Answer Sheet

Q1)

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 abucket in the new country's region.
- Continue to serve customers from the existing region while using Amazon Kinesis to stream transaction data to the regulator.
- ✅ Use CloudFormation to replicate the production application to the new region.
- Use Amazon Cloud Front to serve application content locally in the country; Amazon Cloud Front logs will satisfy the requirement.

Q2)

Your educational institution is setting up a new Amazon ES cluster outside a VPC and must give the university IT staff access to use Kibana so they can query student operations in a computer science lab. You want to configure the Amazon ES to allow access to Kibana from instances inside a public-facing VPC.

How do you allow access for the university IT staff?

- Use the university's Microsoft Active Directory.
- Use a username and password that you create for the professors.
- ✅ Use the Elastic IP address used by the VPC managed NAT gateway.

Explanation:-Using a username and password is not scalable. Using an Elastic IP address with VPC would be secure and would solve the problem and would be scalable to the university. Active Directory is not integrated with ES.

- Use the IAM role used by the Amazon EC2.

Q3)

An IoT vendor is looking to load data into Amazon Redshift in the most efficient way possible.

What is the best way to incrementally load data from SQL Server running on Amazon EC2 instances to Amazon Redshift?

- Use S3 distcp to copy the data to Amazon S3 and EMR to compress and send the data to Redshift.
- Use Amazon Kinesis Data Firehose to copy data from SQL Server to Amazon Redshift.
- ✅ Use AWS DMS to copy data into Amazon S3 and use Amazon Redshift Spectrum to query the data.

Explanation:-Copying data is the most efficient way to put data into Amazon Redshift. Methods that directly insert into Redshift are not as efficient.

- Use AWS Database Migration Service (AWS DMS) to load data directly into Amazon Redshift.

Q4)

A Big Data specialist is looking to perform natural language processing on data that lives in Amazon S3 using Python and is 5 TBs.

What is the best choice?

- Use a Spot instance, install Jupyter, install Sklearn, and then create a model on that instance.
- ✅ Use AWS SageMaker's built-in algorithms.

Explanation:-AWS SageMaker is designed to work with Amazon S3 data and allows for scalable modeling in Python.

- All of these.
- Use Cloud9, install Matplotlib, and then deploy a Lambda function that returns the results of a Sklearn operation to S3.

Q5)

A company uses AWS Batch to run large PCA jobs. The Big Data specialist has been told to optimize costs on these jobs because they have exceeded the planned budget.

What next step is best suited to achieve that goal?

- All of these.
- Set up AWS Batch to run single jobs that span multiple EC2 instances.
- Set up AWS Batch to create fine-grained access control.
- ✅ Set up AWS Batch to dynamically bid on Spot instances.

Explanation:-It optimizes cost savings because AWS Spot instances can save up to 90% compared to on-demand.

Q6)

The VP of Engineering has asked a Big Data specialist to identify the AWS services using serverless technology that the company has in production.

What is the best answer?

- AWS Lambda and EC2
- Comprehend and EMR
- ✅ DeepLens and Step Functions

Explanation:-Both DeepLens and Step Functions have AWS Lambda embedded. Other options are either only partially correct or are not correct.

- Step Functions and EMR

Q7)

The CTO of a computer vision startup has asked you to enable the company to launch thousands of EC2 instances from a command-line tool so the startup can do distributed computer vision modeling.

What options would solve the problem for the CTO and be the cost-efficient?

- ✅ All of these.

Explanation:-All the methods can use Spot instances, which can significantly save costs. They also can run computer vision tasks.

- Use EMR with Spark.
- Use AWS Batch.
- Use Spot instances with SQS.

Q8)

You must create a system to monitor application log files through real-time search plus visualizations, search user tickets, and search support documentation.

What is the best AWS Big Data application architecture for this requirement?

- ✅ AWS Elasticsearch for search with Kibana for visualization

Explanation:-AWS Elasticsearch supports full-text search and comes bundled with Kibana.

- AWS Athena for search integrated with AWS QuickSight for visualization
- AWS Elasticsearch for search integrated with AWS QuickSight for visualization
- AWS Athena for search integrated with Kibana for visualization

---

Q9)

You have 50TB of data that must be queried monthly for a recurring report with only tabular summary data.

Which technical architecture would work and enable the lowest total operating cost?

- ✅ AWS S3 for storage and AWS Athena for reporting

Explanation:-S3 has much lower operating costs than Redshift for storage, and AWS Athena cannot directly query AWS Glacier.

- AWS Glacier for storage and AWS Athena for reporting
- AWS Redshift for storage and AWS Elasticsearch for reporting
- AWS Redshift for storage and reporting

Q10)

A Redshift cluster contains 60TB of weekly PAGEVIEW table from a sales website on a monthly rolling window. The PAGEVIEW data has a JOIN with a PRODUCT data table, and the JOIN is ORDERED BY date and page_id, in priority order. The query also contains a filter on customer_id taken during the visit. Unfortunately, the current query is running too slowly.

What SORTKEY type and key order should be implemented to improve performance?

- INTERLEAVED sort key with unordered keys
- INTERLEAVED sort key with ordered keys
- COMPOUND sort key with keys ordered by date and page_id
- ✅ COMPOUND sort key ordered by date, page_id, and customer_id

Explanation:-This data is time series data. The COMPOUND sort key should be used to speed up JOIN, GROUP BY, and ORDER BY operations. The keys should be listed in weighted order because the problem statement says that date has a higher priority than page_id; and because customer_id is implicitly prioritized last, B is the correct response.

Q11)

A Redshift cluster is in the EU (Ireland) region and consists of 75 ds2.8xlarge nodes for a total storage capacity of 1.2PB. The largest tables have frequent joins on the customer_id in the SALES data tables, and the business_unit and product_category from the PRODUCTS data tables. Both the SALES and PRODUCTS tables are very large.

What distribution style should be used for the SALES and PRODUCTS data tables?

- ALL distribution style for SALES data tables and EVEN distribution style for PRODUCTS data tables
- DISTKEY distribution style for SALES data tables and ALL distribution style for PRODUCTS data tables
- EVEN distribution style for SALES data tables and KEY distribution style for PRODUCTS data tables
- ✅ KEY distribution style for SALES data tables and KEY distribution style for PRODUCTS data tables

Explanation:-The KEY distribution style should be used when a table is used in JOINs.

Q12)

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 S3 bucket and write data into folders by day. Execute the EMR job on the daily folder.
⚪ 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 log data into folders by device. Execute the EMR job on the device folders.

Q13)

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 CloudFormation templates.
✅ Provide the auditor with SSH keys for access to the Amazon EMR cluster.
⚪ Provide the auditor IAM accounts with the Security Audit policy attached to their group.
⚪ Provide the auditor with access to AWS DirectConnect to use their existing tools.

Q14)

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 analyzed at a 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 toanalyze.
⚪ Load all data from a different database/warehouse to S3. Use Redshift COPY command to copydata to Redshift for analysis.
⚪ Install Presto on the EMR cluster where Hive sits. Configure MySQL and PostgreSQL connector toselect from different data sources in a single query.

Q15)

An Amazon EMR cluster using EMRFS has access to petabytes of data on Amazon S3, originating from multiple unique data sources. The customer needs to query common fields across some of the data sets to be able to perform interactive joins and then display results quickly.

Which technology is most appropriate to enable this capability?

⚪ Pig
✅ MicroStrategy
⚪ Presto
⚪ R Studio

Q16)

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 Dynamo DBtable for each week in advance.

How should the administrator accomplish this task?

⚪ Feed the data into Amazon Machine Learning and build a binary classification model.
⚪ Feed the data into Apache Mahout and build a multi-classification model.
✅ Feed the data into Spark Mlib and build a random forest modest.
⚪ Feed the data into Amazon Machine Learning and build a regression model.

Q17)

A data engineer is about to perform a major upgrade to the DDL contained within an Amazon Redshift cluster to support a new data warehouse application. The upgrade scripts will include user permission updates, view and table structure changes as well as additional loading and data manipulation tasks. The data engineer must be able to restore the database to its existing state in the event of issues.

Which action should be taken prior to performing this upgrade task?

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

Q18)

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 metric 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?

- Create an Amazon Machine Learning model and invoke it with AWS Lambda.
- Store the data points in an Amazon DynamoDB table and poll if for peak metrics data from anAmazon EC2 application.
- ✅ Create an AWS IoT rule to generate an Amazon SNS notification.
- Use Amazon Kinesis Streams and a KCL-based application deployed on AWS Elastic Beanstalk.

Q19)

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?

- Binary Classification with a categorical target attribute.
- Regression model with a numeric target attribute.
- Multi-class classification model with a categorical target attribute.
- ✅ K-Nearest Neighbors model with a multi-class target attribute.

Q20)

A data engineer is running a DWH on a 25-node Redshift cluster of a SaaS service. The data engineer needs to build a dashboard that will be used by customers. Five big customers represent 80% of usage, and there is a long tail of dozens of smaller customers.

The data engineer has selected the dashboarding tool.

How should the data engineer make sure that the larger customer workloads do not interfere with the smaller customer workloads?

- Route the largest customers to a dedicated Redshift cluster. Raise the concurrency of the multi-tenant Redshift cluster to accommodate the remaining customers.
- Push aggregations into an RDS for Aurora instance. Connect the dashboard application to Aurorarather than Redshift for faster queries.
- Place the largest customers into a single user group with a dedicated query queue and place therest of the customers into a different query queue.
- ✅ Apply query filters based on customer-id that can NOT be changed by the user and applydistribution keys on customer-id.

Q21)

An Amazon Kinesis stream needs to be encrypted.

Which approach should be used to accomplish this task?

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

Q22)

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 afterimage upload

Which strategy should be used to meet these requirements?

- ✅ 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.
- 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.
- Write image and metadata 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.
- Write images and metadata to Amazon Kinesis. Use a Kinesis Client Library (KCL) application torun the image processing and save the image output to Amazon S3 and metadata to the apprepository DB.

Q23)

A customer needs to determine the optimal distribution strategy for the ORDERS fact table in its Redshift schema.

The ORDERS table has foreign key relationships with multiple dimension tables in this schema.

How should the company determine the most appropriate distribution key for the ORDERS table?

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

Q24)

A customer is collecting clickstream data using Amazon Kinesis and is grouping the events by IP address into 5-minute chunks stored in Amazon S3. Many analysts in the company use Hive on Amazon EMR to analyze this data. Their queries always reference a single IP address.

Data must be optimized for querying based on IP address using Hive running on Amazon EMR.

What is the most efficient method to query the data with Hive?

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

Q25)

An online retailer is using Amazon DynamoDB to store data related to customer transactions. The items in the table contains several string attributes describing the transaction as well as a JSON attribute containing the shopping cart and other details corresponding to the transaction. Average item size is – 250KB, most of which is associated with the JSON attribute. The average customer generates – 3GB of data per month.

Customers access the table to display their transaction history and review transaction details as needed. Ninety percent of the queries against the table are executed when building the transaction history view, with the other 10% retrieving transaction details. The table is partitioned on Customer ID and sorted on transaction date.The client has very high read capacity provisioned for the table and experiences very even utilization, but complains about the cost of Amazon DynamoDB compared to other No SQL solutions.

Which strategy will reduce the cost associated with the client's read queries while not degrading quality?

- Create an LSI sorted on date, project the JSON attribute into the index, and then query the primary table for summary data and the LSI for JSON details.
- ✅ Vertically partition the table, store base attributes on the primary table, and create a foreign key reference to a secondary table containing the JSON data. Query the primary table for summary data and the secondary table for JSON details.
- Change the primary table to partition on TransactionID, create a GSI partitioned on customer and sorted on date, project small attributes into GSI, and then query GSI for summary data and the primary table for JSON details.
- Modify all database calls to use eventually consistent reads and advise customers that transaction history may be one second out-of-date.

Q26)

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.

Q27)

An organization needs a data store to handle the following data types and access pattern -

Faceting

Search

Flexible schema (JSON) and fixed schema

Noise word elimination

Which data store should the organization choose?

- Amazon DynamoDB
- Amazon Redshift
- ✅ Amazon Relational Database Service (RDS)
- Amazon Elasticsearch Service

Q28)

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

Q29)

A system engineer for a company proposes digitalization and backup of large archives for customers.

The systems engineer needs to provide users with a secure storage that makes sure that data will never be tampered with once it has been uploaded.

How should this be accomplished?

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

Q30)

An organization needs to design and deploy a large-scale data storage solution that will be highly durable and highly flexible with respect to the type and structure of data being stored. The data to be stored will be sent or generated from a variety of sources and must be persistently available for access and processing by multiple applications.

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

- Launch an Amazon Relational Database Service (RDS), and use the enterprise grade and capacity of the Amazon Aurora engine for storage, processing, and querying.
- Use Amazon Redshift with data replication to Amazon Simple Storage Service (S3) for comprehensive durable data storage, processing, and querying.
- Deploy a long-running Amazon Elastic MapReduce (EMR) cluster with Amazon Elastic BlockStore (EBS) volumes for persistent HDFS storage and appropriate Hadoop ecosystem tools for processing and querying.
- ✅ Use Amazon Simple Storage Service (S3) as the actual data storage system, coupled with appropriate tools for ingestion/acquisition of data and for subsequent processing and querying.

Q31)

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?

1. Turn on EMRFS consistent view when configuring the EMR cluster.

2. Use AWS Data Pipeline to orchestrate the data processing cycles.

3. Set hadoop.data.consistency = true in the core-site.xml file.

4. Set hadoop.s3.consistency = true in the core-site.xml file.

- ✅ Only 1, 2 and 4
- Only 2, 3 and 4
- Only 1, 2 and 3
- All of these

Q32)

An Amazon Redshift Database is encrypted using KMS. A data engineer needs to use the AWS CLI to create a KMS encrypted snapshot of the database in another AWS region.

Which three steps should the data engineer take to accomplish this task? (Select three options)

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

Q33)

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

Managers must only see information about their direct reports.

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

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

---

Q34)

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 CloudWatch Logs and use an AWS Lambda function to load them into HDFS on an Amazon Elastic MapReduce (EMR) cluster 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 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 directly to a Kinesis Firehose. Configure Kinesis Firehose to load the events into an Amazon Redshift cluster for analysis.

Q35)

An organization uses a custom map reduce application to build monthly reports based on many small data files in an Amazon S3 bucket. The data is submitted from various business units on a frequent but unpredictable schedule. As the dataset continues to grow, it becomes increasingly difficult to process all of the data in one day. The organization has scaled up its Amazon EMR cluster, but other optimizations could improve performance.

The organization needs to improve performance with minimal changes to existing processes and applications.

What action should the organization take?

- ✅ Schedule a daily AWS Data Pipeline process that aggregates content into larger files using S3DistCp.
- Use Amazon S3 Event Notifications and AWS Lambda to index each file into an Amazon Elasticsearch Service cluster.
- Add Spark to the Amazon EMR cluster and utilize Resilient Distributed Datasets in-memory.
- Use Amazon S3 Event Notifications and AWS Lambda to create a quick search file index in DynamoDB.

Q36)

An administrator is processing events in near real-time using Kinesis streams and Lambda.

Lambda intermittently fails to process batches from one of the shards due to a 5-munite time limit.

What is a possible solution for this problem?

- Ignore and skip events that are older than 5 minutes and put them to Dead Letter Queue (DLQ).
- ✅ Reduce the batch size that Lambda is reading from the stream.
- Add more Lambda functions to improve concurrent batch processing.
- Configure Lambda to read from fewer shards in parallel.

Q37)

An organization uses Amazon Elastic MapReduce(EMR) to process a series of extract-transform-load (ETL) steps that run in sequence.

The output of each step must be fully processed in subsequent steps but will not be retained.

Which of the following techniques will meet this requirement most efficiently?

- Define the ETL steps as separate AWS Data Pipeline activities.
- Load the data to be processed into HDFS, and then write the final output to Amazon S3.
- Use the s3n URI to store the data to be processed as objects in Amazon S3.
- ✅ Use the EMR File System (EMRFS) to store the outputs from each step as objects in Amazon Simple Storage Service (S3).

Q38)

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

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

Which solution should the data engineer choose?

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

Q39)

A telecommunications company needs to predict customer churn (i.e., customers who decide to switch to a competitor). The company has historic records of each customer, including monthly consumption patterns, calls to customer service, and whether the customer ultimately quit the service. All of this data is stored in Amazon S3. The company needs to know which customers are likely going to churn soon so that they can win back their loyalty.

What is the optimal approach to meet these requirements?

- Use a Redshift cluster to COPY the data from Amazon S3. Create a User Defined Function in Redshift that computes the likelihood of churn.
- ✅ Use EMR to run the Hive queries to build a profile of a churning customer. Apply a profile to existing customers to determine the likelihood of churn.
- Use AWS QuickSight to connect it to data stored in Amazon S3 to obtain the necessary business insight. Plot the churn trend graph to extrapolate churn likelihood for existing customers.
- Use the Amazon Machine Learning service to build the binary classification model based on the dataset stored in Amazon S3. The model will be used regularly to predict churn attribute for existing customers.

Q40)

A 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 infrequent Access Storage.
- ✅ Write file contents to Amazon ElastiCache.
- Copy files to Amazon S3 Standard Storage.
- Write file contents to an Amazon DynamoDB table.

Q41)

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

Q42)

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?

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

Q43)

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?

1. Compare the S3 Etags from the source and destination objects.

2. Call the S3 CompareObjects API for the source and destination objects.

3. Place a HEAD request against the source and destination objects comparing SIG v4.

4. Compare the size of the source and destination objects.

- Only 1 and 3
- ✅ Only 1 and 4

- Only 2 and 3
- Only 1 and 2

Q44)

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 options)

- The structure of any LSIs that have been defined on the 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
- The structure of any GSIs that have been defined on the table

---

Q45)

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.
- 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 EC-2 based Hadoop cluster with spot instances to run a Spark job that performs a stochastic gradient descent optimization.

Q46)

An administrator tries to use the Amazon Machine Learning service to classify social media posts that mention the administrator's company into posts that require a response and posts that do not. The training dataset of 10,000 posts contains the details of each post including the timestamp, author, and full text of the post.The administrator is missing the target labels that are required for training.

Which Amazon Machine Learning model is the most appropriate for the task?

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

---

Q47)

A medical record filing system for a government medical fund is using an Amazon S3 bucket to archive documents related to patients. Every patient visit to a physician creates a new file, which can add up millions of files each month. Collection of these files from each physician is handled via a batch process that runs ever night using AWS Data Pipeline. This is sensitive data, so the data and any associated metadata must be encrypted at rest.

Auditors review some files on a quarterly basis to see whether the records are maintained according to regulations. Auditors must be able to locate any physical file in the S3 bucket for a given date, patient, or physician. Auditors spend a significant amount of time location such files.

What is the most cost and time-efficient collection methodology in this situation?

- ✅ Use Amazon S3 event notification to populate an Amazon Redshift table with metadata about every file loaded to Amazon S3, and partition them based on the month and year of the file.
- Use Amazon S3 event notification to populate an Amazon DynamoDB table with metadata about every file loaded to Amazon S3, and partition them based on the month and year of the file.
- Use Amazon API Gateway to get the data feeds directly from physicians, batch them using a Spark application on Amazon Elastic MapReduce (EMR), and then store them in Amazon S3 with folders separated per physician.
- Use Amazon Kinesis to get the data feeds directly from physicians, batch them using a Spark application on Amazon Elastic MapReduce (EMR), and then store them in Amazon S3 with folders separated per physician.

Q48)

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.
- 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.

Q49)

A company uses Amazon Redshift for its enterprise data warehouse. A new on-premises PostgreSQL OLTP DB must be integrated into the data warehouse. Each table in the PostgreSQL DB has an indexed last_modified timestamp column.

The data warehouse has a staging layer to load source data into the data warehouse environment for further processing.
The data lag between the source PostgreSQL DB and the Amazon Redshift staging layer should NOT exceed four hours.

What is the most efficient technique to meet these requirements?

- Extract the incremental changes periodically using a SQL query. Upload the changes to a single Amazon Simple Storage Service (S3) object, and run the COPY command to load to the Amazon Redshift staging layer.
- Extract the incremental changes periodically using a SQL query. Upload the changes to multiple Amazon Simple Storage Service (S3) objects, and run the COPY command to load to the Amazon Redshift staging layer.
- Use a PostgreSQL trigger on the source table to capture the new insert/update/delete event and write it to Amazon Kinesis Streams. Use a KCL application to execute the event on the Amazon Redshift staging table.
- ✅ Create a DBLINK on the source DB to connect to Amazon Redshift. Use a PostgreSQL trigger on the source table to capture the new insert/update/delete event and execute the event on the Amazon Redshift staging table.

Q50)

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 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
- Machine Learning on D instance types and ad-hoc queries on I instance types

---

Q51)

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 Amazon Redshift with synchronous replication to Amazon Simple Storage Service (S3) and row-level locking for strong consistency.
- Deploy a NoSQL data store on top of an Amazon Elastic MapReduce (EMR) cluster, and select the HDFS High Durability option.
- Use Amazon DynamoDB as the data store and use strongly consistent reads when necessary.
- ✅ Use an Amazon Relational Database Service (RDS) instance sized to meet the maximum anticipated transaction rate and with the High Availability option enabled.

Q52)

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

Q53)

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

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

On which basis should this binary classification model be built?

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

Q54)

A company with a support organization needs support engineers to be able to search historic cases to provide fast responses on new issues raised.

The company has forwarded all support messages into an Amazon Kinesis Stream. This meets a company objective of using only managed services to reduce operational overhead.The company needs an appropriate architecture that allows support engineers to search on historic cases and find similar issues and their associated responses.

Which AWS Lambda action is most appropriate?

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

Q55)

A solutions architect works for a company that has a data lake based on a central Amazon S3 bucket. The data contains sensitive information.

The architect must be able to specify exactly which files each user can access. Users access the platform through a SAML federation Single.
The architect needs to build a solution that allows fine grained access control, traceability of access to the objects, and usage of the standard tools (AWS Console, AWS CLI) to access the data.

Which solution should the architect build?

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

Q56)

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?

- R Studio
- IPython/Jupyter
- D3.js
- ✅ Hue

Q57)

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

Q58)

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

A data engineer needs to run multiple interactive queries to manipulate the data. The Data Engineer has AWS access to spin up an Amazon EMR cluster. The data engineer needs to use an application on the cluster to process this data and return the results in interactive time frame.

Which application on the cluster should the data engineer use?

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

Q59)

A media advertising company handles a large number of real-time messages sourced from over 200 websites. The company's data engineer needs to collect and process records in real time for analysis using Spark Streaming on Amazon Elastic MapReduce (EMR). The data engineer needs to fulfill a corporate mandate to keep ALL raw messages as they are received as a top priority.

Which Amazon Kinesis configuration meets these requirements?

- Publish messages to Amazon Kinesis Streams, pull messages off with Spark Streaming, and write row data to Amazon Simple Storage Service (S3) before and after processing.
- Publish messages to Amazon Kinesis Firehose backed by Amazon Simple Storage Service (S3). Use AWS Lambda to pull messages from Firehose to Streams for processing with Spark Streaming.
- ✅ Publish messages to Amazon Kinesis Streams. Pull messages off Streams with Spark Streaming in parallel to AWS Lambda pushing messages from Streams to Firehose backed by Amazon Simple Storage Service (S3).
- Publish messages to Amazon Kinesis Firehose backed by Amazon Simple Storage Service (S3). Pull messages off Firehose with Spark Streaming in parallel to persistence to Amazon S3.

Q60)

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 access control and requires the LEAST amount of management work?

- 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 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. 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 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.
