NEW PRACTICE TEST 1 - UPDATED

Attempt	1	**Completed on**	Sunday , 03 February 2019 , 11:42 PM
Marks Obtained	0 / 65	**Time Taken**	00 H 00 M 22 S
Your score is	0.0%	**Result**	Fail

Domains / Topics wise Quiz Performance Report

S.No.	Topic	Total Questions	Correct	Incorrect	Unattempted
1	Collection	10	0	0	10
2	Processing	11	0	0	11
3	Data Security	8	0	0	8
4	Visualization	4	0	0	4
5	Storage	25	0	1	24
6	Analysis	7	0	0	7

65 Questions	0 Correct	1 Incorrect	64 Unattempted

Show Answers: All

QUESTION 1 UNATTEMPTED COLLECTION

An architecture is being considered which would consist of several EC2 Instances hosting a data ingestion application. The application would receive thousands of events per second from various IoT devices. The data from these devices need to be streamed for real time analytics. Which of the following would be the ideal way to ingest the data ensuring high throughput of data of ingestion?

- A. Ensure the application implements the Kinesis API library for ingestion of calls
- B. Ensure the application implements the Kinesis KPL library for ingestion of calls ✔

- **C.** Ensure the application implements the Kinesis KCL library for ingestion of calls
- **D.** Ensure the application ingests the data into a Redshift cluster

Explanation :

Answer – B
The AWS Documentation mentions the following
The KPL can help build high-performance producers. Consider a situation where your Amazon EC2 instances serve as a proxy for collecting 100-byte events from hundreds or thousands of low power devices and writing records into a Kinesis data stream. These EC2 instances must each write thousands of events per second to your data stream. To achieve the throughput needed, producers must implement complicated logic, such as batching or multithreading, in addition to retry logic and record de-aggregation at the consumer side. The KPL performs all of these tasks for you.
Option A is invalid since the Kinesis Producer Library would be more efficient that using the Kinesis API
Option C is invalid since this library is used for consuming records
Option D is invalid since Kinesis needs to be used for real time ingestion of data
For more information on the Kinesis Producer Library, please refer to the below URL

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

Ask our Experts

QUESTION 2 UNATTEMPTED COLLECTION

A company has a set of EC2 Instances that host web applications. The web servers used are NGINX and Apache. The IT Security team need to stream the log files from these servers and perform real time analytics from the log files to check for any abnormal behaviour. Which of the following would be the easiest way to get the log file data and the right storage platform for the streaming data?

- **A.** Use a Lambda function to poll the servers. The Lambda files will then strip the data and send the required streaming data to AWS Kinesis
- **B.** Use a Lambda function to poll the servers. The Lambda files will then strip the data and send the required streaming data to AWS Redshift.
- **C.** Use the AWS Kinesis Agent and install it on the EC2 Instances. Ensure the Kinesis agent file is configured to send data from the log files to a Kinesis stream. ✔
- **D.** Use the AWS Kinesis Agent and install it on the EC2 Instances. Ensure the Kinesis agent file is configured to send data from the log files to Amazon Redshift.

Explanation :

Answer – C
The AWS Documentation mentions the following
Kinesis Agent is a stand-alone Java software application that offers an easy way to collect and send data to Kinesis Data Streams. The agent continuously monitors a set of files and sends new data to your stream. The agent handles file rotation, checkpointing, and retry upon failures. It delivers all of your data in a reliable, timely, and simple manner. It also emits Amazon CloudWatch metrics to help you better monitor and troubleshoot the streaming process.
Option A is incorrect since using the Kinesis Agent would be a more efficient tool rather than using the Lambda functions
Options B and D are incorrect since Amazon Redshift is used as a data warehousing system
For more information on working with agents, please refer to the below URL

- https://docs.aws.amazon.com/streams/latest/dev/writing-with-agents.html (https://docs.aws.amazon.com/streams/latest/dev/writing-with-agents.html)

Ask our Experts

QUESTION 3 UNATTEMPTED PROCESSING

Your development team has created separate applications which implement the KPL and KCL library for writing and reading data from Kinesis streams. The KPL is being used to stream information from thousands of IoT devices. The KCL application is consuming the records and providing real time analytics to the data science team. After a dry run, the KCL based application is getting provisioned throughput errors. Which of the following could should be carried out to resolve this issue?

- A. Ensure the number of shards is increased
- B. Increase the number of streams
- C. Increase the throughput for DynamoDB tables ✔
- D. Ensure the application has the right IAM Role attached

Explanation :

Answer – C
The AWS Documentation mentions the following
If your Amazon Kinesis Data Streams application receives provisioned-throughput exceptions, you should increase the provisioned throughput for the DynamoDB table. The KCL creates the table with a provisioned throughput of 10 reads per second and 10 writes per second, but this might not be

sufficient for your application. For example, if your Amazon Kinesis Data Streams application does frequent checkpointing or operates on a stream that is composed of many shards, you might need more throughput.

- Options A and B are incorrect because the issue is related to a data consuming issue
- Option D is incorrect because this is not a security related issue

For more information on KCL and DynamoDB, please refer to the below URL

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

Note:
For each Amazon Kinesis Data Streams application, the KCL uses a unique Amazon DynamoDB table to keep track of the application's state.

For more detailed info, please read through below link:

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

Ask our Experts

QUESTION 4 UNATTEMPTED PROCESSING

Your development team has created separate applications which implement the KPL and KCL library for writing and reading data from Kinesis streams. The KPL is being used to stream information from thousands of IoT devices. The KCL application is consuming the records and providing real time analytics to the data science team. The KCL application has been programmed to poll the Kinesis stream every 150 milliseconds. During a dry run, the KCL based application is getting a lot of "ProvisionedThroughputExceededException"errors. Which of the following could be underlying issue?

- A. The Kinesis stream is created in the wrong region
- B. The polling interval is too short ✔
- C. The polling interval is too long
- D. You should use the Kinesis API for consuming records

Explanation :

Answer – B
The AWS Documentation mentions the following

Because Kinesis Data Streams has a limit of 5 GetRecords calls per second, per shard, setting the idleTimeBetweenReadsInMillisproperty lower than 200ms may result in your application observing the ProvisionedThroughputExceededException exception. Too many of these exceptions can result in exponential back-offs and thereby cause significant unexpected latencies in processing. If you set this property to be at or just above 200 ms and have more than one processing application, you will experience similar throttling.
Option A is incorrect because the dry run would not work at all if the Kinesis stream was created in the wrong region
Option C is incorrect because the polling interval is too short
Option D is incorrect since you should ideally use the KCL library in conjunction with the KPL library
For more information on Kinesis low latency, please refer to the below URL

- https://docs.aws.amazon.com/streams/latest/dev/kinesis-low-latency.html (https://docs.aws.amazon.com/streams/latest/dev/kinesis-low-latency.html)

Ask our Experts

QUESTION 5 UNATTEMPTED DATA SECURITY

A company is planning on making use of Kinesis streams for analysing user trend data for their ecommerce application. The IT Security department has a requirement all data needs to be encrypted at rest. At the same time, the company does not want to manage the internal cryptography or the keys which is required for encryption of data. Which of the following would be the ideal implementation step for such a sort of requirement?

- A. Use IAM policies for the encryption of data
- B. Enable server-side encryption for Kinesis streams ✔
- C. Encrypt the data before sending it to Kinesis streams
- D. Use a CloudHSM service for managing the encryption

Explanation :

Answer – B
The AWS Documentation mentions the following
Server-side encryption is a feature in Amazon Kinesis Data Streams that automatically encrypts data before it's at rest by using an AWS KMS customer master key (CMK) you specify. Data is encrypted before it's written to the Kinesis stream storage layer, and decrypted after it's retrieved from storage. As a result, your data is encrypted at rest within the Kinesis Data Streams service. This allows you to meet strict regulatory requirements and enhance the security of your data.

With server-side encryption, your Kinesis stream producers and consumers don't need to manage master keys or cryptographic operations. Your data is automatically encrypted as it enters and leaves the Kinesis Data Streams service, so your data at rest is encrypted. AWS KMS provides all the master keys that are used by the server-side encryption feature. AWS KMS makes it easy to use a CMK for Kinesis that is managed by AWS, a user-specified AWS KMS CMK, or a master key imported into the AWS KMS service.
Options C and D are incorrect since the company does not want to manage the encryption process
Option A is incorrect since you need to use the KMS service for encryption of data
For more information on server-side encryption with Kinesis, please refer to the below URL

- https://docs.aws.amazon.com/streams/latest/dev/what-is-sse.html (https://docs.aws.amazon.com/streams/latest/dev/what-is-sse.html)

Ask our Experts

QUESTION 6 UNATTEMPTED DATA SECURITY

A company is planning on hosting an application that will make use of Kinesis Streams. The consumer-based application will be sitting on an EC2 Instance in a private subnet. There is requirement to ensure that the application can connect to the Kinesis stream without passing through the Internet. Which of the following would be ideal for this scenario?

- A. Create a VPN connection and attach it to the VPC
- B. Create a NAT Instance in the public subnet
- C. Create a NAT Gateway in the public subnet
- D. Use a VPC endpoint ✔

Explanation :

Answer – D
The AWS Documentation mentions the following
You can use an interface VPC endpoint to keep traffic between your Amazon VPC and Kinesis Data Streams from leaving the Amazon network. Interface VPC endpoints don't require an internet gateway, NAT device, VPN connection, or AWS Direct Connect connection. Interface VPC endpoints are powered by AWS PrivateLink, an AWS technology that enables private communication between AWS services using an elastic network interface with private IPs in your Amazon VPC
All other options are incorrect since this would mean that the traffic needs to traverse via the Internet
For more information on using streams with VPC endpoints, please refer to the below URL

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

Ask our Experts

QUESTION 7 UNATTEMPTED COLLECTION

A company is planning on using Kinesis Firehose to stream data to an AWS Redshift cluster. The cluster will be hosted in a VPC. Which of the following is required to ensure the data can be sent over from Kinesis to AWS Redshift? Choose 2 answers from the options given below.

- ☐ **A.** Ensure that the Kinesis Stream unblocks the ingress traffic from the AWS Redshift cluster
- ☐ **B.** Ensure that the AWS Redshift cluster unblocks the ingress traffic from the AWS Kinesis stream ✔
- ☐ **C.** Ensure that the Redshift cluster is provided with a private IP
- ☐ **D.** Ensure that the Redshift cluster is provided with a public IP ✔

Explanation :

Answer – B and D
The AWS Documentation mentions the following
If your Amazon Redshift cluster is in a virtual private cloud (VPC), it must be publicly accessible with a public IP address. Also, grant Kinesis Data Firehose access to your Amazon Redshift cluster by unblocking the Kinesis Data Firehose IP addresses.
Since this is clearly mentioned in the AWS Documentation, all other options are incorrect
For more information on controlling access to firehose, please refer to the below URL

- https://docs.aws.amazon.com/firehose/latest/dev/controlling-access.html (https://docs.aws.amazon.com/firehose/latest/dev/controlling-access.html)

Ask our Experts

QUESTION 8 UNATTEMPTED PROCESSING

A company is planning on using Kinesis streams firehose to stream their log data from various web servers that host the Apache web server. An application will then read the data which needs to be in JSON format from the underlying destination bucket. Which of the following ideally needs to be in place to ensure that this flow can be implemented?

- A. Ensure that a Lambda transformation is used along with Kinesis Firehose ✔
- B. Ensure that the KPL library is used to parse the records in JSON format.
- C. Ensure that the KCL library is used to parse the records in JSON format.
- D. Change the configuration of the underlying Kinesis data firehose stream to store JSON formatted data

Explanation :

Answer - A
The AWS Documentation mentions the following
****
Lambda Blueprints
Kinesis Data Firehose provides the following Lambda blueprints that you can use to create a Lambda function for data transformation.

- **General Firehose Processing** — Contains the data transformation and status model described in the previous section. Use this blueprint for any custom transformation logic.
- **Apache Log to JSON** — Parses and converts Apache log lines to JSON objects, using predefined JSON field names.
- **Apache Log to CSV** — Parses and converts Apache log lines to CSV format.
- **Syslog to JSON** — Parses and converts Syslog lines to JSON objects, using predefined JSON field names.
- **Syslog to CSV** — Parses and converts Syslog lines to CSV format.
- **Kinesis Data Firehose Process Record Streams as source** — Accesses the Kinesis Data Streams records in the input and returns them with a processing status.
- **Kinesis Data Firehose CloudWatch Logs Processor** — Parses and extracts individual log events from records sent by CloudWatch Logs subscription filters.

****
Options B and C are incorrect since this needs to be done by an AWS Lambda function
Option D is incorrect since there is no such configuration option
For more information on data transformation, please refer to the below URL

- https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html (https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html)

Ask our Experts

QUESTION 9 UNATTEMPTED PROCESSING

A company is building an application that is going to make use of Kinesis streams. They are going to develop the producer and consumer parts of the application. There is a requirement to ensure the strict ordering of records send and processed in the stream. Which of the following would help ensure this? Choose 2 answers from the options given below

- [ ] **A.** Ensure to use the PutRecord API command ✔
- [ ] **B.** Ensure to use the PutRecords API command
- [ ] **C.** Ensure that records are directed towards a particular shard. ✔
- [ ] **D.** Ensure that records are split across various streams

Explanation :

Answer – A and C
The AWS Documentation mentions the following
The response Records array includes both successfully and unsuccessfully processed records. Kinesis Data Streams attempts to process all records in each PutRecords request. A single record failure does not stop the processing of subsequent records. As a result, PutRecords doesn't guarantee the ordering of records. If you need to read records in the same order they are written to the stream, use PutRecord (https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecord.html) instead of PutRecords, and write to the same shard.
For more information on putting a series of records, please refer to the below URL

- https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecords.html (https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecords.html)

Ask our Experts

QUESTION 10 UNATTEMPTED VISUALIZATION

A company's HR department is planning on storing their data in csv files in different S3 buckets. The development team need to create a serverless solution which could be used to create visualizations from the data stored in the S3 buckets. Which of the following can be used for this purpose? Choose 2 answers from the options given below.

- [ ] **A.** Create an EMR Cluster. Use Hive to query the data and create the visualization

- [ ] **B.** Create Javascript code and use the D3.js library ✔
- [ ] **C.** Use the AWS Quicksight service to create the visualization ✔
- [ ] **D.** Use the AWS Athena service to create the visualization

Explanation :

Answer – B and C
D3.js is a JavaScript library for manipulating documents based on data. D3 helps you bring data to life using HTML, SVG, and CSS. D3's emphasis on web standards gives you the full capabilities of modern browsers without tying yourself to a proprietary framework, combining powerful visualization components and a data-driven approach to DOM manipulation. You can use this to visualize data in an S3 bucket as well.
The AWS Documentation mentions the following
Amazon QuickSight is a fast, cloud-powered BI service that makes it easy to build visualizations, perform ad-hoc analysis, and quickly get business insights from your data. Using our cloud-based service you can easily connect to your data, perform advanced analysis, and create stunning visualizations and rich dashboards that can be accessed from any browser or mobile device
Option A is incorrect since this is a Big Data service and is not a serverless service
Option D is incorrect since this is more of querying service
For more information on D3js and Quicksight, please refer to the below URL

- https://d3js.org/ (https://d3js.org/)
- https://aws.amazon.com/quicksight/ (https://aws.amazon.com/quicksight/)

Ask our Experts

QUESTION 11 UNATTEMPTED STORAGE

A company is planning on using the EMR service for their Big data processing needs. They currently want to experiment with the service and create transient cluster to carry out various data processing jobs. Which of the following would help effectively ensure the clusters are transient in nature? Choose 2 possible answers from the options given below. Each answer is an independent and complete solution.

- [ ] **A.** Create the cluster using the AWS CLI
- [ ] **B.** Create the cluster using the EMR API ✔
- [ ] **C.** Ensure the cluster has auto-termination enabled ✔
- [ ] **D.** Ensure the cluster has auto-termination disabled

Explanation :

Answer – B and C
The AWS Documentation mentions the following
By default, clusters that you create using the console or the AWS CLI continue to run until you shut them down. To have a cluster terminate after running steps, you need to enable auto-termination. In contrast, clusters that you launch using the EMR API have auto-termination enabled by default.
Since the documentation clearly mentions this, all other options are invalid.
For more information on planning for cluster deployment, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-longrunning-transient.html (https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-longrunning-transient.html)

Ask our Experts

QUESTION 12 UNATTEMPTED DATA SECURITY

A company is planning on using the EMR service for their Big data processing needs. The database administrator needs to have the ability to login into the nodes. Which of the following needs to be place for this requirement to be fulfilled.

- A. An IAM Role which allows access to the underlying servers
- B. An IAM User which has access to the underlying servers
- C. SSH connections allowed via the Security Groups ✔
- D. SSH connections allowed via IAM Policies

Explanation :

Answer – C
The AWS Documentation mentions the following
Security groups act as virtual firewalls to control inbound and outbound traffic to your cluster. The default Amazon EMR-managed security groups associated with cluster instances do not allow inbound SSH connections as a security precaution. To connect to cluster nodes using SSH so that you can use the command line and view web interfaces that are hosted on the cluster, you need to add inbound rules that allow SSH traffic from trusted clients.
Options A and B are incorrect since the roles and users don't have an impact on the SSH inbound connections
Option D is incorrect since IAM Policies would not determine the underlying SSH connections
For more information on using SSH for cluster nodes, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs-ssh.html (https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs-ssh.html)

Ask our Experts

QUESTION 13 UNATTEMPTED PROCESSING

A company is planning on using an EMR cluster to process data from their On-premise log files. They need to perform SQL queries on the underlying data. Which of the following can be used along with the EMR cluster to satisfy this requirement? Choose 2 answers from the options given below

- [x] **A.** Hive ✔
- [x] **B.** Presto ✔
- [ ] **C.** HCatalog
- [ ] **D.** Sqoop

Explanation :

Answer – A and 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 a 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.
Presto (https://aws.amazon.com/big-data/what-is-presto/) is a fast SQL query engine designed for interactive analytic queries over large datasets from multiple sources.
Option C is incorrect since this tool is used to access Hive metastore tables
Option D is incorrect since this tool is used for transferring data between Amazon S3, Hadoop, HDFS, and RDBMS databases
For more information on hive and presto, 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)
- https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-presto.html (https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-presto.html)

Ask our Experts

QUESTION 14 UNATTEMPTED STORAGE

A team is building an EMR Cluster and also wants to install Hive as an application on the EMR cluster. They need the hive metastore to persist even after the EMR Cluster is terminated. Which of the following can help fulfil this requirement? Choose 2 answers from the options given below

- [x] **A.** Create a MySQL database to store the metastore records ✔
- [ ] **B.** Create a DynamoDB table to store the metastore records
- [x] **C.** Modify the JDBC configuration in the hive-site.xml file in the configuration for the cluster ✔
- [ ] **D.** Modify the JDBC configuration in the Hue configuration setup

Explanation :

Answer – A and C
The AWS Documentation mentions the following
By default, Hive records metastore information in a MySQL database on the master node's file system. The metastore contains a description of the table and the underlying data on which it is built, including the partition names, data types, and so on. When a cluster terminates, all cluster nodes shut down, including the master node. When this happens, local data is lost because node file systems use ephemeral storage. If you need the metastore to persist, you must create an *external metastore* that exists outside the cluster.
You have two options for an external metastore:

- AWS Glue Data Catalog (Amazon EMR version 5.8.0 or later only).
- Amazon RDS or Amazon Aurora.

For more information on hive metastore, please refer to the below URL

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

Ask our Experts

QUESTION 15 UNATTEMPTED PROCESSING

A team is building an EMR Cluster in AWS. They have their own implementations of the Mapper and Reducer functions developed in python that must be used for the Input data. How would you use this in the EMR Cluster?

- (○) **A.** Create a step for the EMR Cluster ✔

- **B.** Use a custom AMI for the cluster
- **C.** Submit the code to AWS Lambda
- **D.** Submit a job via AWS Data Pipeline

Explanation :

Answer – A
This is mentioned in the AWS Documentation

########

Submit a Streaming Step

This section covers the basics of submitting a Streaming step to a cluster. A Streaming application reads input from standard input and then runs a script or executable (called a mapper) against each input. The result from each of the inputs is saved locally, typically on a Hadoop Distributed File System (HDFS) partition. After all the input is processed by the mapper, a second script or executable (called a reducer) processes the mapper results. The results from the reducer are sent to standard output. You can chain together a series of Streaming steps, where the output of one step becomes the input of another step.
The mapper and the reducer can each be referenced as a file or you can supply a Java class. You can implement the mapper and reducer in any of the supported languages, including Ruby, Perl, Python, PHP, or Bash.
########
Option B is incorrect since this is used if you want a custom image for the nodes in your cluster
Options C and D are incorrect since the ideal approach is to use the inbuilt step functionality
For more information on creating a streaming step, please refer to the below URL

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

Ask our Experts

QUESTION 16 UNATTEMPTED STORAGE

A team is building an EMR Cluster in AWS. Management has requested that costs are optimized when working with the cluster. At the same time, you need to ensure capacity needs are met to ensure that EMR jobs are run as per demand. Which of the following can help you accomplish this? Choose 2 answers from the options given below

- [ ] **A.** Use Spot Instances for the master, core and task nodes
- [ ] **B.** Use an Instance fleet configuration for the EMR Cluster ✔
- [ ] **C.** Use a combination of On-demand and Spot Instances for Core and task nodes. ✔
- [ ] **D.** Use On-Demand Instances for the master, core and task nodes

Explanation :

Answer – B and C
The AWS Documentation mentions the following
The instance fleets configuration for a cluster offers the widest variety of provisioning options for EC2 instances. With instance fleets, you specify target capacities for On-Demand Instances and Spot Instances within each fleet. When the cluster launches, Amazon EMR provisions instances until the targets are fulfilled. You can specify up to five EC2 instance types per fleet for Amazon EMR to use when fulfilling the targets. You can also select multiple subnets for different Availability Zones. When Amazon EMR launches the cluster, it looks across those subnets to find the instances and purchasing options you specify.
While a cluster is running, if Amazon EC2 reclaims a Spot Instance because of a price increase, or an instance fails, Amazon EMR tries to replace the instance with any of the instance types that you specify. This makes it easier to regain capacity during a spike in Spot pricing. Instance fleets allow you to develop a flexible and elastic resourcing strategy for each node type. For example, within specific fleets, you can have a core of On-Demand capacity supplemented with less-expensive Spot capacity if available, and then switch to On-Demand capacity if Spot isn't available at your price.
Option A is incorrect since using Spot Instances for the master node is not recommended
Option D is incorrect since this would not be the most cost effective option
For more information on Instance fleets, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-instance-fleet.html (https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-instance-fleet.html)

Ask our Experts

QUESTION 17 UNATTEMPTED STORAGE

A team is building an EMR Cluster in AWS. The cluster has already been created based on current capacity needs. After a duration of 3 months, based on the new storage requirements, it seems that the cluster does not have the required amount of storage based on these requirements. Which of the following can be used to ensure the storage of the cluster meets the new requirements with the least effect on the cluster. Choose 2 answers from the options given below

- [ ] **A.** If the replication factor is high, you can reduce it on the cluster. ✔
- [ ] **B.** Add more nodes to the cluster ✔
- [ ] **C.** Recreate the cluster with more EBS volumes
- [ ] **D.** Recreate the cluster with more EC2 Instances

Explanation :

Answer – A and B
The AWS Documentation mentions the following
If the calculated HDFS capacity value is smaller than your data, you can increase the amount of HDFS storage in the following ways:

- Creating a cluster with additional EBS volumes or adding instance groups with attached EBS volumes to an existing cluster
- Adding more core nodes
- Choosing an EC2 instance type with greater storage capacity
- Using data compression
- Changing the Hadoop configuration settings to reduce the replication factor

Options C and D are incorrect because this would have a large operational impact on the cluster.
For more information on EMR Instance guidelines, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html (https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-instances-guidelines.html)

Ask our Experts

QUESTION 18 UNATTEMPTED VISUALIZATION

A team is building an EMR Cluster in AWS. The monitoring team has a requirement to ensure that they have a dashboard which can be used to monitor the entire cluster and also individual nodes. Which of the following can be installed with the EMR Cluster to provide the monitoring dashboard?

- **A.** Jupyter Notebook
- **B.** Oozie
- **C.** Sqoop
- **D.** Ganglia ✔

Explanation :

Answer - D
The AWS Documentation mentions the following
The Ganglia open source project is a scalable, distributed system designed to monitor clusters and grids while minimizing the impact on their performance. When you enable Ganglia on your cluster, you can generate reports and view the performance of the cluster as a whole, as well as inspect the performance of individual node instances. Ganglia is also configured to ingest and visualize Hadoop and Spark metrics
Option A is invalid because this tool is used to create and share documents that contain live code, equations, visualizations, and narrative text
Option B is invalid because this tool is used as a workflow scheduler to manage and coordinate Hadoop jobs
Option C is invalid because this tool is used for transferring data between Amazon S3, Hadoop, HDFS, and RDBMS databases
For more information on EMR Ganglia, please refer to the below URL

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

Ask our Experts

QUESTION 19 UNATTEMPTED STORAGE

A company needs to have a data store in AWS. The company is responsible to getting weather data and then performing the required analysis on the data. The amount of data can go into Petabytes. It needs to be ensured that storage is efficient when it comes to storage of sparse data. Which of the following would be the MOST ideal data store?

- A. AWS Redshift
- B. AWS EMR with HBase ✔
- C. AWS RDS
- D. AWS DynamoDB

Explanation :

Answer – B
This is mentioned in one of the whitepapers
For large datasets, such as log data, weather data, product catalogs, and so on, you might already have large amounts of historical data that you want to maintain for historical trend analysis, but need to ingest and batch process current data for predictive purposes. For these types of workloads, Apache HBase is a good choice because of its high read and write throughput and efficient storage of sparse data
Option A is incorrect since this is used for data warehousing purposes
Option C is incorrect since this is used for OLTP types of databases
Option D is partially correct but here EMR with Hbase is better
For more information on DynamoDB vs HBase, please refer to the below URL

- https://d1.awsstatic.com/whitepapers/AWS_Comparing_the_Use_of_DynamoDB_and_HBase_for_NoSQL.p
(https://d1.awsstatic.com/whitepapers/AWS_Comparing_the_Use_of_DynamoDB_and_HBase_for_NoSQL.pdf

Ask our Experts

QUESTION 20 UNATTEMPTED STORAGE

A team is building an EMR Cluster in AWS. One of the major requirements is to ensure that data is available even after the EMR cluster is torn down. Which of the following storage option should be used to fulfil this requirement?

- A. HDFS using Instance store
- B. HDFS using EBS Volumes
- C. Local file system
- D. EMRFS ✔

Explanation :

Answer – D
The AWS Documentation mentions the following

EMRFS is an implementation of the Hadoop file system used for reading and writing regular files from Amazon EMR directly to Amazon S3. EMRFS provides the convenience of storing persistent data in Amazon S3 for use with Hadoop while also providing features like Amazon S3 server-side encryption, read-after-write consistency, and list consistency.
All other options are incorrect because these are all temporary storage options
For more information on EMR File systems, please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-file-systems.html (https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-plan-file-systems.html)

Ask our Experts

QUESTION 21 UNATTEMPTED ANALYSIS

A company is planning on setting up an EMR Cluster in AWS. They need to ensure that the cluster to have machine learning capabilities. The data being ingested will be from various log files from EC2 Instances located in AWS. The data is being used to check for any sort of fraud detection. Which of the following would be the ideal application to use along with the underlying EMR Cluster?

- A. Hive
- B. Presto
- C. Spark ✔
- D. HBase

Explanation :

Answer – C
The AWS Documentation mentions the following
Spark natively supports applications written in Scala, Python, and Java. It also includes several tightly integrated libraries for SQL (Spark SQL (https://spark.apache.org/sql/)), machine learning (MLlib (https://spark.apache.org/mllib/)), stream processing (Spark Streaming (https://spark.apache.org/streaming/)), and graph processing (GraphX (https://spark.apache.org/graphx/)). These tools make it easier to leverage the Spark framework for a wide variety of use cases.
The ideal solution to use which has integrated Machine Learning capabilities is Apache Spark , hence the other options are not the most viable ones.

For more information on EMR spark, please refer to the below URL

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

Ask our Experts

QUESTION 22 UNATTEMPTED PROCESSING

A team is currently making use of Kinesis streams for streaming web clicks for an application. There is now a requirement to enable a data analyst team to perform SQL queries on the live data for analytical purposes. Which of the following can be added to the architecture to achieve this requirement?

- **A.** Create an EMR Cluster with Spark. Stream the data from Kinesis streams to Spark. Use Spark to perform the queries. ✔
- **B.** Use the KCL library to directly perform the SQL queries on the incoming data
- **C.** Embed the SQL queries while developing the application using the KPL Library.
- **D.** Use the Data Pipeline service to transfer the data to AWS RDS. Use normal SQL queries for the analysis.

Explanation :

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

> What if you could use your SQL knowledge to discover patterns directly from an incoming stream of data? Streaming analytics is a very popular topic of conversation around big data use cases. These use cases can vary from just accumulating simple web transaction logs to capturing high volume, high velocity and high variety of data emitted from billions of devices such as Internet of things. Most of these introduce a data stream at some point into your data processing pipeline and there is a plethora of tools that can be used for managing such streams. Sometimes, it comes down to choosing a tool that you can adopt faster with your existing skillset.
>
> In this post, we focus on some key tools available within the Apache Spark application ecosystem for streaming analytics. This covers how features like Spark Streaming, Spark SQL, and HiveServer2 can work together on delivering a data stream as a temporary table that understands SQL queries.

Options B and C are incorrect as these do not give dynamic options for using SQL queries for analysis
Option D is incorrect since AWS RDS should be used to host OLTP database
For more information on this use case, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/querying-amazon-kinesis-streams-directly-with-sql-and-spark-streaming/ (https://aws.amazon.com/blogs/big-data/querying-amazon-kinesis-streams-directly-with-sql-and-spark-streaming/)

Ask our Experts

QUESTION 23 UNATTEMPTED STORAGE

A company is planning on sending and storing historical data for an application in a Redshift cluster. The tables being transferred will consist of a fact table and dimensions table. Which of the following is the ideal distribution style to use for the fact table?

- A. Even
- B. Key ✔
- C. Primary
- D. All

Explanation :

Answer - B
The AWS Documentation mentions the following
########

Choose the Best Distribution Style

When you execute a query, the query optimizer redistributes the rows to the compute nodes as needed to perform any joins and aggregations. The goal in selecting a table distribution style is to minimize the impact of the redistribution step by locating the data where it needs to be before the query is executed.

1. **Distribute the fact table and one dimension table on their common columns.**

   Your fact table can have only one distribution key. Any tables that join on another key aren't collocated with the fact table. Choose one dimension to collocate based on how frequently it is joined and the size of the joining rows. Designate both the dimension table's primary key and the fact table's corresponding foreign key as the DISTKEY.

########

Since this is clearly mentioned in the AWS Documentation, all other options are invalid.
For more information on choosing the best distribution style, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-best-dist-key.html (https://docs.aws.amazon.com/redshift/latest/dg/c_best-practices-best-dist-key.html)

Ask our Experts

QUESTION 24 UNATTEMPTED STORAGE

A company is planning on sending and storing historical data for an application in a Redshift cluster. Below is the table structure. It basically stores the orders received for an application.

Order ID
Product ID
Order Value
Timestamp

Most of the queries fired will try to see the recent orders placed. Which of the following column would be ideal for the sort key for the table?

- A. Order ID
- B. Product ID

- C. Order Value
- D. Timestamp ✔

Explanation :

Answer - D
The best practices are provided in the AWS Documentation

########

Choose the Best Sort Key

Amazon Redshift stores your data on disk in sorted order according to the sort key. The Amazon Redshift query optimizer uses sort order when it determines optimal query plans.

- **If recent data is queried most frequently, specify the timestamp column as the leading column for the sort key.**

  Queries are more efficient because they can skip entire blocks that fall outside the time range.

- **If you do frequent range filtering or equality filtering on one column, specify that column as the sort key.**

  Amazon Redshift can skip reading entire blocks of data for that column. It can do so because it tracks the minimum and maximum column values stored on each block and can skip blocks that don't apply to the predicate range.

- **If you frequently join a table, specify the join column as both the sort key and the distribution key.**

  Doing this enables the query optimizer to choose a sort merge join instead of a slower hash join. Because the data is already sorted on the join key, the query optimizer can bypass the sort phase of the sort merge join.

########
Since the AWS Documentation already provides the recommendations, the best option for the sort key would be the timestamp column.
For more information on best practices for sort key, please refer to the below URL

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

Ask our Experts

QUESTION 25 UNATTEMPTED STORAGE

A company has an existing Redshift table which contains all the order information for a product for historical analysis. Now there is a requirement to add more data to this table. Which of the following is the most efficient way to achieve this?

- A. Use a Batch insert command
- B. Make use of staging table ✔
- C. Execute the merge command with the new rows
- D. Execute the upsert command with the new rows

Explanation :

Answer - B
The AWS Documentation mentions the following
You can efficiently add new data to an existing table by using a combination of updates and inserts from a staging table. While Amazon Redshift does not support a single *merge*, or *upsert*, command to update a table from a single data source, you can perform a merge operation by creating a staging table and then using one of the methods described in this section to update the target table from the staging table.
For more information on using staging tables, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/t_updating-inserting-using-staging-tables-.html (https://docs.aws.amazon.com/redshift/latest/dg/t_updating-inserting-using-staging-tables-.html)

Ask our Experts

QUESTION 26 UNATTEMPTED STORAGE

A company has an existing Redshift table which contains all the order information for a product for historical analysis. Currently the timestamp on the table is being used as the sort key. More batches of data were uploaded on the table, but the performance of the queries on the new batches of data are not up to the mark as the prior queries. What needs to be done to ensure that the queries on the new data is optimized?

- **A.** Run the Query optimizer
- **B.** Use the Analyze Compression command
- **C.** Change the sort key for the table
- **D.** Use the VACUUM command ✔

Explanation :

Answer – D
The AWS Documentation mentions the following
For tables with a sort key, the VACUUM command ensures that new data in tables is fully sorted on disk. When data is initially loaded into a table that has a sort key, the data is sorted according to the SORTKEY specification in the CREATE TABLE (https://docs.aws.amazon.com/redshift/latest/dg/r_CREATE_TABLE_NEW.html) statement. However, when you update the table, using COPY, INSERT, or UPDATE statements, new rows are stored in a separate unsorted region on disk, then sorted on demand for queries as required. If large numbers of rows remain unsorted on disk, query performance might be degraded for operations that rely on sorted data, such as range-restricted scans or merge joins. The VACUUM command merges new rows with existing sorted rows, so range-restricted scans are more efficient and the execution engine doesn't need to sort rows on demand during query execution.
Since this is clearly mentioned in the AWS Documentation , all other options are incorrect.
For more information on reclaiming storage please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/t_Reclaiming_storage_space202.html (https://docs.aws.amazon.com/redshift/latest/dg/t_Reclaiming_storage_space202.html)

Ask our Experts

QUESTION 27 UNATTEMPTED DATA SECURITY

Your team is planning on using the AWS IoT service. During the test phase, there are a number of devices which will be used along with the IoT service. Which of the following is the most secure and ideal way to authenticate IoT devices with AWS?

- A. AWS user names and passwords
- B. AWS Cognito Identities
- C. AWS Federated Identities
- D. X.509 Certificates ✔

Explanation :

Answer - D
The AWS Documentation mentions the following
########
AWS IoT Authentication
AWS IoT supports four types of identity principals for authentication:

- X.509 certificates
- IAM users, groups, and roles
- Amazon Cognito identities
- Federated identities

These identities can be used with mobile applications, web applications, or desktop applications. They can even be used by a user typing AWS IoT CLI commands. Typically, AWS IoT devices use X.509 certificates, while mobile applications use Amazon Cognito identities. Web and desktop applications use IAM or federated identities. CLI commands use IAM.
#######
Because this is clearly mentioned in the AWS Documentation , the other options are incorrect
For more information on IoT authentication, please refer to the below URL

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

Ask our Experts

QUESTION 28 UNATTEMPTED PROCESSING

Your team has setup tables in a Redshift cluster. They are now evaluating the performance of queries to ensure that it is up to the mark when users start using the tables. They are evaluating the query plan. When evaluating the query plan, which of the following results would require the team to re-check on the distribution styles used for the underlying tables. Choose 2 answers from the options given below

- [ ] **A.** DS_DIST_NONE
- [ ] **B.** DS_DIST_INNER ✔
- [ ] **C.** DS_DIST_ALL_INNER ✔
- [ ] **D.** DS_DIST_ALL_NONE

Explanation :

Answer – B and C
The AWS Documentation mentions the following
DS_DIST_NONE and DS_DIST_ALL_NONE are good. They indicate that no distribution was required for that step because all of the joins are collocated.
DS_DIST_INNER means that the step will probably have a relatively high cost because the inner table is being redistributed to the nodes.
DS_DIST_ALL_INNER is not good. It means the entire inner table is redistributed to a single slice because the outer table uses DISTSTYLE ALL, so that a copy of the entire outer table is located on each node.
Since this is clearly mentioned in the AWS Documentation, all other options are invalid.
For more information on data distribution in Redshift, please refer to the below URL

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

Ask our Experts

QUESTION 29 UNATTEMPTED DATA SECURITY

A company currently has a Redshift cluster in place. The cluster consists of tables in a star schema format. Cross region snapshots are also currently configured to ensure the cluster is available in the event of disaster recovery. There is a now a requirement that the cluster data needs to be encrypted at rest. You are currently planning for the implementation of these changes. Which of the following would you need to consider in the implementation plan? Choose 2 answers from the options given below

- [ ] **A.** Ensuring that the data is UNLOADED to S3 prior to the encryption process
- [ ] **B.** Ensuring that an OUTAGE interval is kept in mind for the migration process ✔
- [ ] **C.** Disabling cross-region snapshots ✔
- [ ] **D.** Ensure new encrypted EBS volumes are create for the cluster

Explanation :

Answer – B and C
The AWS Documentation mentions the following
You can modify an unencrypted cluster to use AWS Key Management Service (AWS KMS) encryption, using either an AWS-managed key or a customer-managed key (CMK). When you modify your cluster to enable KMS encryption, Amazon Redshift automatically migrates your data to a new encrypted cluster. You can also migrate an unencrypted cluster to an encrypted cluster by modifying the cluster. During the migration operation, your cluster is available in read-only mode, and the cluster status appears as resizing.
If your cluster is configured to enable cross-region snapshot copy, you must disable it before changing encryption.
Options A and D are incorrect since the service itself will copy the data to an encrypted cluster.
For more information on changing the cluster encryption, please refer to the below URL

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

Ask our Experts

QUESTION 30 UNATTEMPTED PROCESSING

A company is looking towards using a Redshift cluster for hosting their data warehouse. CSV files are being generated from their on-premise location and then will be stored in S3. The company needs to ensure the Loading process into Redshift is as fast and efficient as possible. Which of the following can help ensure this? Choose 2 answers from the options given below

- [ ] **A.** Ensure that all CSV files are compressed into a large file to ensure it gets processed faster.
- [ ] **B.** Ensure that multiple CSV files are uploaded to S3 ✔
- [ ] **C.** Ensure to use a node type with a lower instance size
- [ ] **D.** Ensure to use a node type with a higher instance size ✔

Explanation :

Answer – B and D
These recommendations are given in the AWS Documentation
****
Note
We strongly recommend that you divide your data into multiple files to take advantage of parallel processing.

Split your data into files so that the number of files is a multiple of the number of slices in your cluster. That way Amazon Redshift can divide the data evenly among the slices. The number of slices per node depends on the node size of the cluster. For example, each DS1.XL compute node has two slices, and each DS1.8XL compute node has 32 slices. For more information about the number of slices that each node size has, go to About Clusters and Nodes (https://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html#rs-about-clusters-and-nodes) in the *Amazon Redshift Cluster Management Guide*.
#######
Options A and C are incorrect since these would go against AWS recommendations
For more information on data files for Redshift, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/t_splitting-data-files.html (https://docs.aws.amazon.com/redshift/latest/dg/t_splitting-data-files.html)

Ask our Experts

QUESTION 31 UNATTEMPTED DATA SECURITY

A team is planning on uploading multiple data sets onto AWS. These data sets will be queries using JDBC drivers from existing BI tools. There is a requirement that all data sets are encrypted at rest. They want to ensure that they manage the underlying keys which are used for encryption. Which of the following can be used for this purpose? Choose 3 answers from the options given below

- [ ] **A.** Use S3 server-side encryption with Customer keys ✔
- [ ] **B.** Use S3 client-side encryption ✔
- [ ] **C.** Use S3 server-side encryption with AWS managed keys
- [ ] **D.** Use S3 server-side encryption with AWS KMS keys
- [ ] **E.** E. Use S3 server-side encryption with AWS KMS keys with the keys uploaded by the company to KMS ✔

Explanation :

Answer – A, B and E
The AWS Documentation mentions the following
Server-side encryption is about protecting data at rest. Using server-side encryption with customer-provided encryption keys (SSE-C) allows you to set your own encryption keys. With the encryption key you provide as part of your request, Amazon S3 manages both the encryption, as it writes to disks, and decryption, when you access your objects. Therefore, you don't need to maintain any code to perform data encryption and decryption. The only thing you do is manage the encryption keys you provide
Client-side encryption is the act of encrypting data before sending it to Amazon S3. To enable client-

side encryption, you have the following options:

- Use an AWS KMS-managed customer master key
- Use a client-side master key

Options C and D are incorrect since here you will still not manage the complete lifecycle of the keys
Option E is correct, please check below AWS Docs for more details:

- https://aws.amazon.com/blogs/aws/new-bring-your-own-keys-with-aws-key-management-service/ (https://aws.amazon.com/blogs/aws/new-bring-your-own-keys-with-aws-key-management-service/)

For more information on Server side encryption with customer keys and Client side encryption, please refer to the below URL

- https://docs.aws.amazon.com/AmazonS3/latest/dev/ServerSideEncryptionCustomerKeys.html (https://docs.aws.amazon.com/AmazonS3/latest/dev/ServerSideEncryptionCustomerKeys.html)
- https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingClientSideEncryption.html (https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingClientSideEncryption.html)

Ask our Experts

QUESTION 32 UNATTEMPTED PROCESSING

A company has a Redshift cluster defined in AWS. Different departments currently have tables defined in the cluster. Some of the users are complaining on slow performance for queries fired against the tables in the cluster. After careful investigation it seems that some long running queries are consuming resources are not allowing other queries to run efficiently. Which of the following would have the LEAST impact and also ensure that the issue can be resolved?

- A. Create a new cluster for the slow running queries
- B. Modify the WLM configuration for the cluster ✔
- C. Disable any cross region snapshots for the cluster
- D. Query data using Amazon Redshift Spectrum

Explanation :

Answer – B
The AWS Documentation mentions the following
When you have multiple sessions or users running queries at the same time, some queries might consume cluster resources for long periods of time and affect the performance of other queries. For example, suppose one group of users submits occasional complex, long-running queries that select

and sort rows from several large tables. Another group frequently submits short queries that select only a few rows from one or two tables and run in a few seconds. In this situation, the short-running queries might have to wait in a queue for a long-running query to complete.
You can improve system performance and your users' experience by modifying your WLM configuration to create separate queues for the long-running queries and the short-running queries. At run time, you can route queries to these queues according to user groups or query groups.
Option A is invalid since this would require a completely new setup
Option C is invalid since this is not the underlying reason for the issue
Option D is invalid since this is used for querying nested data in Parquet, ORC, JSON, and Ion file formats
For more information on implementing work load management, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/cm-c-implementing-workload-management.html (https://docs.aws.amazon.com/redshift/latest/dg/cm-c-implementing-workload-management.html)

Ask our Experts

QUESTION 33 UNATTEMPTED COLLECTION

A team is currently sending a number of log files over to S3 from various application sources. The team needs to perform searches on the underlying log files. Which of the following can be part of the implementation steps for having such a solution in place? Choose 2 answers form the options given below

- [ ] **A.** Create a Lambda function and attach it to the Amazon S3 bucket which is used to ingest the log files ✔
- [ ] **B.** Send the logs files from S3 to Amazon Glacier for performing log analytics
- [ ] **C.** Send the data from Amazon S3 over to Amazon ElasticSearch ✔
- [ ] **D.** Create a SNS notification and attach it to the Amazon S3 bucket which is used to ingest the log files

Explanation :

Answer – A 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.

You can use Lambda to send data to your Amazon ES domain from Amazon S3. New data that arrives in an S3 bucket triggers an event notification to Lambda, which then runs your custom code to perform the indexing.
Option B is incorrect since Amazon Glacier is used for archive storage
Option D is incorrect since SNS would just send notifications, but we need to send the data for further analysis
For more information on integrating Amazon ES with other services, please refer to the below URL

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

Ask our Experts

QUESTION 34 UNATTEMPTED STORAGE

A company is planning on using the ElastiSearch service. This needs to be setup in their production environment. They need to come up with the ideal number of dedicated master nodes. What is the recommended number of master nodes that should be setup for an ES domain?

- A. 1
- B. 2
- C. 3 ✔
- D. 4

Explanation :

Answer - C
This is given in the AWS Documentation

########

Dedicated Master Nodes

Amazon Elasticsearch Service uses *dedicated master nodes* to increase cluster stability. A dedicated master node performs cluster management tasks, but does not hold data or respond to data upload requests. This offloading of cluster management tasks increases the stability of your domain.
We recommend that you allocate **three** dedicated master nodes for each production Amazon ES domain:

1. One dedicated master node means that you have no backup in the event of a failure.
2. Two dedicated master nodes means that your cluster does not have the necessary quorum of nodes to elect a new master node in the event of a failure.

   A quorum is Number of Dedicated Master Nodes / 2 + 1 (rounded down to the nearest whole number), which Amazon ES sets to `discovery.zen.minimum_master_nodes` when you create your domain.

   In this case, 2 / 2 + 1 = 2. Because one dedicated master node has failed and only one backup exists, the cluster does not have a quorum and cannot elect a new master.

3. Three dedicated master nodes, the recommended number, provides two backup nodes in the event of a master node failure and the necessary quorum (2) to elect a new master.
4. Four dedicated master nodes is no better than three and can cause issues if you use zone awareness (https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-managedomains.html#es-managedomains-zoneawareness).
    - If one master node fails, you have the quorum (3) to elect a new master. If two nodes fail, you lose that quorum, just as you do with three dedicated master nodes.
    - If each Availability Zone has two dedicated master nodes and the zones are unable to communicate with each other, neither zone has the quorum to elect a new master.

########
Since this is clearly mentioned, all other options are invalid
For more information on dedicated master nodes for Amazon ES, please refer to the below URL

- https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-managedomains-dedicatedmasternodes.html (https://docs.aws.amazon.com/elasticsearch-service/latest/developerguide/es-managedomains-dedicatedmasternodes.html)

Ask our Experts

QUESTION 35 UNATTEMPTED STORAGE

A company is planning on using AWS for their data store. They need to ensure that users in their company can upload different types of data items (size ranging from 16KB to 500 MB) onto AWS. The company does not want the responsibility of managing the underlying capacity of the service. Which of the following would be the ideal data store for this requirement?

- A. AWS S3 ✔
- B. AWS DynamoDB
- C. AWS Redshift
- D. AWS EMR

Explanation :

Answer – A
The AWS Documentation mentions the following
Companies today need the ability to simply and securely collect, store, and analyze their data at a massive scale. Amazon S3 is object storage (https://aws.amazon.com/what-is-cloud-object-storage/) built to store and retrieve any amount of data from anywhere – web sites and mobile apps,

corporate applications, and data from IoT sensors or devices. It is designed to deliver 99.999999999% durability, and stores data for millions of applications used by market leaders in every industry. S3 provides comprehensive security and compliance capabilities that meet even the most stringent regulatory requirements. It gives customers flexibility in the way they manage data for cost optimization, access control, and compliance. S3 provides query-in-place functionality, allowing you to run powerful analytics directly on your data at rest in S3. And Amazon S3 is the most supported cloud storage service available, with integration from the largest community of third-party solutions, systems integrator partners, and other AWS services.
Options C and D are incorrect since here you need to manage the underlying infrastructure
Option B is incorrect since the size of the data items size range is not ideal for DynamoDB
For more information on Amazon S3, please refer to the below URL

- https://aws.amazon.com/s3/ (https://aws.amazon.com/s3/)

Ask our Experts

QUESTION 36 UNATTEMPTED ANALYSIS

A company is planning on using the Machine Learning service to perform a predictive analysis. This is for a ecommerce application wherein, based on various metrics they want to determine if a person would probably buy a product or not. Which of the following ML model would be used for this requirement?

- A. Regression
- B. Multiclass
- C. Binary ✔
- D. Model Size

Explanation :

Answer - C
Since this is a simple case of a Yes or No classification, you should use the Binary classification.
Hence all other options are invalid
The different types of ML models are given in the AWS Documentation

ML Models

An ML model is a mathematical model that generates predictions by finding patterns in your data. Amazon ML supports three types of ML models: binary classification, multiclass classification and regression.
The following table defines terms that are related to ML models.

Term	Definition

Regression	The goal of training a regression ML model is to predict a numeric value.
Multiclass	The goal of training a multiclass ML model is to predict values that belong to a limited, pre-defined set of permissible values.
Binary	The goal of training a binary ML model is to predict values that can only have one of two states, such as true or false.
Model Size	ML models capture and store patterns. The more patterns a ML model stores, the bigger it will be. ML model size is described in Mbytes.
Number of Passes	When you train an ML model, you use data from a datasource. It is sometimes beneficial to use each data record in the learning process more than once. The number of times that you let Amazon ML use the same data records is called the number of passes.
Regularization	Regularization is a machine learning technique that you can use to obtain higher-quality models. Amazon ML offers a default setting that works well for most cases.

For more information on Machine Learning Key concepts, please refer to the below URL

- https://docs.aws.amazon.com/machine-learning/latest/dg/amazon-machine-learning-key-concepts.html (https://docs.aws.amazon.com/machine-learning/latest/dg/amazon-machine-learning-key-concepts.html)

Ask our Experts

QUESTION 37 UNATTEMPTED STORAGE

A company is planning on using the Machine Learning service to perform a predictive analysis. There are various input files which will be used and submitted to Machine Learning. How should you prepare the data to ensure it can be used as Input data for Amazon Machine Learning? Choose 2 answers from the options given below

- [ ] **A.** Ensure that all input files are in JSON format
- [ ] **B.** Ensure that all input files are in csv format ✔
- [ ] **C.** Ensure that all input files have the same data schema ✔
- [ ] **D.** Ensure that all input files have the different data schema's

Explanation :

Answer – B and C
The AWS Documentation mentions the following
Input data is the data that you use to create a datasource. You must save your input data in the comma-separated values (.csv) format
You can provide your input to Amazon ML as a single file, or as a collection of files. Collections must satisfy these conditions:

- All files must have the same data schema.
- All files must reside in the same Amazon Simple Storage Service (Amazon S3) prefix, and the path that you provide for the collection must end with a forward slash ('/') character.

Since this is clearly mentioned in the AWS Documentation , all other options are incorrect
For more information on the data format for Amazon Machine Learning, please refer to the below URL

- https://docs.aws.amazon.com/machine-learning/latest/dg/understanding-the-data-format-for-amazon-ml.html (https://docs.aws.amazon.com/machine-learning/latest/dg/understanding-the-data-format-for-amazon-ml.html)

Ask our Experts

QUESTION 38 UNATTEMPTED ANALYSIS

A company is planning on hosting data sets via files uploaded to S3. Amazon Athena will be used to create tables based on the files in S3. Which of the following must be taken into consideration when carrying out such an implementation?

Choose 2 options.

- [ ] **A.** Ensure that a Lambda function is in place to remove any unwanted files from the S3 bucket ✔
- [ ] **B.** Ensure that all files are in csv file
- [ ] **C.** In the LOCATION clause, use a trailing slash for your folder or bucket ✔
- [ ] **D.** Ensure that versioning is enabled for the S3 bucket

Explanation :

Answer – A and C
The AWS Documentation mentions the following
Use these tips and examples when you specify the location in Amazon S3.

- Athena reads all files in an Amazon S3 location you specify in the CREATE TABLE statement and cannot ignore any files included in the prefix. When you create tables, include in the Amazon S3 path only the files you want Athena to read. Use AWS Lambda functions to scan files in the source location, remove any empty files, and move unneeded files to another location.
- In the LOCATION clause, use a trailing slash for your folder or bucket.
- Option B is incorrect since the file can have other delimiters and it is not necessary that it is in csv format.
- Option D is incorrect since this is not a prime requirement for using Athena with S3

For more information on Table locations for Amazon S3, please refer to the below URL

- https://docs.aws.amazon.com/athena/latest/ug/tables-location-format.html (https://docs.aws.amazon.com/athena/latest/ug/tables-location-format.html)

Ask our Experts

QUESTION 39 UNATTEMPTED STORAGE

A company is planning on hosting data sets via files uploaded to S3. Amazon Athena will be used to create tables based on the files in S3. The files will be in csv format and the tables will be created based on the files. Which of the following needs to be used when creating the tables which works with different formats.

- A. Defining a throughput for the table
- B. Using a SerDe ✔
- C. Using Indexes
- D. Using Primary Keys

Explanation :

Answer – B
This is mentioned in the AWS Documentation
########

Using a SerDe

A SerDe (Serializer/Deserializer) is a way in which Athena interacts with data in various formats.
It is the SerDe you specify, and not the DDL, that defines the table schema. In other words, the SerDe can override the DDL configuration that you specify in Athena when you create your table.

To Use a SerDe in Queries

To use a SerDe when creating a table in Athena, use one of the following methods:

- Use DDL statements to describe how to read and write data to the table and do not specify a `ROW FORMAT`, as in this example. This omits listing the actual SerDe type and the native `LazySimpleSerDe` is used by default.

In general, Athena uses the `LazySimpleSerDe` if you do not specify a `ROW FORMAT`, or if you specify `ROW FORMAT DELIMITED`.

ROW FORMAT
DELIMITED FIELDS TERMINATED BY ','
ESCAPED BY '\\'
COLLECTION ITEMS TERMINATED BY '|'
MAP KEYS TERMINATED BY ':'
########

The other options are more relevant when you start working with other database formats such as DynamoDB or AWS RDS
For more information on using SerDe, please refer to the below URL

- https://docs.aws.amazon.com/athena/latest/ug/serde-about.html (https://docs.aws.amazon.com/athena/latest/ug/serde-about.html)

Ask our Experts

QUESTION 40 UNATTEMPTED STORAGE

A company is planning on using AWS DynamoDB for storing around 10 TB of data. They need to have single-digit milliseconds to data in the table. They also need to ensure that the application sitting on Amazon EC2 Instance uses the right security credentials to access the DynamoDB table. Which of the following implementation steps will help fulfil this requirement. Choose 2 answers from the options given below.

- [ ] **A.** Use an Elastic Cache in front of the DynamoDB table for faster response times.
- [ ] **B.** Use the DAX in memory cache ✔
- [ ] **C.** Use IAM users Access Keys on the EC2 Instance for accessing the DynamoDB table
- [ ] **D.** Attach an IAM Role to the EC2 Instance for accessing the DynamoDB table ✔

Explanation :

Answer – B and D
The AWS Documentation mentions the following
Amazon DynamoDB is designed for scale and performance. In most cases, the DynamoDB response times can be measured in single-digit milliseconds. However, there are certain use cases that require response times in microseconds. For these use cases, *DynamoDB Accelerator (DAX)* delivers fast response times for accessing eventually consistent data.
Instead, you can and should use an IAM role to manage *temporary* credentials for applications that run on an EC2 instance. When you use a role, you don't have to distribute long-term credentials (such as a user name and password or access keys) to an EC2 instance. Instead, the role supplies temporary permissions that applications can use when they make calls to other AWS resources. When you launch an EC2 instance, you specify an IAM role to associate with the instance. Applications that run on the instance can then use the role-supplied temporary credentials to sign API requests.
Option A is partially correct, but you would want to use DAX instead of ElastiCache
Option C is incorrect since Access Keys are a security risk for allowing access
For more information on working with DAX and IAM Roles, please refer to the below URL

- https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/DAX.html (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/DAX.html)
- https://docs.aws.amazon.com/IAM/latest/UserGuide/id_roles_use_switch-role-ec2.html (https://docs.aws.amazon.com/IAM/latest/UserGuide/id_roles_use_switch-role-ec2.html)

Ask our Experts

A company currently has a large data warehouse hosted in their On-premise Oracle database engine. They need to migrate the data warehouse over to AWS Redshift. Which of the following can be used to transfer the data from the on-premise warehouse over to Redshift?

- A. AWS Redshift Copy command
- B. AWS Schema Conversion Tool ✔
- C. AWS VM Migration
- D. AWS Cloudformation

Explanation :

Answer – B
The AWS Documentation mentions the following
You can use the AWS Schema Conversion Tool (AWS SCT) to convert your existing database schema from one database engine to another. You can convert relational OLTP schema, or data warehouse schema. Your converted schema is suitable for an Amazon Relational Database Service (Amazon RDS) MySQL DB instance, an Amazon Aurora DB cluster, an Amazon RDS PostgreSQL DB instance, or an Amazon Redshift cluster. The converted schema can also be used with a database on an Amazon EC2 instance or stored as data on an Amazon S3 bucket.
Option A is incorrect since for this you need to have the data already available for copying into the Redshift cluster
Option C is incorrect since this is used to migrate Virtual Machines
Option D is incorrect since this is used for creating templates for Infrastructure deployment
For more information on the schema conversion tool, please refer to the below URL

- https://docs.aws.amazon.com/SchemaConversionTool/latest/userguide/CHAP_Welcome.html (https://docs.aws.amazon.com/SchemaConversionTool/latest/userguide/CHAP_Welcome.html)

Ask our Experts

QUESTION 42 UNATTEMPTED ANALYSIS

Your company has a requirement to crawl all of the log files generated via Cloudtrail for better analysis of the log files. Which of the following would be part of the Implementation plan for this requirement? Choose 2 answers from the options given below

- A. Use Lambda functions to transform the log files ✔

- [ ] **B.** Use AWS Glue for cataloguing the information ✔
- [ ] **C.** Use AWS Redshift for storing the log files
- [ ] **D.** Use AWS DynamoDB for storing the log files

Explanation :

Answer – A and B
This use case is provided in the AWS Documentation
CloudTrail delivers log files in an Amazon S3 bucket folder. To correctly crawl these logs, you modify the file contents and folder structure using an Amazon S3-triggered Lambda function that stores the transformed files in an S3 bucket single folder. When the files are in a single folder, AWS Glue scans the data, converts it into Apache Parquet format, and catalogs it to allow for querying and visualization using Amazon Athena and Amazon QuickSight.
Options C and D are incorrect since the ideal place to store the logs would be S3.
For more information on Cloud Trail Visualization, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/streamline-aws-cloudtrail-log-visualization-using-aws-glue-and-amazon-quicksight/ (https://aws.amazon.com/blogs/big-data/streamline-aws-cloudtrail-log-visualization-using-aws-glue-and-amazon-quicksight/)

Ask our Experts

QUESTION 43 UNATTEMPTED DATA SECURITY

A company currently has a Hadoop Cluster setup using the AWS EMR service. This is being used to host several tables on which python jobs are run for processing the data. Recently the IT security department have mandated that all data is encrypted in transit within the Hadoop Cluster. Which of the following can be used to fulfil this requirement? Choose 2 answers from the options given below

- [ ] **A.** Hadoop MapReduce Encrypted ✔
- [ ] **B.** VPC Endpoints
- [ ] **C.** Modify the EC2 IAM Roles
- [ ] **D.** Secure Hadoop RPC is set to "Privacy" and use SASL ✔

Explanation :

Answer – A and D
The AWS Documentation mentions the following

Several encryption mechanisms are enabled with in-transit encryption. These are open-source features, are application-specific, and may vary by Amazon EMR release. The following application-specific encryption features can be enabled using security configurations:

- Hadoop (for more information, see Hadoop in Secure Mode (https://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/SecureMode.html) in Apache Hadoop documentation):
  - Hadoop MapReduce Encrypted Shuffle (https://hadoop.apache.org/docs/r2.7.1/hadoop-mapreduce-client/hadoop-mapreduce-client-core/EncryptedShuffle.html) uses TLS.
  - Secure Hadoop RPC (https://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/SecureMode.html#Data_Encryption_on_RPC) is set to "Privacy" and uses SASL (activated in Amazon EMR when at-rest encryption is enabled).
  - Data encryption on HDFS block data transfer (https://hadoop.apache.org/docs/r2.7.2/hadoop-project-dist/hadoop-common/SecureMode.html#Data_Encryption_on_Block_data_transfer.) uses AES 256 (activated in Amazon EMR when at-rest encryption is enabled in the security configuration).

Option B is incorrect since this is normally used when a service in the private subnet needs to access a public AWS service without the traffic moving over the Internet
Option C is incorrect since IAM Roles are used to give access to other AWS services
For more information on Data Encryption options in EMR, please refer to the below URL

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

Ask our Experts

QUESTION 44 UNATTEMPTED VISUALIZATION

A company is planning on using Apache Spark on an EMR Cluster in AWS. They need to have Interactive data analytics which can be performed on the underlying data. Which of the following can be used for this purpose?

- A. Apache Spark SQL
- B. Apache Hue
- C. Apache Zeppelin ✔
- D. Spark Streaming

Explanation :

Answer – C
Apache Zeppelin is a new and incubating multi-purposed web-based notebook which brings data ingestion, data exploration, visualization, sharing and collaboration features to Hadoop and Spark. Zeppelin is the only tool among the above options, which provides visualization and basic reporting capabilities with custom plugin support

Option A is incorrect since this is used to issue SQL queries using the Spark engine
Option B is incorrect since this is used as a web interface for the Hadoop cluster
Option D is incorrect since this is used to stream data into Spark
For more information on Apache Zeppelin, 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)

Note:

- **Apache Zeppelin is** a new and incubating multi-purposed web-based notebook which brings data ingestion, data exploration, visualization, sharing and collaboration features to Hadoop and Spark
- Apache Spark (https://aws.amazon.com/big-data/what-is-spark/) is an open-source, distributed processing system commonly used for big data (https://aws.amazon.com/big-data/what-is-big-data/) workloads. Apache Spark utilizes in-memory caching and optimized execution for fast performance, and it supports general batch processing, streaming analytics, machine learning, graph databases, and ad hoc queries

Ask our Experts

QUESTION 45 UNATTEMPTED COLLECTION

A company is developing an application that will make use of Kinesis streams. They are developing the producer and consumer components. They need to ensure that data is distributed across the shards of the streams. Which of the following aspect of the data record helps achieve this?

- A. Sequence number
- B. Hash Key
- C. Partition key ✔
- D. String Blob

Explanation :

Answer – C
The AWS Documentation mentions the following

The partition key is used by Kinesis Data Streams to distribute data across shards. Kinesis Data Streams segregates the data records that belong to a stream into multiple shards, using the partition key associated with each data record to determine the shard to which a given data record belongs.
Option A is incorrect since this is used to uniquely identify the record in the shard
Option B is incorrect since this is internally generated by AWS Kinesis
Option D is incorrect since this is the data payload which is sent with the streaming data
For more information on Apache Zeppelin, please refer to the below URL

- https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecord.html (https://docs.aws.amazon.com/kinesis/latest/APIReference/API_PutRecord.html)

Ask our Experts

QUESTION 46 UNATTEMPTED STORAGE

A company is planning on using DynamoDB for storing all data related to tweets. The data will go into millions of rows and needs to scale based on demand. The design team needs to ensure that the objects inserted into the DynamoDB tables are uniformly distributed via the partitions created in DynamoDB. Which of the following can help achieve this? Choose 2 answers from the options given below

- [ ] **A.** Place a higher read capacity for the tables
- [ ] **B.** Ensure to choose a sort key when creating the table
- [ ] **C.** Use Random Suffixes as a sharding technique ✔
- [ ] **D.** Use Calculated Suffixes as a sharding technique ✔

Explanation :

Answer – C and D
The AWS Documentation mentions the following as strategies for better partitioning of keys in DynamoDB
Options A and B are incorrect since these are not used for equal distribution of objects in a DynamoDB table
Sharding Using Random Suffixes
One strategy for distributing loads more evenly across a partition key space is to add a random number to the end of the partition key values. Then you randomize the writes across the larger space.
Sharding Using Calculated Suffixes
A randomizing strategy can greatly improve write throughput. But it's difficult to read a specific item because you don't know which suffix value was used when writing the item. To make it easier to read individual items, you can use a different strategy. Instead of using a random number to distribute the items among partitions, use a number that you can calculate based upon something that you want to query on.

For more information on key sharding in DynamoDB, please refer to the below URL

- https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-partition-key-sharding.html (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-partition-key-sharding.html)

Ask our Experts

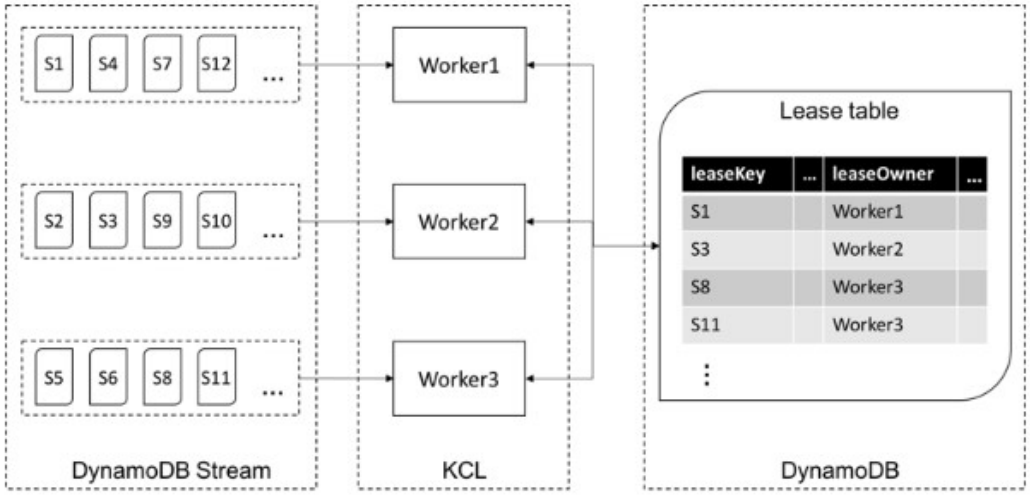
QUESTION 47 UNATTEMPTED COLLECTION

A company is planning on using DynamoDB for storing all data related to tweets. The data will go into millions of rows and needs to scale based on demand. At the same time, they need to ensure that the data from the DynamoDB table with all the tweets gets continually replicated to a DynamoDB table in another region. Which of the following can help effectively manage this requirement? Choose 2 answers from the options given below

- [ ] **A.** Use DynamoDB streams ✔
- [ ] **B.** Use DynamoDB Autoscaling
- [ ] **C.** Use the KCL Library to store the data into the destination table ✔
- [ ] **D.** Use the KPL Library to store the data into the destination table

Explanation :

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



Option B is incorrect since this is an option if you want to ensure that the throughput for the DynamoDB table scales as per demand
Option D is incorrect since you need to use the KCL library to consume the tweets
For more information on working with streams and KCL, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/process-large-dynamodb-streams-using-multiple-amazon-kinesis-client-library-kcl-workers/ (https://aws.amazon.com/blogs/big-data/process-large-dynamodb-streams-using-multiple-amazon-kinesis-client-library-kcl-workers/)

Ask our Experts

QUESTION 48 UNATTEMPTED ANALYSIS

A company currently has large data sets defined in S3 and are using AWS Athena to query the data sets. Since the query time is taking longer than expected , steps need to be taken to improve query performance. Which of the following can be taken ensuring that cost is not increased in the implementation process? Choose 2 answers from the options given below

- [ ] **A.** Consider splitting the data set ✔
- [ ] **B.** Consider using the CREATE TABLE AS SELECT statement ✔
- [ ] **C.** Consider using AWS Quicksight instead
- [ ] **D.** Consider using EMR clusters

Explanation :

Answer – A and B
An example of this is given in the AWS Documentation
#######
Using CTAS statements with Amazon Athena to reduce cost and improve performance
Amazon Athena (https://aws.amazon.com/athena/?nc2=h_m1) is an interactive query service that makes it more efficient to analyze data in Amazon S3 (https://aws.amazon.com/s3/?nc2=h_m1) using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run. Athena recently released support for creating tables using the results of a SELECT query or CREATE TABLE AS SELECT (CTAS) statement (https://docs.aws.amazon.com/athena/latest/ug/ctas.html). Analysts can use CTAS statements to create new tables from existing tables on a subset of data, or a subset of columns. They also have options to convert the data into columnar formats, such as Apache Parquet and Apache ORC, and partition it. Athena automatically adds the resultant table and partitions to the AWS Glue Data Catalog (https://docs.aws.amazon.com/glue/latest/dg/populate-data-catalog.html), making them immediately available for subsequent queries.
CTAS statements help reduce cost and improve performance by allowing users to run queries on smaller tables constructed from larger tables. This post covers three use cases that demonstrate the benefit of using CTAS to create a new dataset, smaller than the original one, allowing subsequent queries to run faster. Assuming our use case requires repeatedly querying the data, we can now query a smaller and more optimal dataset to get the results faster.
#######
Option C is incorrect since this is more of a visualization tool
Option D is incorrect since this would increase the costs of the overall solution
For more information on this use case, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/using-ctas-statements-with-amazon-athena-to-reduce-cost-and-improve-performance/ (https://aws.amazon.com/blogs/big-data/using-ctas-statements-with-amazon-athena-to-reduce-cost-and-improve-performance/)

Ask our Experts

QUESTION 49 UNATTEMPTED DATA SECURITY

A company is planning on using Amazon Athena along with datasets hosted in S3. They want to allow their On-premise Active Directory users to use the AWS Athena service for querying purposes. Which of the following can be used for authentication for the existing users ensuring the least maintenance overhead?

- A. Create IAM Access Keys for the users
- B. Create X.509 Certificates for the users
- C. Use the Secure Token service ✔

D. Use the AD Connect service

Explanation :

Answer - C
An example of this is given in the AWS Documentation
########

Connect to Amazon Athena with federated identities using temporary credentials

Many organizations have standardized on centralized user management, most commonly Microsoft Active Directory or LDAP. Access to AWS resources is no exception. Amazon Athena (https://aws.amazon.com/athena/?nc2=h_m1) is a serverless query engine for data on Amazon S3 (https://aws.amazon.com/s3/?nc2=h_m1) that is popular for quick and cost-effective queries of data in a data lake. To allow users or applications to access Athena, organizations are required to use an AWS access key and an access secret key from which appropriate policies are enforced. To maintain a consistent authorization model across, organizations must enable authentication and authorization for Athena by using federated users.
This blog post shows the process of enabling federated user access with the AWS Security Token Service (AWS STS) (https://docs.aws.amazon.com/STS/latest/APIReference/Welcome.html). This approach lets you create temporary security credentials and provides them to trusted users for running queries in Athena.
########
Since this is clearly mentioned in the AWS Documentation, all other options are incorrect
For more information on using STS with Amazon Athena, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/connect-to-amazon-athena-with-federated-identities-using-temporary-credentials/ (https://aws.amazon.com/blogs/big-data/connect-to-amazon-athena-with-federated-identities-using-temporary-credentials/)

Ask our Experts

QUESTION 50 UNATTEMPTED ANALYSIS

A company wants to start storing their large data sets on S3. They want to use a serverless service for managing the SQL queries. They need to call these queries using JDBC and ODBC drivers. Which of the following can be used for this purpose?

- A. AWS Athena ✔
- B. AWS EMR
- C. AWS EC2
- D. AWS Lambda

Explanation :

Answer – A
The AWS Documentation mentions the following
Athena is an interactive query service that lets you analyze data directly in Amazon S3 by using standard SQL. You can access Athena by using JDBC and ODBC drivers, AWS SDK, or the Athena console.
Options B and C are incorrect since these are not serverless services
Option D is incorrect since this is a compute service and not a query service
For more information on Amazon Athena service, please refer to the below URL

- https://docs.aws.amazon.com/athena/latest/ug/what-is.html (https://docs.aws.amazon.com/athena/latest/ug/what-is.html)

Ask our Experts

QUESTION 51 UNATTEMPTED STORAGE

Your company has just enabled VPC Flow logs for a large number of Network Interfaces. They need to stream the data into an S3 bucket and hence are using Amazon Kinesis Firehose for this purpose. They need to transform the data before it can be used for analysis. Which of the following can be used for the transformation purpose?

- A. AWS Lambda ✔

- B. Kinesis KCL
- C. Kinesis KPL
- D. Amazon SQS

Explanation :

Answer – A
An example of this is given in the AWS Documentation
#######
3. Decompress records with AWS Lambda
There may be situations where you want to transform or enrich streaming data before writing it to its final destination. In this solution, we must decompress the data that is streamed from CloudWatch Logs. With the Amazon Kinesis Data Firehose Data Transformation (https://docs.aws.amazon.com/firehose/latest/dev/data-transformation.html) feature, we can decompress the data with an AWS Lambda (https://aws.amazon.com/lambda/?nc2=h_m1) function. Kinesis Data Firehose manages the invocation of the function. Inside the function, the data is decompressed and returned to Kinesis Data Firehose. The complete source code for the Lambda function can be found here.
#######
Option B is invalid since this is used to consume data from Kinesis streams
Option C is invalid since this is used to send data from Kinesis streams
Option D is invalid since this is a queuing service
For more information on this use case, please refer to the below URL
https://aws.amazon.com/blogs/big-data/analyze-and-visualize-your-vpc-network-traffic-using-amazon-kinesis-and-amazon-athena/ (https://aws.amazon.com/blogs/big-data/analyze-and-visualize-your-vpc-network-traffic-using-amazon-kinesis-and-amazon-athena/)

Ask our Experts

QUESTION 52 UNATTEMPTED COLLECTION

A company has an on-premise data store in Oracle. They need to import the data into an Amazon EMR Cluster which uses HDFS. Which of the following can be used to fulfil this requirement?

- A. Apache Hive
- B. Apache Sqoop ✔
- C. Apache Hue
- D. Jupyter Notebook

Explanation :

Answer – B

#######

Migrate RDBMS or On-Premise data to EMR Hive, S3, and Amazon Redshift using EMR – Sqoop

This blog post shows how our customers can benefit by using the Apache Sqoop tool. This tool is designed to transfer and import data from a Relational Database Management System (RDBMS) into AWS – EMR Hadoop Distributed File System (HDFS), transform the data in Hadoop, and then export the data into a Data Warehouse (e.g. in Hive or Amazon Redshift (https://aws.amazon.com/documentation/redshift/?id=docs_gateway)).

To demonstrate the Sqoop tool, this post uses Amazon RDS (https://aws.amazon.com/documentation/rds/?id=docs_gateway) for MySQL as a source and imports data in the following three scenarios:

- **Scenario 1** – AWS EMR (https://aws.amazon.com/documentation/emr/) (HDFS -> Hive and HDFS)
- **Scenario 2** – Amazon S3 (https://aws.amazon.com/documentation/s3/?id=docs_gateway) (EMFRS), and then to EMR-Hive
- **Scenario 3** – S3 (EMFRS), and then to Redshift

#######

Option A is incorrect since this is an open-source, data warehouse, and analytic package that runs on top of a Hadoop cluster

Option C is incorrect since this is an open-source, web-based, graphical user interface for use with Amazon EMR and Apache Hadoop

Option D is incorrect since this is an open-source web application that you can use to create and share documents that contain live code, equations, visualizations, and narrative text

For more information on a use case that uses Apache Sqoop, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/migrate-rdbms-or-on-premise-data-to-emr-hive-s3-and-amazon-redshift-using-emr-sqoop/ (https://aws.amazon.com/blogs/big-data/migrate-rdbms-or-on-premise-data-to-emr-hive-s3-and-amazon-redshift-using-emr-sqoop/)

Ask our Experts

QUESTION 53 UNATTEMPTED STORAGE

A company has data stores both on their on-premise and AWS Environments. They need to first create a data lake in AWS and then orchestrate several ETL jobs. Which of the following can be used to fulfil this requirement? Choose 2 answers from the options given below

- [ ] **A.** Use AWS S3 for storage of the data lake ✔
- [ ] **B.** Use AWS EMR for storage of the data lake

- [ ] **C.** Use a combination of AWS Lambda and Step Functions ✔
- [ ] **D.** Use SQS for the ETL jobs

Explanation :

Answer – A and C
The AWS Documentation mentions the following
Extract, transform, and load (ETL) operations collectively form the backbone of any modern enterprise data lake. It transforms raw data into useful datasets and, ultimately, into actionable insight. An ETL job typically reads data from one or more data sources, applies various transformations to the data, and then writes the results to a target where data is ready for consumption. The sources and targets of an ETL job could be relational databases in Amazon Relational Database Service (Amazon RDS) (https://aws.amazon.com/rds/?nc2=h_m1) or on-premises, a data warehouse such as Amazon Redshift (https://aws.amazon.com/redshift/?nc2=h_m1), or object storage such as Amazon Simple Storage Service (Amazon S3) (https://aws.amazon.com/s3/?nc2=h_m1) buckets. Amazon S3 as a target is especially commonplace in the context of building a data lake in AWS (https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/).
You can also use AWS Step Functions (https://docs.aws.amazon.com/step-functions/latest/dg/welcome.html) and AWS Lambda (https://docs.aws.amazon.com/lambda/latest/dg/welcome.html) for orchestrating multiple ETL jobs involving a diverse set of technologies in an arbitrarily-complex ETL workflow
Option B is incorrect since AWS S3 will be a more viable option for a data lake
Option D is incorrect since this is a messaging service
For more information on a use case for this, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/orchestrate-multiple-etl-jobs-using-aws-step-functions-and-aws-lambda/ (https://aws.amazon.com/blogs/big-data/orchestrate-multiple-etl-jobs-using-aws-step-functions-and-aws-lambda/)

Ask our Experts

QUESTION 54 UNATTEMPTED COLLECTION

A company wants to build a data lake which will comprise of the following

Logs generated from CloudTrail
VPC Flow Logs
Logs from Application Load Balancers hosted in AWS
They need to stream the logs and also ensure an expansive data lake is available for storage purposes. Which of the following can be used to fulfil this requirement in the most efficient manner? Choose 2 answers from the options given below

- [ ] **A.** Use AWS S3 for storage of the data lake ✔
- [ ] **B.** Use AWS Kinesis Streams for storage of the data lake
- [ ] **C.** Use AWS Kinesis Firehose to stream the various files ✔
- [ ] **D.** Use AWS Kinesis streams

Explanation :

Answer – A and C
The AWS Documentation mentions the following
Amazon Kinesis Data Firehose (https://aws.amazon.com/kinesis/data-firehose/) is the easiest way to capture and stream data into a data lake built on Amazon S3. This data can be anything—from AWS service logs like AWS CloudTrail log files, Amazon VPC Flow Logs, Application Load Balancer logs, and others. It can also be IoT events, game events, and much more. To efficiently query this data, a time-consuming ETL (extract, transform, and load) process is required to massage and convert the data to an optimal file format, which increases the time to insight. This situation is less than ideal, especially for real-time data that loses its value over time.
Option B is incorrect since this option should not be used for persistence of data
Option D is incorrect since this would be less appropriate than Kinesis Firehose. Kinesis Firehose can automatically take the ingestion of data from multiple sources and then directly ingest the data into S3.
For more information on a use case for this, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/ (https://aws.amazon.com/blogs/big-data/analyzing-apache-parquet-optimized-data-using-amazon-kinesis-data-firehose-amazon-athena-and-amazon-redshift/)

Ask our Experts

QUESTION 55 UNATTEMPTED COLLECTION

A company is currently planning on using Redshift to host their data warehouse. Different departments have submitted their files for uploading to various S3 buckets. You need to ensure all the data files are uploaded efficiently to the cluster with the least maintenance overhead. Which of the following method would you incorporate for this scenario?

- ○ **A.** Use a manifest file for the COPY command ✔
- ○ **B.** Ensure all the buckets are made public
- ○ **C.** Copy all the files to a central S3 bucket

○ **D.** Ensure versioning is enabled for the bucket

Explanation :

Answer – A
This is given in the AWS Documentation

########

Using a Manifest to Specify Data Files

You can use a manifest to ensure that the COPY command loads all of the required files, and only the required files, for a data load. Instead of supplying an object path for the COPY command, you supply the name of a JSON-formatted text file that explicitly lists the files to be loaded. The URL in the manifest must specify the bucket name and full object path for the file, not just a prefix. You can use a manifest to load files from different buckets or files that do not share the same prefix. The following example shows the JSON to load files from different buckets and with file names that begin with date stamps.

```
{
 "entries": [
  {"url":"s3://mybucket-alpha/2013-10-04-custdata", "mandatory":true},
  {"url":"s3://mybucket-alpha/2013-10-05-custdata", "mandatory":true},
  {"url":"s3://mybucket-beta/2013-10-04-custdata", "mandatory":true},
  {"url":"s3://mybucket-beta/2013-10-05-custdata", "mandatory":true}
 ]
}
```

########
Option B is incorrect since this is not a requirement and can also pose to be a security issue
Option C is incorrect since this would be in inefficient process
Option D is incorrect since this is not a requirement for the COPY process

For more information on using manifest files, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/loading-data-files-using-manifest.html (https://docs.aws.amazon.com/redshift/latest/dg/loading-data-files-using-manifest.html)

Ask our Experts

QUESTION 56 UNATTEMPTED ANALYSIS

A company has started using the AWS machine learning service and is using Binary classification for its model. After the initial evaluation, the AUC is showing a score of 0.51. What does this indicate?

- A. The evaluation is highly inaccurate
- B. The evaluation is highly accurate
- C. The evaluation is more like taking a guess on the result ✔
- D. You should change the classification model

Explanation :

Answer – C
With the Binary classification model, the default cutoff between a positive or negative is 0.5. Any score close to 1 means that the model is accurate. Having a score of 0.5 is more like taking a guess on whether the result is positive or negative.

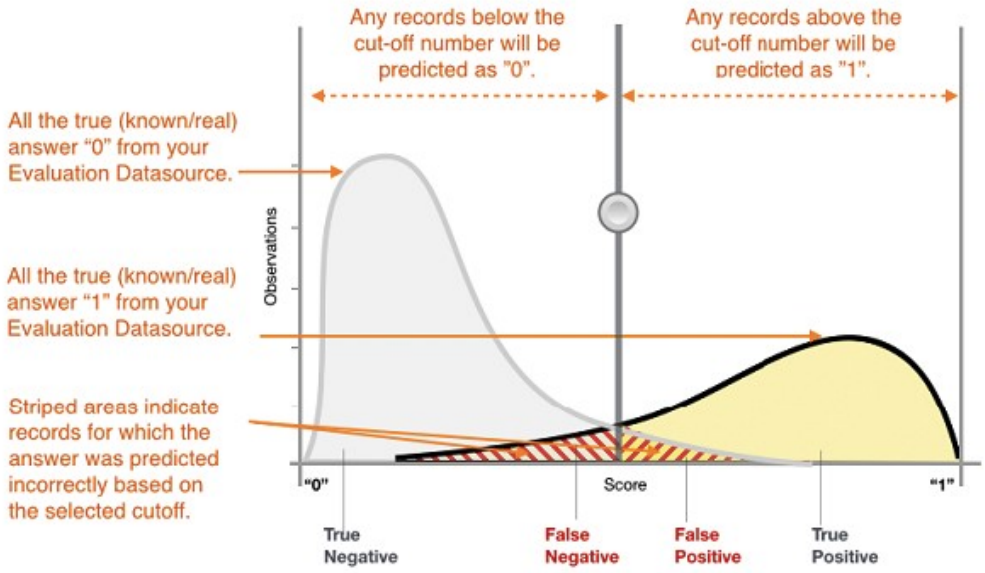


Figure 1: Score Distribution for a Binary Classification Model

For more information on binary classification, please refer to the below URL

- https://docs.aws.amazon.com/machine-learning/latest/dg/binary-classification.html (https://docs.aws.amazon.com/machine-learning/latest/dg/binary-classification.html)

Ask our Experts

QUESTION 57 UNATTEMPTED STORAGE

A company is making use of Kinesis streams for transferring data from various sources. The Consumers will run at different times depending on the priority of data retrieval. Most consumers run within the hour and there are some which run once in 2 days. Which of the following must be implemented on the stream to ensure all data gets processed from within the stream?

- A. Ensure that encryption is enabled on the stream
- B. Ensure that the data retention is changed for the stream ✔
- C. Ensure that Kinesis Firehose is attached to the stream
- D. Ensure that the consumer runs on an EC2 Instance

Explanation :

Answer – B
The AWS Documentation mentions the following
Amazon Kinesis Data Streams supports changes to the data record retention period of your stream. A Kinesis data stream is an ordered sequence of data records meant to be written to and read from in real time. Data records are therefore stored in shards in your stream temporarily. The time period from when a record is added to when it is no longer accessible is called the *retention period*. A Kinesis data stream stores records from 24 hours by default, up to 168 hours.
All of the other options are incorrect since these are not key requirements for ensuring that data gets processed from the stream.
For more information on the retention period, please refer to the below URL

- https://docs.aws.amazon.com/streams/latest/dev/kinesis-extended-retention.html (https://docs.aws.amazon.com/streams/latest/dev/kinesis-extended-retention.html)

Ask our Experts

QUESTION 58 UNATTEMPTED STORAGE

A company is planning on using a plethora of AWS services such as AWS RDS and Amazon Redshift. They need to have a unified metadata repository for all of these data sources. Which of the following is the ideal service to use for this purpose?

- A. AWS Athena
- B. AWS Glue ✔
- C. AWS EMR
- D. AWS Quick Sight

Explanation :

Answer – B
The AWS Documentation mentions the following
AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it simple and cost-effective to categorize your data, clean it, enrich it, and move it reliably between various data stores. The AWS Glue Data Catalog provides a unified metadata repository across a variety of data sources and data formats, integrating with Amazon EMR as well as Amazon RDS, Amazon Redshift, Redshift Spectrum, Athena, and any application compatible with the Apache Hive metastore. AWS Glue crawlers can automatically infer schema from source data in Amazon S3 and store the associated metadata in the Data Catalog.
All other options are incorrect because these cannot be used to catalog the information
For more information on EMR Spark Glue, please refer to the below URL

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

Ask our Experts

QUESTION 59 UNATTEMPTED STORAGE

Your company is going to create a table in a Redshift cluster. Below are the key characteristics for the table:

The data in the table don't change frequently.
There would be less than 10 millions rows.
The table would not have joins with other tables.
Which of the following distribution styles would be ideal for the table?

- A. All ✔
- B. Default
- C. EVEN
- D. Key

Explanation :

Answer – A
The AWS Documentation mentions the following
ALL distribution multiplies the storage required by the number of nodes in the cluster, and so it takes much longer to load, update, or insert data into multiple tables. ALL distribution is appropriate only for relatively slow moving tables; that is, tables that are not updated frequently or extensively. Small dimension tables do not benefit significantly from ALL distribution, because the cost of redistribution is low.
Since the concept of this distribution style is clearly mentioned in the documentation , all other options are incorrect
For more information on Distribution styles, please refer to the below URL

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

Ask our Experts

QUESTION 60 UNATTEMPTED STORAGE

A company is using the IoT service for storing data from multiple IoT devices. They want to ensure that they can get the state of a device even if it is disconnected from the IoT service. Which of the following can help achieve this?

- **A.** Using Device shadows ✔
- **B.** Using Rules
- **C.** Using Jobs
- **D.** Using Device Defender

Explanation :

Answer – A
The AWS Documentation mentions the following
A device's *shadow* is a JSON document that is used to store and retrieve current state information for a device. The Device Shadow service maintains a shadow for each device you connect to AWS IoT. You can use the shadow to get and set the state of a device over MQTT or HTTP, regardless of whether the device is connected to the Internet
Option B is incorrect since this is used for devices to Interact with other AWS services
Option C is incorrect since this is used to define a set of remote operations that are sent to and executed on one or more devices connected to AWS IoT
Option D is incorrect since this is used to audit the configuration of your devices, monitor connected devices to detect abnormal behavior, and to mitigate security risks
For more information on Device shadows, please refer to the below URL

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

The AWS Documentation mentions the following
You use an analysis to create and interact with visuals and stories. You can think of an analysis as a container for a set of related visuals and stories, for example ones that are all applicable to a given business goal or key performance indicator. You can use multiple data sets in an analysis, although any given visual can only use one of those data sets.
For more information on working with Analysis, please refer to the below URL

- https://docs.aws.amazon.com/quicksight/latest/user/working-with-analyses.html (https://docs.aws.amazon.com/quicksight/latest/user/working-with-analyses.html)

Ask our Experts

A company needs to stream logs using the AWS Kinesis Firehose service. They need to decide on a data store. The resulting files on the data store will be heavily queried over a week's period time and after that can be archived for future analysis. Which of the following would be the ideal steps to implement the data store? Please choose 2 correct Options.

- [ ] **A.** Ensure the destination for Kinesis Firehose is marked as S3 ✔
- [ ] **B.** Ensure the destination for Kinesis Firehose is marked as Redshift
- [ ] **C.** Create a Lifecycle policy for S3 to archive older files ✔
- [ ] **D.** Create a Job to move older data from the Redshift table

Explanation :

Answer – A and C

The ideal way is to use Lifecycle policies for S3. The AWS Documentation mentions the following on S3 Lifecycle policies

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

- Transition actions—Define when objects transition to another storage class (https://docs.aws.amazon.com/AmazonS3/latest/dev/storage-class-intro.html). For example, you might choose to transition objects to the STANDARD_IA storage class 30 days after you created them, or archive objects to the GLACIER storage class one year after creating them.

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

The lifecycle expiration costs depend on when you choose to expire objects

For more information on S3 Lifecycle management, please refer to the below URL

- https://docs.aws.amazon.com/AmazonS3/latest/dev/object-lifecycle-mgmt.html (https://docs.aws.amazon.com/AmazonS3/latest/dev/object-lifecycle-mgmt.html)

Ask our Experts

A company has been using DynamoDB tables for 6 months and it contains millions of rows of data. They now need to port the data onto a Redshift table for conducting analysis on historical data. Which of the following needs to be kept in mind when porting data from DynamoDB to Redshift. Choose 2 answers from the options given below

- [ ] **A.** Ensure to enable DynamoDB streams
- [ ] **B.** Ensure that empty attribute values in DynamoDB are properly treated ✔
- [ ] **C.** Ensure the data type matches between engines ✔
- [ ] **D.** Ensure that global tables are enabled in DynamoDB

Explanation :

Answer – B and C
The AWS Documentation mentions the following
Before you can load data from a DynamoDB table, you must first create an Amazon Redshift table to serve as the destination for the data. Keep in mind that you are copying data from a NoSQL environment into a SQL environment, and that there are certain rules in one environment that do not apply in the other. Here are some of the differences to consider:

- DynamoDB table names can contain up to 255 characters, including '.' (dot) and '-' (dash) characters, and are case-sensitive. Amazon Redshift table names are limited to 127 characters, cannot contain dots or dashes and are not case-sensitive. In addition, table names cannot conflict with any Amazon Redshift reserved words.
- DynamoDB does not support the SQL concept of NULL. You need to specify how Amazon Redshift interprets empty or blank attribute values in DynamoDB, treating them either as NULLs or as empty fields.
- DynamoDB data types do not correspond directly with those of Amazon Redshift. You need to ensure that each column in the Amazon Redshift table is of the correct data type and size to accommodate the data from DynamoDB.

Since this is clearly mentioned in the AWS documentation , the other options are invalid
For more information on Redshift for DynamoDB, please refer to the below URL

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

Ask our Experts

QUESTION 63 UNATTEMPTED STORAGE

A company is planning on using DynamoDB as a data store for their application.

Below are the data access patterns for the table

Data is uploaded to the table via an application
The data is heavily used within a week's time from the ingestion of data
After a week's time, the data is no longer used
Which of the following can be used to effectively store the table data in DynamoDB?
Choose 2 answers from the options given below

- [ ] **A.** Create tables based on a weekly basis. Ensure a high read and write capacity for these tables. ✔
- [ ] **B.** Change the Read and write capacity to a lower value after a week's time for the table ✔
- [ ] **C.** Create a table and ensure a datetimestamp is placed as the partition Key
- [ ] **D.** Perform scan operations on the table and delete table data which has an older datetimestamp of more than one week

Explanation :

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

########

Design Pattern for Time-Series Data

Consider a typical time-series scenario, where you want to track a high volume of events. Your write access pattern is that all the events being recorded have today's date. Your read access pattern might be to read today's events most frequently, yesterday's events much less frequently, and then older events very little at all.
The read access pattern is best handled by building the current date and time into the primary key. But that is certain to create one or more hot partitions. The latest one is always the *only* partition that is being written to. All other partitions, including all the partitions from previous days, divert provisioned write capacity from where you need it most.
The following design pattern often handles this kind of scenario effectively:

- Create one table per time period, provisioned with write capacity less than 1,000 write capacity units (WCUs) per partition-key value, and minimum necessary read capacity.
- Before the end of each time period, prebuild the table for the next period. Just as the current period ends, direct event traffic to the new table. You can assign names to these tables that specify the time periods that they have recorded.
- As soon as a table is no longer being written to, reduce its provisioned write capacity to 1 WCU and provision whatever read capacity is appropriate. Reduce the provisioned read capacity of earlier tables as they age, and archive or delete the ones whose contents will rarely or never be needed.

########
Options C and D even though possible are less efficient since we are using the scan operation to delete the data in the tables
For more information on time series data for DynamoDB, please refer to the below URL

- https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-time-series.html (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/bp-time-series.html)

Ask our Experts

QUESTION 64 UNATTEMPTED VISUALIZATION

A company's sales team is planning on using AWS Quicksight for their Visualization needs. They need a way to compare the values of current sales data against the forecasted sales data. Which of the following can be used in Quicksight for this purpose?

- A. Using KPI's ✔
- B. Using Pie Charts
- C. Using Pivot tables

○ **D.** Using Heat Maps

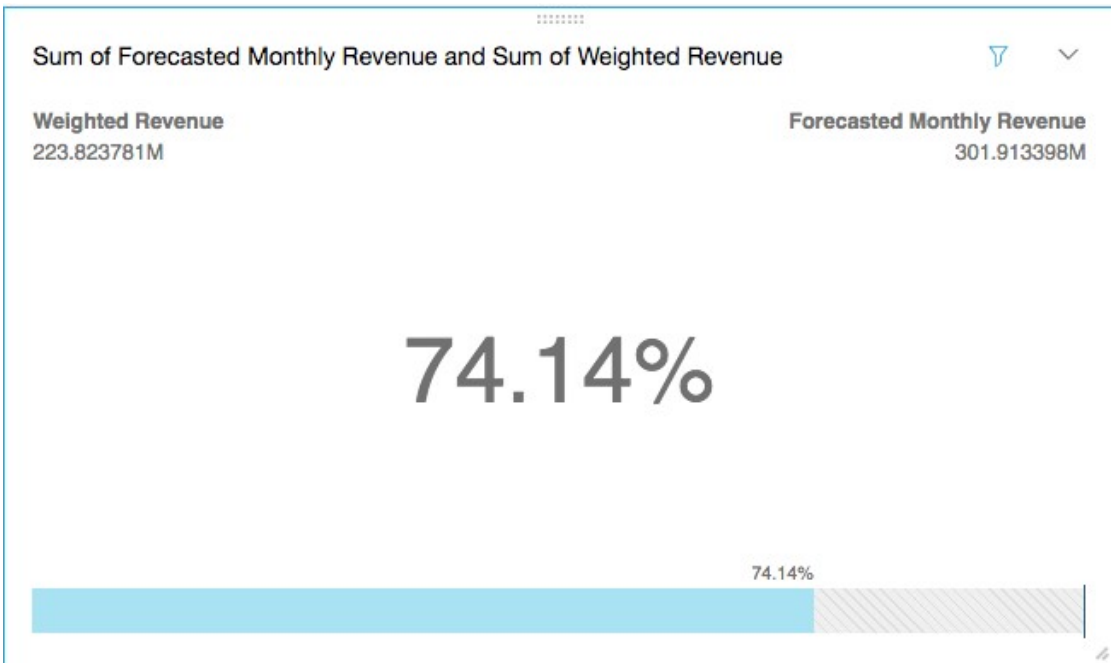
Explanation :

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

########

Using KPIs

Use a KPI to visualize a comparison between a key value and its target value.
A KPI displays a value comparison, the two values being compared, and a progress bar. For example, the following KPI shows how closely revenue is meeting its forecast.



########

Option B is incorrect since this is normally used to compare values for items in a dimension

Option C is incorrect since this is normally used to measure values for the intersection of two dimensions

Option D is incorrect since this is normally used to measure for the intersection of two dimensions, with color-coding to easily differentiate where values fall in the range

For more information on using KPI's in Quicksight, please refer to the below URL

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

Ask our Experts

QUESTION 65 INCORRECT STORAGE

A company is planning on using AWS Cloudsearch. They need to ensure that the data uploaded is searchable. Which of the following format's are allowed for the data to be uploaded into CloudSearch. Choose 2 answers from the options given below

- [ ] A. XML ✔
- [ ] B. JSON ✔
- [x] C. Parquet ✖
- [x] D. SerDe ✖

Explanation :

Answer – A and B
The AWS Documentation mentions the following
To make your data searchable, you need to format it in JSON or XML as described in Preparing Your Data (https://docs.aws.amazon.com/cloudsearch/latest/developerguide/preparing-data.html) and upload it to your search domain for indexing. In most cases, Amazon CloudSearch automatically indexes your data and the changes are visible in search results in just a few minutes. However, certain changes to your domain configuration put the domain in the NEEDS INDEXING state. For those changes to take effect, you must explicitly run indexing to rebuild your index. Currently, you also need to periodically run indexing so your suggesters reflect the most recent data in your index. The following sections describe how to upload data to your domain and run indexing when it's needed.
Since this is clearly mentioned in the documentation, all other options are invalid
For more information on Cloudsearch, please refer to the below URL

- https://docs.aws.amazon.com/cloudsearch/latest/developerguide/uploading-and-indexing-data.html (https://docs.aws.amazon.com/cloudsearch/latest/developerguide/uploading-and-indexing-data.html)

Ask our Experts

Finish Review (https://www.whizlabs.com/learn/course/aws-bds-practice-tests/quiz/14849)

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)