Answer Sheet

Q1)

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?

- Encrypt using S3 Server Side Encryption.
- Encrypt using SSL/TLS for encrypting the data.
- Kinesis does not support encryption
- ✓ Encrypt using Kinesis Server Side Encryption.

Q2)

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

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

It is important to the success of the application that the vast majority of user requests are met in a cost-effective way.

How should this table be created?

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

Q3)

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?

- Ganglia
- QuickSight
- ✓ Presto
- YARN

Q4)

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?

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

Q5)

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

Which type of visualization do you choose?

- ✓ Line Chart
- Pivot Table
- Pie Chart
- Scatter Plot

Q6)

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

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

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

Q7)

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

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

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

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

Q8)

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

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

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

---

Q9)

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

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

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

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

Q10)

Your application requires long-term storage for backups and other data that you need to keep readily available but with lower cost.

Which S3 storage option should you use?

✅ AWS 53 Standard - Infrequent Access
⚪ AWS Glacier
⚪ AWS S3 Standard
⚪ AWS Reduced Redundancy Storage

Q11)

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

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

Which design would you choose to meet these requirements?

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

---

Q12)

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

As a part of a security audit, it was recommended hardening of RDS instance.

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

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

Q13)

There is a requirement to migrate 3TB of data to AWS.

There is a restriction of the time to migrate the data and there is a limitation of only a 1OOMBit line to the AWS Cloud.

What is the best solution to use to migrate the data to the cloud?

- ✓ Amazon Import/Export
- Amazon Direct Connect
- Amazon Storage Gateway
- Amazon 53

Q14)

Which of the following full managed service can be used to deliver real-time streaming data to 53.

Please select:

- EMR
- Red shift
- Amazon Kinesis
- ✓ Amazon Kinesis Fire hose

Q15)

You are planning on using the AWS IoT Rules service to allow IoT enabled devices to write information to AWS Dynamo DB.

Which of the following actions must be done to ensure that the rules will work as intended? Please select:

- Ensure that the IoT devices are defined In the Amazon VPC.
- ✓ Ensure that the right IAM permissions to AWS Dynamo DB is given
- Ensure that the IoT devices are defined in the AWS Config
- Ensure that the devices are recorded as users in IAM

Q16)

An application stores its data as different buckets in S3.

The access to these buckets need to be based on single-sign on which used Microsoft Active Directory.

Which of the following Implementation techniques can be used?

- Create bucket policies based on the groups defined In Microsoft AD
- Use Active Directory connector to ensure the access can be provided based on the user?s permission.
- ✓ Use IAM federation and ensure the access is given based on the user?s group defined in AD 1
- Use Impersonation to ensure the user can access the required bucket.

Q17)

You are managing the development of an application that uses Dynamo DB to store JSON data.

You have already set the Read and Write capacity of the Dynamo DB table. You are unsure of the amount of the traffic that will be received by the application during the deployment time.

How can you ensure that the Dynamo DB is not highly throttled and does not become a bottleneck for the application?

- ✓ Dynamo DBs auto scaling feature will make sure that no read/write throttling will happen due to heavy traffic.
- Create a Cloud watch alarm which would then send a trigger to AWS Lambda to increase the Read and Went capacity of the Dynamo DB table.
- Monitor the System Errors metric using Cloud watch
- Create a Cloud watch alarm which would then send a trigger to AWS Lambda to create a new Dynamo DB table.

Q18)

You have just developed a new mobile application that handles analytics workloads on large scale datasets that are stored on Amazon Red shift.

Consequently, the application needs to access Amazon Red shift tables.

Which of the below methods would be the best, both practically and security-wise, to access the tables?

- ✓ Use roles that allow a web identity federated user to assume a role that allows access to the Red shift table by providing temporary credentials.
- Create a Red shlft read-only access policy In IAM and embed those credentials in the application.

- Create a HSM client certificate in Red shift and authenticate using this certificate
- Create an IAM user and generate encryption keys for that user. Create a policy for Red shift Embed the keys in the application.

Q19) Which of the following API commands can be used to data Into a Kinesis stream for Synchronous processing?

- Set Record
- ✓ Write Record
- Send Record
- Put Record

Q20) Which of the following services can be used for transformation of incoming source data in Amazon Kinesis Data Fire hose ?

- AWS EMR
- ✓ AWS LOT
- AWS Lambda
- AWS Machine Learning

Q21)

There is a requirement to analyze a large set of data updates from Kinesis and Dynamo DB.

Which Big Data tool would be ideal for this scenario?

- Red shift
- EMR
- ✓ Elastic search
- SQS

Q22) Which of the following can be used to enable developers to quickly get started with deep learning in the cloud?

- ✓ Tensor Flow
- Amazon Red shift
- Spark ML on Amazon EMP
- Amazon ML

Q23)

Which of the below components is used to pump data into a Kinesis Stream.

Please select:

- Shards
- Kinesis Fire hose
- Consumers
- ✓ Producers

Q24)

You are currently managing an application that use the Using the Kinesis Client Library to read a Kinesis stream.

You have noticed that you are getting a Provisioned Through put Exceeded Exception in Cloud watch from the stream.

What are possible solutions to rectify this error? Select 2 Please select:

- Add more applications that use the KCL library
- Use more Kinesis streams
- Add more shards to your Kinesis Stream
- ✓ Add retry logic to applications that use the KCL library ?

Q25)

You have a requirement to have a data store in AWS.

The further following requirements have to be in place Showfor this data store

1. Complex SQL queries and transacuons Show

2. Fixed schema Fi

Which of the following would fulfill this requirement?

- ✓ Amazon RDS
- AmazonS3
- Amazon Kinesis
- Amazon Dynamo DB

Q26)

There is a red shift table created in AWS. There is a requirement to ensure that the table can be frequently used in join level

queries.

Which of the following distribution styles would you utilize for the table Please select?

- [x] KEY
- [ ] EVEN
- [ ] PARTITION
- [ ] ALL

Q27)

Which of the following options need to be incorporated in an EMR cluster for better security, for example, fc highly confidential data and for auditing purpose.

Choose 3 answers from the options given below

- [ ] Use SSL for data in transit between nodes in a cluster
- [ ] Use Server .side encryption on 53 using any one of the available encryption mechanisms
- [ ] Use Hive encryption
- [x] Ensure EBS Volumes are encrypted on the EC2 Instances

Q28)

Which of the following methods can be used to disable automated snapshots in Red shift

Please select:

- [ ] Set the retention period to 0 (zero)
- [ ] Set Automated Snapshots to false
- [x] Set the retention period to -1
- [ ] Uncheck Automated Snapshots

Q29) Which of the following is not a node type in Amazon EMR?

- [x] Primary Node
- [ ] Master Node
- [ ] Core Node
- [ ] Task Node

Q30)

What is the default retention period for a Kinesis stream?

Please select:

- [ ] 5 days
- [ ] 3 days
- [ ] 7 days
- [x] 1 day

Q31)

The AWS Documentation mentions the following Amazon Athena Is an interactive query service that makes It easy to analyze data In Amazon S3 using standard SQL.

At Which of the following services can be used as a business analytics service that can be used to build visualizations

- [ ] AWS EMR
- [ ] AWS Kinesis
- [x] AWS Quick sight
- [ ] AWS Visualizes

Q32)

Which of the following are examples of Columnar databases?

Select 2 Answers Please select:

- [ ] Amazon EMR
- [x] Apache H Base
- [x] Amazon Red shift
- [ ] Amazon Dynamo DB

Q33)

You have a series of locations in S3 where files need to be copied onto AWS Red shift.

Which of the following can be used to specify the location of the files that need to be copied?

- Create an event in 53 with the list of flies to be copied
- Define a configuration file in Red shift with the location of the files
- Ensure the metadata for the bucket has the list of files that need to be copied
- ✓ Create a manifest file with the list of files to be copied

Q34) Which of the following commands can be used to see the impact of a query on a Red shift Table?

- PLAN
- EXECUTE
- EXPLAIN
- ✓ TRY

Q35) In order to efficiently insert or update data into a Red shift table, which of the following must be carried out Show?

- ✓ Setup a staging table
- Use the Single merge command
- Use the Up set command
- Issue the Insert command

Q36) Which of the following is an open source deep learning framework?

- ✓ Apache MX Net
- AWS ML
- jupyter
- D3.js

Q37)

You are currently designing a Red shift table in AWS. The data in this table will not be changed frequently. Also, this table will not participate in any joins.

Which of the following distribution style should be preferable for this table Please select:

- Even
- None
- ✓ All
- Key

Q38)

You are currently designing a Red shift table in AWS. You are responsible for ensuring maximum efficiency when queries are run against the tables In Red shift.

The tables will be performing multiple joins based on Customer ID. Also, each table would be storing around a million rows.

Which of the following distribution styles should be used when designing the table?

- ✓ All
- Even
- None
- Key

Q39)

Which of the following services are not supported when creating AWS IoT Rule actions?

Choose 2 answers from the options given below :

- AWS Cloud watch
- ✓ AWS Aurora
- AWS Dynamo DB
- ✓ AWS Red shift

Q40)

Your company has decided to use AWS Quick Sight for visualization. The CEO wants to see the target goals for each department of the company.

During the design phase it has been decided that the UI would have each department represented via a rectangle with the size of the rectangle defining the goal.

In such a case which of the following in Quick sight would help build this visualization properly

- Heat Maps
- KPI
- ✓ Scatter Plots
- Tree Map

Q41)

Which of the following facilitates the sending and receiving of messages with AWS IoT devices?

Please select:

- AWS SQS
- AWS Iot Publisher
- AWS SNS
- ✓ Message Broker

Q42)

A retailer exports data daily from its transactional databases into an S3 bucket in the Sydney region.

The retailer's Data Warehousing team wants to import this data into an existing Amazon Redshift cluster in their VPC at Sydney.

Corporate security policy mandates that data can only be transported within a VPC.

What combination of the following steps will satisfy the security policy? Choose 2 answers

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

Q43)

Your company needs to design a data warehouse for a client in the retail industry. The data warehouse will store historic purchases in Amazon Redshift.

To comply with PCI:DSS requirements and meet data protection standards, the data must be encrypted at rest and have keys managed by a corporate on-premises HSM.

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

- Configure the AWS Key Management Service to point to the corporate HSM device, and then launch the Amazon Redshift cluster with the KMS managing the encryption keys.
- Use the AWS CloudHSM service to establish a trust relationship between the CloudHSM and the corporate HSM over a Direct Connect connection. Configure Amazon Redshift to use the CloudHSM device.
- ✓ Create a VPN connection between a VPC you create in AWS and an on-premises network. Then launch the Redshift cluster in the VPC, and configure it to use your corporate HSM.
- Use AWS Import/Export to import a company HSM device into AWS alongside the Amazon Redshift cluster, and configure Redshift to use the imported HSM.

Q44)

Your client needs to load a 600 GB file into a Redshift cluster from S3, using the Redshift COPY command.

The file has several known (and potentially some unknown) issues that will probably cause the load process to fail.

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

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

Q45)

A company has lot of web applications, databases and data warehouse built on Teradata, NoSQL databases, and other types of data stores.

They have lot of data assets in terms of logs, documents; excel files, CSV files, PDF documents and others. Web Application has different user workloads at different parts of the day.

They are running one of their web application Node.js supported by MongoDB Database. The schema designed is document based. The team wants to migrate the platform on to AWS.

Which NoSQL Managed service provides the document management capability?

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

---

Q46)

A company launched EMR cluster to support their big data analytics requirements.

They have multiple data sources built out of S3, SQL databases, MongoDB, Redis, RDS, other file systems. They are looking for distributed processing framework and programming model that helps you do machine learning, stream processing, or graph analytics using Amazon EMR clusters.

Which EMR Hadoop ecosystem fulfils the requirements?

✅ Apache Spark
⚪ Apache HCatalog
⚪ Apache HBase
⚪ Apache Hive

Q47)

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

The company's data engineer needs to collect and process records in real time for analysis using Spark Streaming on Amazon Elastic MapReduce (EMR).

The data engineer needs to fulfill a corporate mandate to keep ALL raw messages as they are received as a top priority.

Which Amazon Kinesis configuration meets these requirements?

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

Q48)

A customer has an Amazon S3 bucket. Objects are uploaded simultaneously by a cluster of servers from multiple streams of data.

The customer maintains a catalog of objects uploaded in Amazon S3 using an Amazon DynamoDB table. This catalog has the following fields:

StreamName, TimeStamp, and ServerName, from which ObjectName can be obtained.

The customer needs to define the catalog to support querying for a given stream or server within a defined time range.

Which DynamoDB table scheme is most efficient to support these queries?

⚪ Define a Primary Key with ServerName as Partition Key. Define a Local Secondary Index with TimeStamp as Partition Key. Define a Global Secondary Index with StreamName as Partition Key and TimeStamp as Sort Key.
⚪ Define a Primary Key with ServerName as Partition Key. Define a Local Secondary Index with StreamName as Partition Key. Define a Global Secondary Index with TimeStamp as Partition Key.
⚪ Define a Primary Key with ServerName as Partition Key and TimeStamp as Sort Key. Do NOT define a Local Secondary Index or Global Secondary Index.
✅ Define a Primary Key with StreamName as Partition Key and TimeStamp followed by ServerName as Sort Key. Define a Global Secondary Index with ServerName as partition key and TimeStamp followed by StreamName.

Q49)

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

Q50)

A data engineer is running a DWH on a 25-node Redshift cluster of a SaaS service. The data engineer needs to build a dashboard that will be used by customers.

Five big customers represent 80% of usage, and there is a long tail of dozens of smaller customers. The data engineer has selected the dashboarding tool.

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

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

Q51)

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.

Q52)

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".

Q53)

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

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

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

- Setup an Auto Scaling group of EC2 syslogd servers, store the logs on S3 use EMR to apply heuristics on the logs
- Configure Amazon CloudTrail to receive custom logs, use EMR to apply heuristics the logs
- ✅ Send all the log events to Amazon Kinesis develop a client process to apply heuristics on the logs
- Send all the log events to Amazon SQS. Setup an Auto Scaling group of EC2 servers to consume the logs and apply the heuristics.

Q54)

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

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

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

- ✅ In the mobile application, use HLS to display the video stream by using the HLS streaming session URL.
- Transform the stream data to HLS compatible data by using Kinesis Data Analytics or customer code in EC2/Lambda. Then in the mobile application, use HLS protocol to display the video stream by using the converted HLS streaming data.
- Create a Kinesis Data Firehose to ingest, durably store and encrypt the live videos from the users' home devices.
- ✅ Create a Kinesis video stream to capture, store, and index the videos from the camera-equipped home devices.

Q55)

You require the ability to analyze a customer?s clickstream data on a website so they can do behavioral analysis.

Your customer needs to know what sequence of pages and ads their customer clicked on. This data will be used in real time to modify the page layouts as customers click through the site to increase stickiness and advertising click-through.

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

- Publish web clicks by session to an Amazon SQS queue and periodically drain these events to Amazon RDS and analyze with SQL
- Write click events directly to Amazon Redshift and then analyze with SQL
- ✅ Push web clicks by session to Amazon Kinesis and analyze behavior using Kinesis workers
- Log clicks in weblogs by URL store to Amazon S3, and then analyze with Elastic MapReduce

---

Q56)

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

Currently, the on-premises storage holds 900 TB of data and is connected to the Internet over a 100 Mbps link.

Up to 20% of the link's throughput is regularly used in real time by existing systems.

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

- Set up an AWS Direct Connect link to upload the data.
- Use a multipart upload to transfer the data over the existing link.
- ✅ Order multiple AWS Snowball devices to ship the dat

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

Q57)

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
- Unary classification model, where the target class is the require-response post

---

Q58)

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.

Q59)

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.

Q60)

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

- Faceting

- Search

- Flexible schema (JSON) and fixed schema

- Noise word elimination

Which data store should the organization choose?

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