NEW PRACTICE TEST 2 - UPDATED

Attempt	1	**Completed on**	Sunday , 03 February 2019 , 11:43 PM
Marks Obtained	1 / 65	**Time Taken**	00 H 00 M 20 S
Your score is	1.54%	**Result**	Fail

Domains / Topics wise Quiz Performance Report

S.No.	Topic	Total Questions	Correct	Incorrect	Unattempted
1	Storage	15	0	0	15
2	Processing	12	0	0	12
3	Collection	11	0	0	11
4	Data Security	14	1	0	13
5	Analysis	9	0	0	9
6	Visualization	4	0	0	4

65 Questions	1 Correct	0 Incorrect	64 Unattempted

Show Answers: All

QUESTION 1 UNATTEMPTED STORAGE

A company is currently hosting a database in Amazon Redshift. Due to the new requirements in data that needs to be stored in the table, the following changes would occur on one or more of the tables

Changes to existing columns
Uploading of completely new data to the tables
It needs to be ensured that the Redshift database can be restored to its original state, in case there are any issues with the change process. Which of the following can help ensure this in the most efficient manner?

- **A.** Create a manual snapshot of the database ✔
- **B.** Create a copy of the last automated snapshot
- **C.** Use the UNLOAD command to unload all the data to S3
- **D.** Stream all the data to AWS Kinesis as a backup

Explanation :

Answer – A
Snapshots are point-in-time backups of a cluster. There are two types of snapshots: *automated* and *manual*. Amazon Redshift stores these snapshots internally in Amazon S3 by using an encrypted Secure Sockets Layer (SSL) connection. If you need to restore from a snapshot, Amazon Redshift creates a new cluster and imports data from the snapshot that you specify.
It's better to take a manual snapshot which would contain the latest data. This should be done just before the change process
Option B is partially correct, but the automated snapshot may not be the most recent copy of your data
Option C is incorrect since this would be a very inefficient way to manage the backup and restore operation
Option D is incorrect since AWS Kinesis should not be used as a data store
For more information on working with snapshots, please refer to the below URL

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

Ask our Experts

QUESTION 2 UNATTEMPTED PROCESSING

Your company is planning on hosting a set of EMR clusters for the purposes of Machine Learning and ad-hoc query analysis. All the data would be stored on Amazon S3. The underlying instance types needs to be setup for the clusters. Which of the following would you recommend?

- **A.** For the Machine Learning cluster – T and for adhoc query analysis – R
- **B.** For the Machine Learning cluster – T and for adhoc query analysis – G2
- **C.** For the Machine Learning cluster – M and for adhoc query analysis – I
- **D.** For the Machine Learning cluster – C and for adhoc query analysis – R ✔

Explanation :

Answer - D

The C type Instances are good for Machine Learning purposes
High performance web servers, scientific modelling, batch processing, distributed analytics, high-performance computing (HPC), machine/deep learning inference, ad serving, highly scalable multiplayer gaming, and video encoding.
The R Instance types are good for ad-hoc work since they have the capacity for burstable performances
For more information on EC2 Instance types, please refer to the below URL

- https://aws.amazon.com/ec2/instance-types/ (https://aws.amazon.com/ec2/instance-types/)

Ask our Experts

QUESTION 3 UNATTEMPTED COLLECTION

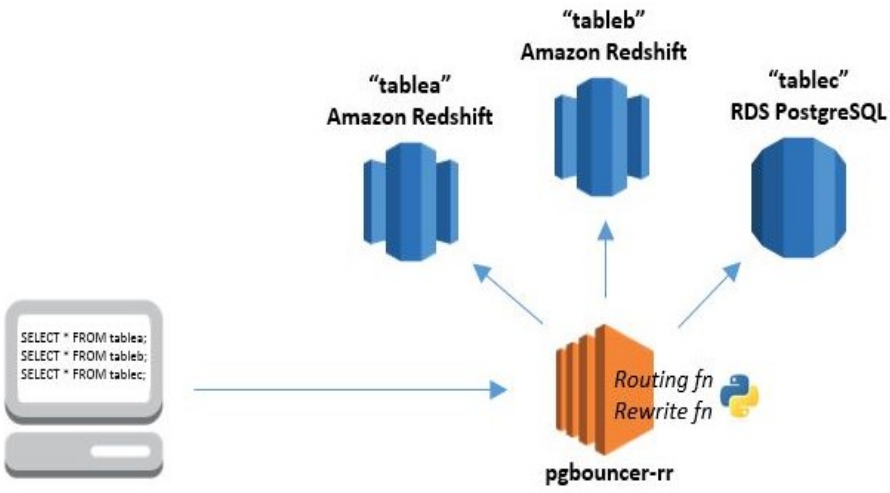
A company currently has a database cluster setup in Redshift. They also have an AWS RDS PostgreSQL database setup in place. A table has been setup in PostgreSQL which stores data based on a timestamp. The requirement now is to ensure that the data from the PostgreSQL table gets stored into the Redshift database. For this a staging table has been setup in Redshift. It needs to be ensured that the data lag between the staging and PostgreSQL tables is not greater than 4 hours. Which of the following is the most efficient implementation step you would use for this requirement?

- **A.** Create a trigger in the PostgreSQL table to send new data to a Kinesis stream. Ensure the data is transferred from the Kinesis Stream to the staging table in Redshift.
- **B.** Create a SQL query that is run every hour to check for new data. Use the query results to send the new data to the staging table.
- **C.** Use the extensions available in PostgreSQL and use the dblink facility ✔
- **D.** Create a trigger in the PostgreSQL table to send new data to a Kinesis Firehose stream. Ensure the data is transferred from the Kinesis Firehose Stream to the staging table in Redshift

Explanation :

Answer – C
Such a use case is clearly mentioned in the documentation

########



For further reading, check out Bob Strahan's blog post Query Routing and Rewrite: Introducing pgbouncer-rr for Amazon Redshift and PostgreSQL (https://blogs.aws.amazon.com/bigdata/post/Tx3G7177U6YHY5I/Query-Routing-and-Rewrite-Introducing-pgbouncer-rr-for-Amazon-Redshift-and-Postg) post.

RDS PostgreSQL includes two extensions to execute queries remotely. The first extension is the PostgreSQL foreign-data wrapper, postgres_fdw. The postgres_fdw module enables the creation of external tables. External tables can be queried in the same way as a local native table, However, the query is not currently executed entirely on the remote side because postgres_fdw doesn't push down aggregate functions and limit clauses. When you perform an aggregate query through an external table, all the data is pulled into PostgreSQL for an aggregation step. This is unacceptably slow for any meaningful number of rows.
The second extension is dblink, which includes a function also called dblink. The dblink function allows the entire query to be pushed to Amazon Redshift. This lets Amazon Redshift do what it does best—query large quantities of data efficiently and return the results to PostgreSQL for further processing.
########
Since this is clearly mentioned in the documentation, all other options are invalid.
For more information on this use case, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/join-amazon-redshift-and-amazon-rds-postgresql-with-dblink/ (https://aws.amazon.com/blogs/big-data/join-amazon-redshift-and-amazon-rds-postgresql-with-dblink/)

Ask our Experts

QUESTION 4 UNATTEMPTED DATA SECURITY

A company stores data in an S3 bucket. Some of the data contains sensitive information. They need to ensure that the bucket complies with PCI DSS compliance standards. Which of the following should be implemented to fulfil this requirement? Choose 2 answers from the options given below

- [ ] **A.** Enable server-side encryption for the bucket ✔
- [ ] **B.** Enable versioning for the bucket
- [ ] **C.** Ensure that access to the bucket is only given to one IAM Role
- [ ] **D.** Ensure that objects from the bucket are requested only via HTTPS ✔

Explanation :

Answer – A and D
Here the main compliance should be to ensure there is encryption of objects at rest and in transit.
Hence one can achieve this with server-side encryption and ensuring that all communication with the S3 bucket is done via HTTPS
For more information on the compliance with AWS, please refer to the below URL

- https://aws.amazon.com/compliance/pci-dss-level-1-faqs/ (https://aws.amazon.com/compliance/pci-dss-level-1-faqs/)

Ask our Experts

QUESTION 5 UNATTEMPTED COLLECTION

A company has developed a platform which consists of multiple components. First is the ability for suppliers to supply data to the platform. And another application component which allows them to perform analysis on that data. The suppliers authenticate using SAML federation. Which of the following will help ensure a complete solution and also ensure a secure way for suppliers to have access to only their set of data?

- **A.** Ingest the data using Kinesis streams. Move the data from Kinesis streams into DynamoDB. Ensure that access is given via STS using temporary access credentials. ✔
- **B.** Ingest the data using Kinesis streams. Move the data from Kinesis streams into DynamoDB. Ensure that access is given via AWS Access Keys.
- **C.** Ingest the data using DynamoDB streams. Stream the data into DynamoDB tables. Ensure that access is given via STS using temporary access credentials.
- **D.** Ingest the data using DynamoDB streams. Stream the data into DynamoDB tables. Ensure that access is given via AWS Access Keys.

Explanation :

Answer – A
The AWS Documentation mentions the following
The AWS Security Token Service (STS) is a web service that enables you to request temporary, limited-privilege credentials for AWS Identity and Access Management (IAM) users or for users that you authenticate (federated users). This guide provides descriptions of the STS API.
Option B is incorrect using AWS Access is not the right approach to giving access
Options C and D are incorrect since DynamoDB streams is not used to ingest data from external sources
For more information on STS and Streams Adaptor for AWS Kinesis , please refer to the below URL

- https://docs.aws.amazon.com/STS/latest/APIReference/Welcome.html (https://docs.aws.amazon.com/STS/latest/APIReference/Welcome.html)
- https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.KCLAdapter.Walkthrough.h (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Streams.KCLAdapter.Walkthrough.htn

Ask our Experts

QUESTION 6 UNATTEMPTED ANALYSIS

A company has an EMR Cluster on AWS. They have some hive job flows running on the EMR cluster. The users need to be able to create dashboards and analyse data in the EMR cluster. Which of the following would be the ideal one to use for this requirement?

- A. Presto
- B. Microstrategy ✔
- C. R Studio
- D. Hue

Explanation :

Answer – B
The AWS Documentation mentions the following
You can use popular business intelligence tools like Microsoft Excel, MicroStrategy, QlikView, and Tableau with Amazon EMR to explore and visualize your data. Many of these tools require an ODBC (Open Database Connectivity) or JDBC (Java Database Connectivity) driver.
Options A and C can be used to perform SQL queries not help in creating BI solutions.
Option D is incorrect since this is the web interface for the EMR cluster
For more information on the EMR BI tools , please refer to the below URL

- https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-bi-tools.html (https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-bi-tools.html)

Ask our Experts

QUESTION 7 UNATTEMPTED STORAGE

A company is planning on hosting their data warehousing solution in Redshift. They are trying to decide on the distribution style for their underlying tables. Which of the following reasons would warrant the use of EVEN distribution style for the underlying tables in Redshift. Choose 2 answers from the options given below

- A. When a table does not participate in joins ✔
- B. When the table participates in multiple joins
- C. When the tables design is new and there is no clear distinction on how the data will be organized ✔
- D. When there is a requirement for queries for keys to be co-located.

Explanation :

Answer – A and C
The AWS Documentation mentions the following
#######
Distribution Styles
When you create a table, you can designate one of three distribution styles; EVEN, KEY, or ALL.
Even distribution
The leader node distributes the rows across the slices in a round-robin fashion, regardless of the values in any particular column. EVEN distribution is appropriate when a table does not participate in joins or when there is not a clear choice between KEY distribution and ALL distribution.
#######
Since this is clearly mentioned in the documentation, the other options are invalid
For more information on choosing the right distribution style , 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 8 UNATTEMPTED STORAGE

A company is planning to load data into a Redshift table from multiple files in an S3 bucket. They want to verify that the data was loaded correctly. How can they verify this?

- A. Check the Cloudwatch log metrics to check if the data was loaded properly
- B. Query the STL_LOAD_COMMITS table ✔
- C. Check the Cloudtrail log metrics to check if the data was loaded properly
- D. Use the Analyze command

Explanation :

Answer – B
The AWS Documentation mentions the following
#######
Verifying That the Data Was Loaded Correctly
After the load operation is complete, query the STL_LOAD_COMMITS (https://docs.aws.amazon.com/redshift/latest/dg/r_STL_LOAD_COMMITS.html) system table to verify that the expected files were loaded. You should execute the COPY command and load verification within the same transaction so that if there is problem with the load you can roll back the entire transaction.
The following query returns entries for loading the tables in the TICKIT database:

```
select query, trim(filename) as filename, curtime, status
from stl_load_commits
where filename like '%tickit%' order by query;

 query |           btrim           |          curtime           | status
-------+---------------------------+----------------------------+--------
 22475 | tickit/allusers_pipe.txt  | 2013-02-08 20:58:23.274186 |      1
 22478 | tickit/venue_pipe.txt     | 2013-02-08 20:58:25.070604 |      1
 22480 | tickit/category_pipe.txt  | 2013-02-08 20:58:27.333472 |      1
 22482 | tickit/date2008_pipe.txt  | 2013-02-08 20:58:28.608305 |      1
 22485 | tickit/allevents_pipe.txt | 2013-02-08 20:58:29.99489  |      1
 22487 | tickit/listings_pipe.txt  | 2013-02-08 20:58:37.632939 |      1
 22489 | tickit/sales_tab.txt      | 2013-02-08 20:58:37.632939 |      1
(6 rows)
```

#######

Option A is incorrect since here metrics about the cluster itself will be recorded and the not the data load process

Option C is incorrect since here the API activity for the service will be recorded and the not the data load process

Option D is incorrect since this is used to update the statistical metadata that the query planner uses to build and choose optimal plan

For more information on verifying whether the data was loaded properly, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/dg/verifying-that-data-loaded-correctly.html (https://docs.aws.amazon.com/redshift/latest/dg/verifying-that-data-loaded-correctly.html)

Ask our Experts

QUESTION 9 UNATTEMPTED COLLECTION

A company is planning on using the Kinesis agent to transfer several logs files from multiple EC2 Instances. There are several log files from various locations that need to be streamed from the servers to Kinesis. Which of the following do you need to configure to ensure that the data gets streamed accordingly?

- **A.** Ensure that the flows section is modified in the agent configuration file ✔
- **B.** Ensure that the KCL is configured to read the different files
- **C.** Ensure to use the Enhanced fan out feature
- **D.** Change the data retention period on the stream

Explanation :

Answer – A

The AWS Documentation mentions the following

########

Monitor Multiple File Directories and Write to Multiple Streams

By specifying multiple flow configuration settings, you can configure the agent to monitor multiple file directories and send data to multiple streams. In the following configuration example, the agent monitors two file directories and sends data to an Kinesis stream and a Kinesis Data Firehose delivery stream respectively. Note that you can specify different endpoints for Kinesis Data Streams and Kinesis Data Firehose so that your Kinesis stream and Kinesis Data Firehose delivery stream don't need to be in the same region.

```
{
  "cloudwatch.emitMetrics": true,
  "kinesis.endpoint": "https://your/kinesis/endpoint",
  "firehose.endpoint": "https://your/firehose/endpoint",
  "flows": [
    {
      "filePattern": "/tmp/app1.log*",
      "kinesisStream": "yourkinesisstream"
    },
    {
      "filePattern": "/tmp/app2.log*",
      "deliveryStream": "yourfirehosedeliverystream"
    }
  ]
}
```

########

Option B is incorrect since the KCL will be used to read data from the streams and not the files

Option C is incorrect since this is used on the consumer side
Option D is incorrect since there is no mention on how long it would be required to maintain the records
For more information on using the Kinesis agent, 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 10 UNATTEMPTED ANALYSIS

A team has a 10 PB data store in EMR. All the data is stored in S3. There is a need for a data engineer to perform interactive analysis. The data engineer already has access to the EMR cluster via the AWS Console. Which of the following can be used by the data engineer for interactive analysis?

- **A.** Presto ✔
- **B.** Apache sqoop
- **C.** Oozie
- **D.** Apache Pig

Explanation :

Answer – A
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 B is incorrect since this is a tool for transferring data between Amazon S3, Hadoop, HDFS, and RDBMS databases
Option C is incorrect since this is used to manage and coordinate Hadoop jobs
Option D is incorrect since this is used to transform large data sets without having to write complex code in a lower level computer language like Java
For more information on AWS Presto, please refer to the below URL

- 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 11 UNATTEMPTED COLLECTION

A company wants to start using AWS Redshift for storing their existing On-premise data warehousing solution. They need to transfer around 10 PB of data from their on-premise environment to AWS. Which of the following would be the ideal solution to transfer the data onto AWS?

- A. Use the AWS Redshift COPY command
- B. Use the AWS Snowball device ✔
- C. Use AWS VPN connections
- D. Use AWS Kinesis

Explanation :

Answer – B
The AWS Documentation mentions the following
AWS provides multiple ways to move data from your datacenter to AWS. To establish a dedicated network connection between your network and AWS, you can use AWS Direct Connect (https://aws.amazon.com/directconnect/). To move petabytes to exabytes of data to AWS using physical appliances, you can use AWS Snowball (https://aws.amazon.com/snowball/) and AWS Snowmobile. (https://aws.amazon.com/snowmobile/?nc2=h_m1) To have your on-premises applications store data directly into AWS, you can use AWS Storage Gateway (https://docs.aws.amazon.com/storagegateway/latest/userguide/WhatIsStorageGateway.html).
Option A is incorrect since this should only used once the data is in AWS
Option C is incorrect since this would not be an adequate enough connection to transfer data onto AWS.
Option D is incorrect since streaming the data would be inconvenient.
For more information on data lakes and analytics, please refer to the below URL

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

Ask our Experts

QUESTION 12 UNATTEMPTED DATA SECURITY

A company is planning on using AWS Kinesis for streaming data from various sources. One of the key requirements is to ensure that data is encrypted at rest in AWS Kinesis? How would you accomplish this?

- A. Use client-side encryption for the data at the consumer side
- B. Use client-side encryption for the data at the Producer side ✔

- C. Create an MD5 hash for the partition key and send it along with the stream
- D. Use VPC Endpoints

Explanation :

Answer – B
From the options given above, the only logical option is to ensure the data is encrypted by the Producer before it reaches the Kinesis Stream endpoint.
Option A is incorrect since here the data would remain unencrypted in the stream
Option C is incorrect since this can be used for encryption
Option D is incorrect since this is used to securely access the Kinesis stream from an EC2 Instance in a private subnet
For more information on using AWS KMS to encrypt data on the client side, please refer to the below URL

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

Ask our Experts

QUESTION 13 UNATTEMPTED DATA SECURITY

You currently have a Redshift cluster which is using KMS key for encryption in a region. You now need to ensure that the cluster can be moved to another region. Which of the following steps would you put as part of the implementation. Choose 2 answers from the options given below

- A. Ensure that the KMS key is copied from the source region to the destination region
- B. Ensure that a new KMS key is created in the new region ✔
- C. Configure a snapshot copy grant for a master key in the destination region ✔
- D. Configure a snapshot copy grant for a master key in the source region

Explanation :

Answer – B and C
The AWS Documentation mentions the following
When you launch an Amazon Redshift cluster, you can choose to encrypt it with a master key from the AWS Key Management Service (AWS KMS). AWS KMS keys are specific to a region. If you want to enable cross-region snapshot copy for an AWS KMS-encrypted cluster, you must configure

a *snapshot copy grant* for a master key in the destination region so that Amazon Redshift can perform encryption operations in the destination region. The following procedure describes the process of enabling cross-region snapshot copy for an AWS KMS-encrypted cluster.
Since this is clearly mentioned in the AWS documentation , the other options are invalid
For more information on using snapshots copy with KMS keys, please refer to the below URL

- https://docs.aws.amazon.com/redshift/latest/mgmt/managing-snapshots-console.html#xregioncopy-kms-encrypted-snapshot (https://docs.aws.amazon.com/redshift/latest/mgmt/managing-snapshots-console.html#xregioncopy-kms-encrypted-snapshot)

Ask our Experts

QUESTION 14 UNATTEMPTED STORAGE

A company is planning on hosting their data warehousing solution in Redshift. They are trying to decide on the distribution style for their underlying tables. Which of the following reasons would warrant the use of KEY distribution style for the underlying tables in Redshift.

- **A.** When traffic between nodes needs to be reduced during the join operation ✔
- **B.** When data for the entire table needs to be placed on each node
- **C.** When you need the rows across the node slices in a round-robin fashion
- **D.** When a table does not participate in join

Explanation :

Answer – A
The AWS Documentation mentions the following
The rows are distributed according to the values in one column. The leader node will attempt to place matching values on the same node slice. If you distribute a pair of tables on the joining keys, the leader node collocates the rows on the slices according to the values in the joining columns so that matching values from the common columns are physically stored together.
Option B is incorrect because this belongs to the ALL distribution style
Options C and D are incorrect because this belongs to the Even distribution style
For more information on the 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 15 UNATTEMPTED DATA SECURITY

There is a Redshift table that is storing information for the employees. The Manager of the department needs to be able to see the information for the employees only in their respective department. How can you accomplish this in Redshift without the need for too much maintenance overheads?

- A. Create a view in Redshift that can be used by the managers. ✔
- B. Create separate tables in Redshift based on the Managers ID.
- C. Create separate clusters in Redshift based on the Managers ID.
- D. Create a work management queue

Explanation :

Answer – A
Options B and C are incorrect since this would add more maintenance overhead to the solution
Option D is incorrect since this is used to define multiple query queues and to route queries to the appropriate queues at runtime
You can use views to hide and show only relevant information
The AWS Documentation mentions the following
To create a standard view, you need access to the underlying tables. To query a standard view, you need select privileges for the view itself, but you don't need select privileges for the underlying tables. To query a late binding view, you need select privileges for the late binding view itself. You should also make sure the owner of the late binding view has select privileges to the referenced objects (tables, views, or user-defined functions).
Options B and C are incorrect since this would add more maintenance overhead to the solution
Option D is incorrect since this is used to define multiple query queues and to route queries to the appropriate queues at runtime
For more information on creating views, please refer to the below URL

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

Ask our Experts

QUESTION 16 UNATTEMPTED DATA SECURITY

You have created a platform wherein companies can place their data in Redshift tables for their order history. There are 5 major players which take up 80% of the data hosted and a small number of other players. You need to ensure that the resources

for querying is distributed properly for the big players since they take up a major percentage of the data on Redshift. How can you achieve this without much additional maintenance overhead?

- **A.** Create a separate user group and make use of Workload Management ✔
- **B.** Segregate the clusters based on the major players
- **C.** Make sure the tables are created based on the EVEN distribution style
- **D.** Make sure the tables are created based on the ALL distribution style

Explanation :

Answer – A
The AWS Documentation mentions the following
Amazon Redshift workload management (WLM) enables users to flexibly manage priorities within workloads so that short, fast-running queries won't get stuck in queues behind long-running queries. Amazon Redshift WLM creates query queues at runtime according to *service classes*, which define the configuration parameters for various types of queues, including internal system queues and user-accessible queues. From a user perspective, a user-accessible service class and a queue are functionally equivalent. For consistency, this documentation uses the term *queue* to mean a user-accessible service class as well as a runtime queue.
Option B is incorrect since this would add a maintenance overhead for the entire solution
Options C and D are incorrect since the distribution styles would assist with ensuring performance for the major players

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

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

Ask our Experts

QUESTION 17 UNATTEMPTED DATA SECURITY

Users are uploaded unencrypted files to different locations in an S3 bucket. You then ensure the files are transferred to a central S3 location. How can you ensure the contents are the same for the object after the files have been transferred to the central S3 bucket?

- **A.** Use the E Tags associated with the objects ✔
- **B.** Use the CompareObjects API command
- **C.** Compare the size of the objects

○ **D.** Use the object key for comparison

Explanation :

Answer – A
You can actually compare the ETag to ensure that the MD5 digest of the object data is the same.
#######
ETag
The entity tag is a hash of the object. The ETag reflects changes only to the contents of an object, not its metadata. The ETag may or may not be an MD5 digest of the object data. Whether or not it is depends on how the object was created and how it is encrypted as described below:

- Objects created by the PUT Object, POST Object, or Copy operation, or through the AWS Management Console, and are encrypted by SSE-S3 or plaintext, have ETags that are an MD5 digest of their object data.
- Objects created by the PUT Object, POST Object, or Copy operation, or through the AWS Management Console, and are encrypted by SSE-C or SSE-KMS, have ETags that are not an MD5 digest of their object data.
- If an object is created by either the Multipart Upload or Part Copy operation, the ETag is not an MD5 digest, regardless of the method of encryption.

Type: String
#######
Option B is incorrect since the API command is not applicable or present for S3
Options C and D are incorrect since the size and key name cannot be solely used to ensure the contents have not been tampered with
For more information on REST response headers, please refer to the below URL

- https://docs.aws.amazon.com/AmazonS3/latest/API/RESTCommonResponseHeaders.html (https://docs.aws.amazon.com/AmazonS3/latest/API/RESTCommonResponseHeaders.html)

Ask our Experts

QUESTION 18 UNATTEMPTED DATA SECURITY

A company has a large number of datasets that are being sent over to S3 for storage. They want their Data science team to query the data using the AWS Athena service. There is the additional requirement for ensuring the data is encrypted at rest. When it comes to the encryption, the following are the key requirements

Custom Keys are used for encryption for one central data set
Generic Encryption is used for all other data sets
Which of the following would you use to fulfil this requirement? Choose 2 answers

from the options given below.

- [ ] **A.** S3 Server-side for the one central data set with S3 Managed Keys
- [ ] **B.** S3 Server-side Encryption for the one central data set with KMS Managed Keys ✔
- [ ] **C.** S3 Server-side Encryption for the one central data set with client-side keys
- [ ] **D.** S3 client-side Encryption for the one central data set with client-side keys
- [ ] **E.** S3 Server-side for the other data sets with S3 Managed Keys ✔
- [ ] **F.** S3 Server-side for the other data sets with client-side Managed Keys

Explanation :

Answer – B and E
The AWS Documentation mentions the following
Athena supports the following Amazon S3 encryption options, both for encrypted datasets in Amazon S3 and for encrypted query results:

- Server side encryption with an Amazon S3-managed key (SSE-S3 (https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingServerSideEncryption.html))
- Server-side encryption with a AWS KMS-managed key (SSE-KMS (https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingKMSEncryption.html)).
- Client-side encryption with a AWS KMS-managed key (CSE-KMS (https://docs.aws.amazon.com/AmazonS3/latest/dev/UsingClientSideEncryption.html#client-side-encryption-kms-managed-master-key-intro))

Option A is incorrect since there is a requirement for custom keys to be used for the central data set
Options C ,D and F are incorrect since Athena does not support either option
For more information on encryption with Athena, please refer to the below URL

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

Ask our Experts

QUESTION 19 UNATTEMPTED PROCESSING

A company has decided to start using AWS Athena for querying data in S3. The amount of data is huge and they need to create queries based on a subset of data. How can they accomplish this in the easiest manner?

- ○ **A.** Split the files in S3
- ○ **B.** Use views in Athena ✔

- C. Create different buckets in S3
- D. Create different types of charts in Athena

Explanation :

Answer B
The AWS Documentation mentions the following
You may want to create views to:

- *Query a subset of data*. For example, you can create a table with a subset of columns from the original table to simplify querying data.
- *Combine multiple tables in one query*. When you have multiple tables and want to combine them with UNION ALL, you can create a view with that expression to simplify queries against the combined tables.

Options A and C are incorrect since this would cause a maintenance overhead
Option D is incorrect since here we need to optimize the query and not the visualization
For more information on when to use views, please refer to the below URL

- https://docs.aws.amazon.com/athena/latest/ug/when-to-use-views.html (https://docs.aws.amazon.com/athena/latest/ug/when-to-use-views.html)

Ask our Experts

QUESTION 20 UNATTEMPTED STORAGE

A company currently is using DynamoDB as their data store. A new application is being introduced wherein the data items vary in size from 1MB to 10 MB. What are the options available for the company for the storage of the data for the new application in the most efficient manner. Choose 2 answers from the options given below

- [x] A. Compress the data item if possible ✔
- [ ] B. Split the data item and store the parts as multiple items in the table
- [x] C. Store the items in S3 and place a link as an attribute in DynamoDB ✔
- [ ] D. Store the items in Redshift and place a link as an attribute in DynamoDB

Explanation :

Answer – A and C
The AWS Documentation mentions the following

Amazon DynamoDB currently limits the size of each item that you store in a table (see Limits in DynamoDB (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Limits.html)). If your application needs to store more data in an item than the DynamoDB size limit permits, you can try compressing one or more large attributes, or you can store them as an object in Amazon Simple Storage Service (Amazon S3) and store the Amazon S3 object identifier in your DynamoDB item.
Option B is incorrect since this would add a maintenance overhead
Option D is incorrect since there are no clear requirements to use Redshift for data storage
For more information on using S3 with DynamoDB, please refer to the below URL

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

Ask our Experts

QUESTION 21 UNATTEMPTED STORAGE

There is a requirement for a data store. It needs to be highly durable and available. Initially the data will be around 10 TB. The data will continuously grow. The data will consist of both structured and un-structed data. Which of the following would be the ideal data store combinations for these requirements?

- A. AWS S3 and DynamoDB ✔
- B. AWS Redshift and Kinesis Streams
- C. AWS Aurora and Kinesis Streams
- D. AWS S3 and Kinesis Stream firehose

Explanation :

Answer – A
The AWS Documentation mentions the following
As mentioned above, you can also take advantage of Amazon Simple Storage Service (Amazon S3) to store large attribute values that cannot fit in a DynamoDB item. You can store them as an object in Amazon S3 and then store the object identifier in your DynamoDB item.
You can also use the object metadata support in Amazon S3 to provide a link back to the parent item in DynamoDB. Store the primary key value of the item as Amazon S3 metadata of the object in Amazon S3. Doing this often helps with maintenance of the Amazon S3 objects.
Option B is incorrect since Redshift is not an ideal case for unstructured data and Kinesis streams is used for streaming and not data persistence
Option C is incorrect since AWS Aurora is not an ideal case for unstructured data and Kinesis streams is used for streaming and not data persistence
Option D is incorrect since there is no requirement for streaming of data
For more information on using S3 with DynamoDB, please refer to the below URL

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

Ask our Experts

QUESTION 22 UNATTEMPTED STORAGE

1. A company wants to have a data store on AWS that can be used to store a large number of log files. Apart from this the service should have the following capabilities

Range searches
Term boosting
Faceting

Which of the below can be used as the data store?

- **A.** AWS S3
- **B.** AWS CloudSearch ✔
- **C.** AWS Redshift
- **D.** AWS DynamoDB

Explanation :

Answer - B
The AWS Documentation mentions the following
You can use Amazon CloudSearch to index and search both structured data and plain text. Amazon CloudSearch features:

- Full text search with language-specific text processing
- Boolean search
- Prefix searches
- Range searches
- Term boosting
- Faceting
- Highlighting
- Autocomplete Suggestions

Since the documentation clearly gives the features of this product , all other options are incorrect

For more information on what is cloudsearch, please refer to the below URL

- https://docs.aws.amazon.com/cloudsearch/latest/developerguide/what-is-cloudsearch.html (https://docs.aws.amazon.com/cloudsearch/latest/developerguide/what-is-cloudsearch.html)

Ask our Experts

QUESTION 23 UNATTEMPTED STORAGE

Currently a company is using AWS DynamoDB since they want a serverless architecture and they don't want to manage the underlying infrastructure. They use the DynamoDB service to ingest and store large amounts of metric related data. They now want to start storing a large number of log files which can be used for log analytics. Which of the following can be used for the storage of the log files?

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

Explanation :

Answer – B
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.
All other options are invalid since the best way to work with search capabilities is to use AWS ElasticSearch
For more information on what is Elasticsearch, please refer to the below URL

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

Ask our Experts

QUESTION 24 UNATTEMPTED STORAGE

A company needs to start using a Redshift cluster. They have 10 GB of data. They don't know the cluster size that should be used for the cluster. They have not done any testing to come to the recommended cluster size. Which of the following would be the ideal way to provision the Redshift cluster in a cost-efficient manner?

- **A.** Create the cluster with a large number of cluster nodes
- **B.** Create the cluster with a small number of cluster nodes ✔
- **C.** Insist on doing the testing to come up with the base requirement for the number of clusters
- **D.** Plan on purchasing Reserved capacity

Explanation :

Answer – B
To be cost efficient you can start with a minimum number of nodes for the cluster
The AWS Documentation mentions the following
If your storage and performance needs change after you initially provision your cluster, you can resize your cluster. You can scale the cluster in or out by adding or removing nodes. Additionally, you can scale the cluster up or down by specifying a different node type.
For example, you can add more nodes, change node types, change a single-node cluster to a multinode cluster, or change a multinode cluster to a single-node cluster. However, you must ensure that the resulting cluster is large enough to hold the data that you currently have or else the resize will fail.
Option A is incorrect since this would not be a cost-effective option
Option C is incorrect since , even though possible , it's is sometimes difficult to come up with a benchmark, that is why you can have a flexible infrastructure setup with the service
Option D is incorrect since we don't the requirement , we cannot plan to buy reserved instances at the start
For more information on working with clusters, please refer to the below URL

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

Ask our Experts

QUESTION 25 UNATTEMPTED STORAGE

A company is planning on creating an EMR cluster for their Big Data needs. To make use of the maximum available space, they want to use Amazon S3 as the underlying data store for the cluster. Which of the following should also be enabled to ensure the cluster always works with the most recent updated data in S3?

- A. Using Hive along with the EMR Cluster
- B. Enable consistent view ✔
- C. Enable Hue on the cluster
- D. Use HDFS as the storage layer

Explanation :

Answer – B
The AWS Documentation mentions the following
EMRFS consistent view is an optional feature available when using Amazon EMR release version 3.2.1 or later. Consistent view allows EMR clusters to check for list and read-after-write consistency for Amazon S3 objects written by or synced with EMRFS. Consistent view addresses an issue that can arise due to the Amazon S3 Data Consistency Model (https://docs.aws.amazon.com/AmazonS3/latest/dev/Introduction.html#ConsistencyModel). For example, if you add objects to Amazon S3 in one operation and then immediately list objects in a subsequent operation, the list and the set of objects processed may be incomplete. This is more commonly a problem for clusters that run quick, sequential steps using Amazon S3 as a data store, such as multi-step extract-transform-load (ETL) data processing pipelines.
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 a web interface for the EMR Cluster
Option D is incorrect since the question clearly states that the underlying storage layer should be S3
For more information on the EMRFS consistent view, please refer to the below URL

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

Ask our Experts

QUESTION 26 UNATTEMPTED PROCESSING

A development team has been instructed to use Kinesis Streams to stream data from various web-based applications. They also need to develop the consumer-based components. Which of the following is used to specify the location from where you can start reading records from a Kinesis stream?

- A. ShardLoc
- B. ShardIterator ✔
- C. ShardNumber
- D. StreamLoc

Explanation :

Answer - B
The AWS Documentation mentions the following
Specify a shard iterator using the ShardIterator parameter. The shard iterator specifies the position in the shard from which you want to start reading data records sequentially. If there are no records available in the portion of the shard that the iterator points to, GetRecords returns an empty list. It might take multiple calls to get to a portion of the shard that contains records.
Since this is clearly mentioned in the AWS Documentation , all other options are incorrect
For more information on the GetRecords command, please refer to the below URL

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

Ask our Experts

QUESTION 27 UNATTEMPTED ANALYSIS

A company wants to start making use of AWS Machine learning. The main aspect of the program would be to determine if the reviews written on their web site was done by a customer or a bot. Which of the following classifications in ML would you make use of?

- A. Single class
- B. Multiclass
- C. Binary ✔
- D. Regression

Explanation :

Answer - C
The AWS Documentation mentions the following
You can use supervised ML approaches for these specific machine learning tasks: binary classification (predicting one of two possible outcomes), multiclass classification (predicting one of more than two outcomes) and regression (predicting a numeric value).
Examples of binary classification problems:

- Will the customer buy this product or not buy this product?
- Is this email spam or not spam?
- Is this product a book or a farm animal?
- Is this review written by a customer or a robot?

For more information on the different classifications, please refer to the below URL

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

Ask our Experts

QUESTION 28 UNATTEMPTED PROCESSING

A company has been using Amazon Redshift and loaded a number of tables in the cluster. After a series of operations over a couple of months, the performance of the queries seems to be deteriorating. Which of the following is an ideal approach to improve the performance of the queries?

- **A.** Carry out the VACUUM command ✔
- **B.** Carry out the COPY command on the table again
- **C.** Enable compression on the columns
- **D.** Disable compression on the columns

Explanation :

Answer – A
The AWS Documentation mentions the following
Amazon Redshift does not automatically reclaim and reuse space that is freed when you delete rows and update rows. To perform an update, Amazon Redshift deletes the original row and appends the updated row, so every update is effectively a delete followed by an insert. When you perform a delete, the rows are marked for deletion, but not removed. The query processor needs to scan the deleted rows as well as undeleted rows, so too many deleted rows can cost unnecessary processing. You should vacuum following a significant number of deletes or updates to reclaim space and improve query performance.
Option B is incorrect since you will first need to get the latest data and then perform the COPY command. And why do this when you can use the VACCUM command to improve performance
Option C is incorrect since this is done when the table is created
Option D is incorrect since this is do with the size and not the performance of queries
For more information on reclaiming storage in Redshift, 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 29 UNATTEMPTED PROCESSING

A company wants to start using Kinesis streams for their ingestion of data. They want their development team to spend less effort when it comes to developing components that send data to the streams. Which of the following would be ideal for such a use case?

- A. Use the KPL library ✔
- B. Use the Kinesis API
- C. Use AWS Lambda
- D. Use AWS SQS

Explanation :

Answer - A
The AWS Documentation mentions the following
The KPL is an easy-to-use, highly configurable library that helps you write to a Kinesis data stream. It acts as an intermediary between your producer application code and the Kinesis Data Streams API actions. The KPL performs the following primary tasks:

- Writes to one or more Kinesis data streams with an automatic and configurable retry mechanism
- Collects records and uses PutRecords to write multiple records to multiple shards per request
- Aggregates user records to increase payload size and improve throughput
- Integrates seamlessly with the Kinesis Client Library (https://docs.aws.amazon.com/kinesis/latest/dev/developing-consumers-with-kcl.html) (KCL) to de-aggregate batched records on the consumer
- Submits Amazon CloudWatch metrics on your behalf to provide visibility into producer performance

Option B is partially correct but the KPL library gives you a lot of benefits
Option C is incorrect since AWS Lambda is normally used at the consumer side
Option D is incorrect since this is a messaging service
For more information on the KPL 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 30 UNATTEMPTED COLLECTION

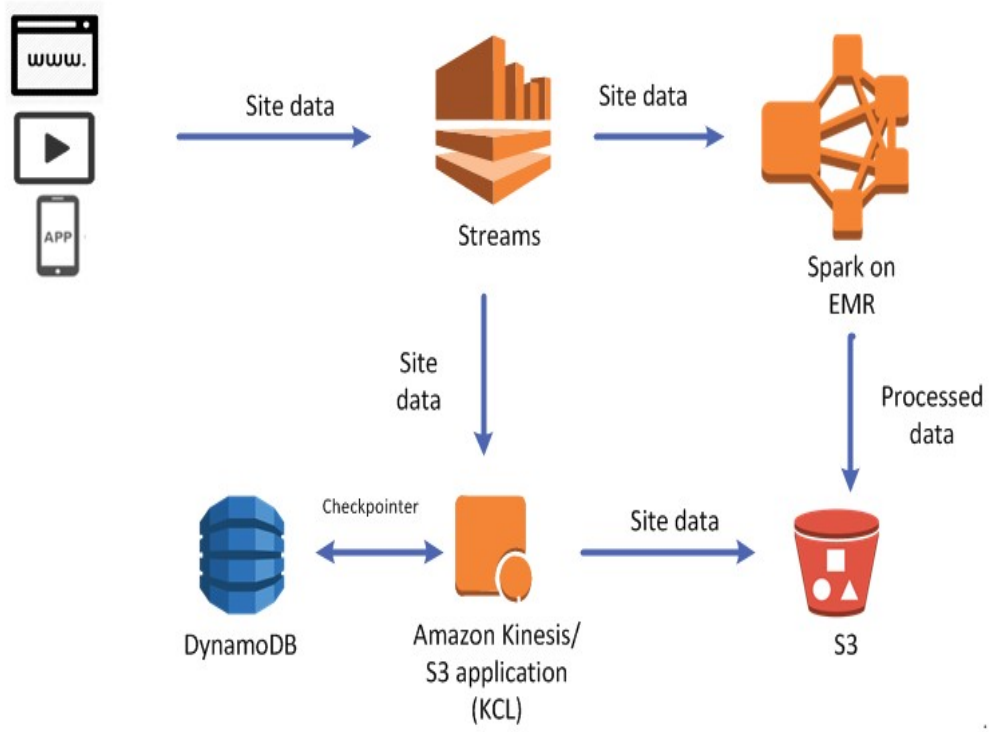
A company wants to have a data store for their log files from various EC2 Instances. These log files need to be streamed from the various servers and then stored for analysis at a later stage. Which of the following can be used for this requirement?

- **A.** Ingest the data using AWS SQS and then store the data in DynamoDB
- **B.** Ingest the data using AWS SQS and then store the data in S3
- **C.** Ingest the data using AWS Kinesis and then store the data in S3 ✔
- **D.** Ingest the data using AWS Kinesis and then store the data in DynamoDB

Explanation :

Answer – C

One example is given in the AWS Blog site. The architecture is given below. Remember you can connect your Kinesis stream to Firehose to directly store the data into S3.



Options A and B are incorrect since here AWS Kinesis is better for ingestion of data
Option D is incorrect since S3 would be better for storage of log files
For more information on this use case scenario, please refer to the below URL

- https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and-aws-lambda/ (https://aws.amazon.com/blogs/big-data/persist-streaming-data-to-amazon-s3-using-amazon-kinesis-firehose-and-aws-lambda/)

Ask our Experts

QUESTION 31 UNATTEMPTED ANALYSIS

A company currently has a Hadoop cluster setup on top of AWS EMR. They now have a requirement to carry out some Machine Learning algorithms on the existing data. Which of the following can be used on top of Hadoop for this purpose?

- **A.** HBase
- **B.** Livy
- **C.** Zeppelin

○ **D.** Mahout ✔

Explanation :

Answer – D
The AWS Documentation mentions the following
Mahout is a machine learning library with tools for clustering, classification, and several types of recommenders, including tools to calculate most-similar items or build item recommendations for users. Mahout employs the Hadoop framework to distribute calculations across a cluster, and now includes additional work distribution methods, including Spark.
Option A is incorrect since this is used as an open source, non-relational, distributed database developed as part of the Apache Software Foundation's Hadoop project
Option B is incorrect since this is used for interaction over a REST interface with an EMR cluster running Spark
Option C is incorrect since this is used as a notebook for interactive data exploration
For more information on Mahout, please refer to the below URL

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

Ask our Experts

QUESTION 32 UNATTEMPTED PROCESSING

A company is planning on using AWS Kinesis streams for streaming large amounts of data. They are going to be using Lambda functions to process the data from the streams. Which of the following can be used to ensure maximum efficiency when it comes to processing of records in the streams. Choose 3 answers from the options given below?

☐ **A.** Memory for AWS Lambda ✔

☐ **B.** Batch size for AWS Lambda ✔

☐ **C.** Number of shards for the Kinesis stream ✔

☐ **D.** Number of shard iterators for the Kinesis stream

Explanation :

Answer – A,B and C
There is a good article which lists down ways in which you can get better performance of your shards and processing data with AWS Lambda.
For more information on this article, please refer to the below URL

- https://tech.trivago.com/2018/07/13/aws-kinesis-with-lambdas-lessons-learned/ (https://tech.trivago.com/2018/07/13/aws-kinesis-with-lambdas-lessons-learned/)

Ask our Experts

QUESTION 33 UNATTEMPTED STORAGE

A company is planning on setting up an EMR cluster in AWS. They need to store the metadata in a central repository. How can they achieve this? Choose 2 answers from the options give below

- [ ] **A.** Create a bucket in S3
- [ ] **B.** Create a MySQL database in AWS RDS ✔
- [ ] **C.** Modify the hiveConfiguration.json file and reference it when you create the cluster. ✔
- [ ] **D.** Modify the Hive setup on the cluster

Explanation :

Answer – B and C
The AWS Documentation mentions the following
To use an external MySQL database or Amazon Aurora as your Hive metastore, you override the default configuration values for the metastore in Hive to specify the external database location, either on an Amazon RDS MySQL instance or an Amazon Aurora instance.
By default, Hive records metastore information in a MySQL database on the master node's file system. But 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 RDS or Amazon Aurora.

So to use an external MySQL database as your Hive metastore, you override the default configuration values for the metastore in Hive to specify the external database locationon an Amazon RDS MySQL instance.
Since the way to accomplish this is clearly given in the documentation , all other options are incorrect.
For more information on how to store the Hive metadata, please refer to the below URL

- https://aws.amazon.com/premiumsupport/knowledge-center/export-metastore-from-emr-to-rds/ (https://aws.amazon.com/premiumsupport/knowledge-center/export-metastore-from-emr-to-rds/)

Ask our Experts

QUESTION 34 UNATTEMPTED VISUALIZATION

A Business Intelligence team have stored their data in S3 and are currently also using AWS Athena. They want to connect their business tools to AWS Athena. How can they accomplish this? Choose 2 answers from the options given below

- [x] **A.** Use JDBC drivers ✔
- [x] **B.** Use ODBC Driver ✔
- [ ] **C.** Use AWS S3 drivers
- [ ] **D.** Use Athena drivers

Explanation :

Answer – A and B
This is given in the AWS Documentation
########

Connecting to Amazon Athena with ODBC and JDBC Drivers

To explore and visualize your data with business intelligence tools, download, install, and configure an ODBC (Open Database Connectivity) or JDBC (Java Database Connectivity) driver.

Topics

- Using Athena with the JDBC Driver (https://docs.aws.amazon.com/athena/latest/ug/connect-with-jdbc.html)
- Connecting to Amazon Athena with ODBC (https://docs.aws.amazon.com/athena/latest/ug/connect-with-odbc.html)

########

Since this is clearly mentioned in the documentation, all other options are invalid
For more information on using BI tools with AWS Athena, please refer to the below URL

- https://docs.aws.amazon.com/athena/latest/ug/athena-bi-tools-jdbc-odbc.html (https://docs.aws.amazon.com/athena/latest/ug/athena-bi-tools-jdbc-odbc.html)

Ask our Experts

QUESTION 35 UNATTEMPTED DATA SECURITY

You need to establish a secure backup and archiving solution for your company, using AWS. Documents should be immediately accessible for three months and available for five years for compliance reasons. Which AWS service fulfills these requirements in the most cost-effective way?

- **A.** Upload data to S3 and use lifecycle policies to move the data into Glacier for long-term archiving. ✔
- **B.** Upload the data on EBS, use lifecycle policies to move EBS snapshots into S3 and later into Glacier for long-term archiving.
- **C.** Use Direct Connect to upload data to S3 and use IAM policies to move the data into Glacier for long-term archiving.
- **D.** Use Storage Gateway to store data to S3 and use lifecycle policies to move the data into Redshift for long-term archiving.

Explanation :

Answer – A
Amazon Glacier is a secure, durable, and extremely low-cost cloud storage service for data archiving and long-term backup. Customers can reliably store large or small amounts of data for as little as $0.004 per gigabyte per month, a significant savings compared to on-premises solutions.
With Amazon lifecycle policies you can create transition actions in which you define when objects transition to another Amazon S3 storage class. For example, you may choose to transition objects to the STANDARD_IA (IA, for infrequent access) storage class 30 days after creation, or archive objects to the GLACIER storage class one year after creation.
Option B is invalid because lifecycle policies are not available for EBS volumes
Option C is invalid because IAM policies cannot be used to move data to Glacier
Option D is invalid because lifecycle policies is not used to move data to Redshif
For more information on S3 lifecycle policies, please visit the url

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

Ask our Experts

QUESTION 36 UNATTEMPTED STORAGE

A company needs to develop a system for a hospital. The application needs to ingest the heart rate recorded for various patients. The requirements for the application are:

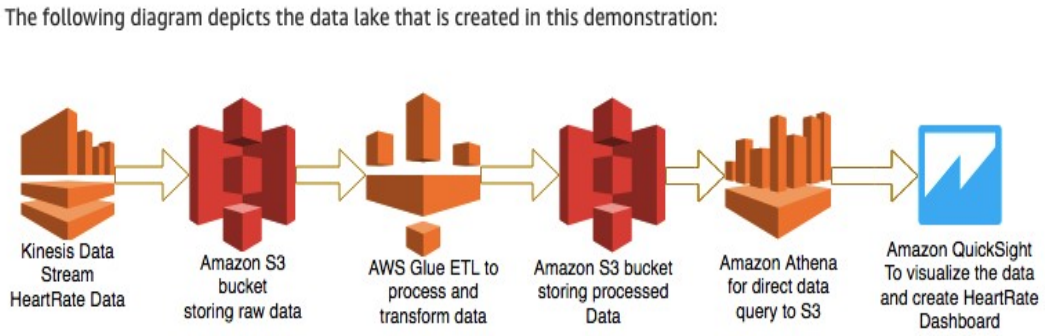
A data lake that can expand on demand to store the heart rate information.
A way to ingest the data and store it in the data lake.
A way to catalogue the information
Which of the following would you use for this requirement?

- **A.** AWS Redshift as the data lake, AWS SQS to ingest data and AWS Glue to catalogue the information
- **B.** AWS S3 as the data lake, AWS SQS to ingest data and AWS Glue to catalogue the information
- **C.** AWS S3 as the data lake, AWS Kinesis to ingest data and AWS Glue to catalogue the information ✔
- **D.** AWS S3 as the data lake, AWS Kinesis to ingest data and AWS Athena to catalogue the information

Explanation :

Answer – C
An example of this architecture is given in the AWS Documentation
#######
How to build a front-line concussion monitoring system using AWS IoT and serverless data lakes – Part 2
In part 1 of this series (https://aws.amazon.com/blogs/big-data/how-to-build-a-front-line-concussion-monitoring-system-using-aws-iot-and-serverless-data-lakes-part-1/), we demonstrated how to build a data pipeline in support of a data lake. We used key AWS services such as Amazon Kinesis Data Streams (https://aws.amazon.com/kinesis/data-streams/), Kinesis Data Analytics (https://aws.amazon.com/kinesis/data-analytics/), Kinesis Data Firehose (https://aws.amazon.com/kinesis/data-firehose/), and AWS Lambda (https://aws.amazon.com/lambda/). In part 2, we discuss how to process and visualize the data by creating a serverless data lake that uses key analytics to create actionable data.
Create a serverless data lake and explore data using AWS Glue, Amazon Athena, and Amazon QuickSight
As we discussed in part 1 (https://aws.amazon.com/blogs/big-data/how-to-build-a-front-line-concussion-monitoring-system-using-aws-iot-and-serverless-data-lakes-part-1/), you can store heart rate data in an Amazon S3 (https://aws.amazon.com/s3/) bucket using Kinesis Data Streams.

However, storing data in a repository is not enough. You also need to be able to catalog and store the associated metadata related to your repository so that you can extract the meaningful pieces for analytics.
For a serverless data lake, you can use AWS Glue (https://aws.amazon.com/glue/), which is a fully managed data catalog and ETL (extract, transform, and load) service. AWS Glue simplifies and automates the difficult and time-consuming tasks of data discovery, conversion, and job scheduling. As you get your AWS Glue Data Catalog data partitioned and compressed for optimal performance, you can use Amazon Athena (https://aws.amazon.com/athena/) for the direct query to S3 data. You can then visualize the data using Amazon QuickSight (https://aws.amazon.com/quicksight/).
The following diagram depicts the data lake that is created in this demonstration:



#######
Options A and B are incorrect since SQS would not be the ideal service to ingest the information
Option D is incorrect since AWS Athena is a querying tool
For more information on this use case, please visit the url

- https://aws.amazon.com/blogs/big-data/how-to-build-a-front-line-concussion-monitoring-system-using-aws-iot-and-serverless-data-lakes-part-2/ (https://aws.amazon.com/blogs/big-data/how-to-build-a-front-line-concussion-monitoring-system-using-aws-iot-and-serverless-data-lakes-part-2/)

Ask our Experts

QUESTION 37 UNATTEMPTED PROCESSING

A company has a large number of spark jobs that need to run on an EMR cluster in AWS. They want some way to orchestrate the series of jobs. Which of the following can be used for this purpose?

- A. AWS Step Functions ✔
- B. AWS SQS
- C. Apache Hive
- D. Apache Pig

Explanation :

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

########

What if you have a simple use case, in which you want to run a few Spark jobs in a specific order, but you don't want to spend time orchestrating those jobs or maintaining a separate application? You can do that today in a serverless fashion using AWS Step Functions (https://aws.amazon.com/step-functions/). You can create the entire workflow in AWS Step Functions and interact with Spark on Amazon EMR (https://aws.amazon.com/emr/) through Apache Livy (https://livy.incubator.apache.org/).
########
Option B is invalid since this is queue-based service
Option C is invalid since this is an open-source, data warehouse, and analytic package that runs on top of a Hadoop cluster
Option D is invalid since this is an open-source Apache library that runs on top of Hadoop, providing a scripting language that you can use to transform large data sets without having to write complex code in a lower level computer language like Java.
For more information on this use case, please visit the url

- https://aws.amazon.com/blogs/big-data/orchestrate-apache-spark-applications-using-aws-step-functions-and-apache-livy/ (https://aws.amazon.com/blogs/big-data/orchestrate-apache-spark-applications-using-aws-step-functions-and-apache-livy/)

Ask our Experts

QUESTION 38 UNATTEMPTED STORAGE

A company needs to process large amounts of data and store them in a data store accordingly. The data consists of all IP addresses which are accessing their website. There would be around billions of rows being stored in the data store. The company have decided to use the AWS EMR service. The company needs to be able to query the data efficiently based on the IP address. Which of the following would be an ideal implementation plan for this?

- **A.** Use S3 as the underlying storage for the EMR cluster. Ensure a bucket is created for each IP address
- **B.** Make use of HBase on EMR. Ensure that the IP address is used as the underlying key ✔

- **C.** Use S3 as the underlying storage for the EMR cluster. Ensure that the prefixes have the IP address attached such as bucketname/IPaddress-filename
- **D.** Post the data from EMR to Redshift for analysis

Explanation :

Answer – B
The best way to handle so much of data is to make use of HBase
The AWS Blog mentions the following
HBase offers a number of powerful features including:

- Strictly consistent reads and writes.
- High write throughput.
- Automatic sharding of tables.
- Efficient storage of sparse data.
- Low-latency data access via in-memory operations.
- Direct input and output to Hadoop jobs.
- Integration with Apache Hive for SQL-like queries over HBase tables, joins, and JDBC support.

Options A and C would be less efficient than using HBase.
Option D would just be an overhead.
For more information on Hbase on EMR, please visit the url

- https://aws.amazon.com/blogs/aws/apache-hbase-on-emr/ (https://aws.amazon.com/blogs/aws/apache-hbase-on-emr/)

Ask our Experts

QUESTION 39 UNATTEMPTED COLLECTION

A company is planning on using AWS Redshift as their data store. They have a lot of files that are going to be dropped into AWS S3 by different departments. They want to have the ability to automate the way the files get loaded into Redshift. How can they accomplish this in an efficient and cost-effective manner?

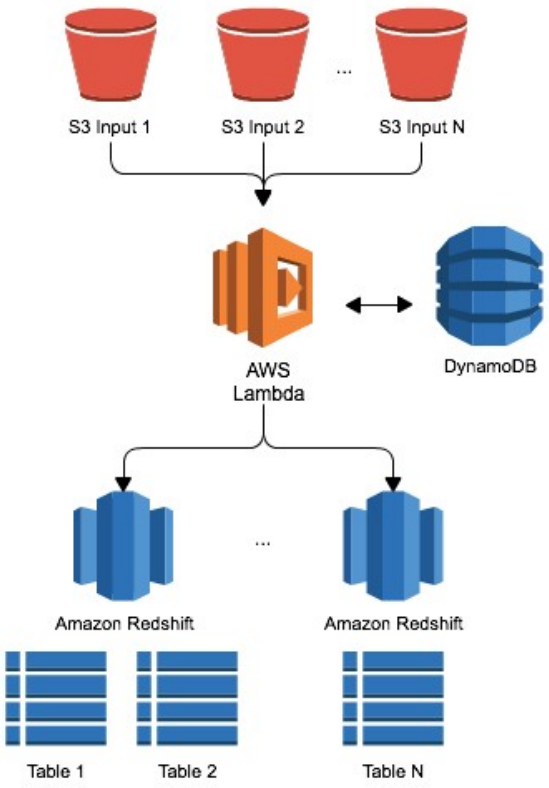
- **A.** Create a cron job on an EC2 Instance to poll the S3 buckets and drop the content onto AWS Redshift
- **B.** Use S3 events to invoke Lambda functions that will transfer the files to AWS Redshift.
- **C.** Use AWS Redshift triggers to poll the S3 buckets and drop the content onto its tables

D. Use AWS S3 events to call SQS and then use the queues to drop the content onto its tables

Explanation :

Answer – B

An example of this is given in the AWS Blog sites. Here you can use AWS Lambda code available in a Gut repository which can be used to automatically transfer files from various S3 buckets to AWS Redshift



Option A is incorrect since this does not cost efficient

Option C is incorrect since there are no triggers as of yet in Redshift

Option D is incorrect since there SQS would be ineffective in this scenario

For more information on this use case scenario, please visit the url

- https://aws.amazon.com/blogs/big-data/a-zero-administration-amazon-redshift-database-loader/ (https://aws.amazon.com/blogs/big-data/a-zero-administration-amazon-redshift-database-loader/)

Ask our Experts

QUESTION 40 UNATTEMPTED COLLECTION

A company wants to store millions of messages being sent from their IoT devices. They need to have a durable and highly available data store in AWS to store the incoming messages. It needs to be ensured that the Infrastructure for the data store can scale without much intervention. Which of the following would be the ideal implementation steps? Choose 2 answers from the options given below

- [ ] **A.** Use IoT Device shadows to store the messages from the devices
- [ ] **B.** Create DynamoDB tables to hold the messages from the IoT devices ✔
- [ ] **C.** Create IoT rules to transfer the data onto DynamoDB ✔
- [ ] **D.** Persist the IoT device shadows to AWS Redshift

Explanation :

Answer – B and C

You can use DynamoDB for storing the messages as this is a high scalable service. Here you don't need to manage the underlying infrastructure for scalability

Then you can use IoT rules to transfer the messages to DynamoDB as shown below

########

To create a DynamoDB rule

1. In the AWS IoT console (https://console.aws.amazon.com/iot/home), in the left navigation pane, choose **Act**.

Option A is incorrect since this is used to maintain the state of the device
Option D is incorrect since AWS DynamoDB would be a better store based on the scalability requirements
For more information on creating a rule for IoT and DynamoDB, please visit the url

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

Ask our Experts

QUESTION 41 UNATTEMPTED PROCESSING

A company has decided to use AWS for their Big Data processing needs. Their first assignment has the following requirements

Use Apache Hadoop to process web server logs
The logs then need to be cleansed

Finally, the logs need to be delivered to Amazon S3

Which of the following can be used to orchestrate this process?

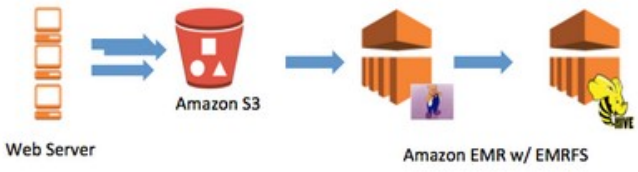
- A. AWS Step Functions
- B. AWS DataPipeline ✔
- C. AWS SQS
- D. AWS Lambda

Explanation :

Answer – B
Such a use case is given in the AWS Documentation
########

ETL Processing Using AWS Data Pipeline and Amazon Elastic MapReduce

This blog post shows you how to build an ETL workflow that uses AWS Data Pipeline to schedule an Amazon Elastic MapReduce (Amazon EMR) cluster to clean and process web server logs stored in an Amazon Simple Storage Service (Amazon S3) bucket. AWS Data Pipeline (http://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/what-is-datapipeline.html) is an ETL service that you can use to automate the movement and transformation of data. It launches an Amazon EMR cluster for each scheduled interval, submits jobs as steps to the cluster, and terminates the cluster after tasks have completed.
In this post, you'll create the following ETL workflow:



########
Option A is incorrect because this is used when you have Lambda functions to orchestrate
Option C is incorrect because this is a queue-based service
Option D is incorrect because this is a serverless compute-based service
For more information on this use case, please visit the url

- https://aws.amazon.com/blogs/big-data/etl-processing-using-aws-data-pipeline-and-amazon-elastic-mapreduce/ (https://aws.amazon.com/blogs/big-data/etl-processing-using-aws-data-pipeline-and-amazon-elastic-mapreduce/)

Ask our Experts

QUESTION 42 UNATTEMPTED DATA SECURITY

A development team is working on an application that will allow customers to view shipment orders placed on their site. The data would be stored in DynamoDB. The application itself would be available on both desktop and mobile based applications. The developers want to use a service which would simplify the authentication process for the application. Which of the following could be used for this purpose?

- A. AWS Cognito ✔
- B. AWS SAML
- C. AWS IAM Access Keys
- D. AWS IAM users

Explanation :

Answer – A
The AWS Documentation mentions the following
Amazon Cognito provides authentication, authorization, and user management for your web and mobile apps. Your users can sign in directly with a user name and password, or through a third party such as Facebook, Amazon, or Google.
The two main components of Amazon Cognito are user pools and identity pools. User pools are user directories that provide sign-up and sign-in options for your app users. Identity pools enable you to grant your users access to other AWS services. You can use identity pools and user pools separately or together.
Option B is incorrect since this should ideally be used for Federated access
Options C and D are incorrect since this is insecure and would result in a large maintenance overhead
For more information on AWS Cognito, please visit the url

- https://docs.aws.amazon.com/cognito/latest/developerguide/what-is-amazon-cognito.html (https://docs.aws.amazon.com/cognito/latest/developerguide/what-is-amazon-cognito.html)

Ask our Experts

QUESTION 43 UNATTEMPTED DATA SECURITY

A company is planning on using AWS Kinesis firehose to stream their log files onto S3. They need to ensure that source data stream for firehorse is encrypted. How can they achieve this? Choose 2 options

- [ ] **A.** Use Kinesis data streams as a source. ✔
- [ ] **B.** Enable encryption at rest in Kinesis Firehose ✔
- [ ] **C.** Enable server-side encryption for S3 using AWS Managed Keys
- [ ] **D.** Enable server-side encryption for S3 using KMS Keys

Explanation :

Answer – A and B

The AWS Documentation mentions the following

If you have sensitive data, you can enable server-side data encryption when you use Amazon Kinesis Data Firehose. But you can only do this if you use a Kinesis data stream as your data source. When you configure a Kinesis data stream as the data source of a Kinesis Data Firehose delivery stream, Kinesis Data Firehose no longer stores the data at rest. Instead, the data is stored in the data stream.

When you send data from your data producers to your Kinesis data stream, Kinesis Data Streams encrypts your data using an AWS KMS key before storing it at rest. When your Kinesis Data Firehose delivery stream reads the data from your Kinesis stream, Kinesis Data Streams first decrypts the data and then sends it to Kinesis Data Firehose.

Using Server-Side Encryption with Amazon Kinesis Data Firehose

If you have sensitive data, you can enable server-side data encryption when you use Amazon Kinesis Data Firehose. How you do this depends on the source of your data.

Server-Side Encryption with Kinesis Data Streams as the Data Source

When you configure a Kinesis data stream as the data source of a Kinesis Data Firehose delivery stream, Kinesis Data Firehose no longer stores the data at rest. Instead, the data is stored in the data stream.

When you send data from your data producers to your data stream, Kinesis Data Streams encrypts your data using an AWS Key Management Service (AWS KMS) key before storing the data at rest. When your Kinesis Data Firehose delivery stream reads the data from your data stream, Kinesis Data Streams first decrypts the data and then sends it to Kinesis Data Firehose. Kinesis Data Firehose buffers the data in memory based on the buffering hints that you specify. It then delivers it to your destinations without storing the unencrypted data at rest.

For information about how to enable server-side encryption for Kinesis Data Streams, see Using Server-Side Encryption in the *Amazon Kinesis Data Streams Developer Guide*.

Server-Side Encryption with Direct PUT or Other Data Sources

If you send data to your delivery stream using PutRecord or PutRecordBatch, or if you send the data using AWS IoT, Amazon CloudWatch Logs, or CloudWatch Events, you can turn on server-side encryption by using the StartDeliveryStreamEncryption operation.

Options C and D are incorrect since the encryption needs to be done at the stream level as per the requirement

For more information on Firehose encryption, please visit the url

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

Ask our Experts

QUESTION 44 UNATTEMPTED STORAGE

A company wants to stream their log files from their EC2 Instances. You are using Kinesis streams and Firehose for this process. The data will be parsed using AWS Lambda and then the resultant data will be stored in AWS Redshift. After the process is complete the amount of data in S3 has increased, and you had to delete the data manually. Since this process will be triggered on a continual basis, you need to ensure the right step is taken to delete the data in S3. How can you accomplish this?

- **A.** Create a Lifecycle policy for the S3 bucket ✔
- **B.** Use Redshift triggers to delete the data after the data has finished loading
- **C.** Use S3 events to delete the data after the data has finished loading
- **D.** Disable S3 logging since this is causing the increase in data

Explanation :

Answer - A
The AWS Documentation mentions the following
Kinesis Data Firehose delivers your data to your S3 bucket first and then issues an Amazon Redshift COPY command to load the data into your Amazon Redshift cluster. Specify an S3 bucket that you own where the streaming data should be delivered. Create a new S3 bucket or choose an existing bucket that you own.
Kinesis Data Firehose doesn't delete the data from your S3 bucket after loading it to your Amazon Redshift cluster. You can manage the data in your S3 bucket using a lifecycle configuration.
Option B is incorrect because triggers don't exist in Redshift
Option C is incorrect sinceS3 events would need to call a Lambda function and cannot conduct the clean-up automatically
Option D is incorrect since the data build is because the intermediate data from the Kinesis firehose process is stored in S3 first
For more information on using various destinations with Firehose, please visit the url

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

Ask our Experts

QUESTION 45 UNATTEMPTED PROCESSING

A team currently maintains a forum-based application. All of the threads and their data are sent to a DynamoDB table. Below is the structure of the DynamoDB table

Thread ID- Partition Key

Author – Sort Key
Number of Replies
LastReplytimestamp
The table currently has around 600 million rows of data. There is a requirement now to get those threads which are not getting any traction and no response has been made in 6 months. This needs to be done on an on-going basis. Which of the following would be the most efficient way to achieve this?

- **A.** Query the data based on the ThreadID
- **B.** Query the data based on the Author
- **C.** Scan the entire table and search for the older records
- **D.** Create a global secondary index on LastReplytimestamp. ✔

Explanation :

Answer – D
Such a similar use case is given in the AWS Documentation:
#######
For example, consider the *Thread* table that is defined in Creating Tables and Loading Sample Data (https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/SampleData.html). This table is useful for an application such as the AWS Discussion Forums (https://forums.aws.amazon.com/). The following diagram shows how the items in the table would be organized. (Not all of the attributes are shown.)

Thread

ForumName	Subject	LastPostDateTime	Replies	
"S3"	"aaa"	"2015-03-15:17:24:31"	12	...
"S3"	"bbb"	"2015-01-22:23:18:01"	3	...
"S3"	"ccc"	"2015-02-31:13:14:21"	4	...
"S3"	"ddd"	"2015-01-03:09:21:11"	9	...
"EC2"	"yyy"	"2015-02-12:11:07:56"	18	...
"EC2"	"zzz"	"2015-01-18:07:33:42"	0	...
"RDS"	"rrr"	"2015-01-19:01:13:24"	3	...
"RDS"	"sss"	"2015-03-11:06:53:00"	11	...
"RDS"	"ttt"	"2015-10-22:12:19:44"	5	...
...	...	...	...	

DynamoDB stores all of the items with the same partition key value contiguously. In this example, given a particular ForumName, a `Query` operation could immediately locate all of the threads for that forum. Within a group of items with the same partition key value, the items are sorted by sort key value. If the sort key (*Subject*) is also provided in the query, DynamoDB can narrow down the results that are returned—for example, returning all of the threads in the "S3" forum that have a *Subject* beginning with the letter "a".
Some requests might require more complex data access patterns. For example:

- Which forum threads get the most views and replies?

- Which thread in a particular forum has the largest number of messages?
- How many threads were posted in a particular forum within a particular time period?

#######
Options A and B are incorrect since here we need to query based on the timestamp
Option C is incorrect because even though this is possible, this is the least efficient way to manage the query
For more information on this use case, please visit the url

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

Ask our Experts

QUESTION 46 UNATTEMPTED PROCESSING

A company is making use of Kinesis streams to process millions of records. Enough shards have been assigned to the stream. The KPL library is being used in an application to submit the requests. A Lambda function is being used to process the requests. During the initial stages of execution, there is a latency of around 150 to 200% being recorded in Cloudwatch for the consumption of records. The AWS Lambda function is showing no throttling errors and the memory and time is being consumed as per standard specifications. Which of the following could be the underlying problem?

- **A.** An incorrect IAM Role has been assigned to the Lambda function
- **B.** The throughput for the DynamoDB table is being exceeded. ✔
- **C.** The right packages have not been installed for the Lambda function
- **D.** The wrong programming language has been chosen for the Lambda function

Explanation :

Answer – B
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.
All other options are incorrect since if this were the cases, then the Lambda function would never run

at all.
For more information on reading records from Kinesis, please visit the 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)

Ask our Experts

QUESTION 47 UNATTEMPTED ANALYSIS

A company is currently using AWS Athena with a large number of data sets in S3. They want to improve the performance of the underlying queries. Which of the following can help achieve this?

- [ ] **A.** Partition the data ✔
- [ ] **B.** Encrypt the data
- [ ] **C.** Compress the data ✔
- [ ] **D.** Version the data

Explanation :

Answer : A and C
The AWS Documentation mentions the following
By partitioning your data, you can restrict the amount of data scanned by each query, thus improving performance and reducing cost. Athena leverages Hive for partitioning (https://cwiki.apache.org/confluence/display/Hive/LanguageManual+DDL#LanguageManualDDL-AlterPartition) data. You can partition your data by any key. A common practice is to partition the data based on time, often leading to a multi-level partitioning scheme. For example, a customer who has data coming in every hour might decide to partition by year, month, date, and hour. Another customer, who has data coming from many different sources but loaded one time per day, may partition by a data source identifier and date.
None of the other options will help in achieving better query performance.
Amazon Athena supports a wide variety of data formats like CSV, TSV, JSON, or Textfiles and also supports open source columnar formats such as Apache ORC and Apache Parquet. Athena also supports compressed data in Snappy, Zlib, LZO, and GZIP formats. By compressing, partitioning, and using columnar formats you can improve performance and reduce your costs.
For more information on using partitions, please visit the url

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

Ask our Experts

QUESTION 48 UNATTEMPTED PROCESSING

A company is currently employing DynamoDB for storing data related to tweets. The number of rows is running in billions. Which of the following are good practises when it comes to querying the data? Choose 2 answers from the options given below

- [ ] **A.** Set a smaller page size for the Scan operation ✔
- [ ] **B.** Try to use the Scan operation with filter expressions
- [ ] **C.** Try to use the Scan operation with certain projections
- [ ] **D.** Try to Query based on Indexes ✔

Explanation :

Answer – A and D
The AWS Documentation mentions the following
Because a Scan operation reads an entire page (by default, 1 MB), you can reduce the impact of the scan operation by setting a smaller page size. The Scan operation provides a *Limit* parameter that you can use to set the page size for your request. Each Query or Scanrequest that has a smaller page size uses fewer read operations and creates a "pause" between each request.
If possible, you should avoid using a Scan operation on a large table or index with a filter that removes many results. Also, as a table or index grows, the Scan operation slows. The Scan operation examines every item for the requested values and can use up the provisioned throughput for a large table or index in a single operation. For faster response times, design your tables and indexes so that your applications can use Queryinstead of Scan. (For tables, you can also consider using the GetItem and BatchGetItem APIs.)
The other options are incorrect as these would slow down the query performance
For more information on the best practices for DynamoDB, please visit the url

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

Ask our Experts

QUESTION 49 UNATTEMPTED COLLECTION

A development team has been requested to create an application that will be used to ingest data. The data consists of various metrics from various devices. The requirement is to create an automated way to store real-time aggregated time series in DynamoDB. How could you accomplish this?

- ○ **A.** Use the Amazon Kinesis Aggregators framework ✔

- **B.** Push the metrics to Cloudwatch for aggregation
- **C.** Aggregate the metrics from the producer side
- **D.** Aggregate the metrics in the stream itself

Explanation :

Answer – A
Amazon Kinesis Aggregators is a Java framework that enables the automatic creation of real-time aggregated time series data from Amazon Kinesis streams.
You can use this data to answer questions such as 'how many times per second has 'x' occurred' or 'what was the breakdown by hour over the day of the streamed data containing 'y'. Using this framework, you simply describe the format of the data on your stream (CSV, JSON, and so on), the granularity of times series that you require (seconds, minutes, hours, and so on), and how the data elements that are streamed should be grouped; the framework handles all the time series calculations and data persistence. You then simply consume the time series aggregates in your application using Amazon DynamoDB, or interact with the time series using Amazon CloudWatch or the Web Query API.
Option B is incorrect since you need to use Kinesis here for ingestion of data
Option C is incorrect since the ingestion of data should go as it is , and you need to aggregate data from the consumer side
Option D is incorrect since the stream is just used to ingest the information
For more information on the Kinesis aggregator library, please visit the url

- https://github.com/awslabs/amazon-kinesis-aggregators (https://github.com/awslabs/amazon-kinesis-aggregators)

Ask our Experts

QUESTION 50 UNATTEMPTED DATA SECURITY

A company currently has a set of DynamoDB tables already defined and items already exist in the tables. As per the company's updated security policy, it is now mandated that all data needs to be encrypted at rest? How can this be achieved?

- **A.** Enable encryption for the DynamoDB tables
- **B.** Enable global tables and then enable encryption
- **C.** Create new tables with encryption enabled. ✔
- **D.** Use the encryption flag for the table

Explanation :

Answer – C
The AWS Documentation mentions the following

Encryption at rest can be enabled only when you are creating a new DynamoDB table. Currently, you can't enable encryption at rest on an existing table. After encryption at rest is enabled, it can't be disabled. We recommend that you enable encryption for any tables that contain sensitive data.
Because this is clearly mentioned in the documentation , all other options are incorrect
For more information on encryption at rest for DynamoDB, please visit the url

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

Ask our Experts

QUESTION 51 UNATTEMPTED VISUALIZATION

A business intelligence development team wants to create visual dashboards and share it with their users. They are planning on using Amazon Quicksight. They want to give the users the ability to customize the data being used in the visual of the dashboard. Which of the following can help you accomplish this?

- A. Parameters ✔
- B. Queries
- C. URL's
- D. Titles

Explanation :

Answer - A
The AWS Documentation mentions the following
Parameters are new variables that you can add to your dashboards. You can use parameters in Amazon QuickSight for filtering, for allowing users to customize what data is being used in the visual, for what-if analysis, and also to drill through from one dashboard to another. After you create a parameter, you nearly always create a control for it. These controls allow your users to change the value of the parameter with intuitive UIs like drop-downs, slider bars, text boxes, and date selectors. A value set for the parameter becomes available in the context of the dashboard.
Since this is clearly mentioned in the documentation, all other options are incorrect
For more information on a use case for this, please visit the url

- https://aws.amazon.com/blogs/big-data/create-amazon-quicksight-dashboards-that-have-impact-with-parameters-on-screen-controls-and-url-actions/ (https://aws.amazon.com/blogs/big-data/create-amazon-quicksight-dashboards-that-have-impact-with-parameters-on-screen-controls-and-url-actions/)

Ask our Experts

QUESTION 52 UNATTEMPTED VISUALIZATION

A business intelligence development team wants to create visual dashboards and share it with their users. They are planning on using Amazon Quicksight. The users would want to create multiple iterations of an analysis. Which of the following feature in Quicksight can help them accomplish this?

- A. Dashboards
- B. Visuals
- C. Stories ✔
- D. Functions

Explanation :

Answer - C
You can use all of these options but in the end to create multiple iterations of an analysis, you need to make use of Stories
The AWS Documentation mentions the following
You can use a story to preserve multiple iterations of an analysis and then play them sequentially to provide a narrative about the analysis data. For example, you might want to see several versions of the analysis, all with the same charts but with different filters applied.
Option A is incorrect since this is used to read-only snapshot of an analysis that you can share with other Amazon QuickSight users for reporting purposes
Option B is incorrect since this is just graphical representation of your data
Option D is incorrect since this can be used to work with your existing data
For more information on working with stories, please visit the url

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

Ask our Experts

QUESTION 53 UNATTEMPTED PROCESSING

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

- A. For various ETL processes which use AWS Redshift commits, use transaction handling ✔

- [ ] **B.** Extract Large results sets from Redshift using the select query
- [ ] **C.** Extract Large results sets from Redshift using the UNLOAD statement ✔
- [ ] **D.** Copy data into Redshift using Insert queries

Explanation :

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

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

Ask our Experts

QUESTION 54 UNATTEMPTED ANALYSIS

A company has a lot of data in many disparate sources such as Hive, Cassandra, Redis, and MongoDB. The company wants to enable its employee to perform fast queries on these underlying data sources using a standard query language. Which of the following can be used for this purpose?

- ○ **A.** Presto ✔
- ○ **B.** SparkSQL
- ○ **C.** Hive
- ○ **D.** Oozie

Explanation :

Answer - A

The AWS Documentation mentions the following
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.
The other components don't have the ability to query multiple data sources
For more information on Presto, please visit the url

- 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

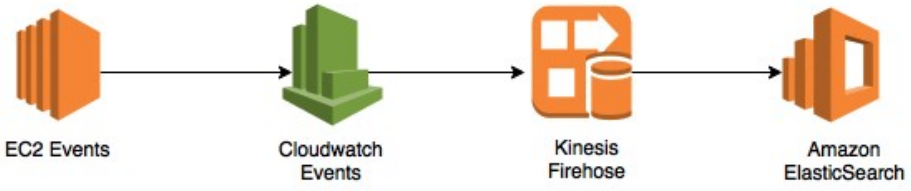
QUESTION 55 UNATTEMPTED COLLECTION

A company wants to a way to ingest all logs from their EC2 Instances into a searchable service. Which of the following implementation steps would you consider fulfilling this requirement? Choose 2 answers from the options given below

- [ ] **A.** Setup Cloudwatch events to stream the data to SQS
- [ ] **B.** Setup Cloudwatch events to stream the data to Kinesis Firehose ✔
- [ ] **C.** Configure the end destination for the log files as Amazon ElasticSearch ✔
- [ ] **D.** Configure the end destination for the log files as Amazon DynamoDB

Explanation :

Answer – B and C
This sort of architecture is given in the AWS documentation
########
Your CloudWatch event target is a Kinesis Firehose delivery stream that delivers this data to an Elasticsearch cluster, where you set up Kibana for visualization. Using this solution, you can easily load and visualize EC2 events in minutes without setting up complicated data pipelines.



########
Option A is incorrect since Kinesis should be used as the streaming service
Option D is incorrect since ElasticSearch is a better option for the final destination for log files and for searching capabilities
For more information on this use case, please visit the url

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

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

Ask our Experts

QUESTION 56 UNATTEMPTED COLLECTION

A company currently has an application that writes data onto AWS RDS-MySQL. They want to push data on a daily basis to archive records from MySQL tables to S3 for future analysis. How can you accomplish this in the easiest way?

- A. Create an S3 event to ingest the data from the MySQL table
- B. Use the AWS DataPipeline service and run a job on a daily basis ✔
- C. Create an EMR Cluster which will run a MapReduce job
- D. Use the database migration service to transfer the data

Explanation :

Answer – B
A sample of this is given in the AWS Documentation
#######
Export MySQL Data to Amazon S3 Using AWS Data Pipeline
This tutorial walks you through the process of creating a data pipeline to copy data (rows) from a table in MySQL database to a CSV (comma-separated values) file in an Amazon S3 bucket and then sending an Amazon SNS notification after the copy activity completes successfully. You will use an EC2 instance provided by AWS Data Pipeline for this copy activity.
#######
Option A is incorrect since the events can only trigger an event but not directly be used to ingest data
Option C is incorrect since this would be not be cost effective option just to try to ingest data
Option D is incorrect since the database migration service should ideally be used for a one time activity
For more information on this tutorial, please visit the url

- https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-copydata-mysql.html (https://docs.aws.amazon.com/datapipeline/latest/DeveloperGuide/dp-copydata-mysql.html)

Ask our Experts

QUESTION 57 UNATTEMPTED DATA SECURITY

A team has setup a Redshift cluster for a company. There is a requirement from the IT Security department to get the list of all IP addresses which establish connections to the Redshift cluster. Which of the following can help achieve this requirement?

- A. Get this information from Cloudtrail logs
- B. Get this information from connection logs in Redshift ✔
- C. Get this information from users logs in Redshift
- D. Get this information from VPC Flow Logs

Explanation :

Answer: B

The AWS Documentation mentions the following

Amazon Redshift logs information in the following log files:

- Connection log — logs authentication attempts, and connections and disconnections.
- User log — logs information about changes to database user definitions.
- User activity log — logs each query before it is run on the database.

The connection and user logs are useful primarily for security purposes. You can use the connection log to monitor information about the users who are connecting to the database and the related connection information, such as their IP address, when they made the request, what type of authentication they used, and so on. You can use the user log to monitor changes to the definitions of database users.

Connection Log

Logs authentication attempts, and connections and disconnect connection log.

Column name	Description
event	Connection or authentication event.
recordtime	Time the event occurred.
remotehost	Name or IP address of remote host.
remoteport	Port number for remote host.
pid	Process ID associated with the statement.
dbname	Database name.
username	User name.
authmethod	Authentication method.
duration	Duration of connection in microseconds.
sslversion	Secure Sockets Layer (SSL) version.
sslcipher	SSL cipher.
mtu	Maximum transmission unit (MTU).
sslcompression	SSL compression type.
sslexpansion	SSL expansion type.

Since the documentation clearly mentions how this can be achieved, the other options are invalid

For more information on database auditing, please visit the url

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

Ask our Experts

QUESTION 58 UNATTEMPTED COLLECTION

A team wants to start using AWS Kinesis Firehose to Ingest data and send them to various locations depending on the type of data being processed. Which of the following is not a supported data destination by Kinesis Firehose?

- **A.** AWS S3
- **B.** AWS Redshift
- **C.** AWS DynamoDB ✔
- **D.** Splunk

Explanation :

Answer - C
The AWS Documentation mentions the current destinations which are available with Kinesis Firehose streams
#####
Choose destination
This topic describes the **Choose destination** page of the **Create Delivery Stream** wizard in Amazon Kinesis Data Firehose.
Kinesis Data Firehose can send records to Amazon Simple Storage Service (Amazon S3), Amazon Redshift, or Amazon Elasticsearch Service (Amazon ES).
Topics

- Choose Amazon S3 for Your Destination (https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html#create-destination-s3)
- Choose Amazon Redshift for Your Destination (https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html#create-destination-redshift)
- Choose Amazon ES for Your Destination (https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html#create-destination-elasticsearch)
- Choose Splunk for Your Destination (https://docs.aws.amazon.com/firehose/latest/dev/create-destination.html#create-destination-splunk)

#####
For more information on working with destinations in Redshift, please visit the url

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

Ask our Experts

QUESTION 59 UNATTEMPTED ANALYSIS

A team wants to stream data into Kinesis Analytics. They need a serverless service which can be used to directly query the data in the stream. Which of the following can be used to achieve this?

- **A.** KPL Aggregators
- **B.** KCL Aggregators
- **C.** AWS Athena
- **D.** Kinesis Data Analytics ✔

Explanation :

Answer - D
The AWS Documentation mentions the following
With Amazon Kinesis Data Analytics, you can process and analyze streaming data using standard SQL. The service enables you to quickly author and run powerful SQL code against streaming sources to perform time series analytics, feed real-time dashboards, and create real-time metrics.
To get started with Kinesis Data Analytics, you create a Kinesis data analytics application that continuously reads and processes streaming data. The service supports ingesting data from Amazon Kinesis Data Streams and Amazon Kinesis Data Firehose streaming sources. Then, you author your SQL code using the interactive editor and test it with live streaming data. You can also configure destinations where you want Kinesis Data Analytics to send the results.
The other options are invalid because they can't directly be used to perform SQL queries on the streaming data.
For more information on Kinesis Analytics, please visit the url

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

Ask our Experts

QUESTION 60 UNATTEMPTED ANALYSIS

A company has a number of EC2 Instances sitting behind an Application Load Balancer. They want their IT Infrastructure team to analyse the client IP addresses for the requests coming into the load balancer. How can you accomplish this? Choose 2 answers from the options given below

- [x] **A.** Enable the logs for the Load Balancer ✔
- [x] **B.** Use the AWS Athena service to query the logs ✔
- [ ] **C.** Use the AWS Load balancer service to send the logs to DynamoDB
- [ ] **D.** Create a view in DynamoDB to check for the client IP address

Explanation :

Answer – A and B

An example of this is given in the AWS Documentation:

#######

Example Queries for ALB Logs

The following query counts the number of HTTP GET requests received by the load balancer grouped by the client IP address:

```
SELECT COUNT(request_verb) AS count, request_verb, client_ip FROM alb_logs GROUP BY req
uest_verb, client_ip LIMIT 100;
```

#######

Options C and D are incorrect since this architecture is not possible

For more information on using AWS Athena with Application Load balancer logs, please visit the url

- https://docs.aws.amazon.com/athena/latest/ug/application-load-balancer-logs.html (https://docs.aws.amazon.com/athena/latest/ug/application-load-balancer-logs.html)

Ask our Experts

QUESTION 61 UNATTEMPTED ANALYSIS

A company decides to use the Amazon Machine Learning service to classify social media posts that mention your company into two categories: posts that require a response and posts that do not. The training dataset of 10,000 posts contains the details of each post, including the timestamp, author, and full text of the post. You are missing the target labels that are required for training. What should be the next step to ensure you get the right results from the Machine Learning analysis?

- ○ **A.** Use the Regression Model of the Machine Learning service to classify the media posts

- **B.** Use the Binary Model of the Machine Learning service to classify the media posts to classify into 2 categories, basically those that require a response and those that don't
- **C.** Ensure a team is assigned to review each post and provide the label. ✔
- **D.** Using the a priori probability distribution of the two classes, use Monte-Carlo simulation to generate the labels

Explanation :

Answer – C
You need to have labelled data to have an accurate representation when working with Machine Learning
The AWS Documentation additionally mentions the following
Often, data is not readily available in a labeled form. Collecting and preparing the variables and the target are often the most important steps in solving an ML problem. The example data should be representative of the data that you will have when you are using the model to make a prediction. For example, if you want to predict whether an email is spam or not, you must collect both positive (spam emails) and negative (non-spam emails) for the machine learning algorithm to be able to find patterns that will distinguish between the two types of email.
All other options are incorrect because you can't directly start working with data which are not labelled properly
For more information on using collecting labelled data, please visit the url

- https://docs.aws.amazon.com/machine-learning/latest/dg/collecting-labeled-data.html (https://docs.aws.amazon.com/machine-learning/latest/dg/collecting-labeled-data.html)

Ask our Experts

QUESTION 62 UNATTEMPTED STORAGE

A company is planning on building a mobile application. The data for the application from various users needs to be stored in a data store that is highly available. The data store should scale on demand and the right authentication mechanism should be in place for the mobile application. Which of the following can help fulfil these requirements? Choose 2 answers from the options given below

- **A.** Make the application store its data on EBS volumes attached to EC2 Instances
- **B.** Make the application store its data on S3 ✔
- **C.** Make the application users authenticate with IAM Access Keys
- **D.** Use the AWS Cognito service for authentication ✔

Explanation :

Answer – B and D
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.
Amazon Cognito provides authentication, authorization, and user management for your web and mobile apps. Your users can sign in directly with a user name and password, or through a third party such as Facebook, Amazon, or Google.
Option A is incorrect since this is not a durable and highly available data store
Option C is incorrect since this is not a secure way to grant access
For more information on AWS Cognito and S3, please visit the url
https://aws.amazon.com/s3/ (https://aws.amazon.com/s3/)

Ask our Experts

QUESTION 63 UNATTEMPTED DATA SECURITY

An IT administrator has been given the task of unloading data from various Redshift tables. The company has mandated that all data which is unloaded is encrypted at rest. Which of the following can help accomplish this? Choose 3 answers from the options given below

- [ ] **A.** Use S3 server-side encryption with KMS Keys ✔
- [ ] **B.** Use S3 server-side encryption with a customer-managed key
- [ ] **C.** Use client-side encryption with a customer-managed key ✔
- [ ] **D.** Use the default AWS Managed encryption ✔

Explanation :

Answer – A, C and D
The AWS Documentation mentions the following
UNLOAD automatically creates files using Amazon S3 server-side encryption with AWS-managed encryption keys (SSE-S3). You can also specify server-side encryption with an AWS Key Management Service key (SSE-KMS) or client-side encryption with a customer-managed key (CSE-CMK). UNLOAD doesn't support Amazon S3 server-side encryption using a customer-supplied key (SSE-C).
For more information on unloading encrypted files, please visit the url

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

Ask our Experts

QUESTION 64 UNATTEMPTED VISUALIZATION

A business intelligence developer is trying to make use of Amazon Quicksight to create different types of visualizations. One of the data sets consists of a date column which is not being recognized in QuickSight. Which of the following can be done to ensure the data comes up properly in Quicksight with the least amount of effort?

- **A.** Change the date values in the data set to reflect the supported data types in Quicksight
- **B.** Change the format of the date column in the data set ✔
- **C.** Create a new data set with the formatted date data type
- **D.** Use a Calculated field for the date column

Explanation :

Answer – B
All of the other options are invalid because they would all require additional effort
The AWS Documentation mentions the following
Amazon QuickSight natively supports a limited number of date formats. However, you can't always control the format of the data provided to you. When your data contains a date in an unsupported format, you can tell Amazon QuickSight how to interpret it.
You can do this by editing the data set, and changing the format of the column from text or numeric to date. A screen appears after you make this change, so you can enter the format. For example, if you are using a relational data source, you can specify MM-dd-yyyy for a text field containing '09-19-2017', so it is interpreted as 2017-09-19T00:00:00.000Z. If you are using a nonrelational data source, you can do the same thing starting with a numeric field or a text field.
All of the other options are invalid because they would all require additional effort
For more information on using unsupported dates, please visit the url

- https://docs.aws.amazon.com/quicksight/latest/user/using-unsupported-dates.html (https://docs.aws.amazon.com/quicksight/latest/user/using-unsupported-dates.html)

Ask our Experts

QUESTION 65 CORRECT DATA SECURITY

A company needs to use a Redshift cluster in AWS. The mandate is that all data is encrypted at rest. It also needs to be ensured that the keys used for encryption for the Redshift cluster are from an on-premise HSM device. Which of the following are most secure and cost-effective solutions? Choose 2 answers from the options given below

- [x] **A.** Create a VPN connection between the VPC holding the cluster and the On-premise network ✔
- [ ] **B.** Create a Direct Connect connection between the VPC holding the cluster and the On-premise network
- [x] **C.** Use client and server certificates to configure a trusted connection between Amazon Redshift and your HSM ✔
- [ ] **D.** Import the keys from the on-premise HSM device to KMS

Explanation :

Answer – A and C
The AWS Documentation mentions the following
Amazon Redshift uses a hierarchy of encryption keys to encrypt the database. You can use either AWS Key Management Service (AWS KMS) or a hardware security module (HSM) to manage the top-level encryption keys in this hierarchy. The process that Amazon Redshift uses for encryption differs depending on how you manage keys. Amazon Redshift automatically integrates with AWS KMS but not with an HSM. When you use an HSM, you must use client and server certificates to configure a trusted connection between Amazon Redshift and your HSM.
Option B is incorrect since using VPN involves encryption and will be more secure for transferring the encryption keys
Option D is incorrect since the KMS service cannot be used for importing keys
For more information on working with DB encryption, please visit the url

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

Ask our Experts

Finish Review (https://www.whizlabs.com/learn/course/aws-bds-practice-tests/quiz/14850)

Certification

- Cloud Certification (https://www.whizlabs.com/cloud-certification-training-courses/)
- Java Certification (https://www.whizlabs.com/oracle-java-certifications/)
- PM Certification (https://www.whizlabs.com/project-management-certifications/)
- Big Data Certification (https://www.whizlabs.com/big-data-certifications/)

Mobile App

- Android Coming Soon
- iOS Coming Soon

- Discussions (http://ask.whizlabs.com/)
- Blog (https://www.whizlabs.com/blog/)

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)