Answer Sheet

Q1)

You have recently joined a startup company building sensors to measure street noise and air quality in urban areas. The company has been running a pilot deployment of around 100 sensors for 3 months. Each sensor uploads 1KB of sensor data every minute to a backend hosted on AWS.

During the pilot, you measured a peak or 10 IOPS on the database, and you stored an average of 3GB of sensor data per month in the database. The current deployment consists of a load-balanced auto scaled Ingestion layer using EC2 instances and a PostgreSQL RDS database with 500GB standard storage. The pilot is considered a success and your CEO has managed to get the attention or some potential investors. The business plan requires a deployment of at least 100K sensors, which needs to be supported by the backend.

You also need to store sensor data for at least two years to be able to compare year over year Improvements. To secure funding, you have to make sure that the platform meets these requirements and leaves room for further scaling.

Which setup will meet the requirements?

- Keep the current architecture but upgrade RDS storage to 3TB and 10K provisioned IOPS
- Replace the RDS instance with a 6 node Redshift cluster with 96TB of storage
- ✅ Ingest data into a DynamoDB table and move old data to a Redshift cluster
- Add an SQS queue to the ingestion layer to buffer writes to the RDS instance

---

Q2)

You have two different groups using Redshift to analyze data of a petabyte-scale data warehouse. Each query issued by the first group takes approximately 1-2 hours to analyze the data while the second group's queries only take between 5-10 minutes to analyze data.

You don't want the second group's queries to wait until the first group's queries are finished. You need to design a solution so that this does not happen.

Which of the following would be the best and cheapest solution to deploy to solve this dilemma?

- Start another Redshift cluster from a snapshot for the second team if the current Redshift cluster is busy processing long queries.
- Pause the long queries when necessary and resume them when there are no queries happening.
- ✅ Create two separate workload management groups and assign them to the respective groups.
- Create a read replica of Redshift and run the second team's queries on the read replica.

Q3)

ABCD has developed a sensor intended to be placed inside of people's shoes, monitoring the number of steps taken every day. ABCD is expecting thousands of sensors reporting in every minute and hopes to scale to millions by the end of the year.

A requirement for the project is it needs to be able to accept the data, run it through ETL to store in warehouse and archive it on Amazon Glacier, with room for a real-time dashboard for the sensor data to be added at a later date.

What is the best method for architecting this application given the requirements? Choose the correct answer:

- ✅ Write the sensor data directly to Amazon Kinesis and output the data into Amazon S3 creating a lifecycle policy for Glacier archiving. Also, have a parallel processing application that runs the data through EMR and sends to a Redshift data warehous
- Write the sensor data to Amazon S3 with a lifecycle policy for Glacier, create an EMR cluster that uses the bucket data and runs it through ETL. It then outputs that data into Redshift data warehouse.
- Write the sensor data directly to a scaleable DynamoDB; create a data pipeline that starts an EMR cluster using data from DynamoDB and sends the data to S3 and Redshift.
- Use Amazon Cognito to accept the data when the user pairs the sensor to the phone, and then have Cognito send the data to DynamodUse Data Pipeline to create a job that takes the DynamoDB tablee and sends it to an EMR cluster for ETL, then outputs to Redshift and S3 while, using S3 lifecycle policies to archive on Glacier.

Q4)

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?

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

---

Q5)

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?

- Use a Redshift cluster to COPY the data from Amazon S3. Create a User Defined Function in Redshift that computes the likelihood of churn.
- 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.
- 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.
- ✅ 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.

Q6)

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?

- Use Amazon S3 Client-Side Encryption with AWS KMS-Managed Keys for storing datUse AWS KMS Grants to allow access to specific elements of the platform. Use AWS CloudTrail for auditing.
- 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.
- ✅ 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.
- Use Amazon S3 Server-Side Encryption with AWS KMS-Managed Keys for storing datUse AWS KMS Grants to allow access to specific elements of the platform. Use AWS CloudTrail for auditing.

Q7)

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 store.

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?

- Launch an Amazon Relational Database Service (RDS), and use the enterprise grade and capacity of the Amazon Aurora engine for storage, processing, and querying.
- Use Amazon Redshift with data replication to Amazon Simple Storage Service (S3) for comprehensive durable data storage, processing, and querying.
- 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.
- ✅ 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.

Q8)

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?

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

Q9)

There are thousands of text files on Amazon S3. The total size of the files is 1 PThe 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?

- Apache Hive
- Apache Pig with Tachyon

- Oozie
- Presto

Q10)

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?

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

Q11)

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 - 250 KB, most of which is associated with the JSON attribute. The average customer generates - 3 GB 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?

- 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.
- Vertically partition the table, store base attributes on the primary table, and create a foreign key reference to a secondary table containing the JSON datQuery the primary table for summary data and the secondary table for JSON details.
- 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.
- Modify all database calls to use eventually consistent reads and advise customers that transaction history may be one second out-of-date.

Q12)

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?

- 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.
- 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.
- 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.
- Write sensor data records to Amazon Kinesis Streams. Process the data using KCL applications for the end-consumer dashboard and anomaly detection workflows.

Q13)

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 date. 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?

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

Q14)

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 scheme.

In which two circumstances would choosing EVEN distribution be most appropriate? (Choose two.)

- ✅ When a new table has been loaded and it is unclear how it will be joined to dimension.
- When data transfer between nodes must be eliminated.
- When data must be grouped based on a specific key on a defined slice.
- ✅ When the tables are highly denormalized and do NOT participate in frequent joins.

Q15)

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?

- Increase the minimum number of instances in the Auto Scaling group.
- Increase the size of the Amazon EC2 instances to increase network throughput.
- Increase the number of shards in the Amazon Kinesis stream to 80 for greater concurrency.
- ✅ Increase Amazon DynamoDB throughput on the checkpoint table

---

Q16)

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 insensitivThe new country contains another AWS region.

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

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

Q17)

A company is storing data on Amazon Simple Storage Service (S3). The company's security policy mandates that data be encrypted at rest.

Which of the following methods can achieve this? Choose 3 answers

- ✅ Encrypt the data on the client-side before ingesting to Amazon S3 using their own master key
- Use Amazon S3 bucket policies to restrict access to the data at rest.
- Use Amazon S3 server-side encryption with EC2 key pair.
- ✅ Use Amazon S3 server-side encryption with customer-provided keys
- ✅ Use Amazon S3 server-side encryption with AWS Key Management Service managed keys.
- Use SSL to encrypt the data while in transit to Amazon S3.

Q18)

You're launching a test Elasticsearch cluster with the Amazon Elasticsearch Service, and you'd like to restrict access to only your office desktop computer that you occasionally share with an intern to allow her to get more experience interacting with Elasticsearch.

What's the easiest way to do this?

- ✅ Create an IP-based resource policy on the Elasticsearch cluster that allows access to requests coming from the IP of the machin
- Create an IAM user and role that allows access to the Elasticsearch cluster.
- Create a username and password combination to allow you to sign into the cluster.
- Create an SSH key and add that to the accepted keys of the Elasticsearch cluster. Then store that SSH key on your desktop and use it to sign in.

Q19)

Your application development team is building a solution with two applications.

The security team wants each application's logs to be captured in two different places because one of the applications produces logs with sensitive data.

How can you meet the requirements with the least risk and effort?

- ✅ Use Amazon CloudWatch logs with two log groups, one for each application, and use an AWS IAM policy to control access to the log groups as

require

- Add logic to the application that saves sensitive data logs on the Amazon EC2 instances' local storage, and write a batch script that logs into the EC2 instances and moves sensitive logs to a secure location.
- Use Amazon CloudWatch logs to capture all logs, write an AWS Lambda function that parses the log file, and move sensitive data to a different log.
- Aggregate logs into one file, then use Amazon CloudWatch Logs and then design two CloudWatch metric filters to filter sensitive data from the logs.

Q20)

A company wants to use Redshift cluster for petabyte-scale data warehousing. Data for processing would be stored on Amazon S3.

As a security requirement, the company wants the data to be encrypted at rest. As a solution architect how would you implement the solution?

- Store the data in S3 with Server Side Encryption. Launch a Redshift cluster, copy the data to cluster and enable encryption on the cluster.
- ✅ Store the data in S3 with Server Side Encryption. Launch an encrypted Redshift cluster and copy the data to the cluster.
- Store the data in S3 with Server Side Encryption and copy the data over to Redshift cluster
- Store the data in S3. Launch an encrypted Redshift cluster, copy the data to the Redshift cluster and store back in S3 in encrypted format

---

Q21)

You have been asked to handle a large data migration from multiple Amazon RDS MySQL instances to a DynamoDB table.

You have been given a short amount of time to complete the data migration.

What will allow you to complete this complex data processing workflow?

- Write a bash script to run on your Amazon RDS instance that will export data into DynamoDB.
- Write a script in your language of choice, install the script on an Amazon EC2 instance, and then use Auto Scaling groups to ensure that the latency of the migration pipelines never exceeds four seconds in any 15-minute period.
- Create an Amazon Kinesis data stream, pipe in all of the Amazon RDS data, and direct the data toward a DynamoDB table.
- ✅ Create a data pipeline to export Amazon RDS data and import the data into DynamoDB.

Q22)

An International company has deployed a multi-tier web application that relies on DynamoDB in a single region. For regulatory reasons they need disaster recovery capability in a separate region with a Recovery Time Objective of 2 hours and a Recovery Point Objective of 24 hours.

They should synchronize their data on a regular basis and be able to provision the web application rapidly using CloudFormation. The objective is to minimize changes to the existing web application, control the throughput of DynamoDB used for the synchronization of data and synchronize only the modified elements.

Which design would you choose to meet these requirements?

- Send each update into an SQS queue in the second region; use an auto-scaling group behind the SQS queue to replay the write in the second region.
- Use AWS Data Pipeline to schedule an export of the DynamoDB table to S3 in the current region once a day then schedule another task immediately after it that will import data from S3 to DynamoDB in the other region.
- Use EMR and write a custom script to retrieve data from DynamoDB in the current region using a SCAN operation and push it to DynamoDB in the second region.
- ✅ Use AWS Data Pipeline to schedule a DynamoDB cross region copy once a day. Create a 'Lastupdated' attribute in your DynamoDB table that would represent the timestamp of the last update and use it as a filter

---

Q23)

A company hosts a web application on AWS which uses RDS instance to store critical data.

As a part of a security audit, it was recommended hardening of RDS instance. What actions would help achieve the same? (Select TWO)

- Use AWS Inspector to apply patches to the RDS instance
- Use AWS CloudTrail to track all the SSH access to the RDS instance
- ✅ Use Secure Socket Layer (SSL) connections with DB instances
- ✅ Use RDS encryption to secure the RDS instances and snapshots at rest.

Q24)

Your company produces customer commissioned one-of-a-kind skiing helmets combining high fashion with custom technical enhancements. Customers can show off their Individuality on the ski slopes and have access to head-up-displays.

GPS rear-view cams and any other technical innovation they wish to embed in the helmet. The current manufacturing process is data rich and complex including assessments to ensure that the custom electronics and materials used to assemble the helmets are to the highest standards.

Assessments are a mixture of human and automated assessments you need to add a new set of assessment to model the failure modes of the custom electronics using GPUs with CUDA across a cluster of servers with low latency networking.

What architecture would allow you to automate the existing process using a hybrid approach and ensure that the architecture can support the evolution of processes over time?

- Use AWS data Pipeline to manage movement of data & meta-data and assessments use auto-scaling group of C3 with SR-IOV (Single Root I/O virtualization)
- Use Amazon Simple Workflow (SWF) to manage assessments, movement of data & meta-datUse an autoscaling group of C3 instances with SR-IOV (Single Root I/O Virtualization).
- ✅ Use Amazon Simple Workflow (SWF) to manage assessments, movement of data & meta-datUse an autoscaling group of G2 instances in a placement group.
- Use AWS Data Pipeline to manage movement of data & meta-data and assessments Use an auto-scaling group of G2 instances in a placement group.

Q25)

You have recently joined a startup company building sensors to measure street noise and air quality in urban areas. The company has been running a pilot deployment of around 100 sensors for 3 months. Each sensor uploads 1KB of sensor data every minute to a backend hosted on AWS.

During the pilot, you measured a peak or 10 IOPS on the database, and you stored an average of 3GB of sensor data per month in the database. The current deployment consists of a load-balanced auto scaled Ingestion layer using EC2 instances and a PostgreSQL RDS database with 500GB standard storage. The pilot is considered a success and your CEO has managed to get the attention or some potential investors.

The business plan requires a deployment of at least 100K sensors, which needs to be supported by the backend. You also need to store sensor data for at least two years to be able to compare year over year Improvements. To secure funding, you have to make sure that the platform meets these requirements and leaves room for further scaling. Which setup will meet the requirements?

- Keep the current architecture but upgrade RDS storage to 3TB and 10K provisioned IOPS
- Replace the RDS instance with a 6 node Redshift cluster with 96TB of storage
- ✅ Ingest data into a DynamoDB table and move old data to a Redshift cluster
- Add an SQS queue to the ingestion layer to buffer writes to the RDS instance

---

Q26)

You have two different groups using Redshift to analyze data of a petabyte-scale data warehouse. Each query issued by the first group takes approximately 1-2 hours to analyze the data while the second group's queries only take between 5-10 minutes to analyze data.

You don't want the second group's queries to wait until the first group's queries are finished. You need to design a solution so that this does not happen. Which of the following would be the best and cheapest solution to deploy to solve this dilemma?

- Start another Redshift cluster from a snapshot for the second team if the current Redshift cluster is busy processing long queries.
- Pause the long queries when necessary and resume them when there are no queries happening.
- ✅ Create two separate workload management groups and assign them to the respective groups.
- Create a read replica of Redshift and run the second team's queries on the read replica.

Q27)

A video-sharing mobile application uploads files greater than 10 GB to an Amazon S3 bucket.

However, when using the application in locations far away from the S3 bucket region, uploads take extended periods of time, and sometimes fail to complete.

Which combination of methods would improve the performance of uploading to the application? (Select TWO.)

- ✅ Enable S3 Transfer Acceleration on the S3 bucket, and configure the application to use the Transfer Acceleration endpoint for uploads.
- Modify the application to add random prefixes to the files before uploading.
- Set up Amazon Route 53 with latency-based routing to route the uploads to the nearest S3 bucket region.
- Configure an S3 bucket in each region to receive the uploads, and use cross-region replication to copy the files to the distribution bucket.
- ✅ Configure the application to break the video files into chunks and use a multipart upload to transfer files to Amazon S3.

Q28)

ABCD has developed a sensor intended to be placed inside of people's shoes, monitoring the number of steps taken every day.

ABCD is expecting thousands of sensors reporting in every minute and hopes to scale to millions by the end of the year. A requirement for the project is it needs to be able to accept the data, run it through ETL to store in warehouse and archive it on Amazon Glacier, with room for a real-time dashboard for the sensor data to be added at a later data.

What is the best method for architecting this application given the requirements? Choose the correct answer:

- ✅ Write the sensor data directly to Amazon Kinesis and output the data into Amazon S3 creating a lifecycle policy for Glacier archiving. Also, have a parallel processing application that runs the data through EMR and sends to a Redshift data warehous
- Write the sensor data to Amazon S3 with a lifecycle policy for Glacier, create an EMR cluster that uses the bucket data and runs it through ETL. It then outputs that data into Redshift data warehouse.
- Write the sensor data directly to a scaleable DynamoDB; create a data pipeline that starts an EMR cluster using data from DynamoDB and sends the data to S3 and Redshift.
- Use Amazon Cognito to accept the data when the user pairs the sensor to the phone, and then have Cognito send the data to DynamodUse Data Pipeline to create a job that takes the DynamoDB tablee and sends it to an EMR cluster for ETL, then outputs to Redshift and S3 while, using S3 lifecycle policies to archive on Glacier.

Q29)

Your social media marketing application has a component written in Ruby running on AWS Elastic Beanstalk. This application component posts messages to social media sites in support of various marketing campaigns.

Your management now requires you to record replies to these social media messages to analyze the effectiveness of the marketing campaign in comparison to past and future efforts.

You've already developed a new application component to interface with the social media site APIs in order to read the replies.

Which process should you use to record the social media replies in a durable data store that can be accessed at any time for analytics of historical data?

- Deploy the new application component as an Amazon Elastic Beanstalk application, read the data from the social media site, store it with Amazon Elastic Block store, and use Amazon Kinesis to stream the data to Amazon CloudWatch for analytics.
- Deploy the new application component in an Auto Scaling group of Amazon EC2 instances, read the data from the social media sites, store it in Amazon Glacier, and use AWS Data Pipeline to publish it to Amazon RedShift for analytics.
- ✅ Deploy the new application component as an Elastic Beanstalk application, read the data from the social media sites, store it in DynamoDB, and use Apache Hive with Amazon Elastic MapReduce for analytics.
- Deploy the new application component in an Auto Scaling group of Amazon EC2 instances, read the data from the social media sites, store it with Amazon Elastic Block Store, and use AWS Data Pipeline to publish it to Amazon Kinesis for analytics.

Q30)

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?

- ✅ Amazon RDS
- Amazon DynamoDB
- Amazon Kinesis
- Amazon S3

Q31)

You have an application that is currently in the development stage but is expected to write 2,400 items per minute to a DynamoDB table, each 2Kb in size or less and then fluctuate to 4,800 writes of items (of the same size) per minute on weekends.

There may be other fluctuations within that range in the future as the application develops.

It is important to the success of the application that the vast majority of user requests are met in a cost-effective way. How should this table be created?

- Provision a base WCU of 160 and then schedule a job that adds 160 more WCUs when a higher load is expected.
- Enabled DynamoDB streams have a Lambda function triggered to review the current capacity on each change to the table.
- ✅ Set up an auto-scaling policy on the DynamoDB table that doesn't let the traffic dip below the usual load and allows it to scale to meet deman
- Provision a base WCU of 80 and then schedule regular increases to 160 WCUs when a higher load is expected.

Q32)

You are receiving traffic flow data in CSV format in an S3 bucket from an external provider, but your application requires the data to be in newline delimited JSON format.

Which of the following would be the best way to transform the data into the correct format?

- Transform the data with JavaScript running in the web browser.
- Create an S3 trigger to call a Lambda function that will transform the data whenever data is added to a bucket.
- ✅ Use AWS Glue to load the data, transform it, and load it back into S3.

Explanation:-AWS Data Pipeline and AWS Glue are both extract, transform, and load products on AWS and either could do the job. However, AWS Glue is a serverless service that requires much less administration and is, therefore, the better choice.

- Use AWS Data Pipeline to load the data, transform it, and load it back into S3.

Q33)

A police department wants to gather information about public opinion of its officers' performance. After an interaction with a police officer, citizens are asked to participate in a survey via text message or web form. Each survey has 10 questions, and there are approximately 100 interactions per day.

Which technologies should you use to gather this data?

- EC2, VPC, and AWS Cognito
- ✅ AWS API Gateway, AWS Lambda, and DynamoDB

Explanation:-This solution is relatively low volume (1,000 data collections per day), so a Kinesis solution would be overkill and more expensive than necessary. API Gateway with Lambda and DynamoDB would provide a cost-effective and scalable solution.

- Kinesis Data Streams, Kinesis Firehose, and Elasticsearch
- IoT Core, SQS, and DynamoDB

Q34)

You operate a primary 20TB Redshift cluster located in Canada (Central), which is replicated to a backup cluster in the EU (Stockholm). On a weekly basis (Saturday night at 0100 GMT), or when the data storage in the primary and backup clusters reaches 17TB, you run a script to move data from the last three business out of the Redshift cluster and into an S3 bucket.

What Redshift command do you use to move the data from Redshift to S3?

- BACKUP
- COPY
- ✅ UNLOAD

Explanation:-UNLOAD is the command that unloads the result of a query to one or more text files on Amazon S3 (see https://docs.aws.amazon.com/redshift/latest/dg/r_UNLOAD.html).

- SELECT

Q35) You are configuring an EMR cluster. What are some security considerations?

- Lambda functions, Redshift, and serverless
- Long-distance charges, foreign exchange rates, and IAM
- Inline encryption, over-the-air access, and S3 buckets
- ✅ At-rest encryption, in-transit encryption, and IAM roles

Explanation:-You must configure all of the items at cluster creation time.

Q36)

Your EMR cluster is unable to access your S3 bucket. You have checked the IAM role and bucket policy and confirmed that they're correct.

What is the most likely cause?

- There are no S3 token available in the EMR larder.
- ✅ An S3 endpoint has not been configured in the EMR VPC.

Explanation:-You must configure an S3 endpoint in the VPC used for EMR instances if they need to access S3.

- EMR does not support S3 access. You must configure EFS to sync the files from S3 and mount them on the EMR instances.
- EMR is not in the same zone as the S3 bucket.

Q37)

A manufacturer has deployed an automated testing system into its manufacturing line. The system has 100 instruments that can each report 100 measurements per second. The results are evaluated in real time to determine whether an individual component meets the manufacturing tolerances.

Which technologies should you employ to collect the data?

- AWS Snowball
- AWS Kinesis Firehose
- ✅ AWS Kinesis Data Streams

Explanation:-The data is analyzed in real time, so a streaming approach is required. SQS isn't called for because messaging semantics aren't required.

- AWS Simple Queue Service (SQS)

Q38)

You are working at a startup that deals with Big Data applications on AWS. Your co-founder mentions that your main investor has asked about the security of the data stored and whether it can be encrypted at rest and whether it can be processed in EMR.

Which of the following would be the best answer for the investor?

- Upload data and private encryption keys to S3. Enable a service role for EMR to access S3.
- Upload your data to Amazon S3 and use a private key stored on EC2 to decrypt it before sending it to EMR. Enable a service role for EMR to access S3, EC2, and private keys.
- Encrypt the data offsite, upload it to S3, and then pass the private key to EMR. Enable a service role for EMR to access S3.
- ✅ Upload your data to Amazon S3 and choose an AWS KMS key to encrypt the data. Enable a service role for EMR to access S3 and the AWS KMS key.

Explanation:-Uploading private keys is a very bad security practice. The only answer that can work involves having the KMS service handle encryption at rest.

Q39)

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?

- Use a Redshift cluster to COPY the data from Amazon S3. Create a User Defined Function in Redshift that computes the likelihood of churn.
- 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.
- 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.
- ✅ 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.

Q40)

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 datWhich solution should the architect build?

- Use Amazon S3 Client-Side Encryption with AWS KMS-Managed Keys for storing datUse AWS KMS Grants to allow access to specific elements of the platform. Use AWS CloudTrail for auditing.
- 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.
- ✅ 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.
- Use Amazon S3 Server-Side Encryption with AWS KMS-Managed Keys for storing datUse AWS KMS Grants to allow access to specific elements of the platform. Use AWS CloudTrail for auditing.

Q41)

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 become 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?

- ✅ Maintain data dependency information in an Amazon DynamoDB tablUse Amazon S3 event notifications to trigger an AWS Lambda function that maps the S3 object to the task associated with it in DynamoDOnce all task dependencies have been resolved, process the data with Amazon EMR.
- 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.
- 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.
- Maintain data dependency information in an Amazon DynamoDB tablUse Amazon SNS and event notifications to publish data to fleet of Amazon EC2 workers. Once the task dependencies have been resolved, process the data with Amazon EMR.

Q42)

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?

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

---

Q43)

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.)

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

Q44)

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?

- Collect both sensor data and emergency services events with Amazon Kinesis Streams and use DynamoDB for analysis.
- Collect both sensor data and emergency services events with Amazon Kinesis Firehose and use Amazon Redshift for analysis.
- 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.
- ✅ 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.

Q45)

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?

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

Q46)

There are thousands of text files on Amazon S3. The total size of the files is 1 PThe files contain retail order information for the past 2 years.

A data engineer needs to run multiple interactive queries to manipulate the datThe 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?

- Apache Hive
- Apache Pig with Tachyon
- Oozie
- ✅ Presto

Q47)

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 15-minute time limit. What is a possible solution for this problem?

- Ignore and skip events that are older than 15 minutes and put them to Dead Letter Queue (DLQ).
- ✅ Reduce the batch size that Lambda is reading from the stream.
- Add more Lambda functions to improve concurrent batch processing.
- Configure Lambda to read from fewer shards in parallel.

Q48)

A company is collected real time sensitive data using Amazon Kinesis. As a security requirement, the Amazon Kinesis stream needs to be encrypted.

Which approach should be used to accomplish this task?

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

Q49)

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 - 250 KB, 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?

- 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.
- Vertically partition the table, store base attributes on the primary table, and create a foreign key reference to a secondary table containing the JSON datQuery the primary table for summary data and the secondary table for JSON details.
- ✅ 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.
- Modify all database calls to use eventually consistent reads and advise customers that transaction history may be one second out-of-date.

Q50)

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?

- 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.
- 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.
- 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.
- ✅ Write sensor data records to Amazon Kinesis Streams. Process the data using KCL applications for the end-consumer dashboard and anomaly detection workflows.

Q51)

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?

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

Q52)

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 date. 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?

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

Q53)

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?

- ✅ Each user time series events in the past 3 months
- Last user session
- User profiles (age, gender, income, occupation)
- Quarterly results

Q54)

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 scheme.

In which two circumstances would choosing EVEN distribution be most appropriate? (Choose two.)

- [x] When a new table has been loaded and it is unclear how it will be joined to dimension.
- [ ] When data transfer between nodes must be eliminated.
- [ ] When data must be grouped based on a specific key on a defined slice.
- [x] When the tables are highly denormalized and do NOT participate in frequent joins.

Q55)

A media advertising company handles a large number of real-time messages sourced from over 200 websites in real timProcessing 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?

- [ ] Increase the minimum number of instances in the Auto Scaling group.
- [ ] Increase the size of the Amazon EC2 instances to increase network throughput.
- [ ] Increase the number of shards in the Amazon Kinesis stream to 80 for greater concurrency.
- [x] Increase Amazon DynamoDB throughput on the checkpoint table

---

Q56)

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?

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

Q57)

You need to filter and transform incoming messages coming from a smart sensor you have connected with AWS.

Once messages are received, you need to store them as time series data in DynamoDB. Which AWS service can you use?

- [ ] Kinesis
- [ ] Redshift
- [ ] IoT Device Shadow Service
- [x] IoT Rules Engine

Q58)

You work for a start-up that tracks commercial delivery trucks via GPS. You receive coordinates that are transmitted from each delivery truck once every 6 seconds.

You need to process these coordinates in real-time from multiple sources and load them into Elasticsearch without significant technical overhead to maintain. Which tool should you use to digest the data?

- [ ] AWS Data Pipeline
- [ ] Amazon EMR
- [x] Amazon Kinesis Firehose
- [ ] Amazon SQS

Q59)

A web application is using Amazon Kinesis Streams for clickstream data that may not be consumed for up to 12 hours.

As a security requirement, how can the data be secured at rest within the Kinesis Streams?

- [ ] Encrypt the data once it is at rest with a Lambda function
- [ ] Use Amazon Kinesis Consumer Library
- [ ] Enable SSL connections to Kinesis
- [x] Enable server-side encryption in Kinesis Streams

Q60)

You're launching a test Elasticsearch cluster with the Amazon Elasticsearch Service, and you'd like to restrict access to only

your office desktop computer that you occasionally share with an intern to allow her to get more experience interacting with Elasticsearch.

What's the easiest way to do this?

- ✅ Create an IP-based resource policy on the Elasticsearch cluster that allows access to requests coming from the IP of the machin
- Create an IAM user and role that allows access to the Elasticsearch cluster.
- Create a username and password combination to allow you to sign into the cluster.
- Create an SSH key and add that to the accepted keys of the Elasticsearch cluster. Then store that SSH key on your desktop and use it to sign in.