Answer Sheet

Q1)

Your company sells consumer devices and needs to record the first activation of all sold devices. Devices are not activated until the information is written on a persistent database. Activation data is very important for your company and must be analyzed daily with a MapReduce job. The execution time of the data analysis process must be less than three hours per day. Devices are usually sold evenly during the year, but when a new device model is out, there is a predictable peak in activation's, that is, for a few days there are 10 times or even 100 times more activation's than in average day.

Which of the following databases and analysis framework would you implement to better optimize costs and performance for this workload?

● Amazon DynamoDB and Amazon Elastic MapReduce with Reserved instances

● Amazon RDS and Amazon Elastic MapReduce with Reserved instances.

✓ Amazon DynamoDB and Amazon Elastic MapReduce with Spot instances.

Explanation:-Key point here is to optimize cost and performance only for the increased workload only, not the existing one
DynamoDB would be preferred over RDS for the throughput supported and Spot instances to reduce cost and handle the temporary workload.

● Amazon RDS and Amazon Elastic MapReduce with Spot instances.

Q2)

You need to perform ad-hoc business analytics queries on well-structured data. Data comes in constantly at a high velocity.

Your business intelligence team can understand SQL.

What AWS service(s) should you look to first?

● EMR using Hive

✓ Kinesis Firehose + Redshift

Explanation:-Key point is perform ad-hoc analytics with data at high velocity
Kinesis Firehose provides a managed service for aggregating streaming data and inserting it into RedShift. RedShift also supports ad-hoc queries over well-structured data using a SQL-compliant wire protocol, so the business team should be able to adopt this system easily.

● Kinesis Firehose + RDS

● EMR running Apache Spark

Q3)

A Company has two batch processing applications that consume financial data about the day's stock transactions. Each transaction needs to be stored durably and guarantee that a record of each application is delivered so the audit and billing batch processing applications can process the data.

However, the two applications run separately and several hours apart and need access to the same transaction information. After reviewing the transaction information for the day, the information no longer needs to be stored.

What is the best way to architect this application? Choose the correct answer from the options below

● Store the transaction information in a DynamoDB table. The billing application can read the rows while the audit application will read the rows them remove the data.

✓ Use Kinesis to store the transaction information. The billing application will consume data from the stream, the audit application can consume the same data several hours later.

Explanation:-The key point here is batch application and message being stored durably and delivery guarantee. Kinesis can store the data durably and allow access to multiple consumers without any dependencies.

● Use SQS for storing the transaction messages. When the billing batch process consumes each message, have the application create an identical message and place it in a different SQS for the audit application to use several hours later.

● Use SQS for storing the transaction messages; when the billing batch process performs first and consumes the message, write the code in a way that does not remove the message after consumed, so it is available for the audit application several hours later. The audit application can consume the SQS message and remove it from the queue when completed.

Q4)

A research scientist is planning for the one-time launch of an Elastic MapReduce cluster and is encouraged by her manager to minimize the costs. The cluster is designed to ingest 200TB of genomics data with a total of 100 Amazon EC2 instances and is expected to run for around four hours.

The resulting data set must be stored temporarily until archived into an Amazon RDS Oracle instance.

Which option will help save the most money while meeting requirements?

● Deploy on-demand master, core and task nodes and store ingest and output files in Amazon S3 RRS

● Store the ingest files in Amazon S3 RRS and store the output files in S3. Deploy Reserved Instances for the master and core nodes and on-demand for the task nodes.

● Optimize by deploying a combination of on-demand, RI and spot-pricing models for the master, core and task nodes. Store ingest and output files in Amazon S3 with a lifecycle policy that archives them to Amazon Glacier.

✓ Store ingest and output files in Amazon S3. Deploy on-demand for the master and core nodes and spot for the task nodes.

Explanation:-Key point here is to save most money while being able to process the huge data.
It follows best practice of using On demand for master and core and spot for task nodes also help reduce cost using spot instances.

Q5)

Your company is storing millions of sensitive transactions across thousands of 100-GB files that must be encrypted in transit and at rest.

Analysts concurrently depend on subsets of files, which can consume up to 5TB of space, to generate simulations that can be used to steer business decisions.

You are required to design an AWS solution that can cost effectively accommodate the long-term storage and in-flight subsets of data.

● Store the full data set in encrypted Amazon Elastic Block Store (EBS) volumes, and regularly capture snapshots that can be cloned to EC2 workstations
● Use HDFS on Amazon Elastic MapReduce (EMR), and run simulations on subsets in-memory on Amazon Elastic Compute Cloud (EC2).
● Use HDFS on Amazon EMR, and run simulations on subsets in ephemeral drives on Amazon EC2.
● Use Amazon S3 with server-side encryption, and run simulations on subsets in-memory on Amazon EC2.
✓ Use Amazon Simple Storage Service (S3) with server-side encryption, and run simulations on subsets in ephemeral drives on Amazon EC2.
Explanation:-The S3 with SSE provides encryption at rest and HTTPS can be used to push data to S3 for encryption in transit. S3 provides an option for cost effective long term storage. Ephemeral drives would help run simulations and the data would lost once the EC2 instance is terminated.

Q6)

A customer's nightly EMR job processes a single 2-TB data file stored on Amazon Simple Storage Service (S3).

The Amazon Elastic Map Reduce (EMR) job runs on two On-Demand core nodes and three On-Demand task nodes.

Which of the following may help reduce the "EMbR job completion time"?

✓ Adjust the number of simultaneous mapper tasks.
Explanation:-Adjusting and tuning the number of simultaneous mapper task would help reduce time
● Launch the core nodes and task nodes within an Amazon Virtual Cloud.
● Use a bootstrap action to present the S3 bucket as a local filesystem.
● Use three Spot Instances rather than three On-Demand instances for the task nodes.
✓ Change the input split size in the MapReduce job configuration.
Explanation:-The split size of the match in memory block size of task and HDFS files will help to complete the job faster.
● Enable termination protection for the job flow.

Q7)

An online gaming company uses DynamoDB to store user activity logs and is experiencing throttled writes on the company's DynamoDB table.

The company is NOT consuming close to the provisioned capacity. The table contains a large number of items and is partitioned on user and sorted by date. The table is 200GB and is currently provisioned at 10K WCU and 20K RCU.

Which two additional pieces of information are required to determine the cause of the throttling? (Choose two.)

✓ The structure of any LSIs that have been defined on the table
Explanation:-An LSI consumes WCU for writes on the primary table.
● Application-level metrics showing the average item size and peak update rates for each attribute
✓ CloudWatch data showing consumed and provisioned write capacity when writes are being throttled
Explanation:-CloudWatch helps shows the stats for consumed vs provisioned throughput capacity.
● The structure of any GSIs that have been defined on the table
● The maximum historical WCU and RCU for the table

Q8)

A city has been collecting data on its public bicycle share program for the past three years. The 5PB dataset currently resides on Amazon S3.

The data contains the following datapoints:

- Bicycle origination points

- Bicycle destination points

- Mileage between the points

- Number of bicycle slots available at the station (which is variable based on the station location)

- Number of slots available and taken at a given time

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

`The new bicycle stations must be located to provide the most riders access to bicycles. How should this task be performed?

✓ Keep the data on Amazon S3 and use an Amazon EMR-based Hadoop cluster with spot instances to run a Spark job that performs a stochastic gradient descent optimization over EMRFS.
Explanation:-The data is already hosted in S3, EMR with EMRFS can be used to perform analysis.
● Persist the data on Amazon S3 and use a transient EMR cluster with spot instances to run a Spark streaming job that will move the data into Amazon Kinesis.
● Use the Amazon Redshift COPY command to move the data from Amazon S3 into Redshift and perform a SQL query that outputs the most popular bicycle stations.
● Move the data from Amazon S3 into Amazon EBS-backed volumes and use an EC2 based Hadoop cluster with spot instances to run a Spark job that performs a stochastic gradient descent optimization.

Q9)

An administrator is deploying Spark on Amazon EMR for two distinct use cases: machine learning algorithms and ad-hoc querying.

All data will be stored in Amazon S3. Two separate clusters for each use case will be deployed. The data volumes on Amazon S3 are less than 10 GB.

How should the administrator align instance types with the cluster's purpose?

● Machine Learning on D instance types and ad-hoc queries on I instance types
● Machine Learning on T instance types and ad-hoc queries on M instance types
● Machine Learning on R instance types and ad-hoc queries on G2 instance types
✓ Machine Learning on C instance types and ad-hoc queries on R instance types

Explanation:-Machine learning are usually compute intensive and adhoc queries are suited for memory optimized.
For memory-intensive applications, prefer R type instances over the other instance types. For compute-intensive applications, prefer C type instances. For applications balanced between memory and compute, prefer M type general-purpose instances.
Compute Optimized - High performance web servers, scientific modelling, batch processing, distributed analytics, high-performance computing (HPC), machine/deep learning inference, ad serving, highly scalable multiplayer gaming, and video encoding.
Memory Optimized - Instances are well suited for memory intensive applications such as high performance databases, distributed web scale in-memory caches, mid-size in-memory databases, real time big data analytics, and other enterprise applications.

Q10)

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

Recently, it has been observed that the file sizes vary, and the EMR jobs take too long.

The distributor needs to tune and optimize the data processing workflow with this limited information to improve the performance of the EMR job.

Which recommendation should an administrator provide?

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

Explanation:-Gzip is not ideal compression for files larger than 1GB and compression technique should be checked with supports splitting like bzip2 or one with higher compression handling like Snappy. Depending on how large your aggregated data files are, the compression algorithm becomes an important choice. For instance, if you are aggregating your data (using the ingest tool of your choice) and the aggregated data files are between 500 MB to 1 GB, GZIP compression is an acceptable data compression type. However, if your data aggregation creates files larger than 1 GB, its best to pick a compression algorithm that supports splitting.

● Use Avro rather than gzip for the archives.

Q11)

Your application generates a 1 KB JSON payload that needs to be queued and delivered to EC2 instances for applications. At the end of the day, the application needs to replay the data for the past 24 hours.

In the near future, you also need the ability for other multiple EC2 applications to consume the same stream concurrently.

What is the best solution for this?

● Kinesis Firehose
✓ Kinesis Data Streams

Explanation:-Kinesis Data Streams allows the ability for replaying the data as well access to the same data to multiple Kinesis client applications.
Amazon Kinesis Data Streams enables you to build custom applications that process or analyze streaming data for specialized needs. You can continuously add various types of data such as clickstreams, application logs, and social media to an Amazon Kinesis data stream from hundreds of thousands of sources. Within seconds, the data will be available for your Amazon Kinesis Applications to read and process from the stream.
Amazon Kinesis Data Streams enables real-time processing of streaming big data. It provides ordering of records, as well as the ability to read and/or replay records in the same order to multiple Amazon Kinesis Applications. The Amazon Kinesis Client Library (KCL) delivers all records for a given partition key to the same record processor, making it easier to build multiple applications reading from the same Amazon Kinesis data stream (for example, to perform counting, aggregation, and filtering).
Amazon Simple Queue Service (Amazon SQS) offers a reliable, highly scalable hosted queue for storing messages as they travel between computers. Amazon SQS lets you easily move data between distributed application components and helps you build applications in which messages are processed independently (with message-level ack/fail semantics), such as automated workflows.

● SNS
● SQS

Q12)

You are deploying an application to track GPS coordinates of delivery trucks in the United States.Coordinates are transmitted from each delivery truck once every three seconds.

You need to design an architecture that will enable real-time processing of these coordinates from multiple consumers.

Which service should you use to implement data ingestion?

● Amazon Simple Queue Service
● Amazon AppStream
● AWS Data Pipeline
✓ Amazon Kinesis

Explanation:-Key point here is address real time data ingestion. Amazon Kinesis is a platform for streaming data on AWS, making it easy to load and analyze streaming data, and also providing the ability for you to build custom streaming data applications for specialized needs. Use Amazon Kinesis Streams to collect and process large streams of data records in real time. Use Amazon Kinesis Firehose to deliver real-time streaming data to destinations such as Amazon S3 and Amazon Redshift. Use Amazon Kinesis Analytics to process and analyze streaming data with standard SQL.

Q13)

You are using IOT sensors to monitor the movement of a group of hikers on a three day trek and send the information into an Kinesis Stream. They each have a sensor in their shoe and you know for certain that there is no problem with mobile coverage so all the data is getting back to the stream. You have used default settings for the stream.

At the end of the third day the data is sent to an S3 bucket. When you go to interpret the data in S3 there is only data for the last day and nothing for the first 2 days.

Which of the following is the most probable cause of this?

● A sensor probably stopped working on the second day. If one sensor fails, no data is sent to the stream until that sensor is fixed
● You cannot send Kinesis data to the same bucket on consecutive days if you do not have versioning enabled on the bucket. If you don't have versioning enabled you would need to define 3 different buckets or else the data is overwritten each day
✓ Data records are only accessible for a default of 24 hours from the time they are added to a stream.
Explanation:-By default, Kinesis stores the records for 24 hours only. By default, Records of a stream are accessible for up to 24 hours from the time they are added to the stream. You can raise this limit to up to 7 days by enabling extended data retention.
● Temporary loss of mobile coverage; although mobile coverage was good in the area, even temporary loss of data will stop the streaming

Q14)

A utility company is building an application that stores data coming from more than 10,000 sensors. Each sensor has a unique ID and will send a datapoint (approximately 1KB) every 10 minutes throughout the day. Each datapoint contains the information coming from the sensor as well as a timestamp. This company would like to query information coming from a particular sensor for the past week very rapidly and want to delete all the data that is older than 4 weeks.

Using Amazon DynamoDB for its scalability and rapidity, how do you implement this in the most cost effective way?

✓ One table for each week, with a primary key that is the sensor ID and a sort key that is the timestamp
Explanation:-Composite key with Sensor ID and timestamp would help for faster queries
● One table for each week, with a primary key that is the concatenation of the sensor ID and timestamp
● One table, with a primary key that is the concatenation of the sensor ID and timestamp
● One table, with a primary key that is the sensor ID and a sort key that is the timestamp

Q15)

You need to provide customers with rich visualizations that allow you to easily connect multiple disparate data sources in S3, Redshift, and several CSV files.

Which tool should you use that requires the least setup?

● Redshift
✓ QuickSight
Explanation:-QuickSight provides visualization capability with integration with RDS, Redshift.
Amazon QuickSight is a fast, cloud-powered business intelligence service that makes it easy to deliver insights to everyone in your organization.
As a fully managed service, QuickSight lets you easily create and publish interactive dashboards that include ML Insights. Dashboards can then be accessed from any device, and embedded into your applications, portals, and websites.
QuickSight allows you to directly connect to and import data from a wide variety of cloud and on-premises data sources. These include SaaS applications such as Salesforce, Square, ServiceNow, Twitter, Github, and JIRA; 3rd party databases such as Teradata, MySQL, Postgres, and SQL Server; native AWS services such as Redshift, Athena, S3, RDS, and Aurora; and private VPC subnets. You can also upload a variety of file types including Excel, CSV, JSON, and Presto.
● Hue on EMR
● Elasticsearch

Q16)

You've been asked by the VP of People to showcase the current breakdown of the headcount for each department within your organization.

What chart do you select to do this to make it easy to compare each department?

● Column chart
✓ Pie chart
Explanation:-Pie charts are best to use when you are trying to compare parts of a whole, which is ideal for the use case. They do not show changes over time.
● Line chart
● Scatter plot

Q17)

Management has requested a comparison of total sales performance in the five North American regions in January.

They're hoping to determine how to allocate a budget to regions based on performance in that single period.

What sort of visualization do you use in Amazon QuickSight?

● Stacked area chart
● Line chart
✓ Bar chart
Explanation:-Bar Chart can be used to represent the data for comparison in sales for each region.

● Histogram

Q18)

A new client is requesting a tool that will provide fast query performance for enterprise reporting and business intelligence workloads, particularly those involving extremely complex SQL with multiple joins and sub-queries.

They also want the ability to give analysts access to a central system through tradition SQL clients that allow them to explore and familiarize themselves with the data.

What solution do you initially recommend they investigate?

● Athena
✓ Redshift
Explanation:-Redshift is a fully managed data warehousing solution providing standard SQL interface and ability to run complex queries. Amazon Redshift is a fast, fully managed data warehouse that makes it simple and cost-effective to analyze all your data using standard SQL and your existing Business Intelligence (BI) tools. It allows you to run complex analytic queries against petabytes of structured data, using sophisticated query optimization, columnar storage on high-performance local disks, and massively parallel query execution.
● SQS
● EMR

Q19)

Your company is in the process of developing a next generation pet collar that collects biometric information to assist families with promoting healthy lifestyles for their pets. Each collar will push 30kb of biometric data In JSON format every 2 seconds to a collection platform that will process and analyze the data providing health trending information back to the pet owners and veterinarians via a web portal Management has tasked you to architect the collection platform ensuring the following requirements are met.

Provide the ability for real-time analytics of the inbound biometric data to ensure processing of the biometric data is highly durable, Elastic and parallel. The results of the analytic processing should be persisted for data mining.

Which architecture outlined below will meet the initial requirements for the collection platform?

● Utilize EMR to collect the inbound sensor data, analyze the data from EMR with Amazon Kinesis and save the results to DynamoDB.
● Utilize SQS to collect the inbound sensor data analyze the data from SQS with Amazon Kinesis and save the results to a Microsoft SQL Server RDS instance.
● Utilize S3 to collect the inbound sensor data analyze the data from S3 with a daily scheduled Data Pipeline and save the results to a Redshift Cluster.
✓ Utilize Amazon Kinesis to collect the inbound sensor data, analyze the data with Kinesis clients and save the results to a Redshift cluster using EMR.
Explanation:-Key point here to architect durable collection platform with real time analytics, data mining storage.
Kinesis to capture the data in a elastic, durable and parallel manner. Analyze data with Kinesis clients and store data to Redshift for data mining using EMR.

Q20)

A company is developing a video application that will emit a log stream. Each record in the stream may contain up to 400 KB of data.

To improve the video-streaming experience, it is necessary to collect a subset of metrics from the stream to be analyzed for trends over time using complex SQL queries. A Solutions Architect will create a solution that allows the application to scale without customer interaction.

Which solution should be implemented to meet these requirements?

● Send the log data to an Amazon SQS standard queue. Make the queue an event source for an AWS Lambda function that transforms the data and stores it in Amazon Redshift. Query the data in Amazon Redshift.
● Send the log data to an Amazon CloudWatch Logs log group. Make the log group an event source for an AWS Lambda function that transforms the data and stores it in an Amazon S3 bucket. Query the data with Amazon Athena.
✓ Send the log data to an Amazon Kinesis data stream. Subscribe an AWS Lambda function to the stream that transforms the data and sends it to a second data stream. Use Amazon Kinesis Data Analytics to query the data in the second stream.
Explanation:-Data can be captured using Kinesis Data Stream and Kinesis Data Analytics can be used to query on the streaming data using time or window queries to generate trend analysis.
Refer AWS documentation - Streaming Analytics Pipeline
Many Amazon Web Services (AWS) customers use streaming data to gain real-time insight into customer activity and immediate business trends. Streaming data, which is generated continuously from thousands of data sources, includes a wide variety of data such as log files from your mobile or web applications, e-commerce purchases, in-game player activity, information from social networks, financial trading floors, or geospatial services, and telemetry from connected devices. This data can help companies make well-informed decisions and proactively respond to changing business conditions.
Amazon Kinesis, a platform for streaming data on AWS, offers powerful services that make it easier to build data processing applications, load massive volumes of streaming data, and analyze it in real time.
● Send the log data to an Amazon Kinesis Data Firehose delivery stream. Use an AWS Lambda function to transform the data. Deliver the data to Amazon Redshift. Query the data in Amazon Redshift.

Q21)

Your website is serving on-demand training videos to your workforce. Videos are uploaded monthly in high resolution MP4 format.

Your workforce is distributed globally often on the move and using company-provided tablets that require the HTTP Live Streaming (HLS) protocol to watch a video.

Your company has no video transcoding expertise and it required you might need to pay for a consultant.

How do you implement the most cost-efficient architecture without compromising high availability and quality of video delivery?

● A video transcoding pipeline running on EC2 using SQS to distribute tasks and Auto Scaling to adjust the number of nodes depending on the length of the queue. EBS volumes to host videos and EBS snapshots to incrementally backup original files after a few days. CloudFront to serve HLS transcoded videos from EC2
● Elastic Transcoder to transcode original high-resolution MP4 videos to HLS EBS volumes to host videos and EBS snapshots to incrementally backup original rues after a few days. CloudFront to serve HLS transcoded videos from EC2.
● A video transcoding pipeline running on EC2 using SQS to distribute tasks and Auto Scaling to adjust the number or nodes depending on the length of the queue S3 to host videos with Lifecycle Management to archive all files to Glacier after a few days CloudFront to serve HLS transcoding videos from Glacier
✓ Elastic Transcoder to transcode original high-resolution MP4 videos to HLS. S3 to host videos with lifecycle Management to archive original flies to Glacier after a few days. CloudFront to serve HLS transcoded videos from S3
Explanation:-Key here the cost efficient solution with company needing video transcoding expertise and needing to hire a consultant with global distribution.
Elastic Transcoder provides and out of box option to transcode videos into any format without any expertise. S3 to host videos and CloudFront to serve HLS transcoded videos for global distribution while being cost efficient

---

Q22)

Your company releases new features with high frequency while demanding high application availability.

As part of the application's A/B testing, logs from each updated Amazon EC2 instance of the application need to be analyzed in near real-time, to ensure that the application is working flawlessly after each deployment.

If the logs show any anomalous behavior, then the application version of the instance is changed to a more stable one.

Which of the following methods should you use for shipping and analyzing the logs in a highly available manner?

● Store the logs locally on each instance and then have an Amazon Kinesis stream pull the logs for live analysis
● Ship the logs to a large Amazon EC2 instance and analyze the logs in a live manner.
● Ship the logs to Amazon CloudWatch Logs and use Amazon EMR to analyze the logs in a batch manner each hour.
✓ Ship the logs to an Amazon Kinesis stream and have the consumers analyze the logs in a live manner.
Explanation:-Data can be ingested into the Kinesis streams using agents and the logs can then be analyzed real time.
Refer AWS documentation - Kinesis Serverless log Analytics
Amazon Kinesis Streams enables you to build custom applications that process or analyze streaming data for specialized needs. Amazon Kinesis Streams can continuously capture and store terabytes of data per hour from hundreds of thousands of sources such as website clickstreams, financial transactions, social media feeds, IT logs, and location-tracking events.
● Ship the logs to Amazon S3 for durability and use Amazon EMR to analyze the logs in a batch manner each hour.

Q23)

A company needs to deploy a data lake solution for their data scientists in which all company data is accessible and stored in a central S3 bucket.

The company segregates the data by business unit, using specific prefixes. Scientists can only access the data from their own business unit.

The company needs a single sign-on identity and management solution based on Microsoft Active Directory (AD) to manage access to the data in Amazon S3.

Which method meets these requirements?

● Use Amazon S3 API integration with AD to impersonate the users on access in a transparent manner.
● Deploy the AD Synchronization service to create AWS IAM users and groups based on AD information.
● Create bucket policies that only allow access to the authorized prefixes based on the users' group name in Active Directory.
✓ Use AWS IAM Federation functions and specify the associated role based on the users' groups in AD.
Explanation:-Identity Federation allows organizations to associate temporary credentials to users authenticated through an external identity provider such as Microsoft Active Directory (AD). These temporary credentials are linked to AWS IAM roles that grant access to the S3 bucket.

Q24)

A web application emits multiple types of events to Amazon Kinesis Streams for operational reporting.

Critical events must be captured immediately before processing can continue, but informational events do not need to delay processing.

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

✓ Log critical events using the PutRecords API method, and log informational events using the Kinesis Producer Library.
Explanation:-The core of this question is how to send event messages to Kinesis at real time. The critical events must be sent without any delay, and the informational events can be sent using batching. The Kinesis Producer Library (KPL) buffers records, so it can be used for the informational messages. PutRecords is a synchronous real time send function, so it must be used for the critical events.
Refer AWS documentation - Developing Producers using KPL
Because the KPL may buffer records before sending them to Kinesis Data Streams, it does not force the caller application to block and wait for a confirmation that the record has arrived at the server before continuing execution. A call to put a record into the KPL always returns immediately and does not wait for the record to be sent or a response to be received from the server. Instead, a Future object is created that receives the result of sending the record to Kinesis Data Streams at a later time. This is the same behavior as asynchronous clients in the AWS SDK.
● Log critical events using the Kinesis Producer Library, and log informational events using the PutRecords API method.
● Log all events using the Kinesis Producer Library.
● Log all events using the PutRecords API method.

Q25)

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

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

● Connect to the source cluster over an SSL client connection, and write data records to Amazon Kinesis Firehose to load into your target data warehouse.
● Run an UNLOAD command that stores the data in an S3 bucket encrypted with an AWS KMS data key; run a COPY command to move the data into the target cluster.
✓ Use AWS KMS data key to run an UNLOAD ENCRYPTED command that stores the data in an unencrypted S3 bucket; run a COPY command to move the data into the target cluster.
Explanation:-The UNLOAD ENCRYPTED command automatically stores the data encrypted using-client side encryption and uses HTTPS to encrypt the data during the transfer to S3.
Refer AWS documentation - Redshift Unloading Data
UNLOAD automatically creates files using Amazon S3 server-side encryption with AWS-managed encryption keys (SSE-S3). You can also specify server-side encryption with an AWS Key Management Service key (SSE-KMS) or client-side encryption with a customer-managed key (CSE-CMK).
● Run an ETL process that connects to the source clusters using SSL to issue a SELECT query for new data, and then write to the target data warehouse using an INSERT command over another SSL secured connection.

Q26)

A data engineer needs to architect a data warehouse for an online retail company to store historic purchases. The data engineer needs to use Amazon Redshift.

To comply with PCI:DSS and meet corporate data protection standards, the data engineer must ensure that data is encrypted at rest and that the keys are managed by a corporate on-premises HSM.

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

● Use AWS Import/Export to import the corporate HSM device into the AWS Region where the Amazon Redshift cluster will launch, and configure Redshift to use the imported HSM.
● 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 VPC, and then establish a VPN connection between the VPC and the on-premises network. Launch the Amazon Redshift cluster in the VPC, and configure it to use your corporate HSM.
Explanation:-Amazon Redshift can use an on-premises HSM for key management over the VPN, which ensures that the encryption keys are locally managed.

Q27)

A data engineer in a manufacturing company is designing a data processing platform that receives a large volume of unstructured data.

The data engineer must populate a well-structured star schema in Amazon Redshift.

What is the most efficient architecture strategy for this purpose?

● Load the unstructured data into Redshift, and use string parsing functions to extract structured data for inserting into the analysis schema.
● When the data is saved to Amazon S3, use S3 Event Notifications and AWS Lambda to transform the file contents. Insert the data into the analysis schema on Redshift.
● Normalize the data using an AWS Marketplace ETL tool, persist the results to Amazon S3, and use AWS Lambda to INSERT the data into Redshift.
✓ Transform the unstructured data using Amazon EMR and generate CSV data. COPY the CSV data into the analysis schema within Redshift.
Explanation:-The data volume is large, it can be processed using EMR to generate structured CSV data and then load the data into Redshift.
Refer AWS documentation - Data Warehousing on AWS
Data in Amazon Redshift must be structured by a defined schema. Amazon Redshift doesn't support an arbitrary schema structure for each row. If your data is unstructured, you can perform extract, transform, and load (ETL) on Amazon EMR to get the data ready for loading into Amazon Redshift. For JSON data, you can store key value pairs and use the native JSON functions in your queries.

Q28)

A company has several teams of analysts. Each team of analysts has their own cluster.

The teams need to run SQL queries using Hive, Spark-SQL, and Presto with Amazon EMR.

The company needs to enable a centralized metadata layer to expose the Amazon S3 objects as tables to the analysts.

Which approach meets the requirement for a centralized metadata layer?

● Naming scheme support with automatic partition discovery from Amazon S3
● s3distcp with the output Manifest option to generate RDS DDL
✓ Bootstrap action to change the Hive Metastore to an Amazon RDS database
Explanation:-The metastore needs to be centralized and accessed from multiple clusters, it can be externalized into RDS. It also supports Spark, Hive and Presto.
Refer AWS documentation - EMR Metastore External Hive @ https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-metastore-external-hive.html
By default, Hive records metastore information in a MySQL database on the master node's file system. The metastore contains a description of the table and the underlying data on which it is built, including the partition names, data types, and so on. When a cluster terminates, all cluster nodes shut down, including the master node. When this happens, local data is lost because node file systems use ephemeral storage. If you need the

metastore to persist, you must create an external metastore that exists outside the cluster.
You have two options for an external metastore:
AWS Glue Data Catalog (Amazon EMR version 5.8.0 or later only). For more information, see Using the AWS Glue Data Catalog as the Metastore for Hive. @ https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive-metastore-glue.html
● EMRFS consistent view with a common Amazon DynamoDB table

Q29)

A Redshift data warehouse has different user teams that need to query the same table with very different query types.

These user teams are experiencing poor performance.

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

● Maintain team-specific copies of the table.
✅ Add interleaved sort keys per team.
Explanation:-Multiple teams query different columns with different queries it would be best to use Interleaved keys to improve performance. Interleaved keys are provided to help with the limitations of compound keys. They are designed to weigh each column in the key evenly, allowing improved performance regardless of which columns in the key you're filtering.
Refer AWS documentation - Redshift Interleaved Sort Keys
An interleaved sort gives equal weight to each column, or subset of columns, in the sort key. If multiple queries use different columns for filters, then you can often improve performance for those queries by using an interleaved sort style. When a query uses restrictive predicates on secondary sort columns, interleaved sorting significantly improves query performance as compared to compound sorting.
● Create custom table views.
● Add support for workload management queue hopping.

Q30)

An administrator needs to design the event log storage architecture for events from mobile devices.

The event data will be processed by an Amazon EMR cluster daily for aggregated reporting and analytics before being archived.

How should the administrator recommend storing the log data?

● Create an Amazon DynamoDB table partitioned on EventID, write log data to table. Execute the EMR job on the table.
● Create an Amazon DynamoDB table partitioned on the device and sorted on date, write log data to table. Execute the EMR job on the Amazon DynamoDB table.
✅ Create an Amazon S3 bucket and write data into folders by day. Execute the EMR job on the daily folder.
Explanation:-EMR jobs needs to process daily data, it would be best to partition the data by day.
Refer AWS documentation - EMR Best Practices
Data partitioning is an essential optimization to your data processing workflow. Without any data partitioning in place, your data processing job needs to read or scan all available data sets and apply additional filters in order to skip unnecessary data. Such architecture might work for a low volume of data, but scanning the entire data set is a very time consuming and expensive approach for larger data sets. Data partitioning lets you create unique buckets of data and eliminate the need for a data processing job to read the entire data set.
Three considerations determine how you partition your data:
• Data type (time series)
• Data processing frequency (per hour, per day, etc.)
• Data access and query pattern (query on time vs. query on geo location)
For instance, if you are processing a time-series data set where you need to process your data once every hour and your data-access pattern is based on time, partitioning your data based on date makes the most sense. An example of such data processing would be processing your daily logs. If you have incoming logs from variety of data sources (web servers, devices etc.), then creating partitions of data based on the hour of the day gives you a date-based partitioning scheme.
The structure of such partitioning scheme will look similar to the following:
/data/logs/YYYY-MM-DD-HH/logfiles for this given hour, where YYYY-MM-DD-HH changes based on the current log ingest time.
● Create an Amazon S3 bucket and write log data into folders by device. Execute the EMR job on the device folders.

Q31)

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

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

● Write them to a file on Amazon Simple Storage Service (S3). Write an AWS Lambda function that runs in response to the S3 event to load the events into Amazon Elasticsearch Service for analysis.
● Write them to the local disk and configure the Amazon CloudWatch Logs agent to load the data into CloudWatch Logs and subsequently into Amazon Elasticsearch Service.
● Write them to CloudWatch Logs and use an AWS Lambda function to load them into HDFS on an Amazon Elastic MapReduce (EMR) cluster for analysis.
✅ Write them directly to a Kinesis Firehose. Configure Kinesis Firehose to load the events into an Amazon Redshift cluster for analysis.
Explanation:-The simplest would be to use Firehose to stream data to collect the logs and load the data to Redshift for analysis.
Refer AWS documentation - Kinesis Data Firehose
Amazon Kinesis Data Firehose is the easiest way to reliably load streaming data into data stores and analytics tools. It can capture, transform, and load streaming data into Amazon S3, Amazon Redshift, Amazon Elasticsearch Service, and Splunk, enabling near real-time analytics with existing business intelligence tools and dashboards you're already using today. It is a fully managed service that automatically scales to match the throughput of your data and requires no ongoing administration. It can also batch, compress, transform, and encrypt the data before loading it, minimizing the amount of storage used at the destination and increasing security.

Q32)

A system needs to collect on-premises application spool files into a persistent storage layer in AWS. Each spool file is 2 KB.

The application generates 1 M files per hour. Each source file is automatically deleted from the local server after an hour.

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

● Copy files to Amazon S3 Standard Storage.
● Write file contents to Amazon ElastiCache.
✓ Write file contents to an Amazon DynamoDB table.
Explanation:-The provisioned throughput required for DynamoDB would be most cost efficient as compared to the PUT requests for S3.
Total Storage = 2KB * 1M/hour = 2GB/hour * 24 * 30 = 1440 GB = 1.4TB
DynamoDB : 1M/hour = 277 Writes per Second * 2KB = 554 WCU = ~$623 (with Storage)
S3 : 1M/hour * 24 * 30 = 720 Million PUT request = ~ $3633.12 (with Storage)
● Copy files to Amazon S3 infrequent Access Storage.

Q33)

Your client has a high-volume DynamoDB table that serves comment information to an internal API.

Currently, the table allows you to query with a composite primary key with postId as a partition key and commentId as a sort key. Application validation ensures that each item has other fields including timestamp, userId, and sentimentScore.

The client has several long-running users, and they would like to provide more effective ways of surfacing posts from them from different time frames.

How might the client enable this sort of functionality?

● Create a Global Secondary Index with a partition key of timestamp and a sort key of userId.
● Create a Local Secondary Index with a partition key of userId and a sort key of timestamp.
● Create a Local Secondary Index with a partition key of timestamp and a sort key of userId.
✓ Create a Global Secondary Index with a partition key of userId and a sort key of timestamp.
Explanation:-They want to query on users in different times, it would be best to create a Global Secondary Index using userId as partition key and timestamp as sort key.

Q34)

You've been asked to select a tool that can easily visualize sales data that comes in as JSON to S3, occasionally as ad-hoc CSV files, and even from the Amazon Redshift data warehouse.

The solution must allow multiple users from the finance department to easily access it and occasionally upload their own Excel spreadsheets to compare with existing data.

What solution do you recommend?

● Use QuickSight and a combination of data source connections with the Redshift cluster and existing S3 JSON documents along with a Lambda function to process the XLSX files and transform them into a QuickSight-readable format.
✓ QuickSight and a combination of data source connections with the Redshift cluster and existing S3 JSON documents while still allowing finance to upload Excel files directly.
Explanation:-QuickSight can provide visualization with out of box integration with Redshift and S3 JSON documents as well as handle Excel files. QuickSight allows you to directly connect to and import data from a wide variety of cloud and on-premises data sources. These include SaaS applications such as Salesforce, Square, ServiceNow, Twitter, Github, and JIRA; 3rd party databases such as Teradata, MySQL, Postgres, and SQL Server; native AWS services such as Redshift, Athena, S3, RDS, and Aurora; and private VPC subnets. You can also upload a variety of file types including Excel, CSV, JSON, and Presto.
● Use Kibana and Amazon Athena to process the S3 data and XLSX files before indexing them in Elasticsearch.
● Use Kibana and a combination of an S3 bucket that accepts the XLSX downloads and processes them with Lambda to transform them into JSON and index them in Elasticsearch.

Q35)

You have a JSON data file in S3 that you are attempting to load into a JavaScript visualization you are writing locally. This visualization makes an HTTP GET request to the S3 location that fails.

However, when you attempt to visit the URL being requested by the JavaScript directly from inside your browser, it seems to be loading fine.

You are also using a private/incognito window and are not signed into the AWS console.

What is the most likely issue?

● The IAM role you used to create and upload the JSON data in the S3 bucket is preventing the JavaScript from loading the file.
● The ACLs on the bucket are preventing the JavaScript from loading the file.
● The bucket policies are preventing the JavaScript from loading the file.
✓ The CORS settings are preventing the JavaScript from loading the file.
Explanation:-CORS in S3 needs to be enabled for the application to be able to access the files. Cross-origin resource sharing (CORS) defines a way for client web applications that are loaded in one domain to interact with resources in a different domain. With CORS support, you can build rich client-side web applications with Amazon S3 and selectively allow cross-origin access to your Amazon S3 resources.

Q36)

You have to design an EMR system where you will be processing highly confidential data.

What can you do to ensure encryption of data at rest?

● TLS

✓ LUKS

Explanation:-SSE-KMS and LUKS can be used for implemented encryption at rest. Amazon S3 encryption works with EMR File System (EMRFS) objects read from and written to Amazon S3. You specify Amazon S3 server-side encryption (SSE) or client-side encryption (CSE) when you enable at-rest encryption. Amazon S3 SSE and CSE encryption with EMRFS are mutually exclusive; you can choose either but not both. Regardless of whether Amazon S3 encryption is enabled, Transport Layer Security (TLS) encrypts the EMRFS objects in-transit between EMR cluster nodes and Amazon S3. SSE-KMS: You use an AWS KMS customer master key (CMK) set up with policies suitable for Amazon EMR. LUKS. In addition to HDFS encryption, the Amazon EC2 instance store volumes and the attached Amazon EBS volumes of cluster instances are encrypted using LUKS.

● VPN

✓ SSE-KMS

Explanation:-SSE-KMS and LUKS can be used for implemented encryption at rest. Amazon S3 encryption works with EMR File System (EMRFS) objects read from and written to Amazon S3. You specify Amazon S3 server-side encryption (SSE) or client-side encryption (CSE) when you enable at-rest encryption. Amazon S3 SSE and CSE encryption with EMRFS are mutually exclusive; you can choose either but not both. Regardless of whether Amazon S3 encryption is enabled, Transport Layer Security (TLS) encrypts the EMRFS objects in-transit between EMR cluster nodes and Amazon S3. SSE-KMS: You use an AWS KMS customer master key (CMK) set up with policies suitable for Amazon EMR. LUKS. In addition to HDFS encryption, the Amazon EC2 instance store volumes and the attached Amazon EBS volumes of cluster instances are encrypted using LUKS.

Q37)

You are provisioning an application using EMR. You have requested 100 instances. You are charged $0.015 per hour, per instance. In the first 10 minutes after your launch request, Amazon EMR starts your cluster. 90 of your instances are available. It takes your cluster one hour to complete.

How much will you be charged for this EMR usage for the first hour?

● $1.50 per hour

✓ $1.35 per hour

Explanation:-EMR starts charging when 90% of the capacity is available, which is $1.35 (90 * $0.015 per hour)

Q: When does billing of my Amazon EMR cluster begin and end?

Billing commences when Amazon EMR starts running your cluster. You are only charged for the resources actually consumed. For example, let's say you launched 100 Amazon EC2 Standard Small instances for an Amazon EMR cluster, where the Amazon EMR cost is an incremental $0.015 per hour. The Amazon EC2 instances will begin booting immediately, but they won't necessarily all start at the same moment. Amazon EMR will track when each instance starts and will check it into the cluster so that it can accept processing tasks.

In the first 10 minutes after your launch request, Amazon EMR either starts your cluster (if all of your instances are available) or checks in as many instances as possible. Once the 10 minute mark has passed, Amazon EMR will start processing (and charging for) your cluster as soon as 90% of your requested instances are available. As the remaining 10% of your requested instances check in, Amazon EMR starts charging for those instances as well.

So, in the above example, if all 100 of your requested instances are available 10 minutes after you kick off a launch request, you'll be charged $1.50 per hour (100 * $0.015) for as long as the cluster takes to complete. If only 90 of your requested instances were available at the 10 minute mark, you'd be charged $1.35 per hour (90 * $0.015) for as long as this was the number of instances running your cluster. When the remaining 10 instances checked in, you'd be charged $1.50 per hour (100 * $0.015) for as long as the balance of the cluster takes to complete. Each cluster will run until one of the following occurs: you terminate the cluster with the TerminateJobFlows API call (or an equivalent tool), the cluster shuts itself down, or the cluster is terminated due to software or hardware failure.

● $0.015

● $0

Q38)

You have been asked to ensure that all AWS API calls are collected across your company's AWS account and that they are kept around for 90 days for analysis. After that, they must be able to be restored for 3 years.

How can you meet these needs in a scalable, cost-effective way?

● Enable CloudTrail logging in all accounts into S3 buckets, and set a lifecycle policy to expire the data in each bucket after 3 years.

● Enable CloudTrail logging to Glacier, and set a lifecycle policy to expire the data after 3 years.

✓ Enable CloudTrail logging to a centralized S3 bucket, set a lifecycle policy to move the data to Glacier after 90 days, and expire the data after 3 years.

Explanation:-The CloudTrail logging can be directed to a centralized S3 bucket. Lifecycle policies on the bucket can help to transition the data to low cost archival storage i.e. Glacier after 90 days and expire after 3 years.

AWS CloudTrail is a web service that records activity made on your account and delivers log files to your Amazon S3 bucket.

Applying a trail to all regions refers to creating a trail that will record AWS account activity in all regions. This setting also applies to any new regions that are added.

You can create and manage a trail across all regions in the partition in one API call or few clicks. You will receive a record of account activity made in your AWS account across all regions to one S3 bucket or CloudWatch logs log group. When AWS launches a new region, you will receive the log files containing event history for the new region without taking any action.

To manage your objects so that they are stored cost effectively throughout their lifecycle, configure their lifecycle. A lifecycle configuration is a set of rules that define actions that Amazon S3 applies to a group of objects. There are two types of actions:

Transition actions—Define when objects transition to another storage class. For example, you might choose to transition objects to the STANDARD_IA storage class 30 days after you created them, or archive objects to the GLACIER storage class one year after creating them.There are costs associated with the lifecycle transition requests.

Expiration actions—Define when objects expire. Amazon S3 deletes expired objects on your behalf.

● Enable AWS CloudTrail logging across all accounts to a centralized Amazon S3 bucket with versioning enabled. Set a lifecycle policy to move the data to Amazon Glacier daily, and expire the data after 90 days.

Q39)

You need to design a solution that can return user profile data to your application with millisecond latency or better and that can store this information for two million users.

Currently, user profile information is capped at 15 KB and typically doesn't exceed 3 KB. It is very important that this profile information always reflects the most recent changes when read. You also want a way to process changes from user profiles

and potentially send out emails when specific types of changes are made but the amount of compute capacity required for this is highly sporadic.

What could you use to help build a cost-effective solution to your application?

● Use separate S3 objects to store user profile information and use AWS Lambda functions that trigger whenever objects are updated. Then send the customized HTML emails out from SNS to users if appropriate.

● Use a DynamoDB table to store profile information as items and then enable DynamoDB streams on the table. Write an application that sits on a cluster of EC2 Reserved Instances and use that to process the DynamoDB streams data and send emails using SES when appropriate.

✓ Use a DynamoDB table to store profile information as items and then enable DynamoDB streams on the table. Whenever changes are made, evaluate them with a Lambda function and send an email with SES if that is appropriate.

Explanation:-DynamoDB can store the data and provide low latency and strongly consistent reads of the user profiles data. DynamoDB streams can help detect any changes on the data and process them using Lambda and notify.

● Use separate S3 objects to store user profile information and run ECS jobs periodically hash the files and compare them to a RDS database of object hashes to determine if they need to be processed and emails might need to be sent out using SES.

Q40)

An enterprise customer is migrating to Redshift and is considering using dense storage nodes in its Redshift cluster.

The customer wants to migrate 50 TB of data. The customer's query patterns involve performing many joins with thousands of rows. The customer needs to know how many nodes are needed in its target Redshift cluster. The customer has a limited budget and needs to avoid performing tests unless absolutely needed.

Which approach should this customer use?

● Insist on performing multiple tests to determine the optimal configuration.

● Start with fewer large nodes.

● Have two separate clusters with a mix of a small and large nodes.

✓ Start with many small nodes.

Explanation:-the customer is planning to use Dense Storage nodes, they can start with more number of small nodes which would be cost-effective as compared to large nodes and easier to improve query performance and storage.

Refer AWS documentation - Redshift Cluster & Nodes

DS2 node types are optimized for large data workloads and use hard disk drive (HDD) storage. Node types are available in different sizes. DS2 nodes are available in xlarge and 8xlarge sizes.

The number of nodes that you choose depends on the size of your dataset and your desired query performance. Using the dense storage node types as an example, if you have 32 TB of data, you can choose either 16 ds2.xlarge nodes or 2 ds2.8xlarge nodes. If your data grows in small increments, choosing the ds2.xlarge node size allows you to scale in increments of 2 TB. If you typically see data growth in larger increments, a ds2.8xlarge node size might be a better choice.

Because Amazon Redshift distributes and executes queries in parallel across all of a cluster's compute nodes, you can increase query performance by adding nodes to your cluster. Amazon Redshift also distributes your data across all compute nodes in a cluster. When you run a cluster with at least two compute nodes, data on each node will always be mirrored on disks on another node and you reduce the risk of incurring data loss.

Q41)

An organization is designing an application architecture. The application will have over 100 TB of data and will support transactions that arrive at rates from hundreds per second to tens of thousands per second, depending on the day of the week and time of the day.

All transaction data must be durably and reliably stored. Certain read operations must be performed with strong consistency.

Which solution meets these requirements?

● Use an Amazon Relational Database Service (RDS) instance sized to meet the maximum anticipated transaction rate and with the High Availability option enabled.

● Deploy a NoSQL data store on top of an Amazon Elastic MapReduce (EMR) cluster, and select the HDFS High Durability option.

● Use Amazon Redshift with synchronous replication to Amazon Simple Storage Service (S3) and row-level locking for strong consistency.

✓ Use Amazon DynamoDB as the data store and use strongly consistent reads when necessary.

Explanation:-DynamoDB can store and handle the transactions. DynamoDB also supports strongly consistent reads. DynamoDB is also a managed AWS service.

Amazon DynamoDB is a key-value and document database that delivers single-digit millisecond performance at any scale. It's a fully managed, multiregion, multimaster database with built-in security, backup and restore, and in-memory caching for internet-scale applications. DynamoDB can handle more than 10 trillion requests per day and support peaks of more than 20 million requests per second.

Q42)

A company stores data in an S3 bucket. Some of the data contains sensitive information.

They need to ensure that the bucket complies with PCI DSS (Payment Card Industry Data Security Standard) compliance standards.

Which of the following should be implemented to fulfill this requirement? (Select TWO)

✓ Ensure that objects from the bucket are request only via HTTPS

Explanation:-One of the requirement is data security with encryption at rest and in transit. PCI DSS helps ensure that companies maintain a secure environment for storing, processing, and transmitting credit card information.

● Ensure that access to the bucket is only given to one IAM role

● Enable versioning for the bucket

✓ Enable server side encryption SSE for a bucket.

Explanation:-One of the requirement is data security with encryption at rest and in transit. PCI DSS helps ensure that companies maintain a secure environment for storing, processing, and transmitting credit card information.

Q43)

A game company needs to properly scale its game application, which is backed by DynamoDB. Amazon Redshift has the past two years of historical data. Game traffic varies throughout the year based on various factors such as season, movie release, and holiday season.

An administrator needs to calculate how much read and write throughput should be provisioned for DynamoDB table for each week in advance.

How should the administrator accomplish this task?

● Feed the data into Spark Mlib and build a random forest model.
● Feed the data into Apache Mahout and build a multi-classification model.
✓ Feed the data into Amazon Machine Learning and build a regression model.
Explanation:-Regression model can help predict or forecast data based on the earlier dataset.
● Feed the data into Amazon Machine Learning and build a binary classification model.

Q44)

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

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

● Use Amazon Kinesis Streams and a KCL-based application deployed on AWS Elastic Beanstalk.
● Create an Amazon Machine Learning model and invoke it with AWS Lambda.
✓ Create an AWS IoT rule to generate an Amazon SNS notification.
Explanation:-IoT rules can help evaluate and send notifications when the peak thresholds are exceeded. The AWS IoT rules engine listens for incoming MQTT messages that match a rule. When a matching message is received, the rule takes some action with the data in the MQTT message (for example, writing data to an Amazon S3 bucket, invoking a Lambda function, or sending a message to an Amazon SNS topic).
Refer AWS documentation - IoT Rules
Rules give your devices the ability to interact with AWS services. Rules are analyzed and actions are performed based on the MQTT topic stream. You can use rules to support tasks like these:
• Augment or filter data received from a device.
• Write data received from a device to an Amazon DynamoDB database.
• Save a file to Amazon S3.
• Send a push notification to all users using Amazon SNS.
• Publish data to an Amazon SQS queue.
• Invoke a Lambda function to extract data.
• Process messages from a large number of devices using Amazon Kinesis.
• Send data to the Amazon Elasticsearch Service.
• Capture a CloudWatch metric.
• Change a CloudWatch alarm.
• Send the data from an MQTT message to Amazon Machine Learning to make predictions based on an Amazon ML model.
• Send a message to a Salesforce IoT Input Stream.
• Send message data to an AWS IoT Analytics channel.
• Start execution of a Step Functions state machine.
• Send message data to an AWS IoT Events input.
● Store the data points in an Amazon DynamoDB table and poll if for peak metrics data from an Amazon EC2 application.

Q45)

A solutions architect for a logistics organization ships packages from thousands of suppliers to end customers. The architect is building a platform where suppliers can view the status of one or more of their shipments.

Each supplier can have multiple roles that will only allow access to specific fields in the resulting information.

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

● Send the tracking data to Amazon Kinesis Firehose. Use Amazon S3 notifications and AWS Lambda to prepare files in Amazon S3 with appropriate data for each supplier's roles. Generate temporary AWS credentials for the suppliers' users with AWS STS. Limit access to the appropriate files through security policies.
● Send the tracking data to Amazon Kinesis Streams. Use Amazon EMR with Spark Streaming to store the data in HBase. Create one table per supplier. Use HBase Kerberos integration with the suppliers' users. Use HBase ACL-based security to limit access for the roles to their specific table and columns.
● Send the tracking data to Amazon Kinesis Firehose. Store the data in an Amazon Redshift cluster. Create views for the suppliers' users and roles. Allow suppliers access to the Amazon Redshift cluster using a user limited to the applicable view.
✓ Send the tracking data to Amazon Kinesis Streams. Use AWS Lambda to store the data in an Amazon DynamoDB Table. Generate temporary AWS credentials for the suppliers' users with AWS STS, specifying fine-grained security policies to limit access only to their applicable data.
Explanation:-DynamoDB can be used to store the data. Access to fields can be controlled using DynamoDB fine grained access control, which can be mapped to IAM role. This solution also requires the least amount of management effort.
Refer AWS documentation - DynamoDB Control Access
In DynamoDB, you have the option to specify conditions when granting permissions using an IAM policy (see Access Control). For example, you can:
• Grant permissions to allow users read-only access to certain items and attributes in a table or a secondary index.
• Grant permissions to allow users write-only access to certain attributes in a table, based upon the identity of that user.

Q46)

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

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

● Provide the auditor with SSH keys for access to the Amazon EMR cluster.
● Provide the auditor with CloudFormation templates.
✓ Provide the auditor IAM accounts with the SecurityAudit policy attached to their group.
Explanation:-The SecurityAudit managed policy can provide the Auditors with the read only access to AWS Services.
● Provide the auditor with access to AWS DirectConnect to use their existing tools.

Q47)

A social media customer has data from different data sources including RDS running MySQL, Redshift, and Hive on EMR.

To support better analysis, the customer needs to be able to analyze data from different data sources and to combine the results.

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

● Write a program running on a separate EC2 instance to run queries to three different systems. Aggregate the results after getting the responses from all three systems.
● Spin up an Elasticsearch cluster. Load data from all three data sources and use Kibana to analyze.
● Load all data from a different database/warehouse to S3. Use Redshift COPY command to copy data to Redshift for analysis.
✓ Install Presto on the EMR cluster where Hive sits. Configure MySQL and PostgreSQL connector to select from different data sources in a single query.
Explanation:-Presto can help query over multiple datasources and also provides connectors to interact directly MySQL, Redshift and Hive. 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.

Q48)

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

What is the appropriate model choice and target attribute combination for this problem?

● K-Nearest Neighbors model with a multi-class target attribute.
● Binary Classification with a categorical target attribute.
● Regression model with a numeric target attribute.
✓ Multi-class classification model with a categorical target attribute.
Explanation:-The blood group types are limited, a multi-class classification model can help classification the result into the blood groups

Q49)

An online photo album app has a key design feature to support multiple screens (e.g, desktop, mobile phone, and tablet) with high-quality displays. Multiple versions of the image must be saved in different resolutions and layouts. The image-processing Java program takes an average of five seconds per upload, depending on the image size and format.

Each image upload captures the following image metadata: user, album, photo label, upload timestamp.

The app should support the following requirements:

- Hundreds of user image uploads per second

- Maximum image upload size of 10 MB

- Maximum image metadata size of 1 KB

Image displayed in optimized resolution in all supported screens no later than one minute after image upload.

Which strategy should be used to meet these requirements?

✓ Upload image with metadata to Amazon S3, use Lambda function to run the image processing and save the images output to Amazon S3 and metadata to the app repository DB.
Explanation:-The images with metadata can be uploaded to S3. S3 can support both the size and request rate. A Lambda function can be triggered to convert and same the images output back to S3 and metadata to app DB.
● Write image and metadata to Amazon Kinesis. Use Amazon Elastic MapReduce (EMR) with Spark Streaming to run image processing and save the images output to Amazon S3 and metadata to app repository DB.
● Write images and metadata to Amazon Kinesis. Use a Kinesis Client Library (KCL) application to run the image processing and save the image output to Amazon S3 and metadata to the app repository DB.
● Write image and metadata to RDS with BLOB data type. Use AWS Data Pipeline to run the image processing and save the image output to Amazon S3 and metadata to the app repository DB.

Q50)

A travel website needs to present a graphical quantitative summary of its daily bookings to website visitors for marketing purposes.

The website has millions of visitors per day, but wants to control costs by implementing the least-expensive solution for this visualization.

What is the most cost-effective solution?

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

● Implement a Jupyter front-end provided by a continuously running EMR cluster leveraging spot instances for task nodes.
● Generate a graph using MicroStrategy backed by a transient EMR cluster.
✓ Generate a static graph with a transient EMR cluster daily, and store it an Amazon S3.
Explanation:-The most cost effective solution is to use a transient cluster to create the stats and use S3 to host the same.

Q51)

Your company operates AWS QuickSight with an Enterprise license for 10 users and has purchased 10GB of additional SPICE capacity. Currently, each user has consumed only 2GB of SPICE capacity. One of the data engineers on your team needs to create a visualization against and AWS Aurora database that requires 100GB of SPICE capacity.

How much additional SPICE capacity does the company need to purchase?

● 100GB
● 20GB
● 90GB
✓ 10GB
Explanation:-The company must purchase an additional 10GB to provide 120GB of SPICE capacity to the shared SPICE pool. The Enterprise license allows 10GB per user, and the company has purchased 10GB more and is currently using 20GB of SPICE capacity. So the existing available capacity = 100 (10GB * 10 per user) + 10GB (purchased) - 20GB used (10 users * 2GB per user) = 90GB available. Since the new requirement is for 100GB, the company must purchase an additional 10GB of SPICE capacity.

Q52)

A healthcare company has sensitive data that needs to be viewed for up to three months in S3. After three months, the data is infrequently accessed for up to one year. A government regulation requires this data be stored for seven years.

How can this healthcare company meet these requirements in the most cost-effective fashion?

● Store the files in S3 Glacier with auto-delete from S3 after three months.
✓ Store the files in S3 Standard. Create a lifecycle policy to transition the storage class to Standard - IA after three months and delete them after one year. Duplicate the files in Amazon Glacier with a Deny Delete vault lock policy for archives less than seven years old.
Explanation:-The requirements are to move the data in S3 based on access patterns and to also store the data long term for seven years to meet government regulations.
● Store the files in S3 Infrequent access with a lifecycle policy to remove them after a year. Duplicate 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 Glacier with a Deny Delete vault lock policy for seven years. Use Infrequent access storage.

Q53)

A film studio that uses AWS is trying to determine the best storage solution for large distributed jobs running on AWS Spot instances. The studio needs to both process Big Data using thousands of Spot instances and also needs to have real-time access to the data (which is often terabytes in size) so that animators can work on the processed files immediately after the jobs run.

Which of the following would be a suitable approach for this film studio?

● Process the data using AWS Spot instances with local EBS storage, unmount the storage after the jobs are completed, and remount all volumes on another host.
✓ Process the data using AWS Spot instances with EFS mounted storage.
Explanation:-The only approach that meets the requirements is to use EFS (Elastic File System). EFS will store the output of the work from the Spot instances and will scale up reads and writes as more storage is added. In addition, the artists can mount the same EFS volume and work on it while jobs are processing the data.
● Process the data using AWS Spot instances and share the Spot instances' local storage via NFS.
● Process the data using AWS Spot instances and do a distributed copy using EMR to artists' workstations.

Q54)

A new data engineer at an energy company has asked a Big Data specialist why Athena is ideal for searching the company's data lake.

What is the correct answer?

● Athena is serverless.
● Athena can be queried with SQL.
● Athena is designed to query data lakes.
✓ All of these.
Explanation:-Athena is designed to query a data lake. It is one of the best options for querying a data lake on AWS because of its serverless nature and its ability to run SQL.

Q55)

You are in charge of a team of data scientists tasked with analyzing data on seismic activity from around the globe in real time. The system must aggregate data from 100,000 sensors at 500 sites. The data are time series data and, therefore, it is important to preserve the order.

Which technology should you select to collect the data?

● IoT Core
✓ Kinesis Data Streams

Explanation:-Two aspects of this scenario indicate that Kinesis Data Streams is the best choice: First, it specifies real-time analysis, and second, it mentions that the order of the data is important.

● S3
● DynamoDB

Q56)

You must create a Redshift cluster to hold 1PB of uncompressed TSV files stored as an archive on AWS S3 Glacier.

What is the minimum size you should make the cluster, assuming below-average compression AND adding an AWS-recommended capacity buffer to absorb typical expected data growth?

● 1500TB
● 313TB
● 330TB
✓ 416TB

Explanation:-This answer requires a simple two-part calculation: (1) apply 3x compression (the lower end of the range), which equals 0.33 * 1PB (1000TB) = 333TB base size; and (2) add 25% to the cluster size (333TB * 125%) = 416TB.

Q57)

A financial services company needs to develop new SaaS applications in a matter of weeks to compete against other emerging financial startups. The software engineering team already knows Python. The concurrency will have to scale to thousands of requests per second during peak times during the day.

Which of the following would be best for developing web SaaS applications of this nature?

● Use EFS with AWS Batch.
✓ Use AWS Lambda, Cognito, DynamoDB, and S3.

Explanation:-The only option here that is a SaaS stack is option C. Other solutions only partially solve the problem or are meant for Big Data applications that are not SaaS backends.

● Use EMR, Beanstalk, and KMS.
● Use Redshift and EC2.

Q58)

The CFO has asked a Big Data specialist to come up with a plan to save costs on training for Machine Learning jobs. The CFO mentions that Machine Learning jobs do not need to be trained quickly. The time to train a job is less important than the overall cost.

What is the best plan the Big Data specialist could recommend to create cost savings?

✓ Use CPU-based instances to train the models.

Explanation:-The best initial recommendation is to use CPU-based hardware to train models. This will be slower, but it can be much less expensive than using GPU-based instances. The other options are either incorrect or not as clearly cost drivers.

● Use a few large instances.
● Use GPU-based instances to train the models.
● Use many small instances.

Q59)

An investor in your company has asked you to explain how your company is safeguarding customer data that is trained using SageMaker.

What do you say to the investor to accurately describe how SageMaker is used?

● Your devops engineer uses SageMaker with a substitution cypher that hides Social Security numbers you use as a unique ID. This was created using Puppet.
● Your company uses encryption algorithms designed by your lead C++ developer; this ensures that hackers will never find out how to decrypt the data.
✓ SageMaker uses Amazon KMS to encrypt all data used during the Machine Learning process.

Explanation:-The best practice is answer A, which ensures that encryption is seamless and secure. The other choices are examples of extreme red flags that naive companies or employees use to secure their data. Such bad practices would be immediately flagged by an experienced security professional.

● All of these.

Q60)

A company is looking to create a stream-based ingestion system for its popular gaming platform. A Big Data specialist working on the project has identified Kinesis Streaming as a solution.

What is an ideal first project to implement?

● Distributed MapReduce
● Key-Value
● Full-text search
✓ Time-series analytics

Explanation:-Time-series analytics is a best-practice Kinesis streaming use case, but the other options would best be served by a different AWS service, such as DynamoDB, Elasticsearch, or EMR.