Answer Sheet

Q1)

You need real-time reporting on logs generated from your applications. In addition, you need anomaly detection. The processing latency needs to be one second or less.

Which option would you choose if your team has no experience with Machine learning libraries and doesn't want to have to maintain any software installations yourself?

- Spark Streaming with SparkSQL and MLlib
- Kinesis Firehose to S3 and Athena
- Kafka
- ✅ Kinesis Streams with Kinesis Analytics

Explanation:-Kinesis Data Streams with Kinesis Data Analytics can provide real time analytics only data while using managed services
Amazon Kinesis Data Analytics is the easiest way to analyze streaming data, gain actionable insights, and respond to your business and customer needs in real time. Amazon Kinesis Data Analytics reduces the complexity of building, managing, and integrating streaming applications with other AWS services. SQL users can easily query streaming data or build entire streaming applications using templates and an interactive SQL editor. Java developers can quickly build sophisticated streaming applications using open source Java libraries and AWS integrations to transform and analyze data in real-time.
Amazon Kinesis Data Analytics takes care of everything required to run your real-time applications continuously and scales automatically to match the volume and throughput of your incoming data.

Q2)

You need to create a recommendation engine for your e-commerce website that sells over 300 items.

The items never change, and the new users need to be presented with the list of all 300 items in order of their interest.

Which option do you use to accomplish this? (choose TWO)

- Amazon Machine Learning
- ✅ Spark/Spark MLlib

Explanation:-Spark's MLlib machine learning library should help with this task. Amazon ML is limited to 100 "categorical' recommendations, so a custom system is required for this purpose.

- ✅ Mahout

Explanation:-Mahout provides recommender engine/collaborative filtering capability

- RDS MySQL

Q3)

You need to create an Amazon Machine Learning model to predict how many inches of rain will fall in an area based on the historical rainfall data.

What type of modelling will you use?

- Binary
- ✅ Regression

Explanation:-Supervised learning using Regression can help build a model to predict rain based on the historical data.

- Categorical
- Unsupervised

Q4)

Your enterprise application requires key-value storage as the database. The data is expected to be about 10 GB the first month and grow to 2 PB over the next two years. There are no other query requirements at this time.

What solution would you recommend?

- RDS MySQL
- ✅ HBase on HDFS

Explanation:-HBase on HDFS provide the ability the store the large amount of data in a non-relational key-value format. HBase is an open source, non-relational, distributed database developed as part of the Apache Software Foundation's Hadoop project. HBase runs on top of Hadoop Distributed File System (HDFS) to provide non-relational database capabilities for the Hadoop ecosystem. HBase works seamlessly with Hadoop, sharing its file system and serving as a direct input and output to the MapReduce framework and execution engine. HBase also integrates with Apache Hive, enabling SQL-like queries over HBase tables, joins with Hive-based tables, and support for Java Database Connectivity (JDBC).

- Hive on HDFS
- Hadoop with Spark

Q5)

You have a customer-facing application running on multiple M3 instances in two AZs. These instances are in an auto-scaling group configured to scale up when load increases. After taking a look at your CloudWatch metrics, you realize that during specific times every single day, the auto-scaling group has a lot more instances than it normally does.

Despite this, one of your customers is complaining that the application is very slow to respond during those time periods every day. The application is reading and writing to a DynamoDB table which has 400 Write Capacity Units and 400 Read Capacity Units. The primary key is the company ID, and the table is storing roughly 20 TB of data.

Which solution would solve the issue in a scalable and cost-effective manner?

Double the number of Read and Write Capacity Units. The DynamoDB table is being throttled when customers from the same company all use the table at the same time.
DynamoDB is not a good solution for this use case. Instead, create a data pipeline to move data from DynamoDB to Amazon RDS, which is more suitable for this.
Add a caching layer in front of the web application with ElastiCache Memcached, or Redis.
✅ Use data pipelines to migrate your DynamoDB table to a new DynamoDB table with a different primary key that evenly distributes the dataset across the table.
Explanation:-A single company is facing the issue and it would be a hot key issue cause of the primary key being Company ID. Data Pipeline can be used to migrate the data.

Q6)

You have to identify potential fraudulent credit card transactions using Amazon Machine Learning.

You have been given historical labeled data that you can use to create your model.

You will also need to the ability to tune the model you pick.

Which model type should you use?

Cannot be done using Amazon Machine Learning
✅ Binary
Explanation:-Binary classification can be used to predict for whether the transaction is fraudulent or not.
Type of ML Problem
Classification
Regression
Clustering
Association rule learning
Structured output
Ranking
Regression
Clustering

Q7)

A company has launched EMR cluster to support their big data analytics requirements. AFS has multiple data sources built out of S3, SQL databases, MongoDB, Redis, RDS, other file systems.

They are looking for a web application to create and share documents that contain live code, equations, visualizations, and narrative text.

Which EMR Hadoop ecosystem fulfils the requirements?

Apache Hue
✅ Jupyter Notebook
Explanation:-Jupyter Notebook is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and narrative text
Apache Hive
Apache Presto

Q8)

Your company uses DynamoDB to support their mobile application and S3 to host the images and other documents shared between users.

DynamoDB has a table with 60 partitions and is being heavily accessed by users. The queries run by users do not fully use the per-partition's throughput.

However there are times when in less than 3 minutes, a heavy load of queries flow in and this happen occasionally.

Sometimes there are many background tasks that are running in background.

How can DynamoDB be configured to handle the workload?

Using Write Sharding to distribute Workloads Evenly
Design Partition Keys to distribute workload evenly
Using Adaptive Capacity
✅ Using Burst Capacity effectively
Explanation:-DynamoDB burst capacity can retain part of unused provisioned capacity, upto 5 minutes, allowing application to burst.
DynamoDB provides some flexibility in your per-partition throughput provisioning by providing burst capacity, as follows. Whenever you are not fully using a partition's throughput, DynamoDB reserves a portion of that unused capacity for later bursts of throughput to handle usage spikes. DynamoDB currently retains up to five minutes (300 seconds) of unused read and write capacity. During an occasional burst of read or write activity, these extra capacity units can be consumed quickly—even faster than the per-second provisioned throughput capacity that you've defined for your table.
DynamoDB can also consume burst capacity for background maintenance and other tasks without prior notice.

Q9)

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?

Publish messages to Amazon Kinesis Streams, pull messages off with Spark Streaming, and write raw data to Amazon Simple Storage Service (S3) before and after processing.
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.
✅ 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).
Explanation:-The data can be capture by Kinesis Streams. Kinesis Streams can feed data to Spark Streaming for analysis and Lambda to move the raw data to S3 for durable storage.
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.

Q10)

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 fields: 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?

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.
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.
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.
✅ 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.
Explanation:-You can use composite primary keys using a combination of (StreamName as the partition key and TimeStamp as the sort key) and (ServerName as the partition key and TimeStamp as the sort key) which would provide the ability to query on both StreamName and ServerName over a time.

Q11)

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?

MicroStrategy
✅ Presto
Explanation:-Presto can help perform interactive analysis and its performance is much better than Pig as it uses a custom query execution engine Presto is an open-source distributed SQL query engine optimized for low-latency, ad-hoc analysis of data. It supports the ANSI SQL standard, including complex queries, aggregations, joins, and window functions. Presto can process data from multiple data sources including the Hadoop Distributed File System (HDFS) and Amazon S3. Presto uses a custom query execution engine with operators designed to support SQL semantics. Different from Hive/MapReduce, Presto executes queries in memory, pipelined across the network between stages, thus avoiding unnecessary I/O. The pipelined execution model runs multiple stages in parallel and streams data from one stage to the next as it becomes available.
Pig
R Studio

Q12)

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?

Route the largest customers to a dedicated Redshift cluster. Raise the concurrency of the multi-tenant Redshift cluster to accommodate the remaining customers.
Push aggregations into an RDS for Aurora instance. Connect the dashboard application to Aurora rather than Redshift for faster queries.
✅ 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.
Explanation:-Redshift allows Workload Management (WLM) to help define queues. In this case, a dedicated Queue can be defined for large customer and other for rest of customers ensuring the queries from them do not interfere with large customers. You can use workload management (WLM) to define multiple query queues and to route queries to the appropriate queues at run time. In some cases, you might have multiple sessions or users running queries at the same time. In these cases, some queries might consume cluster resources for long periods of time and affect the performance of other queries. For example, suppose that one group of users submits occasional complex, long-running queries that select and sort rows from several large tables. Another group frequently submits short queries that select only a few rows from one or two tables and run in a few seconds. In this situation, the short-running queries might have to wait in a queue for a long-running query to complete. Alternatively, you can manage system performance and your users' experience by modifying your WLM configuration to create separate queues for the long-running queries and the short-running queries. At run time, you can route queries to these queues according to user groups or query groups. You can enable this manual configuration using the Amazon Redshift console by switching to Manual WLM. With this choice, you specify the queues used to manage queries, and the Memory and Concurrency on main field values. With a manual configuration, you can configure up to eight query queues and set the number of queries that can run in each of those queues concurrently. You can set up rules to route queries to particular queues based on the user running the query or labels that you specify. You can also configure the amount of memory allocated to each queue, so that large queries run in

queues with more memory than other queues. You can also configure the WLM timeout property to limit long-running queries.

- Apply query filters based on customer-id that can NOT be changed by the user and apply distribution keys on customer-id.

Q13)

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?

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

Explanation:-You should choose the largest and the most frequently joined dimension table and use the key of the dimension table as the distribution key. Using distribution keys is a good way to optimize the performance of Amazon Redshift when you use a star schema. With an EVEN distribution, data is spread equally across all nodes in the cluster to ensure balanced processing. In many cases, simply distributing data equally using EVEN does not optimize performance as the data rows on a node for the table do not have any affinity with each other. Take an example of a fact table for ORDERS where a distribution style of EVEN is chosen. In this case, the orders for a specific customer are potentially spread across many compute nodes in the cluster. However, if the table had a distribution style of KEY and a DISTKEY of customer_id was chosen, then all of the orders for a particular customer would be stored on the same compute node. Using a distribution style of EVEN can lead to more cross-node traffic. A good selection for a distribution key distributes data relatively evenly across nodes while collocating related data on a compute node used in joins or aggregates. When you perform a join on a column that is a distribution key for both tables, Amazon Redshift is able to run the join locally on each node with no inter-node data movement; this is because rows with the same distribution key value reside on the same node for both tables in the join. Similarly, aggregating on a distribution key performs better because the data for the aggregate column value is local to the compute node. In a typical star schema, the fact table has foreign key relationships with multiple dimension tables, so you need to choose one of the dimensions. You would choose the foreign key for the largest frequently joined dimension as a distribution key in the fact table and the primary key in the dimension table. Make sure that the distribution keys chosen result in relatively even distribution for both tables, and if the distribution is skewed, use a different dimension. Then analyze the remaining dimensions to determine if a distribution style of ALL, KEY, or EVEN is appropriate.

Q14)

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?

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

Explanation:-Glacier provides Vault Lock Policy which can be used to prevent an action on the Vault. An Amazon S3 Glacier (Glacier) vault can have one resource-based vault access policy and one Vault Lock policy attached to it. A Vault Lock policy is a vault access policy that you can lock. Using a Vault Lock policy can help you enforce regulatory and compliance requirements. As an example of a Vault Lock policy, suppose that you are required to retain archives for one year before you can delete them. To implement this requirement, you can create a Vault Lock policy that denies users permissions to delete an archive until the archive has existed for one year. You can test this policy before locking it down. After you lock the policy, the policy becomes immutable.

Q15)

Your customer is willing to consolidate their log streams (access logs, application logs, security logs etc.) in one single system.

Once consolidated, the customer wants to analyze these logs in real time based on heuristics. From time to time, the customer needs to validate heuristics, which requires going back to data samples extracted from the last 12 hours.

What is the best approach to meet your customer's requirements?

- Key point here is requiring real time analytics and ability to go back to data samples from last 12 hours
- Setup an Auto Scaling group of EC2 syslogd servers, store the logs on S3 use EMR to apply heuristics on the logs
- Configure Amazon CloudTrail to receive custom logs, use EMR to apply heuristics the logs
- ✅ Send all the log events to Amazon Kinesis develop a client process to apply heuristics on the logs

Explanation:-Kinesis can perform real time analysis and stores data for 24 hours which can be extended to 7 days. Also data is not removed from Kinesis till 24 hours default, can be extended, and can be used to retrieve past data SQS minimum message size is 1,024 bytes (1 KB). The maximum is 262,144 bytes (256 KB). While Kinesis can store upload 1MB

- Send all the log events to Amazon SQS. Setup an Auto Scaling group of EC2 servers to consume the logs and apply the heuristics.

Q16)

Your team is building up a smart home iOS APP. The end users will use your company's camera-equipped home devices such as baby monitors, webcams, and home surveillance systems. Then the videos would be uploaded to AWS.

The users can then play the on-demand or live videos using the format of HTTP Live Streaming (HLS) through the Mobile application.

Which combinations of steps should you use to design the solution? (Select TWO)

- Transform the stream data to HLS compatible data by using Kinesis Data Analytics or customer code in EC2/Lambda. Then in the mobile application, use HLS protocol to display the video stream by using the converted HLS streaming data.

✅ In the mobile application, use HLS to display the video stream by using the HLS streaming session URL.

Explanation:-Kinesis Video Streams can be used to stream, store Videos and these videos can then be streamed back using APIs. Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, machine learning (ML), and other processing. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest streaming video data from millions of devices. It also durably stores, encrypts, and indexes video data in your streams, and allows you to access your data through easy-to-use APIs. Kinesis Video Streams enables you to quickly build computer vision and ML applications through integration with Amazon Rekognition Video and libraries for ML frameworks such as Apache MxNet, TensorFlow, and OpenCV. Kinesis Video Streams is ideal for building computer vision-enabled ML applications that are becoming prevalent in a wide range of use cases such as the following: Smart Home - With Kinesis Video Streams, you can easily stream video and audio from camera-equipped home devices such as baby monitors, webcams, and home surveillance systems to AWS. You can then use the streams to build a variety of smart home applications ranging from simple video playback to intelligent lighting, climate control systems, and security solutions.

⚪ Create a Kinesis Data Firehose to ingest, durably store and encrypt the live videos from the users' home devices.

✅ Create a Kinesis video stream to capture, store, and index the videos from the camera-equipped home devices.

Explanation:-Kinesis Video Streams can be used to stream, store Videos and these videos can then be streamed back using APIs. Amazon Kinesis Video Streams makes it easy to securely stream video from connected devices to AWS for analytics, machine learning (ML), and other processing. Kinesis Video Streams automatically provisions and elastically scales all the infrastructure needed to ingest streaming video data from millions of devices. It also durably stores, encrypts, and indexes video data in your streams, and allows you to access your data through easy-to-use APIs. Kinesis Video Streams enables you to quickly build computer vision and ML applications through integration with Amazon Rekognition Video and libraries for ML frameworks such as Apache MxNet, TensorFlow, and OpenCV. Kinesis Video Streams is ideal for building computer vision-enabled ML applications that are becoming prevalent in a wide range of use cases such as the following: Smart Home - With Kinesis Video Streams, you can easily stream video and audio from camera-equipped home devices such as baby monitors, webcams, and home surveillance systems to AWS. You can then use the streams to build a variety of smart home applications ranging from simple video playback to intelligent lighting, climate control systems, and security solutions.

Q17)

You require the ability to analyze a customer's clickstream data on a website so they can do behavioral analysis. Your customer needs to know what sequence of pages and ads their customer clicked on. This data will be used in real time to modify the page layouts as customers click through the site to increase stickiness and advertising click-through.

Which option meets the requirements for capturing and analyzing this data?

⚪ Publish web clicks by session to an Amazon SQS queue and periodically drain these events to Amazon RDS and analyze with SQL

⚪ Write click events directly to Amazon Redshift and then analyze with SQL

✅ Push web clicks by session to Amazon Kinesis and analyze behavior using Kinesis workers

Explanation:-Key point here is real time data capture and analytics. Kinesis helps to collect real time data capture and analyze using kinesis workers

⚪ Log clicks in weblogs by URL store to Amazon S3, and then analyze with Elastic MapReduce

---

Q18)

A company is performing a full migration of its systems from an on-premises data center to AWS. The company needs to move all the data stored on-premises to Amazon S3 within the next 4 weeks.

Currently, the on-premises storage holds 900 TB of data and is connected to the Internet over a 100 Mbps link. Up to 20% of the link's throughput is regularly used in real time by existing systems.

What is the MOST cost-effective way to perform the data migration in the given time frame?

⚪ Configure a VPN tunnel for the AWS environment to upload the data.

⚪ Set up an AWS Direct Connect link to upload the data.

⚪ Use a multipart upload to transfer the data over the existing link.

✅ Order multiple AWS Snowball devices to ship the data.

Explanation:-With 900TB of data and 80% of 100Mbps line, it would take years to transfer the data. Snowball provides a quick and cost effective option to transfer huge data from on-premises to AWS S3. Snowball is a petabyte-scale data transport solution that uses devices designed to be secure to transfer large amounts of data into and out of the AWS Cloud. Using Snowball addresses common challenges with large-scale data transfers including high network costs, long transfer times, and security concerns. Customers today use Snowball to migrate analytics data, genomics data, video libraries, image repositories, backups, and to archive part of data center shutdowns, tape replacement or application migration projects. Transferring data with Snowball is simple, fast, more secure, and can be as little as one-fifth the cost of transferring data via high-speed Internet.

Q19)

You have been asked to cost optimize a business critical and long-running EMR cluster. The EMR cluster is currently on-demand for the master nodes, core nodes and task nodes. The costs for running the cluster have been steadily increasing as nodes have been added and resized.

What would you suggest the business does "to reduce the costs without requiring any long-term commitment" ?

⚪ Leave all nodes running on-demand instances, the cluster is already cost optimized.

✅ Leave the master and core nodes as on-demand and use spot instances for the task nodes

Explanation:-AWS recommends using reserved or on-demand instances for Master and Core nodes. Task nodes can use spot instances to improve performance and reduce cost. The master node controls and directs the cluster. When it terminates, the cluster ends, so you should only launch the master node as a Spot Instance if you are running a cluster where sudden termination is acceptable. This might be the case if you are testing a new application, have a cluster that periodically persists data to an external store such as Amazon S3, or are running a cluster where cost is more important than ensuring the cluster's completion. Core nodes process data and store information using HDFS. Terminating a core instance risks data loss. For this reason, you should only run core nodes on Spot Instances when partial HDFS data loss is tolerable. The task nodes process data but do not hold persistent data in HDFS. If they terminate because the Spot price has risen above your maximum Spot price, no data is lost and the effect on your cluster is minimal.

⚪ Leave the master node to use on-demand and change the core and task nodes to spot

⚪ Recreate the cluster using spot instances for the master, core and task nodes.

Q20)

You have just joined a new company and have been put in charge of EC2 instances and any other services that use EC2 instances.

You notice that the company has been slow to take advantage of AWS per-second Billing, specifically in the area of EMR and Spot Instances.

What immediate steps can you take on EMR with spot instances to improve cost saving and performance?

- Use on-demand instances instead.
- Run fewer instances for a shorter amount of time.
- Run fewer instances for a longer amount of time.
- ✅ Run more instances for a shorter amount of time.

Explanation:-EMR now supports instances with per second billing, it would be more cost efficient and performant to use more instances for shorter amount of time. Amazon EMR – Our customers add capacity to their EMR clusters in order to get their results more quickly. With per-second billing for the EC2 instances in the clusters, adding nodes is more cost-effective than ever. To learn more, read Amazon EMR Now Supports Per-Second Billing.

Q21)

A retailer exports data daily from its transactional databases into an S3 bucket in the Sydney region. The retailer's Data Warehousing team wants to import this data into an existing Amazon Redshift cluster in their VPC at Sydney. Corporate security policy mandates that data can only be transported within a VPC.

What combination of the following steps will satisfy the security policy?

- ✅ Create and configure an Amazon S3 VPC endpoint.

Explanation:-Redshift Enhanced VPC Routing helps access AWS services including S3 through VPC, without having to route any traffic through internet. Also, note the region is the same.
Refer AWS documentation - Redshift Enhanced VPC Routing
When you use Amazon Redshift Enhanced VPC Routing, Amazon Redshift forces all COPY and UNLOAD traffic between your cluster and your data repositories through your Amazon VPC. You can now use standard VPC features, such as VPC security groups, network access control lists (ACLs), VPC endpoints, VPC endpoint policies, Internet gateways, and Domain Name System (DNS) servers, to tightly manage the flow of data between your Amazon Redshift cluster and other resources. When you use Enhanced VPC Routing to route traffic through your VPC, you can also use VPC flow logs to monitor COPY and UNLOAD traffic.
If Enhanced VPC Routing is not enabled, Amazon Redshift routes traffic through the Internet, including traffic to other services within the AWS network.
VPC Endpoints – For traffic to an Amazon S3 bucket in the same region as your cluster, you can create a VPC endpoint to direct traffic directly to the bucket. When you use VPC endpoints, you can attach an endpoint policy to manage access to Amazon S3.

- Create a NAT gateway in a public subnet to allow the Amazon Redshift cluster to access Amazon S3.
- Create a Cluster Security Group to allow the Amazon Redshift cluster to access Amazon S3.
- ✅ Enable Amazon Redshift Enhanced VPC Routing.

Explanation:-Redshift Enhanced VPC Routing helps access AWS services including S3 through VPC, without having to route any traffic through internet. Also, note the region is the same.
Refer AWS documentation - Redshift Enhanced VPC Routing
When you use Amazon Redshift Enhanced VPC Routing, Amazon Redshift forces all COPY and UNLOAD traffic between your cluster and your data repositories through your Amazon VPC. You can now use standard VPC features, such as VPC security groups, network access control lists (ACLs), VPC endpoints, VPC endpoint policies, Internet gateways, and Domain Name System (DNS) servers, to tightly manage the flow of data between your Amazon Redshift cluster and other resources. When you use Enhanced VPC Routing to route traffic through your VPC, you can also use VPC flow logs to monitor COPY and UNLOAD traffic.
If Enhanced VPC Routing is not enabled, Amazon Redshift routes traffic through the Internet, including traffic to other services within the AWS network.
VPC Endpoints – For traffic to an Amazon S3 bucket in the same region as your cluster, you can create a VPC endpoint to direct traffic directly to the bucket. When you use VPC endpoints, you can attach an endpoint policy to manage access to Amazon S3.

Q22)

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

Explanation:-Data at rest encryption using S3 can be implemented using either Server Side or Client Side encryption. SSE can be implemented using either KMS provided keys (SSE-KMS) or Customer provided keys (SSE-C). CSE can be implemented by encrypting the data before uploading it to S3 and then decrypting the data after downloading it from S3 at client side.

- 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

Explanation:-Data at rest encryption using S3 can be implemented using either Server Side or Client Side encryption. SSE can be implemented using either KMS provided keys (SSE-KMS) or Customer provided keys (SSE-C). CSE can be implemented by encrypting the data before uploading it to S3 and then decrypting the data after downloading it from S3 at client side.

- ✅ Use Amazon S3 server-side encryption with AWS Key Management Service managed keys.

Explanation:-Data at rest encryption using S3 can be implemented using either Server Side or Client Side encryption. SSE can be implemented using either KMS provided keys (SSE-KMS) or Customer provided keys (SSE-C). CSE can be implemented by encrypting the data before uploading it to S3 and then decrypting the data after downloading it from S3 at client side.

- Use SSL to encrypt the data while in transit to Amazon S3.

Q23)

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 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.
- Create an IAM user and role that allows access to the Elasticsearch cluster.
- ✅ Create an IP-based resource policy on the Elasticsearch cluster that allows access to requests coming from the IP of the machine.

Explanation:-IP-based resource policy can restrict access to the specific IP addresses only.
Refer AWS documentation - Elasticsearch Access Control
IP-based Policies - IP-based policies restrict access to a domain to one or more IP addresses or CIDR blocks. Technically, IP-based policies are not a distinct type of policy. Instead, they are just resource-based policies that specify an anonymous principal and include a special Condition element. The primary appeal of IP-based policies is that they allow unsigned requests to an Amazon ES domain, which lets you use clients like curl and Kibana or access the domain through a proxy server.

Q24)

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 required.

Explanation:-Different CloudWatch log groups can be created, which can have separate access control policies.
Refer AWS documentation - CloudWatch Log Groups
A log group is a group of log streams that share the same retention, monitoring, and access control settings. You can define log groups and specify which streams to put into each group. There is no limit on the number of log streams that can belong to one log group.

- 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.
- Aggregate logs into one file, then use Amazon CloudWatch Logs and then design two CloudWatch metric filters to filter sensitive data from the logs.
- 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.

Q25)

Your company needs to design a data warehouse for a client in the retail industry. The data warehouse will store historic purchases in Amazon Redshift. To comply with PCI:DSS requirements and meet data protection standards, the data must be encrypted at rest and have keys managed by a corporate on-premises HSM.

How can you meet these requirements in a cost-effective manner?

- 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.
- 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.
- ✅ Create a VPN connection between a VPC you create in AWS and an on-premises network. Then launch the Redshift cluster in the VPC, and configure it to use your corporate HSM.

Explanation:-Redshift Cluster can integrate with corporate HSM via VPN in a cost-effective way
Refer AWS documentation - Redshift Encryption
In Amazon Redshift, you can enable database encryption for your clusters to help protect data at rest. When you enable encryption for a cluster, the data blocks and system metadata are encrypted for the cluster and its snapshots.
You can enable encryption when you launch your cluster, or you can modify an unencrypted cluster to use AWS Key Management Service (AWS KMS) encryption. To do so, you can use either an AWS-managed key or a customer-managed key (CMK). When you modify your cluster to enable KMS encryption, Amazon Redshift automatically migrates your data to a new encrypted cluster. Snapshots created from the encrypted cluster are also encrypted.
Amazon Redshift uses a hierarchy of encryption keys to encrypt the database. You can use either AWS Key Management Service (AWS KMS) or a hardware security module (HSM) to manage the top-level encryption keys in this hierarchy. The process that Amazon Redshift uses for encryption differs depending on how you manage keys. Amazon Redshift automatically integrates with AWS KMS but not with an HSM. When you use an HSM, you must use client and server certificates to configure a trusted connection between Amazon Redshift and your HSM.

- Use AWS Import/Export to import a company HSM device into AWS alongside the Amazon Redshift cluster, and configure Redshift to use the imported HSM.

Q26)

Your client needs to load a 600 GB file into a Redshift cluster from S3, using the Redshift COPY command. The file has several known (and potentially some unknown) issues that will probably cause the load process to fail.

How should the client most efficiently detect load errors without needing to perform cleanup if the load process fails?

- Split the 600 GB file into smaller 25 GB chunks and load each separately.
- Compress the input file before running COPY.
- Write a script to delete the data from the tables in case of errors.
- ✅ Use the COPY command with the NOLOAD parameter.

Explanation:-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.

Q27)

A company has lot of web applications, databases and data warehouse built on Teradata, NoSQL databases, and other types of data stores. They have lot of data assets in terms of logs, documents; excel files, CSV files, PDF documents and others. Web Application has different user workloads at different parts of the day. They are running one of their web application Node.js supported by MongoDB Database. The schema designed is document based. The team wants to migrate the platform on to AWS.

Which NoSQL Managed service provides the document management capability?

- Amazon Neptune Database, being a graph database support document models and NoSQL requirements
- Amazon RDS Database, being a multi-modal database support document models and NoSQL requirements
- ✅ Amazon DynamoDB Database, being a document database support document models and NoSQL requirements

Explanation:-Amazon DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability.

- Amazon Aurora Database, being a multi-modal database support document models and NoSQL requirements

---

Q28)

A company launched EMR cluster to support their big data analytics requirements. They have multiple data sources built out of S3, SQL databases, MongoDB, Redis, RDS, other file systems.

They are looking for distributed processing framework and programming model that helps you do machine learning, stream processing, or graph analytics using Amazon EMR clusters Which EMR Hadoop ecosystem fulfils the requirements?

- Apache Hive
- Apache Hbase
- Apache Hcatalog
- ✅ Apache Spark

Explanation:-Apache Spark is a distributed processing framework and programming model that helps you do machine learning, stream processing, or graph analytics using Amazon EMR clusters. Similar to Apache Hadoop, Spark is an open-source, distributed processing system commonly used for big data workloads. However, Spark has several notable differences from Hadoop MapReduce. Spark has an optimized directed acyclic graph (DAG) execution engine and actively caches data in-memory, which can boost performance, especially for certain algorithms and interactive queries.

Q29)

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?

- Split the file into 10 files of equal size.
- Convert the file format to AVRO.
- ✅ Split the file into 500 smaller files.

Explanation:-The critical aspect of this question is running the COPY command with the maximum amount of parallelism. It 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) Split Your Load Data into Multiple Files - The COPY command loads the data in parallel from multiple files, dividing the workload among the nodes in your cluster. When you load all the data from a single large file, Amazon Redshift is forced to perform a serialized load, which is much slower. Split your load data files so that the files are about equal size, between 1 MB and 1 GB after compression. For optimum parallelism, the ideal size is between 1 MB and 125 MB after compression. The number of files should be a multiple of the number of slices in your cluster.

- Compress the file using gz compression.

Q30)

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?

- Compress the input file before running COPY.
- Write a script to delete the data from the tables in case of errors.
- ✅ Use COPY with NOLOAD parameter.

Explanation:-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. If you want to validate your data without actually loading the table, use the NOLOAD option with the COPY command.

- Split the input file into 50-GB blocks and load them separately.

Q31)

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?

- Have business units submit data via Amazon Kinesis Firehose to aggregate data hourly into Amazon S3.
- Use Amazon S3 Event Notifications and AWS Lambda to index each file into an Amazon Elasticsearch Service cluster.
- ✅ Schedule a daily AWS Data Pipeline process that aggregates content into larger files using S3DistCp.

Explanation:-The focus is to improve performance with minimal changes. S3DistCp can be used to aggregate smaller files to large ones without any change to the existing applications and processes. Hadoop is optimized for reading a fewer number of large files rather than many small files, whether from S3 or HDFS. You can use S3DistCp to aggregate small files into fewer large files of a size that you choose, which can optimize your analysis for both performance and cost.

- Add Spark to the Amazon EMR cluster and utilize Resilient Distributed Datasets in-memory.
- Use Amazon S3 Event Notifications and AWS Lambda to create a quick search file index in DynamoDB.

Q32)

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?

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

Explanation:-The labels are missing, a Binary classification model with a required response post can be applied. Amazon ML supports three types of ML models: binary classification, multiclass classification, and regression. The type of model you should choose depends on the type of target that you want to predict.?

Q33)

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 ever? 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?

- 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.
- ✅ 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.

Explanation:-The S3 even notification can be used to populate DynamoDB with the metadata of the file like physician, patient and date. This does not impact the current process and provides and easy way for the auditors to query the data. Amazon S3 is a simple key-based object store whose scalability and low cost make it ideal for storing large datasets. Its design enables S3 to provide excellent performance for storing and retrieving objects based on a known key. Finding objects based on other attributes, however, requires doing a linear search using the LIST operation. Because each listing can return at most 1000 keys, it may require many requests before finding the object. Because of these additional requests, implementing attribute-based queries in S3 alone can be challenging. A common solution is to build an external index that maps queryable attributes to the S3 object key. This index can leverage data repositories that are built for fast lookups but might not be great at storing large data blobs. These types of indexes provide an entry point to your data that can be used by a variety of systems.

- 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.
- 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.

Q34)

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?

- 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.
- ✅ 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.

Explanation:-The requirement is not to have a real time change replication, the incremental data can be retrieved and upload to S3 as multiple objects. COPY commands would help load the data into Redshift staging layer taking advantage of parallelism with multiple files. COPY command leverages the Amazon Redshift massively parallel processing (MPP) architecture to read and load data in parallel from files in an Amazon S3 bucket. You can take maximum advantage of parallel processing by splitting your data into multiple files and by setting distribution keys on your tables.

- 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.
- 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.

Q35)

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?

- Using the a priori probability distribution of the two classes, use Monte-Carlo simulation to generate the labels.
- Use the sentiment analysis NLP library to determine whether a post requires a response.
- ✅ Use the Amazon Mechanical Turk web service to publish Human Intelligence Tasks that ask Turk workers to label the posts.

Explanation:-You need accurate data to train the service and get accurate results from future data.

- ✅ Ask the social media handling team to review each post and provide the label.

Explanation:-You need accurate data to train the service and get accurate results from future data.

Q36)

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"?

- Store the events for an IP address as a single file in Amazon S3 and add metadata with keys:Hive_Partitioned_IPAddress.
- Store the data in an HBase table with the IP address as the row key.
- ✅ Store the Amazon S3 objects with the following naming scheme bucket_name/source=ip_address/year=yy/month=mm/day=dd/hour=hh/filename.

Explanation:-You can create an external table with dynamic partitioning enabled and point to S3. Partitioning on ip_address and dates would help in efficient queries.

- Store an index of the files by IP address in the Amazon DynamoDB metadata store for EMRFS.

Q37)

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 or from user feedback.

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

- 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.
- ✅ 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.

Explanation:-There is data already to learn, a machine learning model can be developed to properly classify the goods and run it with DynamoDB Streams.

- 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.
- 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.

Q38)

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?

- Initiate a Python job from AWS Data Pipeline to run directly against the Amazon S3 text files.
- Use Amazon Elasticsearch Service to store the text and then use the Python Elasticsearch Client to run analysis against the text index.
- ✅ Use Amazon EMR to parallelize the text analysis tasks across the cluster using a streaming program step.

Explanation:-The data is huge EMR can be used to parallelly analyse the data using Streaming program which supports python. A Streaming application reads input from standard input and then runs a script or executable (called a mapper) against each input. The result from each of the inputs is saved locally, typically on a Hadoop Distributed File System (HDFS) partition. After all the input is processed by the mapper, a second script or executable (called a reducer) processes the mapper results. The results from the reducer are sent to standard output. You can chain together a series of Streaming steps, where the output of one step becomes the input of another step. The mapper and the reducer can each be referenced as a file or you can supply a Java class. You can implement the mapper and reducer in any of the supported languages, including Ruby, Perl, Python, PHP, or Bash.

- 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.

Q39)

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?

- 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 data. Query 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.

Explanation:-The key requirement is to reduce cost without affecting quality. The issue here is the JSON data is being read always even though it is not needed 90% of time. As the data size for JSON is huge compared to other attributes, the provisioned throughput needed is high. The issue can be resolved by retrieving only the details for history and the transaction details when needed. Creating a GSI for the base transaction history and using the primary for transaction summary would work perfectly.

- Modify all database calls to use eventually consistent reads and advise customers that transaction history may be one second out-of-date.

Q40)

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.

Explanation:-The data can be encrypted using client side encryption. The encryption needs to be done on the producer before the data is pushed to Kinesis Streams.

Q41)

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?

- 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.
- 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.
- ✅ 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.

Explanation:-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. 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. An Amazon S3 Glacier (Glacier) vault can have one resource-based vault access policy and one Vault Lock policy attached to it. A Vault Lock policy is a vault access policy that you can lock. Using a Vault Lock policy can help you enforce regulatory and compliance requirements.

- 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.

Q42)

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.
- ✅ Create a manual snapshot of the Amazon Redshift cluster.

Explanation:-manual snapshot needs to be taken to be able to restore Redshift to the point before upgrade.

Refer AWS documentation - Redshift Snapshots

Snapshots are point-in-time backups of a cluster. There are two types of snapshots: automated and manual. Amazon Redshift stores these snapshots internally in Amazon S3 by using an encrypted Secure Sockets Layer (SSL) connection.

Amazon Redshift automatically takes incremental snapshots that track changes to the cluster since the previous automated snapshot. Automated snapshots retain all of the data required to restore a cluster from a snapshot. You can create a snapshot schedule to control when automated snapshots are taken, or you can take a manual snapshot any time.

When you restore from a snapshot, Amazon Redshift creates a new cluster and makes the new cluster available before all of the data is loaded, so

you can begin querying the new cluster immediately. The cluster streams data on demand from the snapshot in response to active queries, then loads the remaining data in the background.
When you launch a cluster, you can set the retention period for automated and manual snapshots. You can change the retention period for automated and manual snapshots by modifying the cluster. You can change the retention period for a manual snapshot when you create the snapshot or by modifying the snapshot.
You can take a manual snapshot any time. By default, manual snapshots are retained indefinitely, even after you delete your cluster. You can specify the retention period when you create a manual snapshot, or you can change the retention period by modifying the snapshot. If you create a snapshot using the Amazon Redshift console, it defaults the snapshot retention period to 365 days.
If a snapshot is deleted, you can't start any new operations that reference that snapshot. However, if a restore operation is in progress, that restore operation will run to completion.

- Make a copy of the automated snapshot on the Amazon Redshift cluster.
- Run an UNLOAD command for all data in the warehouse and save it to S3.

Q43)

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.

Explanation:-Lambda reads in batches from Kinesis from a single shard, and hence it might timeout if the batch of records is huge.
Refer AWS documentation - Lambda with Kinesis
You can use an AWS Lambda function to process records in an Amazon Kinesis data stream. With Kinesis, you can collect data from many sources and process them with multiple consumers. Lambda supports standard data stream iterators and HTTP/2 stream consumers.
Lambda reads records from the data stream and invokes your function synchronously with an event that contains stream records. Lambda reads records in batches and invokes your function to process records from the batch.
Your Lambda function is a consumer application for your data stream. It processes one batch of records at a time from each shard.
For standard iterators, Lambda polls each shard in your Kinesis stream for records at a base rate of once per second. When more records are available, Lambda keeps processing batches until it receives a batch that's smaller than the configured maximum batch size. The function shares read throughput with other consumers of the shard.

- Add more Lambda functions to improve concurrent batch processing.
- Configure Lambda to read from fewer shards in parallel.

Q44)

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?

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

Explanation:-Presto can help work on Petabytes of data with the interactive ability.
Refer AWS documentation - EMR Presto
Presto is an open-source distributed SQL query engine optimized for low-latency, ad-hoc analysis of data. It supports the ANSI SQL standard, including complex queries, aggregations, joins, and window functions. Presto can process data from multiple data sources including the Hadoop Distributed File System (HDFS) and Amazon S3.
You can quickly and easily create managed Presto clusters from the AWS Management Console, AWS CLI, or the Amazon EMR API. Additionally, you can leverage additional Amazon EMR features, including fast Amazon S3 connectivity, integration with Amazon EC2 Spot instances, choice of a wide variety of Amazon EC2 instances, including the memory optimized instances, and resize commands to easily add or remove instances from your cluster.
Presto uses a custom query execution engine with operators designed to support SQL semantics. Different from Hive/MapReduce, Presto executes queries in memory, pipelined across the network between stages, thus avoiding unnecessary I/O. The pipelined execution model runs multiple stages in parallel and streams data from one stage to the next as it becomes available.
Run interactive queries that directly access data in Amazon S3, save costs using Amazon EC2 Spot instance capacity, use Auto Scaling to dynamically add and remove capacity, and launch long-running or ephemeral clusters to match your workload. You can also add other Hadoop ecosystem applications on your cluster.

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?

- 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.

Explanation:-Elasticsearch provides full text search capability and is a fully managed AWS service.

Refer AWS documentation - Elasticsearch Integration
You can load streaming data into your Amazon Elasticsearch Service domain from many different sources. Some sources, like Amazon Kinesis Data Firehose and Amazon CloudWatch Logs, have built-in support for Amazon ES. Others, like Amazon S3, Amazon Kinesis Data Streams, and Amazon DynamoDB, use AWS Lambda functions as event handlers. The Lambda functions respond to new data by processing it and streaming it to your domain.
- Aggregate feedback in Amazon S3 using a columnar format with partitioning.

Q46)

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?

- 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.

Explanation:-S3 can provide the most cost-effective solution to store data while providing highly durable and highly flexible storage option with respect to the type and structure of data.

Q47)

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 table. Use Amazon S3 event notifications to trigger an AWS Lambda function that maps the S3 object to the task associated with it in DynamoDB. Once all task dependencies have been resolved, process the data with Amazon EMR.

Explanation:-The data dependency can be managed in DynamoDB. S3 event notifications can trigger Lambda functions to map the objects and check dependency. Once all satisfied, EMR job can be triggered.
- 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 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 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.

Q48)

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 data. Use 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.

Explanation:-S3 Server Side Encryption with S3 Managed Keys provide encryption. S3 ACLs allows fine grained control access and S3 to access logs would help provide traceability across all tools.
Use Server-Side Encryption with Amazon S3-Managed Keys (SSE-S3) – Each object is encrypted with a unique key. As an additional safeguard, it encrypts the key itself with a master key that it regularly rotates. Amazon S3 server-side encryption uses one of the strongest block ciphers available, 256-bit Advanced Encryption Standard (AES-256), to encrypt your data.
- 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.

Q49)

Your data warehouse tracks application errors over time. You must use AWS QuickSight to email a visualization of the trend of error events over a two-week rolling window.

What is the correct type of visualization to use for this type of data?

- A scatter plot showing the application errors over time
- A KPI showing the number of errors compared to expected value
- A heat map showing the application errors over time
- ✅ A line chart showing the application errors over time

Explanation:-Using a line chart is the correct way to visualize changes in values over time. The other answers do not show trend data over time.

Q50)

You manage a large data warehouse for sales data for a global company in the US, UK, Germany, and Japan. Your data warehouse consists of a 30-node cluster of ds2.xlarge Dense Storage nodes with a 2TB HDD, for a total capacity of 60TB. Your company has a single globally accessible instance of Microsoft Active Directory for user authentication and authorization. You must provide AWS QuickSight access to the sales data for the 40 users working in all four countries and implement the correct security model. The security model needs to prevent users from seeing any data outside their own region and ensure that users can see only row-level data for the sales events created within their own user group.

What is the least expensive and correct licensing model for this scenario?

- AWS QuickSight Standard license with Active Directory groups and row-level security
- ✅ AWS QuickSight Enterprise license with Active Directory groups and row-level security

Explanation:-A QuickSight Enterprise license is required to use Active Directory groups. Option A is incorrect because the Standard license does not provide row-level security. Option B is incorrect for the same reason and also because the Standard license does not provide access to Active Directory groups. Option D is incorrect because secure data encryption at rest does not provide row-level security.

- AWS QuickSight Standard license with IAM Credentials access, Reader role, and row-level security
- AWS QuickSight Enterprise license with Active Directory groups and secure data encryption at rest

---

Q51)

Your data warehouse tracks detailed customer geographic location of users of the company website over time. You must use AWS QuickSight to create a story showing the locations of users for all users during a week and the relative fraction of users for the top five countries for a four-week rolling window.

Which of the following pairs of visualizations should you use for this requirement?

- Combo chart and pie chart
- Combo chart and geospatial chart
- Heat map and scatter plot
- ✅ Geospatial chart and pie chart

Explanation:-A geospatial chart is the only QuickSight visualization that shows geographic location, and a pie chart shows relative amounts when precision is not important. The other charts listed either do not show geographic location or do not show relative amounts.

Q52)

You must create multiple visualizations with AWS QuickSight to show a company standardized analysis of recent log content in a Redshift data warehouse, plus historic log content stored in S3. The visualizations must be shared with end users in the business on a weekly basis.

What architecture and AWS QuickSight embedded analytics feature should be implemented?

- QuickSight sheets showing visualizations from Redshift and S3
- A QuickSight dashboard showing visualizations from Redshift and S3
- ✅ QuickSight sheets showing visualizations from an AWS Athena query against Redshift and S3

Explanation:-AWS QuickSight sheets deliver a set of visualizations, but they must be from the same data source, so AWS Athena is required to integrate data from Redshift and S3 before AWS QuickSight can use it.

- A QuickSight dashboard showing visualizations from an AWS Athena query against Redshift and S3

Q53)

Your company stores 1,100 TSV files in an S3 bucket. Each file has a unique name, and all the files are the same size (2.5MB).

What is the lowest-cost method to load this data into AWS QuickSight?

- Query the S3 bucket with AWS Athena and use Athena as the QuickSight data source.
- ✅ Set the S3 bucket as the single data source for AWS QuickSight.

Explanation:-AWS QuickSight can load all data from an S3 bucket, as long as the total data size is less than 25GB. Option is not possible because AWS QuickSight limits the maximum number of files on a manifest to 1,000—not 1,100. Option B is technically incorrect, and Option C is more expensive than simply loading the data directly into AWS QuickSight.

- List the TSV files in a JSON manifest and load the data into AWS QuickSight.
- Set the source of the dataset to the JSON manifest and load the data into AWS QuickSight.

Q54)

Your company collects clickstream data and dumps it on an hourly basis into an S3 bucket a uncompressed CSV files. The S3 bucket has a folder structure organized by site/YYYY/MM/DD. The bucket retains all the data for a rolling 365-day window before old data is archived to Glacier. In S3, the filename for each log file has the prefix clickstream_results followed by a timestamp in the format HH in 24-hour format, and each log file is approximately 850MB. You must design an AWS QuickSight system to show the hourly trend of clickstream history for three specific dates.

What is the simplest design to get the data into an AWS QuickSight embedded analytics story?

- Create an AWS Athena query to the S3 source, filtered on the three dates, and use that as input to the AWS QuickSight visualization.
- ✅ Set the AWS QuickSight source to S3 with a manifest file specifying only the required log files.

Explanation:-Defining an AWS manifest file is the simplest, lowest-cost, and fastest way to achieve this result because QuickSight can load only specific files from an S3 bucket when provided a simple JSON manifest with the S3 location and list of unique files. Options A and D are unnecessarily complex. Option B is not possible because QuickSight has a maximum data set limit of 25GB and the total size of the S3 bucket is 310GB (365 days' worth of data at 0.85GB file size per day).

- Set the AWS QuickSight source to S3 and create a QuickSight filter on the required dates.
- Create an AWS Athena query to the S3 source, use that as an input to the AWS QuickSight visualization, and then create a simple filter in AWS QuickSight to display only the required dates.

Q55)

A CEO asked a Big Data specialist about the best architecture for an ad hoc SQL-based pipeline that can both be queried by the team and used to feed data into a Big Data application.

The data lives in Amazon S3. What is the best answer?

- Use AWS SageMaker and AWS Lambda.
- ✅ Use AWS Athena and AWS Glue.

Explanation:-The best solution to support both ad hoc querying of data via SQL and to allow that same data to be sent to a Big Data pipeline is to use AWS Athena and AWS Glue. AWS Athena can perform ad hoc queries, and AWS Glue can do the ETL.

- Use AWS Glue and AWS DynamoDB.
- Use AWS Elasticsearch and AWS Lambda.

Q56)

You have more than 25PB of data in a private cloud. This quarter, you want to use AWS Snowball to move 12PB of data to AWS. This data will be stored in S3, and EMR will be used to process the data. The ideal average file size after compression will be 9GB.

What method would accomplish this task while optimizing for cost, bandwidth, and storage?

- ✅ Partition the data by day, month, and year and compress the data using the bzip2 algorithm.

Explanation:-It handles all the requirements, including splitting and compressing the data. Other options either do not support file splitting or do not support compression.

- Partition the data by year and compress the data using snappy in high-compression mode.
- Partition the data by month and year and use Avro in high-compression mode.
- Partition the data by day, month, and year and compress the data using gzip and KMS.

Q57)

A data engineer is building a financial services analytics store on Amazon EMR and is concerned about the Amazon S3 eventual consistency model. The application has multiple S3ServiceException responses per minute. The exception response, when queried in CloudWatch, says that the specified key does not exist.

How can this issue be solved as quickly as possible?

- Enable consistent bucket view in S3 to fix issues with distributed writes and reads.
- The wrong version of Hadoop is running on EMR. Upgrade it by sshing into the deployed nodes.
- This is an application exception in the Python code. Have a developer fix it.
- ✅ This may be an eventual consistency issue in Amazon S3. Enable the Amazon EMRFS Consistent View feature.

Explanation:-Amazon S3 has eventual consistency, so use the Amazon EMRFS Consistent View feature.

Q58)

A SaaS backup company is using AWS Glacier as part of its technology stack. The company needs to meet the requirements of the terms of service it created, including securely storing the data for 10 years and making it possible to recall the data in minutes.

What workflow describes the best solution?

- Store the data in S3 and archive it immediately to AWS Glacier. Retrieve the data with Data Pipeline.
- Store the data in S3 and archive it immediately to AWS Glacier. Retrieve the data with EMR.
- Store the data in S3 and archive it after 10 years to AWS Glacier. Retrieve the data with AWS Glue.
- ✅ Store the data in S3 and archive it after one year to AWS Glacier. Retrieve the data using Expedited retrieval if data is older than one year.

Explanation:-The scenario that fits the requirement is to retrieve the data using the Expedited retrieval tier from Glacier.

Q59)

A Big Data specialist notices that 20 instances at a time are being launched in AWS regions after a co-worker gives a tech talk on GitHub coding practices.

Within minutes of the talk, hundreds of instances are running. How can this be fixed?

- All of these.
- ✅ Look in CloudWatch and observe scripted behavior by hackers who have stolen AWS credentials. Deactivate the API keys and accounts of compromised users.

Explanation:-This behavior is common when API keys are compromised. It is best to validate the behavior in CloudWatch and then disable API keys and compromised accounts.

- Terminate all API keys.

- Terminate all EC2 instances.

Q60)

A Big Data specialist is looking to create a data lake on AWS for a healthcare company. An executive has asked for a justification of this approach.

What should the Big Data specialist highlight?

- Ability to work on data where it resides
- ✔ All of these

Explanation:-A data lake can store structured, unstructured, and semi-structured data. It can be used for analytics as well as Big Data and Machine Learning (ML). It can also work on data without data movement. Finally, it is low-cost storage.

- Possible use for analytics, Big Data, and Machine Learning
- Ability to store structured, semi-structured, and unstructured data