NEW PRACTICE TEST 3 - UPDATED

Attempt 1
Marks Obtained 1 / 48
Your score is 2.08%

Completed on Sunday , 03 February 2019 , 11:45 PM
Time Taken 00 H 00 M 14 S
Result Fail

Domains / Topics wise Quiz Performance Report

S.No.	Topic	Total Questions	Correct	Incorrect	Unattempted
1	Processing	11	0	0	11
2	Storage	9	1	0	8
3	Analysis	11	0	0	11
4	Collection	10	0	0	10
5	Data Security	5	0	0	5
6	Visualization	2	0	0	2

48 Questions	1 Correct	0 Incorrect	47 Unattempted

Show Answers: All

QUESTION 1 UNATTEMPTED PROCESSING

A company is planning on using Amazon Redshift as part of their ETL ecosystem. They want to ensure that they use the recommended practices for using AWS Redshift for the various process. Which of the following are recommended from AWS? Choose 2 answers from the options given below

- [ ] **A.** For various ETL processes which use AWS Redshift commits, use transaction handling ✔
- [ ] **B.** Extract Large results sets from Redshift using the select query
- [ ] **C.** Extract Large results sets from Redshift using the UNLOAD statement ✔

☐ **D.** Copy data into Redshift using Insert queries

Explanation :

Answer – A and C
The AWS Documentation mentions the following
ETL transformation logic often spans multiple steps. Because commits in Amazon Redshift are expensive, if each ETL step performs a commit, multiple concurrent ETL processes can take a long time to execute.
To minimize the number of commits in a process, the steps in an ETL script should be surrounded by a BEGIN...END statement so that a single commit is performed only after all the transformation logic
Use UNLOAD to extract large results sets directly to S3. After it's in S3, the data can be shared with multiple downstream systems. By default, UNLOAD writes data in parallel to multiple files according to the number of slices in the cluster. All the compute nodes participate to quickly offload the data into S3.
The other options are invalid since for large data sets use the UNLOAD command and for copying data use the COPY command
For more information on high performance for Redshift, please visit the url

- https://aws.amazon.com/blogs/big-data/top-8-best-practices-for-high-performance-etl-processing-using-amazon-redshift/ (https://aws.amazon.com/blogs/big-data/top-8-best-practices-for-high-performance-etl-processing-using-amazon-redshift/)

Ask our Experts

QUESTION 2 UNATTEMPTED STORAGE

A company has a large number of data sets defined in S3. They want to be able to load data in S3 as tables. You also need to be able to load table partitions automatically from Amazon S3. Which of the following combinations will allow to fulfil this requirement

○ **A.** EMR and Pig
○ **B.** EMR and Hive ✔
○ **C.** Redshift and Athena
○ **D.** Redshift and SQL

Explanation :

Answer – B
The AWS Documentation mentions the following

Hive is an open-source, data warehouse, and analytic package that runs on top of a Hadoop cluster. Hive scripts use an SQL-like language called Hive QL (query language) that abstracts programming models and supports typical data warehouse interactions. Hive enables you to avoid the complexities of writing Tez jobs based on directed acyclic graphs (DAGs) or MapReduce programs in a lower level computer language, such as Java.
Option A is incorrect since Pig is used in providing a scripting language that you can use to transform large data sets without having to write complex code in a lower level computer language like Java
Options C and D are incorrect because here there is no mention on exporting the data to a warehouse
For more information on EMR Hive, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive.html (https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hive.html)

Ask our Experts

QUESTION 3 UNATTEMPTED ANALYSIS

A company has a set of DynamoDB tables which is being used to store click stream data for a web application. There are millions and millions of rows of data in the tables. There is a requirement to perform joins to understand the most popular web pages. Which of the following can be used for this purpose?

- A. Data Pipeline and EMR
- B. EMR and Hive ✔
- C. Kinesis and Hive
- D. SQS

Explanation :

Answer – B
The AWS Documentation mentions the following
DynamoDB is a fully managed NoSQL database service that provides fast and predictable performance with seamless scalability. Developers can create a database table and grow its request traffic or storage without limit. DynamoDB automatically spreads the data and traffic for the table over a sufficient number of servers to handle the request capacity specified by the customer and the amount of data stored, while maintaining consistent, fast performance. Using Amazon EMR and Hive you can quickly and efficiently process large amounts of data, such as data stored in DynamoDB
Option A is invalid since you would not use EMR alone
Option C is invalid since EMR needs to be used in accordance with Hive
Option D is invalid since using SQS as a queue service will not help fulfil the requirement
For more information on EMR for DynamoDB, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ReleaseGuide/EMRforDynamoDB.html (https://docs.aws.amazon.com/emr/latest/ReleaseGuide/EMRforDynamoDB.html)

Note:

Kinesis Data Firehose doesn't support EMR as destination i.e it cannot deliver records to EMR .

Kinesis Data Firehose can send records to following services/tools :

- Amazon Simple Storage Service (Amazon S3)
- **Amazon Redshift**
- Amazon Elasticsearch Service (Amazon ES)
- Splunk

Attaching screenshot wherein you get an option to choose Kinesis Firehose delivery stream

Detailed Guide to Kinesis firehose delivery destinations :

- https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html (https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html)

Alternatively, we do have a option to send data to s3 and processing it using EMR using spark in real time(in micro batches)

Redshift would be the most appropriate answer as per options provided.

Ask our Experts

QUESTION 4 UNATTEMPTED COLLECTION

A company is planning on setting up an Ingestion stream for an application. The application will be sending records with sizes ranging from 2 MB to 5 MB. The data then needs to be processed and stored in S3. Which of the following can you use for the Ingestion process with the least amount of implementation?

- **A.** SQS
- **B.** Kafka ✔
- **C.** Kinesis Streams
- **D.** Kinesis Firehose

Explanation :

Answer – B
The Apache Kafta documentation mentions the following
Kafka is generally used for two broad classes of applications:

- Building real-time streaming data pipelines that reliably get data between systems or applications
- Building real-time streaming applications that transform or react to the streams of data

All other options will not work because of the limitations in the size of records which can be ingested. You would either need to do additional work of compressing or splitting the records and then ingesting it.
For more information on the producer aspect for Kafta, please refer to the below URL

- https://kafka.apache.org/documentation/#producerapi (https://kafka.apache.org/documentation/#producerapi)

Ask our Experts

QUESTION 5 UNATTEMPTED PROCESSING

A company has a set of users from different departments that process workloads on AWS Redshift. One set of users are issuing queries which take around 30 minutes to complete. Some users who want to issue queries which generally take 1-2 minutes sometimes have to wait for the long-standing queries to finish first. How can you ensure that all queries get processed at the same time?

- A. Add more nodes to the Redshift cluster
- B. Make use of Workload management ✔
- C. Change the Key distribution for the table
- D. Add a sort key for the table

Explanation :

Answer – B
The AWS documentation mentions the following
Amazon Redshift workload management (WLM) enables users to flexibly manage priorities within workloads so that short, fast-running queries won't get stuck in queues behind long-running queries. Amazon Redshift WLM creates query queues at runtime according to *service classes*, which define the configuration parameters for various types of queues, including internal system queues and user-accessible queues. From a user perspective, a user-accessible service class and a queue are functionally equivalent.
Option A would just add more cost and not necessarily allow the slow running queries to run
Options C and D are incorrect since changing the table structure will not allow the slow running queries to run at the same time as the long running queries
For more information on workload management, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/c_workload_mngmt_classification.html (https://docs.aws.amazon.com/redshift/latest/dg/c_workload_mngmt_classification.html)

Ask our Experts

QUESTION 6 UNATTEMPTED STORAGE

A company currently has a Redshift Cluster in their staging AWS account. They need to test out some business intelligence tools using the table data in the cluster. The IT Security department has advised to copy the data onto the test AWS account for conducting the tests. How can this be achieved? Choose 2 answers from the options given below

- [ ] **A.** Create a manual snapshot of the Redshift database ✔
- [ ] **B.** Enable cross-region snapshot copy
- [ ] **C.** Share the existing manual snapshot with the test account ✔
- [ ] **D.** Create an IAM user for the test account and give access to the manual snapshot in the staging account

Explanation :

Answer – A and C
The AWS documentation mentions the following
You can share an existing manual snapshot with other AWS customer accounts by authorizing access to the snapshot. You can authorize up to 20 for each snapshot and 100 for each AWS Key Management Service (AWS KMS) key. That is, if you have 10 snapshots that are encrypted with a single KMS key, then you can authorize 10 AWS accounts to restore each snapshot, or other combinations that add up to 100 accounts and do not exceed 20 accounts for each snapshot. A person logged in as a user in one of the authorized accounts can then describe the snapshot or restore it to create a new Amazon Redshift cluster under their account. For example, if you use separate AWS customer accounts for production and test, a user can log on using the production account and share a snapshot with users in the test account
Option B is incorrect since this is used for ensuring snapshots are available across regions for the same AWS account
Option D is incorrect since this is a security issue to give this sort of access
For more information on working with snapshots, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-snapshots.html (https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-snapshots.html)

Ask our Experts

QUESTION 7 UNATTEMPTED PROCESSING

A company is currently using an EMR cluster which runs 10 On-demand Instances. The jobs itself run only during the day time. The EMR cluster is only used for processing and reporting during business hours. Which of the following can be used to reduce the cost of the cluster? Choose 2 answers from the options given below

- [ ] **A.** Consider using Spot Instances ✔
- [ ] **B.** Consider using Reserved Instances
- [ ] **C.** Shutdown the cluster when not in use ✔
- [ ] **D.** Enable termination protection for the cluster

Explanation :

Answer – A and C
An example of this is given in the AWS Blogs

It is critical that we keep our Amazon EMR costs down as we scale up. To that end, we've adopted the following strategies

1. Use AWS Spot Instances rather than On-Demand Instances whenever possible. Amazon Elastic Cloud Compute (Amazon EC2) Spot Instances are unused Amazon EC2 capacity that you bid on; the price you pay is determined by the supply and demand for Spot Instances. The cost of using Spot Instances can be 80% less than using On-Demand Instances. It's important to manage Spot Instances because they can be terminated if the Spot market price exceeds your bid price. At BloomReach, we have written an orchestration system that schedules jobs on Amazon EMR. The system implements a Hartmann pipeline that can run a variety of jobs both locally and on Amazon EMR. It can also detect failures such as Spot Instance termination and reschedule jobs on different clusters as needed.

2. Create a system that shares clusters among several small jobs rather than launching a separate cluster for every job. Remember, whether your job takes 10 minutes or 60 minutes, you're paying for an hour of access. If you have four 10-minute jobs, you could share one cluster to do them all and be charged for one hour. Or you could employ one cluster for each and be charged for four hours. Sharing clusters among jobs also allows you to save the time and cost of bootstrapping a new cluster. The time savings alone can be a significant factor for real-time jobs.

3. Use Amazon EMR tags for cost tracking. Using EMR tags lets you track the cost of your cloud usage by project or by department, which gives you deeper insight into return on investment and provides transparency for budgeting purposes.

4. Create a lifecycle management system that allows you to track clusters and eliminate idle clusters.

5. Use the right instance types for your jobs. For example, use c3 instance type for compute-heavy jobs. This can significantly reduce waste and costs based on the scale of your jobs. Below is an algorithm we have found useful for selecting the instance type with the best value for compute capacity based on its Spot price:

Option B is incorrect since we don't know the long-term usage of the cluster to decide on using Reserved Instances
Option D is incorrect because ideally this is used to prevent accidental termination of the cluster
For more information on reducing EMR costs, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/strategies-for-reducing-your-amazon-emr-costs/ (https://aws.amazon.com/blogs/big-data/strategies-for-reducing-your-amazon-emr-costs/)

Ask our Experts

QUESTION 8 UNATTEMPTED COLLECTION

A company needs to send a 50 TB data set onto AWS which will then be ported onto AWS Redshift. Which of the following can help you achieve the transfer in a cost-effective manner?

- A. Upload over a VPN connection
- B. S3 Transfer Acceleration
- C. Upload over a Direct Connect
- D. Use AWS Snowball ✔

Explanation :

Answer – D
The AWS documentation mentions the following
AWS Snowball is a service that accelerates transferring large amounts of data into and out of AWS using physical storage devices, bypassing the Internet. Each AWS Snowball device type can transport data at faster-than internet speeds. This transport is done by shipping the data in the devices through a regional carrier. The devices are rugged shipping containers, complete with E Ink shipping labels.
Options A and B is incorrect since the bandwidth would be a limitation
Option C is incorrect since this would not be cost effective
For more information on AWS Snowball, please refer to the below URL

- https://docs.aws.amazon.com/snowball/latest/ug/whatissnowball.html (https://docs.aws.amazon.com/snowball/latest/ug/whatissnowball.html)

Ask our Experts

QUESTION 9 UNATTEMPTED STORAGE

A company needs a large data warehouse which can be used to store millions and millions of images uploaded by users. The data store itself should be able to scale on demand. Which of the following combinations would you use for this purpose?

- A. Store the Images in DynamoDB along with an image_id as the partition key
- B. Store the Images in Redshift along with an image_id as the distribution key
- C. Store the Images in S3 and store the S3 reference link in DynamoDB ✔
- D. Store the Images in S3 and store the S3 reference link in Glacier

Explanation :

Answer – C
The AWS documentation mentions the following
Amazon DynamoDB currently limits the size of each item that you store in a table (see Limits in DynamoDB (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Limits.html)). If your application needs to store more data in an item than the DynamoDB size limit permits, you can try compressing one or more large attributes, or you can store them as an object in Amazon Simple Storage Service (Amazon S3) and store the Amazon S3 object identifier in your DynamoDB item.
Options A and B are incorrect as binary objects such as files should not be stored in Redshift or DynamoDB
Option D is incorrect since Glacier is used for archive storage
For more information on using AWS S3 with DynamoDB, please refer to the below URL

- https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-use-s3-too.html (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-use-s3-too.html)

Ask our Experts

QUESTION 10 UNATTEMPTED ANALYSIS

A team needs a recommendation for an ideal data store for log files. The team also needs to perform analytics of the log files after it has been ingested. They also need a visualization tool which can be given to their IT administrators department to visualize the log files. Which of the following services can be used as the data store and for the visualization?

- A. AWS Kinesis
- B. AWS ElasticSearch ✔
- C. Kibana ✔
- D. Hive

Explanation :

Answer – B and C
The AWS documentation mentions the following
Amazon Elasticsearch Service (Amazon ES) is a managed service that makes it easy to deploy, operate, and scale Elasticsearch clusters in the AWS Cloud. Elasticsearch is a popular open-source search and analytics engine for use cases such as log analytics, real-time application monitoring, and clickstream analysis. With Amazon ES, you get direct access to the Elasticsearch APIs; existing code and applications work seamlessly with the service.
Kibana is a popular open source visualization tool designed to work with Elasticsearch. Amazon ES provides an installation of Kibana with every Amazon ES domain. You can find a link to Kibana on your domain dashboard on the Amazon ES console

Option A is incorrect since this is used for streaming and not as a data store.
Option D is incorrect since this is an open-source, data warehouse, and analytic package that runs on top of a Hadoop cluster.
For more information on ElasticSearch and Kibana, please refer to the below URL

- https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/what-is-amazon-elasticsearch-service.html (https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/what-is-amazon-elasticsearch-service.html)
- https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-kibana.html (https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-kibana.html)

Ask our Experts

QUESTION 11 UNATTEMPTED PROCESSING

A team has currently developed components for an application that makes use of Kinesis streams. The producer and consumer components have been developed using the KPL and KCL libraries respectively. While consuming the records there are throughput provisioned exceptions which are being recorded. Which of the following change can be made to alleviate this issue?

- A. Consider using Kinesis Firehose
- B. Consider changing the partition key for the records
- C. Increase the number of shards ✔
- D. Decrease the number of shards

Explanation :

Answer - C
The AWS documentation mentions the following
The throughput of a stream is provisioned at the shard level. Each shard has a read throughput of up to 5 transactions per second for reads, up to a maximum total data read rate of 2 MB per second. If an application (or a group of applications operating on the same stream) attempts to get data from a shard at a faster rate, Kinesis Data Streams throttles the corresponding Get operations.
In an Amazon Kinesis Data Streams application, if a record processor is processing data faster than the limit — such as in the case of a failover — throttling occurs. Because the Kinesis Client Library (https://docs.aws.amazon.com/streams/latest/dev/developing-consumers-with-kcl.html#kinesis-record-processor-overview-kcl) manages the interactions between the application and Kinesis Data Streams, throttling exceptions occur in the KCL code rather than in the application code. However, because the KCL logs these exceptions, you see them in the logs.
If you find that your application is throttled consistently, you should consider increasing the number of shards for the stream.
Since this is clearly mentioned in the documentation , all other options are incorrect
For more information on processing records in Kinesis, please refer to the below URL

- https://docs.aws.amazon.com/streams/latest/dev/kinesis-record-processor-additional-considerations.html (https://docs.aws.amazon.com/streams/latest/dev/kinesis-record-processor-additional-considerations.html)

Ask our Experts

QUESTION 12 UNATTEMPTED STORAGE

Your company has a requirement for a data store on AWS. The main key requirements are

Storage of large volumes of structured data
Query using standard SQL and existing Business Intelligence tools
Which of the following would you use for this purpose?

- A. Kinesis
- B. Redshift ✔
- C. Data Pipeline
- D. DynamoDB

Explanation :

Answer – B
The AWS documentation mentions the following
Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud. You can start with just a few hundred gigabytes of data and scale to a petabyte or more. This enables you to use your data to acquire new insights for your business and customers.
The first step to create a data warehouse is to launch a set of nodes, called an Amazon Redshift cluster. After you provision your cluster, you can upload your data set and then perform data analysis queries. Regardless of the size of the data set, Amazon Redshift offers fast query performance using the same SQL-based tools and business intelligence applications that you use today.
Option A is invalid since this service is used for streaming data
Option C is invalid since this service is not used to store any data
Option D is invalid since this service is a NoSQL database solution
For more information on Amazon Redshift, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/mgmt/welcome.html (https://docs.aws.amazon.com/redshift/latest/mgmt/welcome.html)

Ask our Experts

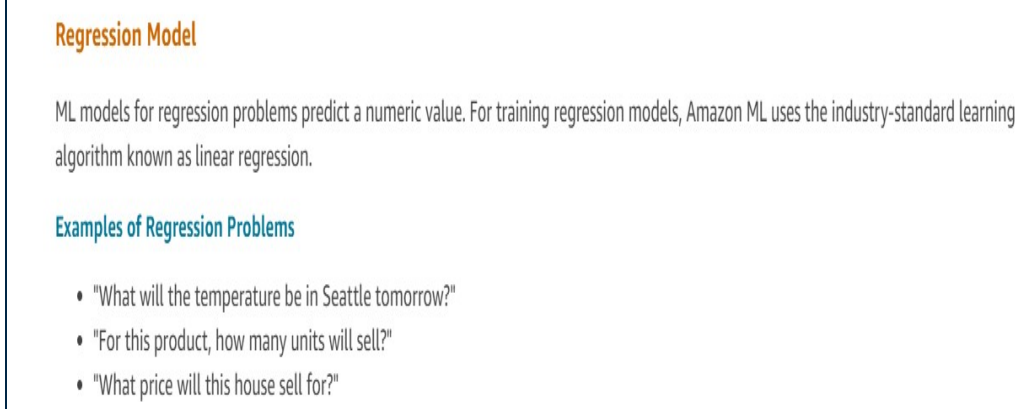
QUESTION 13 UNATTEMPTED ANALYSIS

Your company needs to build a machine learning model for a real estate agent. Based on the data they want to predict how much houses could be sold for in different areas. Which of the following model could be used for this purpose?

- A. Regression ✔
- B. Binary
- C. Unsupervised
- D. Categorial

Explanation :

Answer – A
An example of this is given in the AWS Documentation



Regression Model

ML models for regression problems predict a numeric value. For training regression models, Amazon ML uses the industry-standard learning algorithm known as linear regression.

Examples of Regression Problems

- "What will the temperature be in Seattle tomorrow?"
- "For this product, how many units will sell?"
- "What price will this house sell for?"

Since this is clearly given as an example, all other options are incorrect
For more information on types of machine learning models, please refer to the below URL

- https://docs.aws.amazon.com/machine-learning/latest/dg/types-of-ml-models.html (https://docs.aws.amazon.com/machine-learning/latest/dg/types-of-ml-models.html)

Ask our Experts

QUESTION 14 UNATTEMPTED PROCESSING

A team is building an IoT system which will make use of the AWS IoT services. They need to ensure that data ingested by the IoT service can be forwarded to AWS ElasticSearch. Which of the following could they use for this purpose?

- A. Kinesis
- B. Redshift

- C. IoT Rules Engine ✔
- D. IoT Device Shadow

Explanation :

Answer – C
The AWS documentation mentions the following
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.

Option A is incorrect since this is a streaming service
Option B is incorrect since this is a data ware housing solution
Option D is incorrect since this is used to maintain the state of a device
For more information on IoT Rules engine, please refer to the below URL

- https://docs.aws.amazon.com/iot/latest/developerguide/iot-rules.html (https://docs.aws.amazon.com/iot/latest/developerguide/iot-rules.html)

Ask our Experts

QUESTION 15 UNATTEMPTED DATA SECURITY

A team is building an IoT system which will make use of the AWS IoT services. They need to ensure that custom code is used to authorize IoT devices. Which of the following can be used to host the code for the custom authorizer?

- A. AWS Lambda ✔
- B. AWS SQS
- C. AWS IAM
- D. AWS EMR

Explanation :

Answer – A
The AWS documentation mentions the following
AWS IoT allows you to define custom authorizers that allow you to manage your own authentication and authorization strategy using a custom authentication service and a Lambda function. Custom authorizers allow AWS IoT to authenticate your devices and authorize operations using bearer token authentication and authorization strategies.
Option B is incorrect since this is a queue service
Option C is incorrect since you would manage users and groups here
Option D is incorrect since this is a Big data service
For more information on custom authentication, please refer to the below URL

- https://docs.aws.amazon.com/iot/latest/developerguide/iot-custom-authentication.html (https://docs.aws.amazon.com/iot/latest/developerguide/iot-custom-authentication.html)

Ask our Experts

QUESTION 16 UNATTEMPTED STORAGE

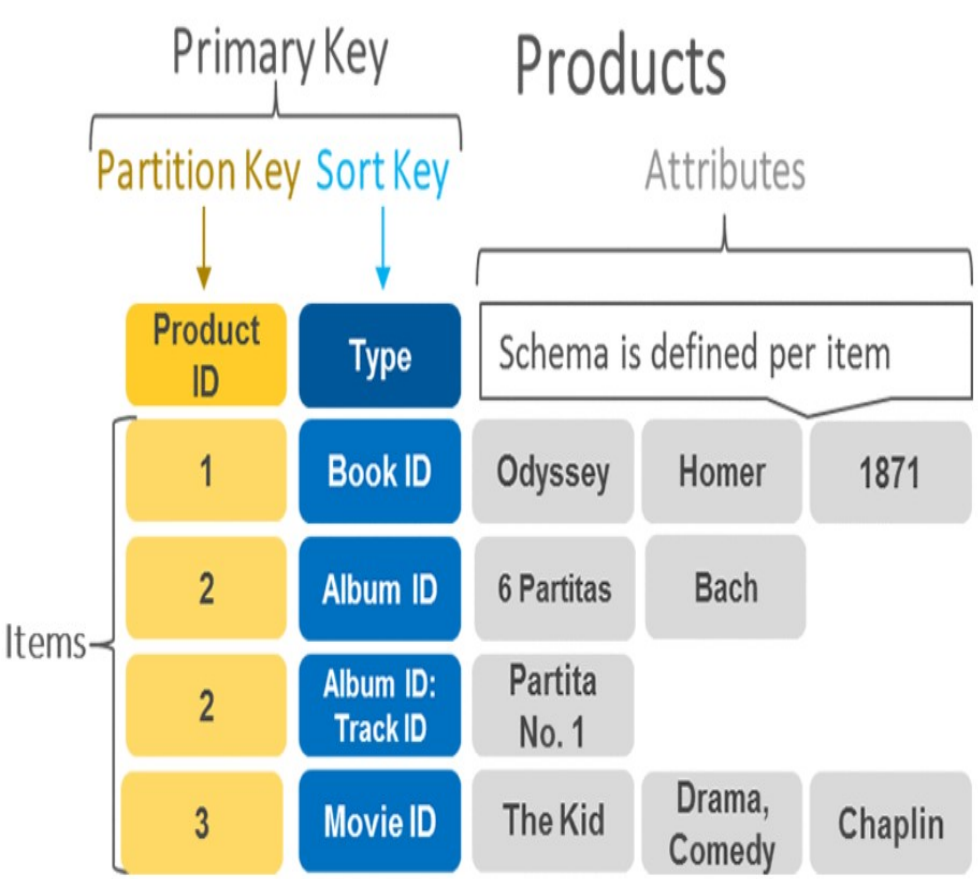
A company needs to create a DynamoDB table which will be used by an application. The table will ingest millions of rows per month. Below are the various attributes for the table

Order ID
Order Type
Order Value
Order Description
Which of the following would be the ideal partition key that should be used for the underlying table?

- **A.** Order ID ✔
- **B.** Order Type
- **C.** Order Value
- **D.** Order Description

Explanation :

Answer - A
Some example of the best practises for Partition keys are given in the AWS Blogs. Using the ID would help in a better distribution of data.



The other options would not guarantee an ideal placement of records across partitions in DynamoDB. For more information on this blog post, please refer to the below URL

- https://aws.amazon.com/blogs/database/choosing-the-right-dynamodb-partition-key/ (https://aws.amazon.com/blogs/database/choosing-the-right-dynamodb-partition-key/)

Ask our Experts

QUESTION 17 UNATTEMPTED STORAGE

A company has the requirement to store a large number of social media tweets. The tweets need to be stored in a data store that can grow on scale. You should have the ability to perform analytics of historical data. Which of the following implementation steps can help achieve this?

- **A.** Store the incoming data in AWS SQS. Push the data into Redshift for future analysis.
- **B.** Store the incoming data on EC2 Instances. Use programs on the EC2 Instance to push the data to DynamoDB for future analysis

- C. Store the incoming data in DynamoDB and use Apache Hive for future analysis ✔
- D. Store the incoming data in AWS RDS and use AWS Data Pipeline for future analysis

Explanation :

Answer – C
The AWS Documentation mentions the following
Amazon DynamoDB is integrated with Apache Hive, a data warehousing application that runs on Amazon EMR. Hive can read and write data in DynamoDB tables, allowing you to:

- Query live DynamoDB data using a SQL-like language (HiveQL).
- Copy data from a DynamoDB table to an Amazon S3 bucket, and vice-versa.
- Copy data from a DynamoDB table into Hadoop Distributed File System (HDFS), and vice-versa.
- Perform join operations on DynamoDB tables.

Option A is incorrect since using SQS for taking the incoming tweets may not be the ideal solution
Option B is incorrect since using EC2 is not the ideal solution for scaling to meet the data ingestion needs in this case
Option D is incorrect since AWS Data Pipeline is not used to store data
For more information on EMR for DynamoDB, please refer to the below URL

- https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/EMRforDynamoDB.html (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/EMRforDynamoDB.html)

Ask our Experts

QUESTION 18 UNATTEMPTED STORAGE

You are an architect for a company. There is a requirement to create an application that has the following key requirements

Take data from various sources and store it in a large data lake solution
The data from various sources can be in various data format
The data needs to be processed so that a common repository with a common structure can be in place for future analysis
Which of the following would you use for this purpose? Choose 2 answers from the options given below

- A. Use DynamoDB to store the incoming data
- B. Use S3 to store the incoming data ✔
- C. Use EMR Clusters to process the data ✔
- D. Use DynamoDB streams to process the data

Explanation :

Answer – B and C
The simple storage service can be used to store data sets that can be of different formats and types. So, this would an ideal data lake store. And then EMR Clusters can be used to process the data.
Option A is incorrect since we don't know the ideal item size for the data. The data seems to come from many different disparate sources. So S3 would better suit for this purpose.
Option D is incorrect since DynamoDB streams is used to relay changes from DynamoDB table items
For more information on big data and analytics, please refer to the below URL

- https://aws.amazon.com/big-data/datalakes-and-analytics/ (https://aws.amazon.com/big-data/datalakes-and-analytics/)

Ask our Experts

QUESTION 19 UNATTEMPTED COLLECTION

Your company currently has an application hosted on an EC2 Instance that is used to ingest and process data from various log files. These log files are generated from various sources. Due to a service interruption on the EC2 Instance, the log files were not being processed. How can you modify the architecture to ensure reliable collection of data?

- A. Make use of DynamoDB streams
- B. Make use of Kinesis streams ✔
- C. Make use of the AWS Data Pipeline service
- D. Make use of S3 events

Explanation :

Answer – B
You can use Amazon Kinesis Data Streams to collect and process large streams (https://aws.amazon.com/streaming-data/) of data records in real time. You can create data-processing applications, known as Kinesis Data Streams applications.
Option A is incorrect since DynamoDB streams is used to relay changes from DynamoDB table items
Option C is incorrect since this service is more of a orchestration service
Option D is incorrect since this feature is used as a trigger for changes occurring to objects in S3 buckets
For more information on Kinesis data streams, please refer to the below URL

- https://docs.aws.amazon.com/streams/latest/dev/introduction.html (https://docs.aws.amazon.com/streams/latest/dev/introduction.html)

Ask our Experts

QUESTION 20 UNATTEMPTED COLLECTION

A development team has been instructed with developing producing and consuming components for a system. These components will interact with Kinesis streams. There is a requirement to ensure that maximum efficiency is reached when sending data to the streams. Which of the following should be implemented by the development team?

- A. Ensure to batch records and send them with the PutRecord API command
- B. Ensure that a loop is constructed to send multiple records at a time with the PutRecord API command
- C. Make the producing application send multiple items via the PutRecords (https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecords.html) API command ✔
- D. Make the producing application send multiple items via the AddRecordsAPI command

Explanation :

Answer - C
The AWS Documentation mentions the following
The PutRecords (https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecords.html) operation sends multiple records to Kinesis Data Streams in a single request. By using PutRecords, producers can achieve higher throughput when sending data to their Kinesis data stream. Each PutRecords request can support up to 500 records. Each record in the request can be as large as 1 MB, up to a limit of 5 MB for the entire request, including partition keys. As with the single PutRecord operation described below, PutRecords uses sequence numbers and partition keys. However, the PutRecord parameter SequenceNumberForOrdering is not included in a PutRecords call. The PutRecords operation attempts to process all records in the natural order of the request.
Since this is clearly mentioned in the documentation , all other options are invalid.
For more information on developing producers, please refer to the below URL

- https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-sdk.html (https://docs.aws.amazon.com/streams/latest/dev/developing-producers-with-sdk.html)

Ask our Experts

QUESTION 21 UNATTEMPTED DATA SECURITY

A company is planning on provisioning a Redshift cluster. The IT Security department has mandated that all data that gets transferred to Redshift does not pass via the Internet. Also, the IT Administrators team needs to monitor all the traffic during the copy operation. How can you accomplish this? Choose 2 answers from the options given below?

- [ ] **A.** Enable Redshift Enhanced VPC Routing ✔
- [ ] **B.** Create a Direct Connect connection to the Redshift cluster
- [ ] **C.** Use VPC Flow logs to monitor the traffic ✔
- [ ] **D.** Use Cloudwatch metrics to monitor the traffic

Explanation :

Answer – A and C
The AWS Documentation mentions the following
When you use Amazon Redshift Enhanced VPC Routing, Amazon Redshift forces all COPY (https://docs.aws.amazon.com/redshift/latest/dg/r_COPY.html) and UNLOAD (https://docs.aws.amazon.com/redshift/latest/dg/r_UNLOAD.html) traffic between your cluster and your data repositories through your Amazon VPC. By using Enhanced VPC Routing, you can use standard VPC features, such as VPC security groups (https://docs.aws.amazon.com/vpc/latest/userguide/VPC_SecurityGroups.html), network access control lists (ACLs) (https://docs.aws.amazon.com/vpc/latest/userguide/VPC_ACLs.html), VPC endpoints (https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-s3.html), VPC endpoint policies (https://docs.aws.amazon.com/vpc/latest/userguide/vpc-endpoints-s3.html#vpc-endpoints-policies-s3), internet gateways (https://docs.aws.amazon.com/vpc/latest/userguide/VPC_Internet_Gateway.html), and Domain Name System (DNS) (https://docs.aws.amazon.com/vpc/latest/userguide/vpc-dns.html) servers, as described in the *Amazon VPC User Guide*. You use these features to tightly manage the flow of data between your Amazon Redshift cluster and other resources. When you use Enhanced VPC Routing to route traffic through your VPC, you can also use VPC flow logs (https://docs.aws.amazon.com/vpc/latest/userguide/flow-logs.html) to monitor COPY and UNLOAD traffic.
Option B is incorrect since we don't know from the copy of data will originate
Option D is incorrect since this service can only provide metrics and not the actual log data
For more information on VPC Routing, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/mgmt/enhanced-vpc-routing.html (https://docs.aws.amazon.com/redshift/latest/mgmt/enhanced-vpc-routing.html)

Ask our Experts

QUESTION 22 UNATTEMPTED PROCESSING

A company is planning on setting up Spark on an EMR Cluster. They are trying to understand the best Instance type to use for the underlying cluster. Which of the following would be ideal for the underlying EMR Cluster Instances?

- A. Compute Optimized
- B. Memory Optimized ✔
- C. Storage Optimized
- D. GPU Optimized

Explanation :

Answer – B
This is clearly mentioned in the AWS Documentation

Memory-Optimized Instance Types

We recommend memory-optimized Amazon Elastic Compute Cloud (Amazon EC2) instance types for Apache Spark workloads because Spark attempts to process as much data in memory as possible. By default, this solution deploys an r3.xlarge instance for the Amazon EMR cluster nodes to deliver optimal performance.

Since this is clearly mentioned, all other options are invalid
For more information on considerations for hosting SPARK on EMR, please refer to the below URL

- https://docs.aws.amazon.com/solutions/latest/real-time-analytics-spark-streaming/considerations.html (https://docs.aws.amazon.com/solutions/latest/real-time-analytics-spark-streaming/considerations.html)

Ask our Experts

QUESTION 23 UNATTEMPTED ANALYSIS

A company has a set of data sources which vary from Redshift to MySQL to Hive on EMR and PostgreSQL. There is a requirement to have a single tool that can run queries on all the different platform for your daily ad-hoc analysis. Which of the following can be used for this requirement?

- A. Apache Pig
- B. Ganglia
- C. Presto ✔
- D. YARN

Explanation :

Answer – C
The AWS Documentation mentions the following
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 (https://aws.amazon.com/s3/), Cassandra, MongoDB, and HBase (https://aws.amazon.com/emr/details/hbase/), and relational data sources such as MySQL, PostgreSQL, Amazon Redshift (https://aws.amazon.com/redshift/), Microsoft SQL Server, and Teradata.
Option A is incorrect since this is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs
Option B is incorrect since this is an open source project is a scalable, distributed system designed to monitor clusters
Option D is incorrect since this is used to centrally manage cluster resources for multiple data-processing frameworks
For more information on Presto, please refer to the below URL

- https://aws.amazon.com/big-data/what-is-presto/ (https://aws.amazon.com/big-data/what-is-presto/)

Ask our Experts

QUESTION 24 UNATTEMPTED COLLECTION

A team needs to create a system that can ingest data with each data item around 400KB. The records need to be processed as they arrive with the least amount of delay. Which service would help achieve this requirement?

- **A.** Kinesis Stream ✔
- **B.** Apache Presto
- **C.** AWS SQS
- **D.** AWS EMR

Explanation :

Answer – A
The AWS Documentation mentions the following
You can use Kinesis Data Streams for rapid and continuous data intake and aggregation. The type of data used can include IT infrastructure log data, application logs, social media, market data feeds, and web clickstream data. Because the response time for the data intake and processing is in real time, the processing is typically lightweight.
Option B is incorrect since this is an open source, distributed SQL query engine
Option C is incorrect because of the high processing required and item size
Option D is incorrect since you can't use EMR alone for data ingestion

For more information on Kinesis streams, please refer to the below URL

- https://docs.aws.amazon.com/streams/latest/dev/introduction.html (https://docs.aws.amazon.com/streams/latest/dev/introduction.html)

Ask our Experts

QUESTION 25 UNATTEMPTED COLLECTION

A company has a large Oracle Data warehouse in their On-premise infrastructure. They need to migrate this data warehouse to Redshift since they are not able to expand on their current infrastructure. Which of the following tools can in the migration process? Choose 2 answers from the options given below

- [ ] **A.** AWS Schema Conversion Tool ✔
- [ ] **B.** AWS Redshift Conversion Tool
- [ ] **C.** AWS Database Migration Service ✔
- [ ] **D.** AWS Oracle Migration Service

Explanation :

Answer - A and C
The AWS Documentation mentions the following
Convert the data warehouse schema and code from a sample Oracle data warehouse running on Amazon Relational Database Service (Amazon RDS) using the AWS Schema Conversion Tool (https://docs.aws.amazon.com/SchemaConversionTool/latest/userguide/Welcome.html) (AWS SCT). AWS SCT helps you automatically convert the source schema and majority of the custom code to a format compatible with Amazon Redshift. AWS SCT clearly marks any code that it cannot convert so that you can manually convert it.
Migrate data from the Oracle data warehouse to Amazon Redshift using the AWS Database Migration Service (https://aws.amazon.com/dms/)(AWS DMS). With AWS DMS, you can begin the data migration with just a few clicks in the AWS Management Console. The source data warehouse remains fully operational during the migration.
Since this is clearly mentioned in the documentation , the other options are invalid
For more information on migrating from Oracle to Amazon Redshift, please refer to the below URL

- https://aws.amazon.com/getting-started/projects/migrate-oracle-to-amazon-redshift/ (https://aws.amazon.com/getting-started/projects/migrate-oracle-to-amazon-redshift/)

Ask our Experts

QUESTION 26 UNATTEMPTED COLLECTION

A company currently has a large data store wherein the data is stored in Apache Parquet. They need to transfer this to Amazon Redshift. How can this be achieved in the easiest way possible? Choose 2 answers from the options below

- [ ] **A.** Copy the data onto DynamoDB
- [ ] **B.** Copy the data onto S3 ✔
- [ ] **C.** Use the COPY command to transfer the data to Redshift ✔
- [ ] **D.** Use the AWS DataPipeline service to transfer the data from DynamoDB to Redshift

Explanation :

Answer – B and C
The AWS Documentation mentions the following
You can now COPY (https://docs.aws.amazon.com/redshift/latest/dg/r_COPY.html) Apache Parquet and Apache ORC file formats from Amazon S3 to your Amazon Redshift (https://aws.amazon.com/redshift/) cluster. Apache Parquet and ORC are columnar data formats that allow users to store their data more efficiently and cost-effectively. With this update, Redshift now supports COPY from six file formats: AVRO, CSV, JSON, Parquet, ORC and TXT.
Options A and D are incorrect since copying the data files onto DynamoDB is not recommended.
For more information on this feature, please refer to the below URL

- https://aws.amazon.com/about-aws/whats-new/2018/06/amazon-redshift-can-now-copy-from-parquet-and-orc-file-formats/ (https://aws.amazon.com/about-aws/whats-new/2018/06/amazon-redshift-can-now-copy-from-parquet-and-orc-file-formats/)

Ask our Experts

QUESTION 27 UNATTEMPTED ANALYSIS

A company currently deals in a bike sharing service. They want to adopt Machine Learning on AWS to determine how many bikes would be required every hour to ensure that they keep up with demand. Which of the following Machine learning technique would they use for this purpose?

- ○ **A.** Regression ✔
- ○ **B.** Multi-class
- ○ **C.** Categorical
- ○ **D.** Binary

Explanation :

Answer - A

An example of this use case is given in the AWS documentation

Building a Numeric Regression Model with Amazon Machine Learning

by Guy Ernest | on 28 APR 2015 | in Amazon Machine Learning | Permalink | Comments | Share

Guy Ernest is a Solutions Architect with AWS

We need to predict future values in our businesses. These predictions are important for better planning of resource allocation and making other business decisions. Often, we settle for a simplified heuristic of average values from the past and some change assumption because more accurate alternatives are too complex or expensive. The new Amazon Machine Learning (Amazon ML) service changes this equation by providing a simple and inexpensive way of building and using models such as numeric regression.

This post uses the example of a bike share program where you need to know how many bikes are required at each hour of each day in a specific city. In this scenario, you need a machine learning model that predicts a number based on a set of features or predictors. You will build a regression model based on a data set that is publicly available in Kaggle, a large community site of data scientists that compete against each other to solve data science problems. By building the model, you will explore a few concepts around the successful application of machine learning to solve similar problems in your domain.

The other options are invalid since these are not the ideal ML models to be used for this scenario

For more information on this use case, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/building-a-numeric-regression-model-with-amazon-machine-learning/ (https://aws.amazon.com/blogs/big-data/building-a-numeric-regression-model-with-amazon-machine-learning/)

Ask our Experts

QUESTION 28 UNATTEMPTED COLLECTION

A company needs to transform and move a series of data sets from their on-premise infrastructure on to AWS. They have decided to use AWS S3 as their data lake store. Which of the following can be used to transform and move the data?

- A. AWS Glue ✔
- B. AWS SQS
- C. AWS EMR
- D. AWS Kinesis

Explanation :

Answer – A

An example of this use case is given in the AWS documentation

> In this post, I describe a solution for transforming and moving data from an on-premises data store to Amazon S3 using AWS Glue that simulates a common data lake ingestion pipeline. AWS Glue can connect to Amazon S3 and data stores in a virtual private cloud (VPC) such as Amazon RDS, Amazon Redshift, or a database running on Amazon EC2. For more information, see Adding a Connection to Your Data Store. AWS Glue can also connect to a variety of on-premises JDBC data stores such as PostgreSQL, MySQL, Oracle, Microsoft SQL Server, and MariaDB.
>
> AWS Glue ETL jobs can use Amazon S3, data stores in a VPC, or on-premises JDBC data stores as a source. AWS Glue jobs extract data, transform it, and load the resulting data back to S3, data stores in a VPC, or on-premises JDBC data stores as a target.

Option B is incorrect because this is a queue-based service

Option C is incorrect since AWS Glue can easily connect to on-premise data stores via JDBC connectors

Option D is incorrect since we don't know the type of data sets that need to be streamed.

For more information on this use case, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/ (https://aws.amazon.com/blogs/big-data/how-to-access-and-analyze-on-premises-data-stores-using-aws-glue/)

Ask our Experts

QUESTION 29 UNATTEMPTED PROCESSING

A company is planning on using Hive on EMR to query their DynamoDB tables. They want to ensure they achieve the best response for the queries fired against the table. Which of the following can be done to achieve the best performance? Choose 2 answers from the options given below

- ☐ **A.** Ensure streaming is enabled on DynamoDB
- ☐ **B.** Ensure the read capacity is properly set for the DynamoDB table ✔
- ☐ **C.** Increase the number of instances in the cluster ✔
- ☐ **D.** Enable Enhanced VPC Routing

Explanation :

Answer – B and C

The AWS Documentation mentions the following

When you run Hive queries against a DynamoDB table, you need to ensure that you have provisioned a sufficient amount of read capacity units.

The mapper daemons that Hadoop launches to process your requests to export and query data stored in DynamoDB are capped at a maximum read rate of 1 MiB per second to limit the read capacity used. If you have additional provisioned throughput available on DynamoDB, you can improve the performance of Hive export and query operations by increasing the number of mapper daemons. To do this, you can either increase the number of EC2 instances in your cluster or increase the number of mapper daemons running on each EC2 instance.
Option A is incorrect since streams will not help in this case
Option D is incorrect since this is a setting in AWS Redshift
For more information on optimizing EMR Hive, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ReleaseGuide/EMR_Hive_Optimizing.html (https://docs.aws.amazon.com/emr/latest/ReleaseGuide/EMR_Hive_Optimizing.html)

Ask our Experts

QUESTION 30 UNATTEMPTED VISUALIZATION

A company currently has Spark and Hive running on an EMR Cluster. They need an option to perform interactive and collaborative notebook for data exploration. Which of the following can be used for this purpose?

- A. Kinesis Analytics
- B. D3.js
- C. Zeppelin ✔
- D. Hue

Explanation :

Answer – C
The AWS Documentation mentions the following

Apache Zeppelin

Use Apache Zeppelin as a notebook for interactive data exploration. For more information about Zeppelin, see https://zeppelin.apache.org/. Zeppelin is included in Amazon EMR release version 5.0.0 and later. Earlier release versions include Zeppelin as a sandbox application. For more information, see Amazon EMR 4.x Release Versions.

To access the Zeppelin web interface, set up an SSH tunnel to the master node and a proxy connection. For more information, see View Web Interfaces Hosted on EMR Clusters. Zeppelin is included in Amazon EMR release version 5.0.0 and later. Earlier release versions include Zeppelin as a sandbox application. For more information, see Amazon EMR 4.x Release Versions.

Option A is incorrect since this is used for interactive analysis on Kinesis streams
Option B is incorrect since this is a Javascript library for visualization
Option D is incorrect since this is a web interface for EMR Clusters

For more information on Apache Zepplin, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-zeppelin.html (https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-zeppelin.html)

Ask our Experts

QUESTION 31 UNATTEMPTED PROCESSING

A company has an application that is processing and sending tweet data to a DynamoDB table. After viewing the Cloudwatch logs over a period of time, you can see that the table is using 80% on both read and write throughput. A promotional period is coming up and you need to ensure that your application is ready to take in the new influx of data. Which of the following would you consider doing?

- A. Increasing the size of the DynamoDB tables
- B. Increasing the read and write capacity on the table ✔
- C. Creating a Local secondary index
- D. Creating a Global secondary index

Explanation :

Answer – B
Here the major concern is that the read and write capacity are already being heavily used so it would be better to provision more capacity.

Option A is incorrect because you don't need to manage the infrastructure for DynamoDB
Options C and D are incorrect since creating indexes may not help ensure the DynamoDB tables will be able to handle the new influx of traffic.
For more information on DynamoDB throughput, please refer to the below URL

- https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/HowItWorks.ProvisionedThroughput. (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/HowItWorks.ProvisionedThroughput.ht

Ask our Experts

QUESTION 32 UNATTEMPTED COLLECTION

A company needs to ingest a large amount of raw data. This data needs to be transformed and sent to S3. The original raw data also needs to be sent to another bucket in S3. Which of the following service would you use for this requirement which would ensure the least amount of implementation effort?

- A. AWS Kinesis
- B. AWS Kinesis Firehose ✔
- C. AWS Redshift
- D. AWS SQS

Explanation :

Answer – B
You can this entirely with the Kinesis Firehose service

To choose Amazon S3 for your destination

- On the **Choose destination** page, enter values for the following fields:

Destination

Choose **Amazon S3**.

Destination S3 bucket

Choose an S3 bucket that you own where the streaming data should be delivered. You can create a new S3 bucket or choose an existing one.

Destination S3 bucket prefix

(Optional) To use the default prefix for Amazon S3 objects, leave this option blank. Kinesis Data Firehose automatically uses a prefix in "YYYY/MM/DD/HH" UTC time format for delivered Amazon S3 objects. You can add to the start of this prefix. For more information, see Amazon S3 Object Name Format.

Source record S3 backup

Choose **Disabled** to disable source record backup. If you enable data transformation with AWS Lambda, you can enable source record backup to deliver untransformed incoming data to a separate S3 bucket. You can add to the start of the "YYYY/MM/DD/HH" UTC time prefix that is generated by Kinesis Data Firehose. You cannot disable source record backup after you enable it.

Option A is incorrect because here you would need to implement producers and consumers to perform the required operations
Option C is incorrect because this is a data ware housing solution
Option D is incorrect because this is a queue-based service
For more information on the various destinations for Kinesis Firehose, please refer to the below URL

- https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html (https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html)

Ask our Experts

QUESTION 33 UNATTEMPTED PROCESSING

A company wants to build a system that will do the following

Store advertising campaign information in a DynamoDB table
Stream click stream data using Amazon Kinesis
They now need some way to write queries that would join the data from the stream and the table. Which of the following could be used for this purpose?

- A. Amazon EMR ✔
- B. Amazon Redshift
- C. Amazon Kinesis Firehose
- D. Amazon Quicksight

Explanation :

Answer - A
This example is given in the AWS Documentation

The other options are incorrect because none of these actually have the capability to perform these sort of join queries without performing additional activities in between.
For more information on EMR and Kinesis, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-kinesis.html (https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-kinesis.html)

Ask our Experts

QUESTION 34 UNATTEMPTED PROCESSING

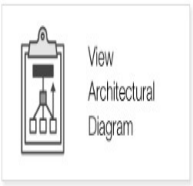
A company wants to build an end to end log analytics solution which could collect, ingest and make a searchable index based on the ingested data. Which of the following can be used to fulfil these requirements? Choose 2 answers from the options given below

- [x] **A.** Use AWS Kinesis for ingesting the data ✔
- [ ] **B.** Use AWS Redshift for ingesting the data
- [ ] **C.** Use AWS ElastiCache for indexing the data
- [x] **D.** Use AWS ElasticSearch for indexing the data ✔

Explanation :

Answer – A and D
This example is given in the AWS Documentation

Log analytics is a common big data use case that allows you to analyze log data from websites, mobile devices, servers, sensors, and more for a wide variety of applications such as digital marketing, application monitoring, fraud detection, ad tech, gaming, and IoT. In this project, you will use Amazon Web Services to build an end-to-end log analytics solution that collects, ingests, processes, and loads both batch data and streaming data, and makes the processed data available to your users in analytics systems they are already using and in near real-time. The solution is highly reliable, cost-effective, scales automatically to varying data volumes, and requires almost no IT administration.



What you'll accomplish:

Set up a Kinesis Agent on data sources to collect data and send it continuously to Amazon Kinesis Firehose.

Create an end-to-end data delivery stream using Kinesis Firehose. The delivery stream will transmit your data from the agent to destinations including Amazon Kinesis Analytics, Amazon Redshift, Amazon Elasticsearch Service, and Amazon S3.

Process incoming log data using SQL queries in Amazon Kinesis Analytics.

Load processed data from Kinesis Analytics to Amazon Elasticsearch Service to index the data.

Analyze and visualize the processed data using Kibana.

Option B is incorrect since this is used as a data ware house solution
Option C is incorrect since this is a cache service
For more information on this use case, please refer to the below URL

- https://aws.amazon.com/getting-started/projects/build-log-analytics-solution/ (https://aws.amazon.com/getting-started/projects/build-log-analytics-solution/)

Ask our Experts

QUESTION 35 UNATTEMPTED VISUALIZATION

A company wants to make use of AWS Quicksight for their Visualization needs. They need to gross sales for a product over a month on month period. Which of the following chart type would they draw up in AWS Quicksight?

- A. PieChart
- B. Line Chart ✔
- C. Pivot Tables
- D. KPI's

Explanation :

Answer - B
The AWS Documentation mentions the following
Use line charts to compare changes in measure values over period of time, for the following scenarios:

- One measure over a period of time, for example gross sales by month.
- Multiple measures over a period of time, for example gross sales and net sales by month.
- One measure for a dimension over a period of time, for example number of flight delays per day by airline.

Line charts show the individual values of a set of measures or dimensions against the range displayed by the Y axis. Area line charts differ from regular line charts in that each value is represented by a colored area of the chart instead of just a line, to make it easier to evaluate item values relative to each other.
Option A is incorrect since this type of chart is used to compare values for items in a dimension.
Option C is incorrect since this type of chart is used to show measure values for the intersection of two dimensions.
Option D is incorrect since this type of chart is used to visualize a comparison between a key value and its target value
For more information on line charts, please refer to the below URL

- https://docs.aws.amazon.com/quicksight/latest/user/line-charts.html (https://docs.aws.amazon.com/quicksight/latest/user/line-charts.html)

Ask our Experts

QUESTION 36 UNATTEMPTED PROCESSING

A company has setup an EMR cluster in AWS. They want to submit interactive jobs to the cluster. Which of the following are ways in which this can be accomplished? Choose 2 answers from the options given below

- [ ] **A.** Add a step to the cluster ✔
- [ ] **B.** Connect to the task node and submit the Hadoop job
- [ ] **C.** Connect to the core node and submit the Hadoop job
- [ ] **D.** Connect to the master node and submit the Hadoop job ✔

Explanation :

Answer – A and D
The AWS Documentation mentions the following
In addition to adding steps to a cluster, you can connect to the master node using an SSH client or the AWS CLI and interactively submit Hadoop jobs. For example, you can use PuTTY to establish an SSH connection with the master node and submit interactive Hive queries which are compiled into one or more Hadoop jobs.
Since this is clearly given in the documentation , all other options are incorrect
For more information on interactive jobs, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/interactive-jobs.html (https://docs.aws.amazon.com/emr/latest/ManagementGuide/interactive-jobs.html)

Ask our Experts

QUESTION 37 UNATTEMPTED ANALYSIS

A company has enabled VPC Flow logs to have the ability to capture all network traffic to their EC2 Instances. They need to process the data to understand from where the most traffic is coming from. Which of the following could ideally be used to process the data?

- ○ **A.** AWS EMR ✔
- ○ **B.** AWS Kinesis
- ○ **C.** AWS Kinesis Firehose

- D. AWS Redshift

Explanation :

Answer – A

An example of this is given in the AWS Documentation

> It's easy to understand network patterns in small AWS deployments where *software stacks* are well defined and managed. But as teams and usage grow, its gets harder to understand which systems communicate with each other, and on what ports. This often results in overly permissive security groups.
>
> In this post, I show you how to gain valuable insight into your network by using Amazon EMR and Amazon VPC Flow Logs. The walkthrough implements a pattern often found in network equipment called *'Top Talkers'*, an ordered list of the heaviest network users, but the model can also be used for many other types of network analysis. Customers have successfully used this process to lock down security groups, analyze traffic patterns, and create network graphs.

Options B and C are incorrect since these are used for ingestion of data

Option D is incorrect since this is a data warehouse store

For more information on this use case, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/processing-vpc-flow-logs-with-amazon-emr/ (https://aws.amazon.com/blogs/big-data/processing-vpc-flow-logs-with-amazon-emr/)

Ask our Experts

QUESTION 38 UNATTEMPTED ANALYSIS

A company is planning to host a large number of data sets in S3. The Simple storage service will act as their data lake. They need their business departments to have a serverless service to perform queries on the data in the data lake. Which of the following could easily fulfil this requirement? Choose 2 answers from the options given below

- [x] A. AWS Athena ✔
- [x] B. AWS Redshift Spectrum ✔
- [ ] C. AWS Lambda
- [ ] D. AWS ECS

Explanation :

Answer – A and B

A use case of this is mentioned in the AWS Blog site

We built Redshift Spectrum to end this "tyranny of OR." With Redshift Spectrum, Amazon Redshift customers can easily query their data in Amazon S3. Like Amazon EMR, you get the benefits of open data formats and inexpensive storage, and you can scale out to thousands of nodes to pull data, filter, project, aggregate, group, and sort. Like Amazon Athena (https://aws.amazon.com/athena/), Redshift Spectrum is serverless and there's nothing to provision or manage. You just pay for the resources you consume for the duration of your Redshift Spectrum query. Like Amazon Redshift itself, you get the benefits of a sophisticated query optimizer, fast access to data on local disks, and standard SQL. And *like nothing else*, Redshift Spectrum can execute highly sophisticated queries against an exabyte of data or more—in just minutes.
Option C is incorrect since this is a serverless compute service
Option D is incorrect since this is a serverless container orchestration service
For more information on this use case, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/amazon-redshift-spectrum-extends-data-warehousing-out-to-exabytes-no-loading-required/ (https://aws.amazon.com/blogs/big-data/amazon-redshift-spectrum-extends-data-warehousing-out-to-exabytes-no-loading-required/)

Ask our Experts

QUESTION 39 UNATTEMPTED DATA SECURITY

A company is planning on setting up an EMR cluster which will store data in S3. The requirement is to also ensure that data is encrypted at rest. The company wants to manage the lifecycle of the keys that are used for the underlying encryption process. Which of the following steps need to be implemented to ensure this works? Choose 2 answers from the options given below

- A. Ensure to choose AWS S3 server-side managed encryption keys
- B. Ensure to create KMS keys to be used in the encryption process ✔
- C. Ensure that the IAM role EMR_EC2_DefaultRole has access to the keys ✔
- D. Ensure that the IAM role EMR_DefaultRole has access to the keys

Explanation :

Answer – B and C
The AWS Documentation mentions the following
The AWS KMS encryption key must be created in the same region as your Amazon EMR cluster instance and the Amazon S3 buckets used with EMRFS. If the key that you specify is in a different account from the one that you use to configure a cluster, you must specify the key using its ARN.
The role for the Amazon EC2 instance profile must have permissions to use the CMK you specify. The default role for the instance profile in Amazon EMR is EMR_EC2_DefaultRole.
Option A is incorrect since the question states that you need to manage the lifecycle of the keys so you should use KMS keys

Option D is incorrect since this role is used for the EMR itself to use other services. But the permission to use CMK keys is required by the underlying EC2 Instances
For more information on EMR encryption, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-emrfs-encryption.html (https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-emrfs-encryption.html)

Ask our Experts

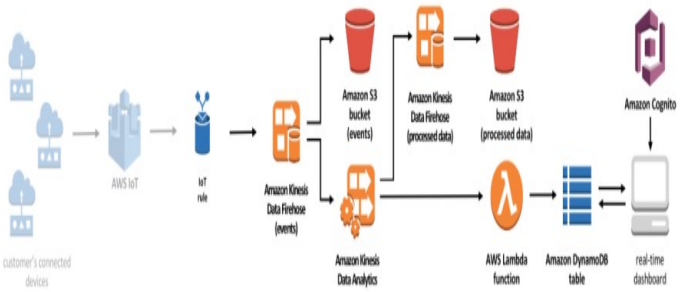
QUESTION 40 UNATTEMPTED COLLECTION

A company wants to create a system that ingests and helps analyse data in real time from various IoT devices. Which of the following combination of services can help fulfil this requirement?

- A. AWS Kinesis Firehose for Ingestion and Kinesis Analytics for analysing the data ✔
- B. AWS Kinesis Firehose for Ingestion and AWS Lambda for analysing the data
- C. AWS SQS for Ingestion and AWS Lambda for analysing the data
- D. AWS Lambda for ingestion and Kinesis Analytics for analysing the data

Explanation :

Answer – A
An example of this is given in the AWS documentation



1. When AWS IoT ingests data from your connected devices, an AWS IoT rule sends the data to a Kinesis data delivery stream.
2. The delivery stream archives the events in an Amazon S3 bucket and sends the data to a Kinesis Data Analytics application for processing.
3. The application sends the data to a Lambda function that sends it in real-time to a DynamoDB table to be stored. The application also sends processed data to a second Kinesis data delivery stream which archives it in an Amazon S3 bucket.
4. The solution also creates an Amazon Cognito user pool, an Amazon S3 bucket, and a real-time dashboard to securely read and display the account activity stored in the DynamoDB table.

Options B and D are incorrect since AWS Lambda can be used for transforming the data
Option C is incorrect since SQS is a queue service and would not be ideal for ingestion of data
For more information on this use case, please refer to the below URL

- https://aws.amazon.com/answers/iot/real-time-iot-device-monitoring-with-kinesis/ (https://aws.amazon.com/answers/iot/real-time-iot-device-monitoring-with-kinesis/)

Ask our Experts

QUESTION 41 UNATTEMPTED STORAGE

Your team has a series of tables defined in Redshift. They need to copy data from one table to another. Which of the following are effectives ways of carrying this out? Choose 2 answers from the options given below.

- [ ] **A.** Use the Insert statement ✔
- [ ] **B.** UNLOAD the data from the source table
- [ ] **C.** LOAD the data from to the destination table
- [ ] **D.** Use the CREATE TABLE AS statement ✔

Explanation :

Answer – A and D
This is mentioned in the AWS documentation

Use a Bulk Insert

Use a bulk insert operation with a SELECT clause for high-performance data insertion.

Use the INSERT and CREATE TABLE AS commands when you need to move data or a subset of data from one table into another.

For example, the following INSERT statement selects all of the rows from the CATEGORY table and inserts them into the CATEGORY_STAGE table.

```
insert into category_stage
(select * from category);
```

The following example creates CATEGORY_STAGE as a copy of CATEGORY and inserts all of the rows in CATEGORY into CATEGORY_STAGE.

```
create table category_stage as
select * from category;
```

Now B and C are also ways of accomplishing this , but when it comes to cross table data copying , AWS recommends the Insert or Create table command

For more information on bulk inserts, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-bulk-inserts.html (https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-bulk-inserts.html)

Ask our Experts

QUESTION 42 UNATTEMPTED DATA SECURITY

A company is currently using AWS ElasticSearch for storing and analysing their log data. <annual They have to ensure that data is encrypted at rest. If there are any manual snapshots created, the data on these snapshots also have to be encrypted at rest. How can you achieve this? Choose 2 answers from the options given below

- [ ] **A.** Use AWS KMS to encrypt the data in the ES domains ✔
- [ ] **B.** Use AWS Default encryption to encrypt the data in the ES domains
- [ ] **C.** Use AWS KMS to encrypt the data in the manual snapshots
- [ ] **D.** Use server-side encryption with S3-managed keys (https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingServerSideEncryption.html) for the manual snapshots ✔

Explanation :

Answer – A and D
The AWS Documentation mentions the following
Amazon ES domains offer encryption of data at rest, a security feature that helps prevent unauthorized access to your data. The feature uses AWS Key Management Service (AWS KMS) to store and manage your encryption keys.
Manual snapshots: Currently, you can't use KMS master keys to encrypt manual snapshots. You can, however, use server-side encryption with S3-managed keys (https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingServerSideEncryption.html) to encrypt the bucket that you use as a snapshot repository
Since this is clearly mentioned in the AWS documentation , all other options are incorrect
For more information on encryption at rest for elasticsearch, please refer to the below URL

- https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/encryption-at-rest.html (https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/encryption-at-rest.html)

Ask our Experts

QUESTION 43 UNATTEMPTED ANALYSIS

A company has setup a Redshift cluster with multiple tables already loaded using the COPY command. They want to allow developers to run queries against the tables in the cluster. Which of the following are 2 ways developers can connect to the cluster?

- [ ] **A.** Download the Redshift client from the AWS Console
- [ ] **B.** Use SQL client tools with JDBC connectivity to the cluster ✔
- [ ] **C.** Use SQL client tools with ODBC connectivity to the cluster ✔
- [ ] **D.** Use the Data Explorer in RDS to query the data

Explanation :

Answer – B and C
The AWS Documentation mentions the following
You can connect to Amazon Redshift clusters from SQL client tools over Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC) connections. Amazon Redshift does not provide or install any SQL client tools or libraries, so you must install them on your client computer or Amazon EC2 instance to use them to work with data in your clusters. You can use most SQL client tools that support JDBC or ODBC drivers.
Since this is clearly mentioned in the documentation , all other options are incorrect
For more information on connecting to the cluster, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/mgmt/connecting-to-cluster.html (https://docs.aws.amazon.com/redshift/latest/mgmt/connecting-to-cluster.html)

Ask our Experts

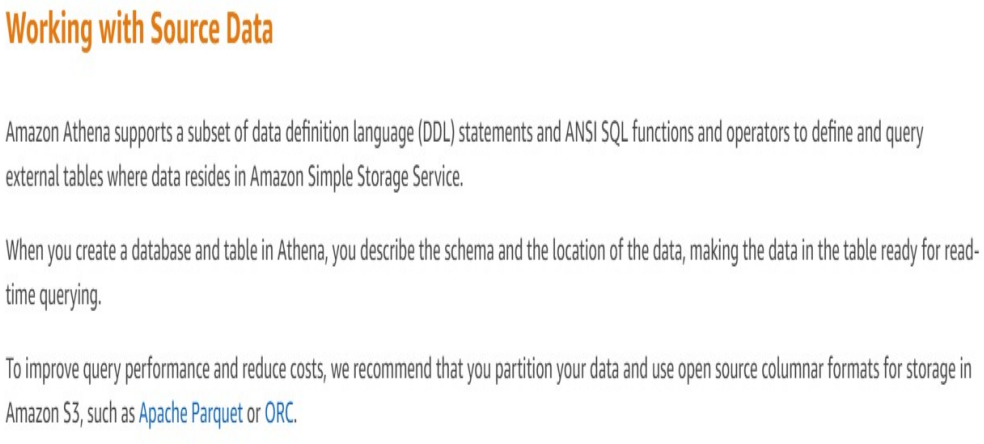
QUESTION 44 UNATTEMPTED ANALYSIS

A set of developers have been requested to analyse different data sets hosted on S3 using AWS Athena. As an architect you need to ensure that the queries are optimized via AWS Athena. Which of the following can help in this regard? Choose 3 answers from the options given below

- [ ] **A.** Partition the data in S3 ✔
- [ ] **B.** Save the data sets as .csv files
- [ ] **C.** Store the files in S3 in Apache Parquet format ✔
- [ ] **D.** Store the files in S3 in ORC format ✔

Explanation :

Answer – A,C and D
This is given in the AWS Documentation



Since this is clearly mentioned in the documentation , storing the files in .csv format is not recommended.
For more information on working with data, please refer to the below URL

- https://docs.aws.amazon.com/athena/latest/ug/work-with-data.html (https://docs.aws.amazon.com/athena/latest/ug/work-with-data.html)

Ask our Experts

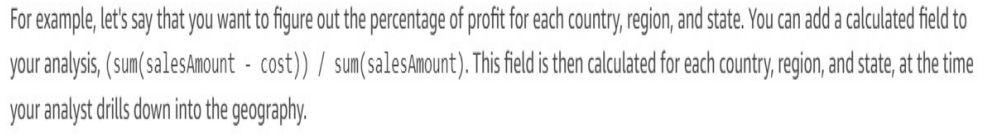
QUESTION 45 UNATTEMPTED ANALYSIS

You've setup AWS Quicksight to analyse data from a csv file stored in S3. The data basically contains the sales figures for the products on a daily basis. You need to also have a field to get the total of the sales for all products on a weekly basis. Which of the following can you add to the analysis for this?

- A. A parameter
- B. A calculated field ✔
- C. Add filters
- D. Use Stories

Explanation :

Answer – B
An example of this is given in the AWS Documentation



Option A is incorrect because these are named variables that can provide values to an action or an object.
Option C is incorrect because these are used to define the data in a data set.
Option D is incorrect because these are used to preserve multiple iterations of an analysis and then play them sequentially to provide a narrative about the analysis data
For more information on calculated fields, please refer to the below URL

- https://docs.aws.amazon.com/quicksight/latest/user/adding-a-calculated-field-analysis.html (https://docs.aws.amazon.com/quicksight/latest/user/adding-a-calculated-field-analysis.html)

Ask our Experts

QUESTION 46 UNATTEMPTED ANALYSIS

A company has a large set of EC2 Instances hosted in an AWS account. They want their IT administrative department to analyse all the EC2 events being generated in near real time. Which of the following can be part of the possible implementations for this requirement? Choose 2 answers from the options given below.

- [ ] **A.** Stream the Cloudwatch events onto Kinesis Firehose ✔
- [ ] **B.** Configure the AWS Config tool on each EC2 Instance to send the EC2 data to Kinesis streams
- [ ] **C.** Stream the data from Kinesis Firehose to Kinesis Analytics ✔
- [ ] **D.** Stream the data from Kinesis Firehose to DynamoDB

Explanation :

Answer – A and C
This is mentioned in one of the AWS Blog posts

- CloudWatch Events offers a near real-time stream of system events that describe changes in AWS resources. CloudWatch Events now supports Kinesis Firehose as a target.
- Kinesis Firehose is a fully managed service for continuously capturing, transforming, and delivering data in minutes to storage and analytics destinations such as Amazon S3, Amazon Kinesis Analytics, Amazon Redshift, and Amazon Elasticsearch Service.

Option B is incorrect since Cloudwatch events should be used to drive the event data from EC2 Instances
Option D is incorrect since data needs to be analysed in real time, hence Kinesis Analytics would be more preferable rather than a data store such as DynamoDB
For more information on this blog post, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/visualize-and-monitor-amazon-ec2-events-with-amazon-cloudwatch-events-and-amazon-kinesis-firehose/ (https://aws.amazon.com/blogs/big-

data/visualize-and-monitor-amazon-ec2-events-with-amazon-cloudwatch-events-and-amazon-kinesis-firehose/)

Ask our Experts

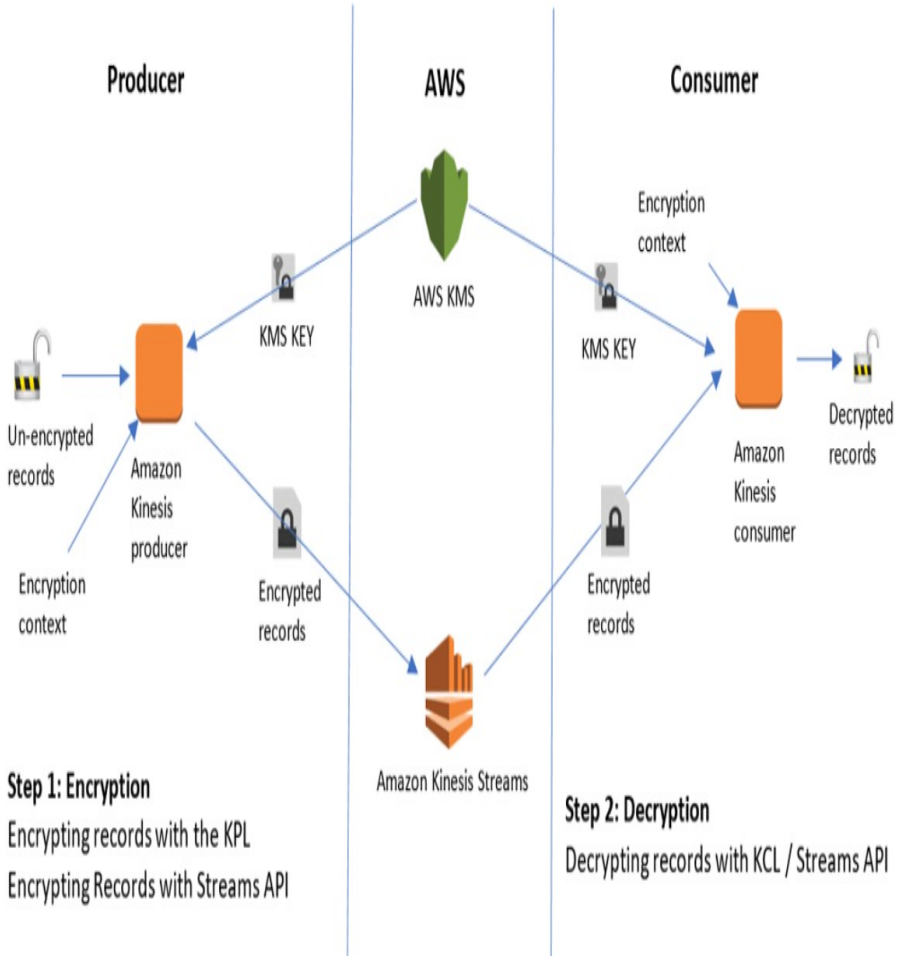
QUESTION 47 UNATTEMPTED DATA SECURITY

A development team has been instructed to implement data producers and consumers which will ingest and consume data using AWS Kinesis Data streams. A key requirement is to ensure that all data sent to Kinesis streams is encrypted. How can you achieve this?

- **A.** Enable encryption at rest in Kinesis streams
- **B.** Use KMS keys to encrypt the data at the consumer side
- **C.** Use KMS keys to encrypt the data at the producer side ✔
- **D.** Embed the KMS keys in the Kinesis streams to encrypt the data

Explanation :

Answer – C
An example of this is given in the AWS Blog posts



Option A is incorrect since this will only encrypt the data at rest within the stream whereas the question states that the data should be encrypted before entering the stream
Option B is incorrect since the encryption needs to happen at the producer side
Option D is incorrect since this is not a possible option for encryption
For more information on this blog post, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/encrypt-and-decrypt-amazon-kinesis-records-using-aws-kms/ (https://aws.amazon.com/blogs/big-data/encrypt-and-decrypt-amazon-kinesis-records-using-aws-kms/)

Ask our Experts

QUESTION 48 CORRECT STORAGE

A company is currently running an EMR cluster with various Apache Hive jobs. They are using Spark SQL on top of the EMR cluster. They need to have a metadata store for Apache Hive and Spark SQL applications that are running on Amazon EMR. Which of the following can be used for this purpose?

- A. Apache Presto
- B. Apache Hue
- C. AWS Glue ✔
- D. Apache Oozie

Explanation :

Answer – C
The following is mentioned in one of the AWS blog posts
AWS Glue Data Catalog provides this essential capability, allowing you to automatically discover and catalog metadata about your data stores in a central repository. Since Amazon EMR 5.8.0, customers have been using the AWS Glue Data Catalog as a metadata store for Apache Hive and Spark SQL applications that are running on Amazon EMR. Starting with Amazon EMR 5.10.0, you can catalog datasets using AWS Glue and run queries using Presto on Amazon EMR from the Hue (Hadoop User Experience) and Apache Zeppelin UIs.
Option A is incorrect since this is an open-source distributed SQL query engine optimized for low-latency, ad-hoc analysis of data.
Option B is incorrect since this is a web interface for the EMR Cluster
Option D is incorrect since this is a workflow Scheduler to manage and coordinate Hadoop jobs
For more information on this blog post, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/easily-manage-table-metadata-for-presto-running-on-amazon-emr-using-the-aws-glue-data-catalog/ (https://aws.amazon.com/blogs/big-data/easily-manage-table-metadata-for-presto-running-on-amazon-emr-using-the-aws-glue-data-catalog/)

Ask our Experts

Finish Review (https://www.whizlabs.com/learn/course/aws-bds-practice-tests/quiz/14851)

- PM Certification (https://www.whizlabs.com/project-management-certifications/)
- Big Data Certification (https://www.whizlabs.com/big-data-certifications/)

Mobile App

- Android Coming Soon
- iOS Coming Soon

Follow us

(https://www.facebook.com/whizlabs.software/)

(https://in.linkedin.com/company/whizlabs-software)

(https://twitter.com/whizlabs?lang=en)

(https://plus.google.com/+WhizlabsSoftware)