001

A data engineer is configuring an AWS Glue job to read data from an Amazon S3 bucket. The data engineer has set up the necessary AWS Glue connection details and an associated IAM role. However, when the data engineer attempts to run the AWS Glue job, the data engineer receives an error message that indicates that there are problems with the Amazon S3 VPC gateway endpoint. The data engineer must resolve the error and connect the AWS Glue job to the S3 bucket. Which solution will meet this requirement?

A. Update the AWS Glue security group to allow inbound traffic from the Amazon S3 VPC gateway endpoint. B. Configure an S3 bucket policy to explicitly grant the AWS Glue job permissions to access the S3 bucket. C. Review the AWS Glue job code to ensure that the AWS Glue connection details include a fully qualified domain name. D. Verify that the VPC's route table includes inbound and outbound routes for the Amazon S3 VPC gateway endpoint. Most Voted

Answer

D. Verify that the VPC's route table includes inbound and outbound routes for the Amazon S3 VPC gateway endpoint.

Concept

- **VPC gateway endpoint** : : AWS

2

A retail company has a customer data hub in an Amazon S3 bucket. Employees from many countries use the data hub to support company-wide analytics. A governance team must ensure that the company's data analysts can access data only for customers who are within the same country as the analysts. Which solution will meet these requirements with the LEAST operational effort?

A. Create a separate table for each country's customer data. Provide access to each analyst based on the country that the analyst serves.

B. Register the S3 bucket as a data lake location in AWS Lake Formation. Use the Lake Formation row-level security features to enforce the company's access policies.

C. Move the data to AWS Regions that are close to the countries where the customers are. Provide access to each analyst based on the country that the analyst serves.

D. Load the data into Amazon Redshift. Create a view for each country. Create separate IAM roles for each country to provide access to data from each country. Assign the appropriate roles to the analysts.

Answer

least operational effort 의미

Least Operational Effort: Once the policies are defined within Lake Formation, they can be centrally managed and applied to the data in the S3 bucket without the need for creating separate tables or views for each

country, as in options A, C, and D. This reduces operational overhead and complexity.

Concept

- **AWS Lake Formation** : : 데이터 레이크의 구축, 보안 설정 관리를 자동화해주는 AWS 서비스. 데이터 레이크 구축시 단순하고 손이 많이 가는 관리작업들을 자동화한다. : AWS
- **AWS DataSync** : a managed data transfer service that simplifies and accelerates moving large amounts of data online between on-premises storage and Amazon S3, EFS, or FSx for Windows File Server. DataSync is optimized for efficient, incremental, and reliable transfers of large datasets, making it suitable for transferring 5 TB of data with daily updates. : AWS
- **AWS DataExchange** : : AWS
- **Data Mesh** : : AWS
- **Lambda Layers** : : AWS
- **AWS GLUE Workflows** : ETL을 위한 가장 Cost Effective 한 Solution : AWS
- **AWS Redshift Data API** : : AWS
- **Athena workgroups** : ETL을 위한 가장 Cost Effective 한 Solution : AWS
- **Glue excutionClass** : : AWS
- **Amazon Resource Name** : : AWS
- **Amazon Managed Service for Apache Flink** : : AWS
- **Amazon TimeStream DB** : : AWS
- **AWS Database Migration Service** : : AWS
- **Workload Management** : : AWS
- **AWS Glue Studio** : : AWS
- **Redshift Data Sharing** : : AWS
- **Graviton** : : AWS
- **materialized view**

---

obfuscate proliferate

AWS Database Migration Service (DMS) is specifically designed for migrating data from various sources, including on-premises databases, to AWS with minimal downtime and disruption to applications. It supports homogeneous migrations (e.g., SQL Server to SQL Server) as well as heterogeneous migrations (e.g., SQL Server to Amazon RDS for SQL Server).

003

A media company wants to improve a system that recommends media content to customer based on user behavior and preferences. To improve the recommendation system, the company needs to incorporate insights from third-party datasets into the company's existing analytics platform. **The company wants to minimize the effort and time required to incorporate third-party datasets.** Which solution will meet these requirements with the LEAST operational overhead?

A. Use API calls to access and integrate third-party datasets from AWS Data Exchange. B. Use API calls to access and integrate third-party datasets from AWS DataSync. C. Use Amazon Kinesis Data Streams to access and integrate third-party datasets from AWS CodeCommit repositories. D. Use Amazon Kinesis Data Streams to access and integrate third-party datasets from Amazon Elastic Container Registry (Amazon ECR).

Answer

AWS DataSync is primarily used for data transfer services designed to simplify, automate, and accelerate moving data between on-premises storage systems and AWS storage services, as well as between different AWS storage services. Its primary role is not for accessing third-party datasets but for efficiently transferring large volumes of data. In contrast, AWS Data Exchange is designed specifically for discovering and subscribing to third-party data in the cloud, providing direct API access to these datasets, which aligns perfectly with the company's need to integrate this data into their recommendation systems with minimal overhead.

004

A financial company wants to implement a data mesh. The data mesh must support centralized data governance, data analysis, and data access control. The company has decided to use AWS Glue for data catalogs and extract, transform, and load (ETL) operations. Which combination of AWS services will implement a data mesh? (Choose two.)

A. Use Amazon Aurora for data storage. Use an Amazon Redshift provisioned cluster for data analysis. B. Use Amazon S3 for data storage. Use Amazon Athena for data analysis. C. Use AWS Glue DataBrew for centralized data governance and access control. D. Use Amazon RDS for data storage. Use Amazon EMR for data analysis. E. Use AWS Lake Formation for centralized data governance and access control.

Answer

The data mesh implementation uses Amazon S3 and Athena for data storage and analysis, and AWS Lake Formation for centralized data governance and access control. When combined with AWS Glue, you can efficiently manage your data.

005

A data engineer maintains custom Python scripts that perform a data formatting process that many AWS Lambda functions use. When the data engineer needs to modify the Python scripts, the data engineer must manually update all the Lambda functions. The data engineer requires a less manual way to update the Lambda functions. Which solution will meet this requirement?

A. Store a pointer to the custom Python scripts in the execution context object in a shared Amazon S3 bucket. B. Package the custom Python scripts into Lambda layers. Apply the Lambda layers to the Lambda functions. C. Store a pointer to the custom Python scripts in environment variables in a shared Amazon S3 bucket. D. Assign the same alias to each Lambda function. Call reach Lambda function by specifying the function's alias.

Answer

Lambda layers allow you to centrally manage shared code and dependencies across multiple Lambda functions. By packaging the custom Python scripts into a Lambda layer, you can simply update the layer whenever changes are made to the scripts, and all the Lambda functions that use the layer will automatically inherit the updates. This approach reduces manual effort and ensures consistency across the functions.

006

A company created an extract, transform, and load (ETL) data pipeline in AWS Glue. A data engineer must crawl a table that is in Microsoft SQL Server. The data engineer needs to extract, transform, and load the output of the crawl to an Amazon S3 bucket. The data engineer also must orchestrate the data pipeline. Which AWS service or feature will meet these requirements MOST cost-effectively?

A. AWS Step Functions B. AWS Glue workflows C. AWS Glue Studio D. Amazon Managed Workflows for Apache Airflow (Amazon MWAA)

Answer

Glue workflows are the easiest solution here:

https://aws.amazon.com/blogs/big-data/orchestrate-an-etl-pipeline-using-aws-glue-workflows-triggers-and-crawlers-with-custom-classifiers/

https://aws.amazon.com/blogs/big-data/extracting-multidimensional-data-from-microsoft-sql-server-analysis-services-using-aws-glue/

007

A financial services company stores financial data in Amazon Redshift. A data engineer wants to run real-time queries on the financial data to support a web-based trading application. The data engineer wants to run the queries from within the trading application. Which solution will meet these requirements with the LEAST operational overhead?

A. Establish WebSocket connections to Amazon Redshift. B. Use the Amazon Redshift Data API. C. Set up Java Database Connectivity (JDBC) connections to Amazon Redshift. D. Store frequently accessed data in Amazon S3. Use Amazon S3 Select to run the queries.

Answer

The Amazon Redshift Data API enables you to painlessly access data from Amazon Redshift with all types of traditional, cloud-native, and containerized, serverless web service-based applications and event-driven applications.

008

A company uses Amazon Athena for one-time queries against data that is in Amazon S3. The company has several use cases. The company must implement permission controls to separate query processes and access to query history among users, teams, and applications that are in the same AWS account. Which solution will meet these requirements?

A. Create an S3 bucket for each use case. Create an S3 bucket policy that grants permissions to appropriate individual IAM users. Apply the S3 bucket policy to the S3 bucket. B. Create an Athena workgroup for each use case. Apply tags to the workgroup. Create an IAM policy that uses the tags to apply appropriate permissions to the workgroup. C. Create an IAM role for each use case. Assign appropriate permissions to the role for each use case. Associate the role with Athena. D. Create an AWS Glue Data Catalog resource policy that grants permissions to appropriate individual IAM users for each use case. Apply the resource policy to the specific tables that Athena uses.

Answer

Athena workgroups allow you to isolate and manage different workloads, users, and permissions. By creating a separate workgroup for each use case, you can control access to query history, manage permissions, and enforce resource usage limits independently for each workload. Applying tags to workgroups allows you to categorize and organize them based on the use case, which simplifies policy management.

009

A data engineer needs to schedule a workflow that runs a set of AWS Glue jobs every day. The data engineer does not require the Glue jobs to run or finish at a specific time. Which solution will run the Glue jobs in the MOST cost-effective way?

A. Choose the FLEX execution class in the Glue job properties. Most Voted B. Use the Spot Instance type in Glue job properties. C. Choose the STANDARD execution class in the Glue job properties. D. Choose the latest version in the GlueVersion field in the Glue job properties

Answer

The FLEX execution class leverages spare capacity within the AWS infrastructure to run Glue jobs at a discounted price compared to the standard execution class. Since the data engineer doesn't have specific time constraints, utilizing spare capacity is ideal for cost savings. Today's date its a checkbox in order to spare capacity and will mean we dont know when is going to finish, which is recommended to increase a timeout

010

A data engineer needs to create an AWS Lambda function that converts the format of data from .csv to Apache Parquet. The Lambda function must run only if a user uploads a .csv file to an Amazon S3 bucket. Which solution will meet these requirements with the LEAST operational overhead?

A. Create an S3 event notification that has an event type of s3:ObjectCreated:*. Use a filter rule to generate notifications only when the suffix includes .csv. Set the Amazon Resource Name (ARN) of the Lambda function as the destination for the event notification.

B. Create an S3 event notification that has an event type of s3:ObjectTagging:* for objects that have a tag set to .csv. Set the Amazon Resource Name (ARN) of the Lambda function as the destination for the event notification.

C. Create an S3 event notification that has an event type of s3:*. Use a filter rule to generate notifications only when the suffix includes .csv. Set the Amazon Resource Name (ARN) of the Lambda function as the destination for the event notification.

D. Create an S3 event notification that has an event type of s3:ObjectCreated:*. Use a filter rule to generate notifications only when the suffix includes .csv. Set an Amazon Simple Notification Service (Amazon SNS) topic as the destination for the event notification. Subscribe the Lambda function to the SNS topic.

11

A data engineer needs Amazon Athena queries to finish faster. The data engineer notices that all the files the Athena queries use are currently stored in uncompressed .csv format. The data engineer also notices that

users perform most queries by selecting a specific column. Which solution will MOST speed up the Athena query performance?

A. Change the data format from .csv to JSON format. Apply Snappy compression. B. Compress the .csv files by using Snappy compression. C. Change the data format from .csv to Apache Parquet. Apply Snappy compression. D. Compress the .csv files by using gzip compression.

12

A manufacturing company collects sensor data from its factory floor to monitor and enhance operational efficiency. The company uses Amazon Kinesis Data Streams to publish the data that the sensors collect to a data stream. Then Amazon Kinesis Data Firehose writes the data to an Amazon S3 bucket. The company needs to display a real-time view of operational efficiency on a large screen in the manufacturing facility. Which solution will meet these requirements with the LOWEST latency?

A. Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to process the sensor data. Use a connector for Apache Flink to write data to an Amazon Timestream database. Use the Timestream database as a source to create a Grafana dashboard. B. Configure the S3 bucket to send a notification to an AWS Lambda function when any new object is created. Use the Lambda function to publish the data to Amazon Aurora. Use Aurora as a source to create an Amazon QuickSight dashboard. C. Use Amazon Managed Service for Apache Flink (previously known as Amazon Kinesis Data Analytics) to process the sensor data. Create a new Data Firehose delivery stream to publish data directly to an Amazon Timestream database. Use the Timestream database as a source to create an Amazon QuickSight dashboard. D. Use AWS Glue bookmarks to read sensor data from the S3 bucket in real time. Publish the data to an Amazon Timestream database. Use the Timestream database as a source to create a Grafana dashboard.

13

A company stores daily records of the financial performance of investment portfolios in .csv format in an Amazon S3 bucket. A data engineer uses AWS Glue crawlers to crawl the S3 data. The data engineer must make the S3 data accessible daily in the AWS Glue Data Catalog. Which solution will meet these requirements?

A. Create an IAM role that includes the AmazonS3FullAccess policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Create a daily schedule to run the crawler. Configure the output destination to a new path in the existing S3 bucket. B. Create an IAM role that includes the AWSGlueServiceRole policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Create a daily schedule to run the crawler. Specify a database name for the output. C. Create an IAM role that includes the AmazonS3FullAccess policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Allocate data processing units (DPUs) to run the crawler every day. Specify a database name for the output. D. Create an IAM role that includes the AWSGlueServiceRole policy. Associate the role with the crawler. Specify the S3 bucket path of the source data as the crawler's data store. Allocate data processing units (DPUs) to run the crawler every day. Configure the output destination to a new path in the existing S3 bucket.

14

A company loads transaction data for each day into Amazon Redshift tables at the end of each day. The company wants to have the ability to track which tables have been loaded and which tables still need to be loaded. A data engineer wants to store the load statuses of Redshift tables in an Amazon DynamoDB table. The data engineer creates an AWS Lambda function to publish the details of the load statuses to DynamoDB. How should the data engineer invoke the Lambda function to write load statuses to the DynamoDB table?

A. Use a second Lambda function to invoke the first Lambda function based on Amazon CloudWatch events. B. Use the Amazon Redshift Data API to publish an event to Amazon EventBridge. Configure an EventBridge rule to invoke the Lambda function. Most Voted C. Use the Amazon Redshift Data API to publish a message to an Amazon Simple Queue Service (Amazon SQS) queue. Configure the SQS queue to invoke the Lambda function. D. Use a second Lambda function to invoke the first Lambda function based on AWS CloudTrail events.

15

A data engineer needs to securely transfer 5 TB of data from an on-premises data center to an Amazon S3 bucket. Approximately 5% of the data changes every day. Updates to the data need to be regularly proliferated to the S3 bucket. The data includes files that are in multiple formats. The data engineer needs to automate the transfer process and must schedule the process to run :ggeriodically. Which AWS service should the data engineer use to transfer the data in the **MOST operationally efficient way?**

A. AWS DataSync B. AWS Glue C. AWS Direct Connect D. Amazon S3 Transfer Acceleration

16

A company uses an on-premises Microsoft SQL Server database to store financial transaction data. The company migrates the transaction data from the on-premises database to AWS at the end of each month. The company has noticed that the cost to migrate data from the on-premises database to an Amazon RDS for SQL Server database has increased recently. The company requires a cost-effective solution to migrate the data to AWS. The solution must cause minimal downtown for the applications that access the database. Which AWS service should the company use to meet these requirements?

A. AWS Lambda B. AWS Database Migration Service (AWS DMS) C. AWS Direct Connect D. AWS DataSync

17

A data engineer is building a data pipeline on AWS by using AWS Glue extract, transform, and load (ETL) jobs. The data engineer needs to process data from Amazon RDS and MongoDB, perform transformations, and load the transformed data into Amazon Redshift for analytics. The data updates must occur every hour. Which combination of tasks will meet these requirements with the LEAST operational overhead? (Choose two.)

A. Configure AWS Glue triggers to run the ETL jobs every hour. B. Use AWS Glue DataBrew to clean and prepare the data for analytics. C. Use AWS Lambda functions to schedule and run the ETL jobs every hour. D. Use AWS Glue connections to establish connectivity between the data sources and Amazon Redshift. E. Use the Redshift Data API to load transformed data into Amazon Redshift.

Answer

A. Configure AWS Glue triggers to run the ETL jobs every hour. Reduced Code Complexity: Glue triggers eliminate the need to write custom code for scheduling ETL jobs. This simplifies the pipeline and reduces maintenance overhead. Scalability and Integration: Glue triggers work seamlessly with Glue ETL jobs, ensuring efficient scheduling and execution within the Glue ecosystem. D. Use AWS Glue connections to establish connectivity between the data sources and Amazon Redshift. Pre-Built Connectors: Glue connections offer pre-built connectors for various data sources like RDS and Redshift. This eliminates the need for manual configuration and simplifies data source access within the ETL jobs. Centralized Management: Glue connections are centrally managed within the Glue service, streamlining connection management and reducing operational overhead.

18

A company uses an Amazon Redshift cluster that runs on RA3 nodes. The company wants to scale read and write capacity to meet demand. A data engineer needs to identify a solution that will turn on concurrency scaling. Which solution will meet this requirement?

A. Turn on concurrency scaling in workload management (WLM) for Redshift Serverless workgroups. B. Turn on concurrency scaling at the workload management (WLM) queue level in the Redshift cluster. C. Turn on concurrency scaling in the settings during the creation of any new Redshift cluster. D. Turn on concurrency scaling for the daily usage quota for the Redshift cluster.

Answer

B. Turn on concurrency scaling at the workload management (WLM) queue level in the Redshift cluster.

Explanation: Concurrency scaling in Amazon Redshift allows the cluster to automatically add and remove compute resources in response to workload demands. Enabling concurrency scaling at the workload management (WLM) queue level allows you to specify which queues can benefit from concurrency scaling based on the query workload.

19

A data engineer must orchestrate a series of Amazon Athena queries that will run every day. Each query can run for more than 15 minutes. Which combination of steps will meet these requirements MOST cost-effectively? (Choose two.)

Lambda call Athena query; Step function orchestrate query workflow

A. Use an AWS Lambda function and the Athena Boto3 client start_query_execution API call to invoke the Athena queries programmatically. B. Create an AWS Step Functions workflow and add two states. Add the first state before the Lambda function. Configure the second state as a Wait state to periodically check whether the Athena query has finished using the Athena Boto3 get_query_execution API call. Configure the workflow to invoke the next query when the current query has finished running. C. Use an AWS Glue Python shell job and the Athena Boto3 client start_query_execution API call to invoke the Athena queries programmatically. D. Use an AWS Glue Python shell script to run a sleep timer that checks every 5 minutes to determine whether the current Athena query has finished running successfully. Configure the Python shell script to invoke the next query when the current query has finished running. E. Use Amazon Managed Workflows for Apache Airflow (Amazon MWAA) to orchestrate the Athena queries in AWS Batch.

Answer

AWS Lambda can be effectively used to trigger Athena queries. By using the start_query_execution API from the Athena Boto3 client, you can programmatically start Athena queries. Lambda functions are cost-effective as they charge based on the compute time used, and there's no charge when the code is not running. However, Lambda has a maximum execution timeout of 15 minutes, which means it's not suitable for long-running operations but can be used to trigger or start queries. AWS Step Functions can orchestrate multiple AWS services in workflows. By using a Wait state, the workflow can periodically check the status of the Athena query, and proceed to the next step once the query is complete. This approach is more scalable and reliable compared to continuously running a Lambda function, as Step Functions can handle long-running processes better and can maintain the state of each step in the workflow.

20

A company is migrating on-premises workloads to AWS. The company wants to reduce overall operational overhead. The company also wants to explore serverless options. The company's current workloads use Apache Pig, Apache Oozie, Apache Spark, Apache Hbase, and Apache Flink. The on-premises workloads process petabytes of data in seconds. The company must maintain similar or better performance after the migration to AWS. Which extract, transform, and load (ETL) service will meet these requirements?

A. AWS Glue B. Amazon EMR C. AWS Lambda D. Amazon Redshift

21

A data engineer must use AWS services to ingest a dataset into an Amazon S3 data lake. The data engineer profiles the dataset and discovers that the dataset contains personally identifiable information (PII). The data engineer must implement a solution to profile the dataset and obfuscate the PII. Which solution will meet this requirement with the LEAST operational effort?

A. Use an Amazon Kinesis Data Firehose delivery stream to process the dataset. Create an AWS Lambda transform function to identify the PII. Use an AWS SDK to obfuscate the PII. Set the S3 data lake as the target for the delivery stream. B. Use the Detect PII transform in AWS Glue Studio to identify the PII. Obfuscate the PII. Use an AWS Step Functions state machine to orchestrate a data pipeline to ingest the data into the S3 data lake. C. Use the Detect PII transform in AWS Glue Studio to identify the PII. Create a rule in AWS Glue Data Quality to obfuscate the PII. Use an AWS Step Functions state machine to orchestrate a data pipeline to ingest the data into the S3 data lake. D. Ingest the dataset into Amazon DynamoDB. Create an AWS Lambda function to identify and obfuscate the PII in the DynamoDB table and to transform the data. Use the same Lambda function to ingest the data into the S3 data lake.

22

A company maintains multiple extract, transform, and load (ETL) workflows that ingest data from the company's operational databases into an Amazon S3 based data lake. The ETL workflows use AWS Glue and Amazon EMR to process data. The company wants to improve the existing architecture to provide automated orchestration and to require minimal manual effort. Which solution will meet these requirements with the LEAST operational overhead?

A. AWS Glue workflows B. AWS Step Functions tasks C. AWS Lambda functions D. Amazon Managed Workflows for Apache Airflow (Amazon MWAA) workflows

Ansser

B - because AWS Glue can't trigger EMR

23

A company currently stores all of its data in Amazon S3 by using the S3 Standard storage class. A data engineer examined data access patterns to identify trends. During the first 6 months, most data files are accessed several times each day. Between 6 months and 2 years, most data files are accessed once or twice each month. After 2 years, data files are accessed only once or twice each year. The data engineer needs to use an S3 Lifecycle policy to develop new data storage rules. The new storage solution must continue to provide high availability. Which solution will meet these requirements in the MOST cost-effective way?

A. Transition objects to S3 One Zone-Infrequent Access (S3 One Zone-IA) after 6 months. Transfer objects to S3 Glacier Flexible Retrieval after 2 years. B. Transition objects to S3 Standard-Infrequent Access (S3 Standard-IA) after 6 months. Transfer objects to S3 Glacier Flexible Retrieval after 2 years. C. Transition objects to S3 Standard-Infrequent Access (S3 Standard-IA) after 6 months. Transfer objects to S3 Glacier Deep Archive after 2 years. D. Transition objects to S3 One Zone-Infrequent Access (S3 One Zone-IA) after 6 months. Transfer objects to S3 Glacier Deep Archive after 2 years.

24

A company maintains an Amazon Redshift provisioned cluster that the company uses for extract, transform, and load (ETL) operations to support critical analysis tasks. A sales team within the company maintains a Redshift cluster that the sales team uses for business intelligence (BI) tasks. The sales team recently requested access to the data that is in the ETL Redshift cluster so the team can perform weekly summary analysis tasks. The sales team needs to join data from the ETL cluster with data that is in the sales team's BI cluster. The company needs a solution that will share the ETL cluster data with the sales team without interrupting the critical analysis tasks. The solution must minimize usage of the computing resources of the ETL cluster. Which solution will meet these requirements?

A. Set up the sales team BI cluster as a consumer of the ETL cluster by using **Redshift data sharing.** B. Create materialized views based on the sales team's requirements. Grant the sales team direct access to the ETL cluster. C. Create database views based on the sales team's requirements. Grant the sales team direct access to the ETL cluster. D. Unload a copy of the data from the ETL cluster to an Amazon S3 bucket every week. Create an Amazon Redshift Spectrum table based on the content of the ETL cluster.

25

A data engineer needs to join data from multiple sources to perform a one-time analysis job. The data is stored in Amazon DynamoDB, Amazon RDS, Amazon Redshift, and Amazon S3. Which solution will meet this requirement MOST cost-effectively?

A. Use an Amazon EMR provisioned cluster to read from all sources. Use Apache Spark to join the data and perform the analysis. B. Copy the data from DynamoDB, Amazon RDS, and Amazon Redshift into Amazon S3. Run Amazon Athena queries directly on the S3 files. C. Use Amazon Athena Federated Query to join the data from all data sources. D. Use Redshift Spectrum to query data from DynamoDB, Amazon RDS, and Amazon S3 directly from Redshift.

26

A company is planning to use a provisioned Amazon EMR cluster that runs Apache Spark jobs to perform big data analysis. The company requires high reliability. A big data team must follow best practices for running cost-optimized and long-running workloads on Amazon EMR. The team must find a solution that will maintain the company's current level of performance. Which combination of resources will meet these requirements MOST cost-effectively? (Choose two.)

A. Use Hadoop Distributed File System (HDFS) as a persistent data store. B. Use Amazon S3 as a persistent data store. C. Use x86-based instances for core nodes and task nodes. D. Use Graviton instances for core nodes and task nodes. E. Use Spot Instances for all primary nodes.

27

A company wants to implement real-time analytics capabilities. The company wants to use Amazon Kinesis Data Streams and Amazon Redshift to ingest and process streaming data at the rate of several gigabytes per second. The company wants to derive near real-time insights by using existing business intelligence (BI) and analytics tools. Which solution will meet these requirements with the LEAST operational overhead?

A. Use Kinesis Data Streams to stage data in Amazon S3. Use the COPY command to load data from Amazon S3 directly into Amazon Redshift to make the data immediately available for real-time analysis. B. Access the data from Kinesis Data Streams by using SQL queries. Create materialized views directly on top of the stream. Refresh the materialized views regularly to query the most recent stream data. C. Create an external schema in Amazon Redshift to map the data from Kinesis Data Streams to an Amazon Redshift object. Create a materialized view to read data from the stream. Set the materialized view to auto refresh. D. Connect Kinesis Data Streams to Amazon Kinesis Data Firehose. Use Kinesis Data Firehose to stage the data in Amazon S3. Use the COPY command to load the data from Amazon S3 to a table in Amazon Redshift.

28

A company uses an Amazon QuickSight dashboard to monitor usage of one of the company's applications. The company uses AWS Glue jobs to process data for the dashboard. The company stores the data in a single Amazon S3 bucket. The company adds new data every day. A data engineer discovers that dashboard queries are becoming slower over time. The data engineer determines that the root cause of the slowing queries is long-running AWS Glue jobs. Which actions should the data engineer take to improve the performance of the AWS Glue jobs? (Choose two.)

A. Partition the data that is in the S3 bucket. Organize the data by year, month, and day. B. Increase the AWS Glue instance size by scaling up the worker type. C. Convert the AWS Glue schema to the DynamicFrame schema class. D. Adjust AWS Glue job scheduling frequency so the jobs run half as many times each day. E. Modify the IAM role that grants access to AWS glue to grant access to all S3 features.

29

A data engineer needs to use AWS Step Functions to design an orchestration workflow. The workflow must parallel process a large collection of data files and apply a specific transformation to each file. Which Step Functions state should the data engineer use to meet these requirements?

A. Parallel state B. Choice state C. Map state D. Wait state

30

A company is migrating a legacy application to an Amazon S3 based data lake. A data engineer reviewed data that is associated with the legacy application. The data engineer found that the legacy data contained some duplicate information. The data engineer must identify and remove duplicate information from the legacy application data. Which solution will meet these requirements with the LEAST operational overhead?

A. Write a custom extract, transform, and load (ETL) job in Python. Use the DataFrame.drop_duplicates() function by importing the Pandas library to perform data deduplication. B. Write an AWS Glue extract, transform, and load (ETL) job. Use the FindMatches machine learning (ML) transform to transform the data to perform data deduplication. C. Write a custom extract, transform, and load (ETL) job in Python. Import the Python dedupe library. Use the dedupe library to perform data deduplication. D. Write an AWS Glue extract, transform, and load (ETL) job. Import the Python dedupe library. Use the dedupe library to perform data deduplication.

31

A company is building an analytics solution. The solution uses Amazon S3 for data lake storage and Amazon Redshift for a data warehouse. The company wants to use Amazon Redshift Spectrum to query the data that is in Amazon S3. Which actions will provide the FASTEST queries? (Choose two.)

A. Use gzip compression to compress individual files to sizes that are between 1 GB and 5 GB. B. Use a columnar storage file format. C. Partition the data based on the most common query predicates. D. Split the data into files that are less than 10 KB. E. Use file formats that are not splittable.

32

A company uses Amazon RDS to store transactional data. The company runs an RDS DB instance in a private subnet. A developer wrote an AWS Lambda function with default settings to insert, update, or delete data in the DB instance. The developer needs to give the Lambda function the ability to connect to the DB instance privately without using the public internet. Which combination of steps will meet this requirement with the LEAST operational overhead? (Choose two.)

A. Turn on the public access setting for the DB instance. B. Update the security group of the DB instance to allow only Lambda function invocations on the database port. C. Configure the Lambda function to run in the same subnet that the DB instance uses. D. Attach the same security group to the Lambda function and the DB instance. Include a self-referencing rule that allows access through the database port. E. Update the network ACL of the private subnet to include a self-referencing rule that allows access through the database port.

33

A company has a frontend ReactJS website that uses Amazon API Gateway to invoke REST APIs. The APIS perform the functionality of the website. A data engineer needs to write a python script that can be occasionally invoked through API Gateway. The Code must return results to API Gateway. Which Solution will meet these requirements with the LEAST operational overhead?