Answer Sheet

Q1)

The department of transportation for a major metropolitan area has placed sensors on roads at key locations around the city. The goal is to analyze the flow of traffic and notifications from emergency services to identify potential issues and to help planners correct trouble spots.

A data engineer needs a scalable and fault-tolerant solution that allows planners to respond to issues within 30 seconds of their occurrence.

Which solution should the data engineer choose?

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

Explanation:-We need to tackle 2 issues. First is to capture real time sensor data and store it for analysis. Second is to respond to emergency notifications events with low latency. First can be handled using Kinesis Firehose to load data in Redshift for analysis. Second can be handled using SQS for notifications and DynamoDB for quick analysis or processing.
Refer AWS documentation - Kinesis Firehose FAQs
Amazon Kinesis Data Firehose buffers incoming streaming data to a certain size or for a certain period of time before delivering it to destinations. You can configure buffer size and buffer interval while creating your delivery stream. Buffer size is in MBs and ranges from 1MB to 128MB for Amazon S3 destination and 1MB to 100MB for Amazon Elasticsearch Service destination. Buffer interval is in seconds and ranges from 60 seconds to 900 seconds. Please note that in circumstances where data delivery to destination is falling behind data writing to delivery stream, Firehose raises buffer size dynamically to catch up and make sure that all data is delivered to the destination.

Q2)

A data engineer chooses Amazon DynamoDB as a data store for a regulated application.

This application must be submitted to regulators for review. The data engineer needs to provide a control framework that lists the security controls from the process to follow to add new users down to the physical controls of the data center, including items like security guards and cameras.

How should this control mapping be achieved using AWS?

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

Explanation:-These are AWS specific and not accessible directly. AWS provides access to third party audit reports to confirm the same.
Refer AWS documentation - Risk Compliance Whitepaper
AWS and its customers share control over the IT environment, both parties have responsibility for managing the IT environment. AWS' part in this shared responsibility includes providing its services on a highly secure and controlled platform and providing a wide array of security features customers can use. The customers' responsibility includes configuring their IT environments in a secure and controlled manner for their purposes. While customers don't communicate their use and configurations to AWS, AWS does communicate its security and control environment relevant to customers. AWS does this by doing the following:
• Obtaining industry certifications and independent third-party attestations described in this document
• Publishing information about the AWS security and control practices in whitepapers and web site content
• Providing certificates, reports, and other documentation directly to AWS customers under NDA (as required)

Q3)

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

What architecture securely meets these requirements?

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

Explanation:-It is essential when writing mobile applications that you consider the security of both how the application authenticates and how it stores credentials. Amazon Cognito gives you the ability to securely authenticate pools of users on any type of device at scale.

Q4)

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

- Key-value access pattern

- Complex SQL queries and transactions

- Consistent reads

- Fixed schema

Which data store should the organization choose?

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

Explanation:-Amazon RDS handles all these requirements, and although Amazon RDS is not typically thought of as optimized for key-value based access, a schema with a good primary key selection can provide this functionality.

Q5)

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

Your management now requires you to record replies to these social media messages to analyze the effectiveness of the marketing campaign in comparison to past and future efforts. You've already developed a new application component to interface with the social media site APIs in order to read the replies.

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

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

Explanation:-The point here is durable data store with any time analytics the best option is to store the data in DynamoDB and use Apache Hive with Amazon Elastic MapReduce for analytics.

- Deploy the new application component in an Auto Scaling group of Amazon EC2 instances, read the data from the social media sites, store it with Amazon Elastic Block Store, and use AWS Data Pipeline to publish it to Amazon Kinesis for analytics.

Q6)

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

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

What is the best method for architecting this application given the requirements?

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

Explanation:-the requirement is real time data ingestion and analytics, the best option is to use Kinesis for storing the real time incoming data. The data can then be moved to S3 and analyzed using EMR and Redshift. Data can then be moved to Glacier for archival.

Refer AWS documentation - Kinesis

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.

Q7)

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

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

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

- ✅ Configure the application to break the video files into chunks and use a multipart upload to transfer files to Amazon S3.

Explanation:-multipart upload helps provide better recoverability.

Depending on the size of the data you are uploading, Amazon S3 offers the following options:

• Upload objects in a single operation—With a single PUT operation, you can upload objects up to 5 GB in size.
• Upload objects in parts—Using the multipart upload API, you can upload large objects, up to 5 TB.The multipart upload API is designed to improve the upload experience for larger objects. You can upload objects in parts. These object parts can be uploaded independently, in any order, and in parallel. You can use a multipart upload for objects from 5 MB to 5 TB in size.

We recommend that you use multipart uploading in the following ways:

• If you're uploading large objects over a stable high-bandwidth network, use multipart uploading to maximize the use of your available bandwidth by uploading object parts in parallel for multi-threaded performance.
• If you're uploading over a spotty network, use multipart uploading to increase resiliency to network errors by avoiding upload restarts. When using multipart uploading, you need to retry uploading only parts that are interrupted during the upload. You don't need to restart uploading your object

from the beginning.

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

Explanation:-S3 Transfer Acceleration helps speed up the upload performance. Amazon S3 Transfer Acceleration enables fast, easy, and secure transfers of files over long distances between your client and an S3 bucket. Transfer Acceleration takes advantage of Amazon CloudFront's globally distributed edge locations. As the data arrives at an edge location, data is routed to Amazon S3 over an optimized network path.S3 Transfer Acceleration helps speed up the upload performance. Amazon S3 Transfer Acceleration enables fast, easy, and secure transfers of files over long distances between your client and an S3 bucket. Transfer Acceleration takes advantage of Amazon CloudFront's globally distributed edge locations. As the data arrives at an edge location, data is routed to Amazon S3 over an optimized network path.

Q8)

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

You don't want the second group's queries to wait until the first group's queries are finished.

You need to design a solution so that this does not happen.

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

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

Explanation:-Redshift workload management allows proper usage of cluster.
Refer to the AWS Blog for Redshift to run mixed workloads
Amazon Redshift Workload Management allows you to manage workloads of various sizes and complexity for specific environments. Parameter groups contain WLM configuration, which determines how many query queues are available for processing and how queries are routed to those queues. Following settings are available
• How many queries can run concurrently in each queue
• How much memory is allocated among the queues
• How queries are routed to queues, based on criteria such as the user who is running the query or a query label
• Query timeout settings for a queue

- Create a read replica of Redshift and run the second team's queries on the read replica.

Q9)

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

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

To secure funding, you have to make sure that the platform meets these requirements and leaves room for further scaling. Which setup will meet the requirements?

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

Explanation:-Key point here is backend supporting the data with 2 years retention and architecture being scalable
DynamoDB can be used to support the ingestion throughput via autoscaled instances and later store data into Redshift for analysis

- Add an SQS queue to the ingestion layer to buffer writes to the RDS instance

---

Q10)

Your company produces customer commissioned one-of-a-kind skiing helmets combining high fashion with custom technical enhancements. Customers can show off their Individuality on the ski slopes and have access to head-up-displays. GPS rear-view cams and any other technical innovation they wish to embed in the helmet. The current manufacturing process is data rich and complex including assessments to ensure that the custom electronics and materials used to assemble the helmets are to the highest standards.

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

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

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

Explanation:-Key point here hybrid work flow with both automated and manual tasks and ability to replay also needing GPUs with CUD instances with low latency networking
SWF provides an ability to have both human and automated assessments with G2 instances in a placement group providing GPU and low latency

networking.
- Use AWS Data Pipeline to manage movement of data & meta-data and assessments Use an auto-scaling group of G2 instances in a placement group.

Q11)

A company has two different types of reporting needs on their 200-GB data warehouse;

– Data scientists run a small number of concurrent adhoc SQL queries that can take several minutes each to run.

– Display screens throughout the company run many fast SQL queries to populate dashboards,

Which design would meet these requirements with the LEAST cost?

- Use Amazon Redshift for Data Scientists; Run automated dashboard queries against Redshift and store the results in Amazon ElastiCache, Dashboards query ElastiCache.
- Configure auto-replication between Amazon Redshift and Amazon RDS. Data scientists use Redshift and Dashboards use RDS
- ✅ Use Amazon Redshift for both requirements, with separate query queues configured in workload management.

Explanation:-Redshift provides workload management which can help prioritize the interactive and long running jobs. Storing the data in a single storage service would also help keep the costs to minimum.

- Replicate relevant data between Amazon Redshift and Amazon DynamoDB. Data scientists use Redshift. Dashboards use DynamoDB

Q12)

A Solutions Architect is designing a weather forecast application. Every hour, the application will receive a new set of raw data from weather stations. The application will analyze this data and produce a set of local weather forecasts available for users to download. The analysis takes 50 minutes to run on 2,000 vCPUs. The analysis must complete before the next set of data is available.

Each local weather forecast is typically 10 GB in size. The forecasts are accessed heavily during the first hour they are available, with usage dropping rapidly as newer forecasts become available.

Which combination of steps is the MOST cost-effective architecture? (Select TWO.)

- Store weather forecast data in Amazon S3 Standard-Infrequent Access (S3 Standard-IA). Configure a lifecycle policy to transition the data to Amazon Glacier after 90 days.
- ✅ Store weather forecast data in Amazon S3 One Zone-Infrequent Access (S3 One Zone-IA). Configure a lifecycle policy to transition the data to Amazon Glacier after 90 days.

Explanation:-The focus is on most cost effective architecture, S3 One Zone-IA would be an ideal option as the data is used only during the first hour.

- Store weather forecast data in Amazon S3 Standard. Configure a lifecycle policy to transition the data to Amazon S3 Standard-Infrequent Access (S3 Standard-IA) after 30 days.
- Conduct the analysis on an Amazon EC2-based cluster using 1-hour Spot blocks in multiple AWS Regions.
- ✅ Conduct the analysis on a cluster of Amazon EC2 instances using Reserved Instances in a single AWS Region.

Explanation:-The job runs for 50 mins within an hour with a sustained usage of around 20 hours of the 24 hours.

Q13)

Your company will be launching its SaaS in five additional territories. Your CTO has asked you to provide resources for understanding the regulatory requirements for each area.

Which of the following is an AWS resource that is available to help?

- ✅ AWS Compliance Center

Explanation:-The AWS Compliance Center provides details about international regulations related to handling personally identifiable information and other topics.

- AWS International Business Services Portal
- AWS ICS
- ISO9001

Q14) Which AWS services can manage encryption keys? (Choose two.)

- AWS Secrets Manager
- AWS Config
- ✅ AWS CloudHSM

Explanation:-AWS Key Management Service is a serverless key-hosting service on AWS. CloudHSM requires a cluster to be provisioned.

- ✅ AWS Key Management Service

Explanation:-AWS Key Management Service is a serverless key-hosting service on AWS. CloudHSM requires a cluster to be provisioned.

Q15) What steps must be taken to secure a new AWS account?

- Disable API access for the root account
- Remove all IAM accounts
- Enable MFA for the root account
- ✅ All of these

Explanation:-The root account should always be secured with MFA and should not be used for day-to-day administrative tasks.

Q16)

You will be provisioning an AWS Redshift cluster to analyze the outcomes of patients involved in a drug trial. Because the data contains sensitive and personally identifiable information, it must be encrypted both in transit and at rest.

Which technologies could you employ to ensure that your solution complies with all potential regulatory requirements?

- Provision a CloudHSM cluster, use client-side encryption to load data into RedShift, and then delete the encryption keys.
- Use SSL.
- Create a key pair with CloudKMS, generate certificates for TLS, and configure Redshift to encrypt data with CloudKMS.
- ✅ Provision a CloudHSM cluster, generate key pairs to encrypt data in transit, and configure Redshift to use CloudHSM to encrypt data at rest.

Explanation:-Only CloudHSM can be used to create asymmetric key pairs for TLS.

Q17)

A movie studio has an image library of 75TB that it wants to make available to its EMR cluster.

What is the best tool for making this data available in AWS?

- ✅ AWS Snowball

Explanation:-AWS Snowball allows you to transfer large quantities of data at a low price, and for large data sets it is often faster than using Direct Connect. Snowball comes in 50TB or 80TB variants, and 75TB will fit on a single Snowball device.

- AWS Direct Connect
- AWS VPC
- AWS Storage Gateway

Q18) Which protocols can collect data with IoT Core?

- IFTTT, SQS, and S3
- SQS, MQTT, and HTTPS
- TLS, AMQP, and STOMP
- ✅ MQTT, HTTPS, and MQTT over WebSockets

Explanation:-IoT Core primarily uses MQTT for data collection, but it can also be configured to use HTTPS or MQTT over WebSockets.

Q19) Which services can you use to integrate on-premises data storage with AWS S3? (Choose two.)

- ✅ AWS Storage Gateway

Explanation:-AWS Storage Gateway can be used in conjunction with AWS Direct Connect to create a secure data pipeline between on-premises data storage systems and AWS S3. AWS Storage Gateway supports the CIFS, NFS, and iSCSI protocols.

- ✅ AWS Direct Connect

Explanation:-AWS Storage Gateway can be used in conjunction with AWS Direct Connect to create a secure data pipeline between on-premises data storage systems and AWS S3. AWS Storage Gateway supports the CIFS, NFS, and iSCSI protocols.

- AWS Glacier
- AWS Integrator

Q20)

You are gathering sales data from an ecommerce website to create a monthly report about product performance. Your CEO tells you that you must provide the sales team with a real-time dashboard for an upcoming promotion. Your current collection pipeline periodically uploads CSV files into S3, and they are then processed by an ETL process.

What is the most efficient way to architect a system to meet the needs of the sales team?

- Ask an operator to monitor sales and enter them into a spreadsheet as they come in. Store the spreadsheet in S3 so that the sales team can access it.
- ✅ Create a parallel data collection pipeline to feed the real-time dashboard. Use Kinesis Data Streams to gather data and Kinesis Firehose to forward data to Elasticsearch.

Explanation:-Kinesis Data Streams and Kinesis Firehose are specifically designed for real-time data collection. Kinesis Data Streams can be used to collect the data, while Kinesis Firehose can be used to forward data to destinations such as Elasticsearch.

- Instrument the application with collectd to emit measurements. Set up a time series database to report on current sales.
- Increase the frequency of the ETL run so that the statistics are more up to date.

Q21)

You are the lead architect for a new smart truck. The CEO asks you to collect as much data as possible to improve safety and predictive maintenance.

What tool would you use to collect the data that the trucks generate?

- AWS S3
- ✅ AWS IoT Core

Explanation:-AWS IoT Core is designed to collect data from devices, vehicles, and other sensors. It has the ability to gather data even when a network connection is intermittent, which makes it an ideal solution for this scenario. Kinesis Data Streams may also be a part of this solution, but the question is specifically about collection.

- AWS Storage Gateway
- AWS Kinesis Streams

Q22)

The head of HR has informed you that your AWS accounts must use the company's Active Directory as a source of identity moving forward.

How do you configure this?

- Use MARCOS synchronization to unify the logins on the whitfield axis.
- Manually create an IAM user for each user in Active Directory.
- ✅ Use SAML federation and assign roles to users based on group membership.

Explanation:-SAML federation can use an external account provider, such as Active Directory, to allow login to AWS.

- Write a script to sync Active Directory accounts to AWS IAM accounts. Run the script on a cron schedule.

Q23)

The compliance officer in your organization informs you that moving forward you must maintain a log of all access to your AWS account.

How can you achieve this?

- The compliance officer must personally make all changes to the AWS account to ensure that he knows everything that happens.
- Assign a very strong password for the root account and store it in an envelope kept by the security guard. Employees must sign a form agreeing to comply with all company policies before being handed the envelope.
- ✅ Create IAM user accounts for all employees that need to access the AWS account and configure AWS Cloud Trail to send reports to the compliance officer.

Explanation:-All users who need access should have individual accounts. All AWS account activity is logged by CloudTrail, and reports can be sent.

- Require all employees to log their activity on a clipboard in the manager's office before accessing AWS.

Q24)

Your company needs to create a data-driven, interactive visualization for the company website. The data for the website is created from a recurring query on a Redshift data warehouse that populates and updates a single table used by the website. You lead the engineering team, which manages the entire website frontend and the data warehouse in the backend. You have to create a customized, interactive visualization that includes complex filters and geospatial views and is available directly on the public website.

What technology would be most appropriate?

- An AWS QuickSight combo chart embedded on the website
- An AWS QuickSight geospatial chart embedded on the website
- An AWS QuickSight dashboard containing at least a geospatial chart, pivot tables, and tree maps embedded on the website
- ✅ A custom d3.js visualization

Explanation:-d3.js is the correct technology for customized data-driven visualizations that update based on changes to the underlying source data and are interactive.

Q25)

An IT auditor has asked the Big Data specialist in a Fortune 500 company to identify AWS services used in production that have the ability to create snapshots.

Which of the following services should the specialist list?

- RDS, S3, Athena, and Presto
- RDS, S3, Amazon and ElastiCache
- RDS, S3, EMR, and Hadoop
- ✅ RDS, S3, and DynamoDB

Explanation:-RDS, S3, and DynamoDB all have the capability to take snapshots. The ElastiCache service does not offer snapshots. Snapshots do not make sense for EMR or Athena because they are transitory services.

Q26)

You manage a web advertising platform on a single AWS account. This platform produces real-time ad-click data that you store as objects in an Amazon S3 bucket called "click-data".

Your advertising partners want to use Amazon Elastic MapReduce in their own AWS accounts to do analytics on the ad-click data.

They've asked for immediate access to the ad-dick data so that they can run analytics.

Which two choices are required to facilitate secure access to this data? Choose 2 answers.

- Configure AWS Data Pipeline in the partner AWS accounts to use the web Identity Federation API to access data in the "click-data" bucket.
- Configure AWS Data Pipeline to transfer the data from the "click-data" bucket to the partner's Amazon Elastic MapReduce cluster.
- Configure the Amazon S3 bucket access control list to allow access to the partners Amazon Elastic MapReduce cluster.
- Create a new IAM group for AWS Data Pipeline users with a trust policy that contains partner AWS account IDs.
- ✅ Configure an Amazon S3 bucket policy for the "click-data" bucket that allows Read-Only access to the objects and associate this policy with an IAM role.

Explanation:-The access needs to secure, the data should be sent to the partner account. An IAM cross account role can be created for the AWS partner account with a external ID for security and an S3 bucket policy can be created to allow only read access and associate it with an IAM role.

- ✅ Create a cross-account IAM role with a trust policy that contains partner AWS account IDs and a unique external ID

Explanation:-The access needs to secure, the data should be sent to the partner account. An IAM cross account role can be created for the AWS partner account with a external ID for security and an S3 bucket policy can be created to allow only read access and associate it with an IAM role.

Q27)

Managers in a company need access to the human resources database that runs on Amazon Redshift, to run reports about their employees.

Managers must only see information about their direct reports.

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

✅ Define a view that uses the employee's manager name to filter the records based on current user names.

Explanation:-You can create a view in Redshift which filters the records based on the current user name and show only those results for the logged in user.

- Define a key for each manager in AWS KMS and encrypt the data for their employees with their private keys.
- Use Amazon Redshift snapshot to create one cluster per manager. Allow the manager to access only their designated clusters.
- Define an IAM group for each manager with each employee as an IAM user in that group, and use that to limit the access.

Q28)

A customer has a machine learning workflow that consists of multiple quick cycles of reads-writes-reads on Amazon S3.

The customer needs to run the workflow on EMR but is concerned that the reads in subsequent cycles will miss new data critical to the machine learning from the prior cycles.

How should the customer accomplish this?

- Set hadoop.data.consistency=true in the core-site.xml file.
- ✅ Turn on EMRFS consistent view when configuring the EMR cluster.

Explanation:-EMRFS Consistent View helps provide a view of the objects in S3 and also tracks the consistency.

Refer AWS documentation - EMRFS Consistent View

EMRFS consistent view is an optional feature available when using Amazon EMR release version 3.2.1 or later. Consistent view allows EMR clusters to check for list and read-after-write consistency for Amazon S3 objects written by or synced with EMRFS. Consistent view addresses an issue that can arise due to the Amazon S3 Data Consistency Model. For example, if you add objects to Amazon S3 in one operation and then immediately list objects in a subsequent operation, the list and the set of objects processed may be incomplete. This is more commonly a problem for clusters that run quick, sequential steps using Amazon S3 as a data store, such as multi-step extract-transform-load (ETL) data processing pipelines. When you create a cluster with consistent view enabled, Amazon EMR uses an Amazon DynamoDB database to store object metadata and track consistency with Amazon S3. If consistent view determines that Amazon S3 is inconsistent during a file system operation, it retries that operation according to rules that you can define.

With consistent view enabled, EMRFS returns the set of objects listed in an EMRFS metadata store and those returned directly by Amazon S3 for a given path. Because Amazon S3 is still the "source of truth" for the objects in a path, EMRFS ensures that everything in a specified Amazon S3 path is being processed regardless of whether it is tracked in the metadata. However, EMRFS consistent view only ensures that the objects in the folders that you track are checked for consistency.

- Use AWS Data Pipeline to orchestrate the data processing cycles.
- Set hadoop.s3.consistency=true in the core-site.xml file.

Q29)

A company that manufactures and sells smart air conditioning units also offers add-on services so that customers can see real-time dashboards in a mobile application or a web browser. Each unit sends its sensor information in JSON format every two seconds for processing and analysis.

The company also needs to consume this data to predict possible equipment problems before they occur. A few thousand pre-purchased units will be delivered in the next couple of months. The company expects high market growth in the next year and needs to handle a massive amount of data and scale without interruption.

Which ingestion solution should the company use?

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

Explanation:-Kinesis Data Streams can help handle the streaming data. Kinesis Streams provides you with the ability to build custom applications to process and analyze streaming data using KCL which can be used for anomaly detection and processing.

Although you can use Kinesis Data Streams to solve a variety of streaming data problems, a common use is the real-time aggregation of data followed by loading the aggregate data into a data warehouse or map-reduce cluster.

Data is put into Kinesis data streams, which ensures durability and elasticity. The delay between the time a record is put into the stream and the time it can be retrieved (put-to-get delay) is typically less than 1 second. In other words, a Kinesis Data Streams application can start consuming the data from the stream almost immediately after the data is added. The managed service aspect of Kinesis Data Streams relieves you of the operational burden of creating and running a data intake pipeline. You can create streaming map-reduce–type applications. The elasticity of Kinesis Data Streams enables you to scale the stream up or down, so that you never lose data records before they expire.

Multiple Kinesis Data Streams applications can consume data from a stream, so that multiple actions, like archiving and processing, can take place concurrently and independently. For example, two applications can read data from the same stream. The first application calculates running aggregates and updates an Amazon DynamoDB table, and the second application compresses and archives data to a data store like Amazon Simple Storage Service (Amazon S3). The DynamoDB table with running aggregates is then read by a dashboard for up-to-the-minute reports.

The Kinesis Client Library enables fault-tolerant consumption of data from streams and provides scaling support for Kinesis Data Streams applications.

Q30)

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

The production application is latency insensitive. The new country contains another AWS region.

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

✅ Use Amazon S3 cross-region replication to copy and persist production transaction logs to a bucket in the new country's region.
Explanation:-Only the logs need to be maintained in the new country, S3 cross region replication can be used to copy the data to the AWS region within the new Country.
- Continue to serve customers from the existing region while using Amazon Kinesis to stream transaction data to the regulator.
- Use CloudFormation to replicate the production application to the new region.
- Use Amazon CloudFront to serve application content locally in the country; Amazon CloudFront logs will satisfy the requirement.

Q31)

You need to filter and transform incoming messages coming from a smart sensor you have connected with AWS. Once messages are received, you need to store them as time series data in DynamoDB.

Which AWS service can you use?

- IoT Device Shadow Service
- Redshift
- Kinesis
✅ IoT Rules Engine
Explanation:-IoT Rules Engine can be used to capture data from Sensor and data received from the device can be inserted into DynamoDB.
Refer AWS documentation - AWS 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.

Q32)

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

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

- Enable SSL connections to Kinesis
- Use Amazon Kinesis Consumer Library
- Encrypt the data once it is at rest with a Lambda function
✅ Enable server-side encryption in Kinesis Streams
Explanation:-Kinesis support Server Side Encryption with which the data can be encrypted at rest.
Refer AWS documentation - Kinesis Server Side Encryption
Server-side encryption is a feature in Amazon Kinesis Data Streams that automatically encrypts data before it's at rest by using an AWS KMS customer master key (CMK) you specify. Data is encrypted before it's written to the Kinesis stream storage layer, and decrypted after it's retrieved from storage. As a result, your data is encrypted at rest within the Kinesis Data Streams service. This allows you to meet strict regulatory requirements and enhance the security of your data.
With server-side encryption, your Kinesis stream producers and consumers don't need to manage master keys or cryptographic operations. Your data is automatically encrypted as it enters and leaves the Kinesis Data Streams service, so your data at rest is encrypted. AWS KMS provides all the master keys that are used by the server-side encryption feature. AWS KMS makes it easy to use a CMK for Kinesis that is managed by AWS, a user-specified AWS KMS CMK, or a master key imported into the AWS KMS service.

Q33)

A company is using Kinesis data streams to store the log data, which is processed by an application every 12 hours. As the data needs to reside in Kinesis data streams for 12 hours, the Security team wants the data to be encrypted at rest.

How can it be secured in a most efficient way?

- Kinesis does not support encryption
- Encrypt using SSL/TLS for encrypting the data.
- Encrypt using S3 Server Side Encryption.
✅ Encrypt using Kinesis Server Side Encryption.
Explanation:-Kinesis support Server Side Encryption with which the data can be encrypted at rest.
Refer AWS documentation - Kinesis Server Side Encryption
Server-side encryption is a feature in Amazon Kinesis Data Streams that automatically encrypts data before it's at rest by using an AWS KMS customer master key (CMK) you specify. Data is encrypted before it's written to the Kinesis stream storage layer, and decrypted after it's retrieved from storage. As a result, your data is encrypted at rest within the Kinesis Data Streams service. This allows you to meet strict regulatory requirements and enhance the security of your data.
With server-side encryption, your Kinesis stream producers and consumers don't need to manage master keys or cryptographic operations. Your

data is automatically encrypted as it enters and leaves the Kinesis Data Streams service, so your data at rest is encrypted. AWS KMS provides all the master keys that are used by the server-side encryption feature. AWS KMS makes it easy to use a CMK for Kinesis that is managed by AWS, a user-specified AWS KMS CMK, or a master key imported into the AWS KMS service.

Q34)

You have an application that is currently in the development stage but is expected to write 2,400 items per minute to a DynamoDB table, each 2Kb in size or less and then fluctuate to 4,800 writes of items (of the same size) per minute on weekends. There may be other fluctuations within that range in the future as the application develops. It is important to the success of the application that the vast majority of user requests are met in a cost-effective way.

How should this table be created?

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

Explanation:-DynamoDB Auto Scaling can help scale as per the demand.
Refer AWS documentation - DynamoDB AutoScaling
Many database workloads are cyclical in nature or are difficult to predict in advance. For example, consider a social networking app where most of the users are active during daytime hours. The database must be able to handle the daytime activity, but there's no need for the same levels of throughput at night. Another example might be a new mobile gaming app that is experiencing rapid adoption. If the game becomes too popular, it could exceed the available database resources, resulting in slow performance and unhappy customers. These kinds of workloads often require manual intervention to scale database resources up or down in response to varying usage levels.
DynamoDB auto scaling uses the AWS Application Auto Scaling service to dynamically adjust provisioned throughput capacity on your behalf, in response to actual traffic patterns. This enables a table or a global secondary index to increase its provisioned read and write capacity to handle sudden increases in traffic, without throttling. When the workload decreases, Application Auto Scaling decreases the throughput so that you don't pay for unused provisioned capacity.

- Provision a base WCU of 80 and then schedule regular increases to 160 WCUs when a higher load is expected.

Q35)

Your company recently purchased five different companies that run different backend databases that include Redshift, MySQL, Hive on EMR and PostgreSQL.

You need a single tool that can run queries on all the different platform for your daily ad-hoc analysis.

Which tool enables you to do that?

- ✅ Presto

Explanation:-Presto allows ad hoc query analysis over multiple data sources.
Refer AWS documentation - Presto
Presto (or PrestoDB) is an open source, distributed SQL query engine, designed from the ground up for fast analytic queries against data of any size. It supports both non-relational sources, such as the Hadoop Distributed File System (HDFS), Amazon S3, Cassandra, MongoDB, and HBase, and relational data sources such as MySQL, PostgreSQL, Amazon Redshift, Microsoft SQL Server, and Teradata.
Presto can query data where it is stored, without needing to move data into a separate analytics system. Query execution runs in parallel over a pure memory-based architecture, with most results returning in seconds.

- QuickSight
- Ganglia
- YARN

Q36)

You need to visualize data from Spark and Hive running on an EMR cluster.

Which of the options is best for an interactive and collaborative notebook for data exploration?

- Hive
- D3.js
- Kinesis Analytics
- ✅ Zeppelin

Explanation:-Zeppelin provides data ingestion, data discovery, data analytics and data visualization & collaboration.
Refer documentation - Zeppelin
Zeppelin is a web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala and more.
Apache Zeppelin interpreter concept allows any language/data-processing-backend to be plugged into Zeppelin. Currently Apache Zeppelin supports many interpreters such as Apache Spark, Python, JDBC, Markdown and Shell.
Some basic charts are already included in Apache Zeppelin. Visualizations are not limited to SparkSQL query, any output from any language backend can be recognized and visualized.

Q37)

You are using QuickSight to identify demand trends over multiple months for your top five product lines.

Which type of visualization do you choose?

- Pivot Table
- ✅ Line Chart

Explanation:-You need to represent time driven data for demand trends over multiple months, Line chart would be an ideal choice.

- Pie Chart
- Scatter Plot

Q38)

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

As a security requirement, the company wants the data to be encrypted at rest.

As a solution architect how would you implement the solution?

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

Explanation:-The need is for data at rest encryption. S3 with SSE will help store the data in S3 in encrypted format.
Refer AWS documentation - Redshift Encryption & S3 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.
Encryption is an optional, immutable setting of a cluster. If you want encryption, you enable it during the cluster launch process. To go from an unencrypted cluster to an encrypted cluster or the other way around, unload your data from the existing cluster and reload it in a new cluster with the chosen encryption setting.

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

---

Q39)

You have been asked to handle a large data migration from multiple Amazon RDS MySQL instances to a DynamoDB table. You have been given a short amount of time to complete the data migration.

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

- ✓ Create a data pipeline to export Amazon RDS data and import the data into DynamoDB.

Explanation:-Data Pipeline can be used to import the data from MySQL and Export it to DynamoDB as batch.

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

Q40)

An International company has deployed a multi-tier web application that relies on DynamoDB in a single region. For regulatory reasons they need disaster recovery capability in a separate region with a Recovery Time Objective of 2 hours and a Recovery Point Objective of 24 hours. They should synchronize their data on a regular basis and be able to provision the web application rapidly using CloudFormation.

The objective is to minimize changes to the existing web application, control the throughput of DynamoDB used for the synchronization of data and synchronize only the modified elements.

Which design would you choose to meet these requirements?

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

Explanation:-The key requirement here is DR with RTO of 2 hours and a RPO of 24 hours with only the changed items to be replicated.
DynamoDB cross region copy would help for DR with required RPO and RTO with Lastupdated time would help replicate only updated items.

Q41)

A company hosts a web application on AWS which uses RDS instance to store critical data. As a part of a security audit, it was recommended hardening of RDS instance.

What actions would help achieve the same? (Select TWO)

- ✓ Use RDS encryption to secure the RDS instances and snapshots at rest.

Explanation:-The RDS security can tightened using SSL connection and encryption.
Refer AWS documentation - RDS Security
• Run your DB instance in an Amazon Virtual Private Cloud (VPC) for the greatest possible network access control.
• Use AWS Identity and Access Management (IAM) policies to assign permissions that determine who is allowed to manage RDS resources. For example, you can use IAM to determine who is allowed to create, describe, modify, and delete DB instances, tag resources, or modify security groups.
• Use security groups to control what IP addresses or Amazon EC2 instances can connect to your databases on a DB instance. When you first create a DB instance, its firewall prevents any database access except through rules specified by an associated security group.
• Use Secure Socket Layer (SSL) connections with DB instances running the MySQL, MariaDB, PostgreSQL, Oracle, or Microsoft SQL Server database engines.
• Use RDS encryption to secure your RDS instances and snapshots at rest. RDS encryption uses the industry standard AES-256 encryption algorithm to encrypt your data on the server that hosts your RDS instance.
• Use network encryption and transparent data encryption with Oracle DB instances
• Use the security features of your DB engine to control who can log in to the databases on a DB instance, just as you do if the database was on your local network.

- Use AWS Inspector to apply patches to the RDS instance

Use AWS CloudTrail to track all the SSH access to the RDS instance

Use Secure Socket Layer (SSL) connections with DB instances

Explanation:-The RDS security can tightened using SSL connection and encryption.

Refer AWS documentation - RDS Security

• Run your DB instance in an Amazon Virtual Private Cloud (VPC) for the greatest possible network access control.

• Use AWS Identity and Access Management (IAM) policies to assign permissions that determine who is allowed to manage RDS resources. For example, you can use IAM to determine who is allowed to create, describe, modify, and delete DB instances, tag resources, or modify security groups.

• Use security groups to control what IP addresses or Amazon EC2 instances can connect to your databases on a DB instance. When you first create a DB instance, its firewall prevents any database access except through rules specified by an associated security group.

• Use Secure Socket Layer (SSL) connections with DB instances running the MySQL, MariaDB, PostgreSQL, Oracle, or Microsoft SQL Server database engines.

• Use RDS encryption to secure your RDS instances and snapshots at rest. RDS encryption uses the industry standard AES-256 encryption algorithm to encrypt your data on the server that hosts your RDS instance.

• Use network encryption and transparent data encryption with Oracle DB instances

• Use the security features of your DB engine to control who can log in to the databases on a DB instance, just as you do if the database was on your local network.

Q42)

You need to perform ad-hoc SQL queries on massive amounts of well-structured data.

Additional data comes in constantly at a high velocity, and you don't want to have to manage the infrastructure processing it if possible.

Which solution should you use?

EMR running Apache Spark

Kinesis Firehose and Redshift

Explanation:-Kinesis Firehose can capture the data and store the in data in Redshift. Redshift can provide processing of huge structured data.

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.

Kinesis Firehose and RDS

EMR using Hive

Q43)

A media advertising company handles a large number of real-time messages sourced from over 200 websites in real time. Processing latency must be kept low. Based on calculations, a 60-shard Amazon Kinesis stream is more than sufficient to handle the maximum data throughput, even with traffic spikes.

The company also uses an Amazon Kinesis Client Library (KCL) application running on Amazon Elastic Compute Cloud (EC2) managed by an Auto Scaling group. Amazon CloudWatch indicates an average of 25% CPU and a modest level of network traffic across all running servers. The company reports a 150% to 200% increase in latency of processing messages from Amazon Kinesis during peak times. There are NO reports of delay from the sites publishing to Amazon Kinesis.

What is the appropriate solution to address the latency?

Increase the number of shards in the Amazon Kinesis stream to 80 for greater concurrency.

Increase the size of the Amazon EC2 instances to increase network throughput.

Increase the minimum number of instances in the Auto Scaling group.

Increase Amazon DynamoDB throughput on the checkpoint table.

Explanation:-KCL uses DynamoDB for checkpointing and there would be throttling on the DynamoDB.

Refer AWS documentation - Kinesis Record Processor DynamoDB

For each Amazon Kinesis Data Streams application, the KCL uses a unique Amazon DynamoDB table to keep track of the application's state. Because the KCL uses the name of the Amazon Kinesis Data Streams application to create the name of the table, each application name must be unique.

If your Amazon Kinesis Data Streams application receives provisioned-throughput exceptions, you should increase the provisioned throughput for the DynamoDB table. The KCL creates the table with a provisioned throughput of 10 reads per second and 10 writes per second, but this might not be sufficient for your application. For example, if your Amazon Kinesis Data Streams application does frequent checkpointing or operates on a stream that is composed of many shards, you might need more throughput.

Q44)

An administrator needs to design a strategy for the schema in a Redshift cluster.

The administrator needs to determine the optimal distribution style for the tables in the Redshift schema.

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

When a new table has been loaded and it is unclear how it will be joined to dimension.

Explanation:-EVEN distribution distributes the data across slides in a round robin fashion and does not participate in joins.

Refer AWS documentation - Redshift Distribution Style

EVEN distribution - The leader node distributes the rows across the slices in a round-robin fashion, regardless of the values in any particular column. EVEN distribution is appropriate when a table does not participate in joins or when there is not a clear choice between KEY distribution and ALL distribution.

When data transfer between nodes must be eliminated.

When data must be grouped based on a specific key on a defined slice.

When the tables are highly denormalized and do NOT participate in frequent joins.

Explanation:-EVEN distribution distributes the data across slides in a round robin fashion and does not participate in joins.
Refer AWS documentation - Redshift Distribution Style
EVEN distribution - The leader node distributes the rows across the slices in a round-robin fashion, regardless of the values in any particular column. EVEN distribution is appropriate when a table does not participate in joins or when there is not a clear choice between KEY distribution and ALL distribution.

Q45)

A company that provides economics data dashboards needs to be able to develop software to display rich, interactive, data-driven graphics that run in web browsers and leverages the full stack of web standards (HTML, SVG, and CSS).

Which technology provides the most appropriate support for this requirements?

- IPython/Jupyter
- ✅ D3.js

Explanation:-D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation.

- R Studio
- Hue

Q46)

A company needs a churn prevention model to predict which customers will NOT renew their yearly subscription to the company's service.

The company plans to provide these customers with a promotional offer. A binary classification model that uses Amazon Machine Learning is required.

On which basis should this binary classification model be built?

- Last user session
- ✅ Each user time series events in the past 3 months

Explanation:-The time series data regarding the usage of the customer can give insights and help build and train the model

- User profiles (age, gender, income, occupation)
- Quarterly results

Q47)

A company develops a tool whose coverage includes blogs, news sites, forums, videos, reviews, images and social networks such as Twitter and Facebook. Users can search data by using Text and Image Search, and use charting, categorization, sentiment analysis and other features to provide further information and analysis. They have access to over 80 million sources.

They want to provide Image and text analysis capabilities to the applications which includes identify objects, people, text, scenes, and activities and also provides highly accurate facial analysis and facial recognition.

What service can provide this capability?

- Amazon Polly
- ✅ Amazon Rekognition

Explanation:-Amazon Rekognition makes it easy to add image and video analysis to your applications. You just provide an image or video to the Rekognition API, and the service can identify objects, people, text, scenes, and activities. It can detect any inappropriate content as well. Amazon Rekognition also provides highly accurate facial analysis and facial recognition. You can detect, analyze, and compare faces for a wide variety of use cases, including user verification, cataloging, people counting, and public safety.

- Amazon Comprehend
- Amazon SageMaker

Q48)

A company is looking to collect and process the log files in near real time that are generated from thousands of applications in their AWS cloud.

They are also collecting stock pricing information from stock price publishing data providers and using the information to recommend stocks to customers. They are looking at querying streams and using Kinesis Analytics application to process all the stocks for recommendation if price changes greater than 10 percent.

What kind of Queries will help fulfil the requirement?

- Stagger Windows queries
- Tumbling Windows queries
- Sliding windows queries
- ✅ Continuous queries

Explanation:-Continuous Query can provide the ability to monitor, query and generate alerts on a stream. Continuous Query is a query over a stream executes continuously over streaming data. This continuous execution enables scenarios, such as the ability for applications to continuously query a stream and generate alerts.

Q49)

You company has launched an EMR cluster to support their big data analytics requirements. They are planning to build an application running on EMR which supports both OLTP and operational analytics allowing you to use standard SQL queries and JDBC APIs to work with an Apache HBase backing store. Also data transfer tool between Amazon S3, Hadoop, HDFS, and RDBMS databases.

Which EMR Hadoop ecosystem fulfils the requirements? (Select TWO)

✅ Apache Sqoop

Explanation:-Apache Sqoop is a tool for transferring data between Amazon S3, Hadoop, HDFS, and RDBMS databases.

✅ Apache Phoenix

Explanation:-Apache Phoenix is used for OLTP and operational analytics, allowing you to use standard SQL queries and JDBC APIs to work with an Apache HBase backing store.

- Apache Flink
- Apache Hue
- Apache Ganglia

Q50)

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

- Faceting

- Flexible schema (JSON) and fixed schema

- Noise word elimination

Which data store should the organization choose?

- Amazon Relational Database Service (RDS)

✅ Amazon Elasticsearch Service

Explanation:-Elasticsearch provides all the listed features. It provides full text search, faceting, noise words or also known as stopwords and a flexible schema option. Amazon Elasticsearch Service is a fully managed service that makes it easy for you to deploy, secure, and operate Elasticsearch at scale with zero down time. Provide a low-latency, high-throughput, personalized search experience for your users across e-commerce applications, website, data lake catalogs, and other curated application data. Amazon Elasticsearch Service provides direct access to all of Elasticsearch's rich search APIs, supporting natural language search across free text, Boolean combinations of text and metadata search, auto-completion, faceted search, location-aware search, and much more.

- Amazon DynamoDB
- Amazon Redshift

Q51)

An administrator needs to design a distribution strategy for a star schema in a Redshift cluster.

The administrator needs to determine the optimal distribution style for the tables in the Redshift schema.

In which three circumstances would choosing key-based distribution be most appropriate? (Select three.)

✅ When the administrator needs to take advantage of data locality on a local node for joins and aggregates.

Explanation:-With key-based distribution, the rows are distributed according to the values in one column. The leader node places matching values on the same node slice. If you distribute a pair of tables on the joining keys, the leader node collocates the rows on the slices according to the values in the joining columns so that matching values from the common columns are physically stored together. 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. For slowly changing dimensions of reasonable size, DISTSTYLE ALL is a good choice for the dimension (reasonable size in this case means up to a few million rows, and that the number of rows in the dimension table is fewer than the filtered fact table for a typical join). If you have very frequent updates to a dimension table, then DISTSTYLE ALL may not be appropriate. In this case, it is better to use a distribution style of KEY and distribute on a column that distributes data relatively evenly rather than using DISTSTYLE EVEN. Joins will be more efficient when the joined tables have a distribution style of KEY even if the join column does not use the distribution key.

- When the administrator needs to optimize the fact table for parity with the number of slices.

✅ When the administrator needs to balance data distribution and collocation data.

Explanation:-With key-based distribution, the rows are distributed according to the values in one column. The leader node places matching values on the same node slice. If you distribute a pair of tables on the joining keys, the leader node collocates the rows on the slices according to the values in the joining columns so that matching values from the common columns are physically stored together. 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. For slowly changing dimensions of reasonable size, DISTSTYLE ALL is a good choice for the dimension (reasonable size in this case means up to a few million rows, and that the number of rows in the dimension table is fewer than the filtered fact table for a typical join). If you have very frequent updates to a dimension table, then DISTSTYLE ALL may not be appropriate. In this case, it is better to use a distribution style of KEY and distribute on a column that distributes data relatively evenly rather than using DISTSTYLE EVEN. Joins will be more efficient when the joined tables have a distribution style of KEY even if the join column does not use the distribution key.

- When the administrator needs to optimize a large, slowly changing dimension table.

✅ When the administrator needs to reduce cross-node traffic.

Explanation:-With key-based distribution, the rows are distributed according to the values in one column. The leader node places matching values on the same node slice. If you distribute a pair of tables on the joining keys, the leader node collocates the rows on the slices according to the values in the joining columns so that matching values from the common columns are physically stored together. 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. For slowly changing dimensions of reasonable size, DISTSTYLE ALL is a good choice for the dimension (reasonable size in this case means up to a few million rows, and that the number of rows in the dimension table is fewer than the filtered fact table for a typical join). If you have very frequent updates to a dimension table, then DISTSTYLE ALL may not be appropriate. In this case, it is better to use a distribution style of KEY and distribute on a column that distributes data relatively evenly

rather than using DISTSTYLE EVEN. Joins will be more efficient when the joined tables have a distribution style of KEY even if the join column does not use the distribution key.

Q52)

An administrator receives about 100 files per hour into Amazon S3 and will be loading the files into Amazon Redshift.

Customers who analyze the data within Redshift gain significant value when they receive data as quickly as possible.

The customers have agreed to a maximum loading interval of 5 minutes.

Which loading approach should the administrator use to meet this objective?

- Load each file as it arrives because getting data into the cluster as quickly as possibly is the priority.
- Load the cluster as soon as the administrator has the same number of files as nodes in the cluster.
- ✅ Load the cluster when the administrator has the number of files as multiple of files relative to Cluster Slice Count, or 5 minutes, whichever comes first.

Explanation:-The maximum lag allowed is 5 minutes, the target should to load the files once they reach multiple of the slice count of 5 minutes whichever comes first. Loading the files in multiple of slice count provides best performance. Amazon Redshift is an MPP (massively parallel processing) database, where all the compute nodes divide and parallelize the work of ingesting data. Each node is further subdivided into slices, with each slice having one or more dedicated cores, equally dividing the processing capacity. The number of slices per node depends on the node type of the cluster. For example, each DS2.XLARGE compute node has two slices, whereas each DS2.8XLARGE compute node has 16 slices. When you load data into Amazon Redshift, you should aim to have each slice do an equal amount of work. When you load the data from a single large file or from files split into uneven sizes, some slices do more work than others. As a result, the process runs only as fast as the slowest, or most heavily loaded, slice. When splitting your data files, ensure that they are of approximately equal size – between 1 MB and 1 GB after compression. The number of files should be a multiple of the number of slices in your cluster. Also, I strongly recommend that you individually compress the load files using gzip, lzop, or bzip2 to efficiently load large datasets. When loading multiple files into a single table, use a single COPY command for the table, rather than multiple COPY commands. Amazon Redshift automatically parallelizes the data ingestion. Using a single COPY command to bulk load data into a table ensures optimal use of cluster resources, and quickest possible throughput.

- Load the cluster when the number of files is less than the Cluster Slice Count.

Q53)

A company is centralizing a large number of unencrypted small files from multiple Amazon S3 buckets.

The company needs to verify that the files contain the same data after centralization.

Which method meets the requirements?

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

Explanation:-S3 stores the MD5 digest of the object data which can be verify to ensure the object contents have not changed.

Q54)

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

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

- Use an EMR cluster to aggregate the patient sensor data each night and provide Zeppelin notebooks that look at the new data residing on the cluster each morning for the physician to review.
- Use Spark streaming on EMR to aggregate the patient sensor data in every 15 minutes and generate a QuickSight visualization on the new data each morning for the physician to review.
- ✅ Use a transient EMR cluster that shuts down after use to aggregate the patient sensor data each night and generate a QuickSight visualization on the new data each morning for the physician to review.

Explanation:-A transient cluster can be used to aggregate the data and use QuickSight for visualization.

- Use Kinesis Aggregators Library to generate reports for reviewing the patient sensor data and generate a QuickSight visualization on the new data each morning for the physician to review.

Q55)

A company generates a large number of files each month and needs to use AWS import/export to move these files into Amazon S3 storage.

To satisfy the auditors, the company needs to keep a record of which files were imported into Amazon S3.

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

- Use a script to iterate over files in Amazon S3 to generate a log after each import/export job.
- Use the log file checksum in the import/export manifest files to create a unique log file in Amazon S3 for each import.
- ✅ Use the log file prefix in the import/export manifest files to create a unique log file in Amazon S3 for each import.

Explanation:-Creating a unique log file for each import would help

The AWS Import/Export process generates a log file. The log file name always ends with the phrase import-log- followed by your JobId. There is a remote chance that you already have an object with this name.To avoid a key collision, you can add an optional prefix to the log file by adding the logPrefix option in the manifest. AWS Import/Export takes the string value specified for this option and inserts it between the bucket name and log report name

The log file is a UTF-8 encoded CSV file that contains, among other things, information about each file loaded to or from your storage device. With Amazon S3 import jobs, AWS Import/Export saves the log to the same Amazon S3 bucket as your data.

For an import job, the log name ends with the phrase import-log- followed by your JOBID. For example, if the import JOBID is 53TX4, the log name

ends in import-log-53TX4. By default, if you do not set logPrefix in the manifest file, a job loaded to mybucket with the JOBID of 53TX4 loads the logs to http://mybucket.s3.amazonaws.com/import-log-53TX4. If you set logPrefix to logs/, the log location is http://s3.amazonaws.com/mybucket/logs/import-log-53TX4. Note If you have a log object with the same key name as an existing Amazon S3 object, the new log overwrites the existing object.You can use the logPrefix option to prevent object collisions.

- Use the same log file prefix in the import/export manifest files to create a versioned log file in Amazon S3 for all imports.

Q56)

A company hosts a portfolio of e-commerce websites across the Oregon, N. Virginia, Ireland, and Sydney AWS regions. Each site keeps log files that capture user behavior. The company has built an application that generates batches of product recommendations with collaborative filtering in Oregon.

Oregon was selected because the flagship site is hosted there and provides the largest collection of data to train machine learning models against. The other regions do NOT have enough historic data to train accurate machine learning models.

Which set of data processing steps improves recommendations for each region?

- Use the e-commerce application in Oregon to write replica log files in each other region.
- Use Amazon S3 bucket replication to consolidate log entries and build a single model in Oregon.
- Use Kinesis as a buffer for web logs and replicate logs to the Kinesis stream of a neighboring region.
- ✅ Use the CloudWatch Logs agent to consolidate logs into a single CloudWatch Logs group.

Explanation:-The resources can point to a CloudWatch logs destination to consolidate the logs using Subscription filters. You can collaborate with an owner of a different AWS account and receive their log events on your AWS resources, such as a Amazon Kinesis stream (this is known as cross-account data sharing). For example, this log event data can be read from a centralized Amazon Kinesis stream to perform custom processing and analysis. Custom processing is especially useful when you collaborate and analyze data across many accounts. For example, a company's information security group might want to analyze data for real-time intrusion detection or anomalous behaviors so it could conduct an audit of accounts in all divisions in the company by collecting their federated production logs for central processing. A real-time stream of event data across those accounts can be assembled and delivered to the information security groups who can use Kinesis to attach the data to their existing security analytic systems.

Q57)

A company's social media manager requests more staff on the weekends to handle an increase in customer contacts from a particular region.

The company needs a report to visualize the trends on weekends over the past 6 months using QuickSight.

How should the data be represented?

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

Explanation:-The main requirement is to track the customer contacts over a period of 6 months, a line graph with customer contacts vs. time would meet the requirement.

Q58)

Company A operates in Country X. Company A maintains a large dataset of historical purchase orders that contains personal data of their customers in the form of full names and telephone numbers. The dataset consists of 5 text files, 1TB each.

Currently the dataset resides on-premises due to legal requirements of storing personal data in-country. The research and development department needs to run a clustering algorithm on the dataset and wants to use Elastic Map Reduce service in the closest AWS region. Due to geographic distance, the minimum latency between the on-premises system and the closet AWS region is 200 ms.

Which option allows Company A to do clustering in the AWS Cloud and meet the legal requirement of maintaining personal data in-country?

- Use AWS Import/Export Snowball device to securely transfer the data to the AWS region and copy the files onto an EBS volume. Have the EMR cluster read the dataset using EMRFS.
- Encrypt the data files according to encryption standards of Country X and store them on AWS region in Amazon S3. Have the EMR cluster read the dataset using EMRFS.
- Establish a Direct Connect link between the on-premises system and the AWS region to reduce latency. Have the EMR cluster read the data directly from the on-premises storage system over Direct Connect.
- ✅ Anonymize the personal data portions of the dataset and transfer the data files into Amazon S3 in the AWS region. Have the EMR cluster read the dataset using EMRFS.

Explanation:-The latency is high it would be ideal to transfer the data to AWS and process using EMR and EMRFS. Also, anonymizing the data would help meet the legal requirement.

Q59)

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

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

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

algorithm for retries until a successful response is received.

Explanation:-You can use Kinesis PutRecord to insert data into Kinesis. To handle failure of PutRecord, AWS recommends using Error Retries and Exponential Backoff in AWS. The request rate for the stream is too high, or the requested data is too large for the available throughput. Reduce the frequency or size of your requests. For more information, see Streams Limits in the Amazon Kinesis Data Streams Developer Guide, and Error Retries and Exponential Backoff in AWS in the AWS General Reference.