1) A company needs to deploy a data lake solution for their data scientists in which all company data is accessible and stored in a central S3 bucket. The company segregates the data by business unit, using specific prefixes. Scientists can only access the data from their own business unit. The company needs a single sign-on identity and management solution based on Microsoft Active Directory (AD) to manage access to the data in Amazon S3.

Which method meets these requirements?

A) Use AWS IAM Federation functions and specify the associated role based on the users' groups in AD.
B) Create bucket policies that only allow access to the authorized prefixes based on the users' group name in Active Directory.
C) Deploy the AD Synchronization service to create AWS IAM users and groups based on AD information.
D) Use Amazon S3 API integration with AD to impersonate the users on access in a transparent manner.

2) An administrator has a 500-GB file in Amazon S3. The administrator runs a nightly COPY command into a 10-node Amazon Redshift cluster. The administrator wants to prepare the data to optimize performance of the COPY command.

How should the administrator prepare the data?

A) Compress the file using gz compression.
B) Split the file into 500 smaller files.
C) Convert the file format to AVRO.
D) Split the file into 10 files of equal size.

3) A customer needs to load a 550-GB data file into an Amazon Redshift cluster from Amazon S3, using the COPY command. The input file has both known and unknown issues that will probably cause the load process to fail. The customer needs the most efficient way to detect load errors without performing any cleanup if the load process fails.

Which technique should the customer use?

A) Split the input file into 50-GB blocks and load them separately.
B) Use COPY with NOLOAD parameter.
C) Write a script to delete the data from the tables in case of errors.
D) Compress the input file before running COPY.

4) An organization needs a data store to handle the following data types and access patterns:

- **Key-value access pattern**
- **Complex SQL queries and transactions**
- **Consistent reads**
- **Fixed schema**

Which data store should the organization choose?

A) Amazon S3
B) Amazon Kinesis
C) Amazon DynamoDB
D) Amazon RDS

5) A web application emits multiple types of events to Amazon Kinesis Streams for operational reporting. Critical events must be captured immediately before processing can continue, but informational events do not need to delay processing.

What is the most appropriate solution to record these different types of events?

A) Log all events using the Kinesis Producer Library.
B) Log critical events using the Kinesis Producer Library, and log informational events using the PutRecords API method.
C) Log critical events using the PutRecords API method, and log informational events using the Kinesis Producer Library.
D) Log all events using the PutRecords API method.

6) An administrator decides to use the Amazon Machine Learning service to classify social media posts that mention your company into two categories: 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. You are missing the target labels that are required for training.

Which two options will create valid target label data?

A) Ask the social media handling team to review each post and provide the label.
B) Use the sentiment analysis NLP library to determine whether a post requires a response.
C) Use the Amazon Mechanical Turk web service to publish Human Intelligence Tasks that ask Turk workers to label the posts.
D) Using the *a priori* probability distribution of the two classes, use Monte-Carlo simulation to generate the labels.

7) A mobile application collects data that must be stored in multiple Availability Zones within five minutes of being captured in the app.

What architecture securely meets these requirements?

A) The mobile app should write to an S3 bucket that allows anonymous PutObject calls.
B) The mobile app should authenticate with an Amazon Cognito identity that is authorized to write to an Amazon Kinesis Firehose with an Amazon S3 destination.
C) The mobile app should authenticate with an embedded IAM access key that is authorized to write to an Amazon Kinesis Firehose with an Amazon S3 destination.
D) The mobile app should call a REST-based service that stores data on Amazon EBS. Deploy the service on multiple EC2 instances across two Availability Zones.

8) A data engineer needs to collect data from multiple Amazon Redshift clusters within a business and consolidate the data into a single central data warehouse. Data must be encrypted at all times while at rest or in flight.

What is the most scalable way to build this data collection process?

A) Run an ETL process that connects to the source clusters using SSL to issue a SELECT query for new data, and then write to the target data warehouse using an INSERT command over another SSL secured connection.
B) Use AWS KMS data key to run an UNLOAD ENCRYPTED command that stores the data in an unencrypted S3 bucket; run a COPY command to move the data into the target cluster.
C) Run an UNLOAD command that stores the data in an S3 bucket encrypted with an AWS KMS data key; run a COPY command to move the data into the target cluster.
D) Connect to the source cluster over an SSL client connection, and write data records to Amazon Kinesis Firehose to load into your target data warehouse.

9) A company logs data from its application in large files and runs regular analytics of these logs to support internal reporting for three months after the logs are generated. After three months, the logs are infrequently accessed for up to a year. The company also has a regulatory control requirement to store application logs for seven years.

Which course of action should the company take to achieve these requirements in the most cost-efficient way?

A) Store the files in S3 Glacier with a Deny Delete vault lock policy for archives less than seven years old and a vault access policy that restricts read access to the analytics IAM group and write access to the log writer service role.
B) Store the files in S3 Standard with a lifecycle policy to transition the storage class to Standard - IA after three months. After a year, transition the files to Glacier and add a Deny Delete vault lock policy for archives less than seven years old.
C) Store the files in S3 Standard with lifecycle policies to transition the storage class to Standard – IA after three months and delete them after a year. Simultaneously store the files in Amazon Glacier with a Deny Delete vault lock policy for archives less than seven years old.
D) Store the files in S3 Standard with a lifecycle policy to remove them after a year. Simultaneously store the files in Amazon S3 Glacier with a Deny Delete vault lock policy for archives less than seven years old.

10) A data engineer needs to architect a data warehouse for an online retail company to store historic purchases. The data engineer needs to use Amazon Redshift. To comply with PCI:DSS and meet corporate data protection standards, the data engineer must ensure that data is encrypted at rest and that the keys are managed by a corporate on-premises HSM.

Which approach meets these requirements in the most cost-effective manner?

A) Create a VPC, and then establish a VPN connection between the VPC and the on-premises network. Launch the Amazon Redshift cluster in the VPC, and configure it to use your corporate HSM.
B) Use the AWS CloudHSM service to establish a trust relationship between the CloudHSM and the corporate HSM over a Direct Connect connection. Configure Amazon Redshift to use the CloudHSM device.
C) Configure the AWS Key Management Service to point to the corporate HSM device, and then launch the Amazon Redshift cluster with the KMS managing the encryption keys.
D) Use AWS Import/Export to import the corporate HSM device into the AWS Region where the Amazon Redshift cluster will launch, and configure Redshift to use the imported HSM.

Answers

1) A - Identity Federation allows organizations to associate temporary credentials to users authenticated through an external identity provider such as Microsoft Active Directory (AD). These temporary credentials are linked to AWS IAM roles that grant access to the S3 bucket. Option B does not work because bucket policies are linked to IAM principles and cannot recognize AD attributes. Option C does not work because AD Synchronization will not sync directly with AWS IAM, and custom synchronization would not result in Amazon S3 being able to see group information. D isn't possible because there is no feature to integrate Amazon S3 directly with external identity providers.

2) B - The critical aspect of this question is running the COPY command with the maximum amount of parallelism. The two options that will increase parallelism are B and D. Option D will load one file per node in parallel, which will increase performance, but option B will have a greater effect because it will allow Amazon Redshift to load multiple files per instance in parallel (COPY can process one file per slice on each node). Compressing the files (option A) is a recommended practice and will also increase performance, but not to the same extent as loading multiple files in parallel.

3) B - From the AWS Documentation for NOLOAD: *NOLOAD checks the integrity of all of the data without loading it into the database. The NOLOAD option displays any errors that would occur if you had attempted to load the data*. All other options will require subsequent processing on the cluster which will consume resources.

4) D - Amazon RDS handles all these requirements, and although Amazon RDS is not typically thought of as optimized for key-value based access, a schema with a good primary key selection can provide this functionality. Amazon S3 provides no fixed schema and does not have consistent read after PUT support. Amazon Kinesis supports streaming data that is consistent as of a given sequence number but doesn't provide key/value access. Finally, although Amazon DynamoDB provides key/value access and consistent reads, it does not support SQL-based queries.

5) C – The core of this question is how to send event messages to Kinesis synchronously vs. asynchronously. The critical events must be sent synchronously, and the informational events can be sent asynchronously. The Kinesis Producer Library (KPL) implements an asynchronous send function, so it can be used for the informational messages. PutRecords is a synchronous send function, so it must be used for the critical events.

6) A, C - You need accurate data to train the service and get accurate results from future data. The options described in B and D would end up training an ML model using the output from a different machine learning model and therefore would significantly increase the possible error rate. It is extremely important to have a very low error rate (if any!) in your training set, and therefore human-validated or assured labels are essential.

7) B – It is essential when writing mobile applications that you consider the security of both how the application authenticates and how it stores credentials. Option A uses an anonymous Put, which may allow other apps to write counterfeit data; Option B is the right answer, because using Amazon Cognito gives you the ability to securely authenticate pools of users on any type of device at scale. Option C would put credentials directly into the application, which is strongly discouraged because applications can be decompiled which can compromise the keys. Option D does not meet our availability requirements: although the EC2 instances are running in different Availability Zones, the EBS volumes attached to each instance only store data in a single Availability Zone.

8) B - The most scalable solutions are the UNLOAD/COPY solutions because they will work in parallel, which eliminates A and D as answers. Option C is incorrect because the data would not be encrypted in flight, and you cannot encrypt an entire bucket with a KMS key. Option B meets the encryption requirements, the UNLOAD ENCRYPTED command automatically stores the data encrypted using-client side encryption and uses HTTPS to encrypt the data during the transfer to S3.

9) C – There are two aspects to this question: setting up a lifecycle policy to ensure that objects are stored in the most cost-effective storage, and ensuring that the regulatory control is met. The lifecycle policy will store the objects on S3 Standard during the three months of active use, and then move the objects to S3 Standard – IA when access will be infrequent. That narrows the possible answer set to B and C. The Deny Delete vault lock policy will ensure that the regulatory policy is met, but that policy must be applied over the entire lifecycle of the object, not just after it is moved to Glacier after the first year. Option C has the Deny Delete vault lock applied over the entire lifecycle of the object and is the right answer.

10) A - Amazon Redshift can use an on-premises HSM for key management over the VPN, which ensures that the encryption keys are locally managed. Option B is possible: CloudHSM can cluster to an on-premises HSM. But then key management could be performed on either the on-premises HSM or CloudHSM, and that doesn't meet the design goal. Option C does not describe a valid feature of KMS and violates the requirement for the corporate HSM to manage the keys requirement, even if it were possible. Option D is not possible because you cannot put hardware into an AWS Region.