spark Streaming.
D. Publish messages to Amazon Kinesis Streams, pull messages off with Spark Streaming, and write row data to Amazon Simple Storage Service (S3) before and after processing.

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference:
QUESTION 64
A solutions architect for a logistics organization ships packages from thousands of suppliers to end customers. The architect is building a platform where suppliers can view the status of one or more of their shipments. Each supplier can have multiple roles that will only allow access to specific fields in the resulting information.

Which strategy allows the appropriate level of access control and requires the LEAST amount of management work?

A. Send the tracking data to Amazon Kinesis Streams. Use AWS Lambda to store the data in an Amazon DynamoDB Table. Generate temporary AWS credentials for the suppliers' users with AWS STS, specifying fine-grained securitypolicies to limit access only to their applicable data.
B. Send the tracking data to Amazon Kinesis Firehose. Use Amazon S3 notifications and AWS Lambda to prepare files in Amazon S3 with appropriate data for each supplier's roles. Generate temporary AWS credentials for the suppliers'users with AWS STS. Limit access to the appropriate files through security policies.
C. Send the tracking data to Amazon Kinesis Streams. Use Amazon EMR with Spark Streaming to store the data in HBase. Create one table per supplier. Use HBase Kerberos integration with the suppliers' users. Use HBase ACL-basedsecurity to limit access for the roles to their specific table and columns.
D. Send the tracking data to Amazon Kinesis Firehose. Store the data in an Amazon Redshift cluster. Create views for the suppliers' users and roles. Allow suppliers access to the Amazon Redshift cluster using a user limited to theapplicable view.

Correct Answer: B
Section: (none)
Explanation

Explanation/Reference:

QUESTION 65
A company's social media manager requests more staff on the weekends to handle an increase in customer contacts from a particular region. The company needs a report to visualize the trends on weekends over the past 6 months using QuickSight.

How should the data be represented?

A. A line graph plotting customer contacts vs. time, with a line for each region
B. A pie chart per region plotting customer contacts per day of week
C. A map of regions with a heatmap overlay to show the volume of customer contactsD. A bar graph plotting region vs. volume of social media contacts

Correct Answer: C
Section: (none)
Explanation

Explanation/Reference: