

# AWS Certified Data Analytics - Specialty

1. A financial services company needs to aggregate daily stock trade data from the exchanges into a data store. The company requires that data be streamed directly into the data store, but also occasionally allows data to be modified using SQL. The solution should integrate complex, analytic queries running with minimal latency. The solution must provide a business intelligence dashboard that enables viewing of the top contributors to anomalies in stock prices.

Which solution meets the company's requirements?

A. Use Amazon Kinesis Data Firehose to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.

B. Use Amazon Kinesis Data Streams to stream data to Amazon Redshift. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.

C. Use Amazon Kinesis Data Firehose to stream data to Amazon Redshift. Use Amazon Redshift as a data source for Amazon QuickSight to create a business intelligence dashboard.

D. Use Amazon Kinesis Data Streams to stream data to Amazon S3. Use Amazon Athena as a data source for Amazon QuickSight to create a business intelligence dashboard.

**Answer(s): C**

---

2. A financial company hosts a data lake in Amazon S3 and a data warehouse on an Amazon Redshift cluster. The company uses Amazon QuickSight to build dashboards and wants to secure access from its on-premises Active Directory to Amazon QuickSight.

How should the data be secured?

A. Use an Active Directory connector and single sign-on (SSO) in a corporate network environment.

B. Use a VPC endpoint to connect to Amazon S3 from Amazon QuickSight and an IAM role to authenticate Amazon Redshift.

C. Establish a secure connection by creating an S3 endpoint to connect Amazon QuickSight and a VPC endpoint to connect to Amazon Redshift.

D. Place Amazon QuickSight and Amazon Redshift in the security group and use an Amazon S3 endpoint to connect Amazon QuickSight to Amazon S3.

**Answer(s): A**

---

3. A real estate company has a mission-critical application using Apache HBase in Amazon EMR. Amazon EMR is configured with a single master node. The company has over 5 TB of data stored on an Hadoop Distributed File System (HDFS). The company wants a cost-effective solution to make its HBase data highly available.

Which architectural pattern meets company's requirements?

A. Use Spot Instances for core and task nodes and a Reserved Instance for the EMR master node. Configure the EMR cluster with multiple master nodes. Schedule automated snapshots using Amazon EventBridge.

B. Store the data on an EMR File System (EMRFS) instead of HDFS. Enable EMRFS consistent view. Create an EMR HBase cluster with multiple master nodes. Point the HBase root directory to an Amazon S3 bucket.

C. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view. Run two separate EMR clusters in two different Availability Zones. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

D. Store the data on an EMR File System (EMRFS) instead of HDFS and enable EMRFS consistent view. Create a primary EMR HBase cluster with multiple master nodes. Create a secondary EMR HBase read-replica cluster in a separate Availability Zone. Point both clusters to the same HBase root directory in the same Amazon S3 bucket.

**Answer(s): B**

---

4. A software company hosts an application on AWS, and new features are released weekly. As part of the application testing process, a solution must be developed that analyzes logs from each

Amazon EC2 instance to ensure that the application is working as expected after each deployment. The collection and analysis solution should be highly available with the ability to display new information with minimal delays.

Which method should the company use to collect and analyze the logs?

A. Enable detailed monitoring on Amazon EC2, use Amazon CloudWatch agent to store logs in Amazon S3, and use Amazon Athena for fast, interactive log analytics.

B. Use the Amazon Kinesis Producer Library (KPL) agent on Amazon EC2 to collect and send data to Kinesis Data Streams to further push the data to Amazon OpenSearch Service (Amazon Elasticsearch Service) and visualize using Amazon QuickSight.

C. Use the Amazon Kinesis Producer Library (KPL) agent on Amazon EC2 to collect and send data to Kinesis Data Firehose to further push the data to Amazon OpenSearch Service (Amazon Elasticsearch Service) and OpenSearch Dashboards (Kibana).

D. Use Amazon CloudWatch subscriptions to get access to a real-time feed of logs and have the logs delivered to Amazon Kinesis Data Streams to further push the data to Amazon OpenSearch Service (Amazon Elasticsearch Service) and OpenSearch Dashboards (Kibana).

**Answer(s): C**

---

5. A data analyst is using AWS Glue to organize, cleanse, validate, and format a 200 GB dataset. The data analyst triggered the job to run with the Standard worker type. After 3 hours, the AWS Glue job status is still RUNNING. Logs from the job run show no error codes. The data analyst wants to improve the job execution time without overprovisioning.

Which actions should the data analyst take?

A. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the pro led metrics, increase the value of the executor- cores job parameter.

B. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the pro led metrics, increase the value of the maximum capacity job parameter.

C. Enable job metrics in AWS Glue to estimate the number of data processing units (DPUs). Based on the pro led metrics, increase the value of the spark.yarn.executor.memoryOverhead job parameter.

D. Enable job bookmarks in AWS Glue to estimate the number of data processing units (DPUs). Based on the pro led metrics, increase the value of the num- executors job parameter.

**Answer(s): B**

---

6. A company has a business unit uploading .csv files to an Amazon S3 bucket. The company's data platform team has set up an AWS Glue crawler to do discovery, and create tables and schemas. An AWS Glue job writes processed data from the created tables to an Amazon Redshift database. The AWS Glue job handles column mapping and creating the Amazon Redshift table appropriately. When the AWS Glue job is rerun for any reason in a day, duplicate records are introduced into the Amazon Redshift table.

Which solution will update the Redshift table without duplicates when jobs are rerun?

A. Modify the AWS Glue job to copy the rows into a staging table. Add SQL commands to replace the existing rows in the main table as postactions in the DynamicFrameWriter class.

B. Load the previously inserted data into a MySQL database in the AWS Glue job. Perform an upsert operation in MySQL, and copy the results to the Amazon Redshift table.

C. Use Apache Spark's DataFrame dropDuplicates() API to eliminate duplicates and then write the data to Amazon Redshift.

D. Use the AWS Glue ResolveChoice built-in transform to select the most recent value of the column.

**Answer(s): A**

---

7. A streaming application is reading data from Amazon Kinesis Data Streams and immediately writing the data to an Amazon S3 bucket every 10 seconds. The application is reading data from hundreds of shards. The batch interval cannot be changed due to a separate requirement. The data is being accessed by Amazon

Athena. Users are seeing degradation in query performance as time progresses.

Which action can help improve query performance?

A. Merge the files in Amazon S3 to form larger files.

B. Increase the number of shards in Kinesis Data Streams.

C. Add more memory and CPU capacity to the streaming application.

D. Write the files to multiple S3 buckets.

**Answer(s): A**

---

8. A company uses Amazon OpenSearch Service (Amazon Elasticsearch Service) to store and analyze its website clickstream data. The company ingests 1 TB of data daily using Amazon Kinesis Data Firehose and stores one day's worth of data in an Amazon ES cluster. The company has very slow query performance on the Amazon ES index and occasionally sees errors from Kinesis Data Firehose when attempting to write to the index. The Amazon ES cluster has 10 nodes running a single index and 3 dedicated master nodes. Each data node has 1.5 TB of Amazon EBS storage attached and the cluster is configured with 1,000 shards. Occasionally, JVMMemoryPressure errors are found in the cluster logs. Which solution will improve the performance of Amazon ES?

A. Increase the memory of the Amazon ES master nodes.

B. Decrease the number of Amazon ES data nodes.

C. Decrease the number of Amazon ES shards for the index.

D. Increase the number of Amazon ES shards for the index.

**Answer(s): C**

---

9. A manufacturing company has been collecting IoT sensor data from devices on its factory floor for a year and is storing the data in Amazon Redshift for daily analysis. A data analyst has determined that, at an expected ingestion rate of about 2 TB per day, the cluster will be undersized in less than 4 months. A long-term solution is needed. The data analyst has indicated that most queries only reference the most recent 13 months of data, yet there are also quarterly reports that need to query all the data generated from the past 7 years. The chief technology officer (CTO) is concerned about the costs, administrative effort, and performance of a long-term solution. Which solution should the data analyst use to meet these requirements?

A. Create a daily job in AWS Glue to UNLOAD records older than 13 months to Amazon S3 and delete those records from Amazon Redshift. Create an external table in Amazon Redshift to point to the S3 location. Use Amazon Redshift Spectrum to join to data that is older than 13 months.

B. Take a snapshot of the Amazon Redshift cluster. Restore the cluster to a new cluster using dense storage nodes with additional storage capacity.

C. Execute a CREATE TABLE AS SELECT (CTAS) statement to move records that are older than 13 months to quarterly partitioned data in Amazon Redshift Spectrum backed by Amazon S3.

D. Unload all the tables in Amazon Redshift to an Amazon S3 bucket using S3 Intelligent-Tiering. Use AWS Glue to crawl the S3 bucket location to create external tables in an AWS Glue Data Catalog. Create an Amazon EMR cluster using Auto Scaling for any daily analytics needs, and use Amazon Athena for the quarterly reports, with both using the same AWS Glue Data Catalog.

**Answer(s):** A

---

**10.** An insurance company has raw data in JSON format that is sent without a predefined schedule through an Amazon Kinesis Data Firehose delivery stream to an Amazon S3 bucket. An AWS Glue crawler is scheduled to run every 8 hours to update the schema in the data catalog of the tables stored in the S3 bucket. Data analysts analyze the data using Apache Spark SQL on Amazon EMR set up with AWS Glue Data Catalog as the metastore. Data analysts say that, occasionally, the data they receive is stale. A data engineer needs to provide access to the most up-to-date data. Which solution meets these requirements?

A. Create an external schema based on the AWS Glue Data Catalog on the existing Amazon Redshift cluster to query new data in Amazon S3 with Amazon Redshift Spectrum.

B. Use Amazon CloudWatch Events with the rate (1 hour) expression to execute the AWS Glue crawler every hour.

C. Using the AWS CLI, modify the execution schedule of the AWS Glue crawler from 8 hours to 1 minute.

D. Run the AWS Glue crawler from an AWS Lambda function triggered by an S3:ObjectCreated:\* event notification on the S3 bucket.

**Answer(s):** D

---

**11.** A company that produces network devices has millions of users. Data is collected from the devices on an hourly basis and stored in an Amazon S3 data lake. The company runs analyses on the last 24 hours of data on logs for abnormality detection and to troubleshoot and resolve user issues. The company also analyzes historical logs dating back 2 years to discover patterns and look for improvement opportunities. The data on logs contain many metrics, such as date, timestamp, source IP, and target IP. There are about 10 billion events every day. How should this data be stored for optimal performance?

A. In Apache ORC partitioned by date and sorted by source IP

B. In compressed .csv partitioned by date and sorted by source IP

C. In Apache Parquet partitioned by source IP and sorted by date

D. In compressed nested JSON partitioned by source IP and sorted by date

**Answer(s): A**

---

**12.** A banking company is currently using an Amazon Redshift cluster with dense storage (DS) nodes to store sensitive data. An audit found that the cluster is unencrypted. Compliance requirements state that a database with sensitive data must be encrypted through a hardware security module (HSM) with automated key rotation.

Which combination of steps is required to achieve compliance? (Choose two.)

A. Set up a trusted connection with HSM using a client and server certificate with automatic key rotation.

B. Modify the cluster with an HSM encryption option and automatic key rotation.

C. Create a new HSM-encrypted Amazon Redshift cluster and migrate the data to the new cluster.

D. Enable HSM with key rotation through the AWS CLI.

E. Enable Elliptic Curve Diffie-Hellman Ephemeral (ECDHE) encryption in the HSM.

**Answer(s): A C**

---

**13.** A company is planning to do a proof of concept for a machine learning (ML) project using Amazon SageMaker with a subset of existing on-premises data hosted in the company's 3 TB data warehouse. For part of the project, AWS Direct Connect is established and tested. To prepare the data for ML, data analysts are performing data curation. The data analysts want to perform multiple steps, including mapping, dropping nulls, resolving choices, and splitting fields. The company needs the fastest solution to curate the data for this project.

Which solution meets these requirements?

A. Ingest data into Amazon S3 using AWS DataSync and use Apache Spark scripts to curate the data in an Amazon EMR cluster. Store the curated data in Amazon S3 for ML processing.

B. Create custom ETL jobs on-premises to curate the data. Use AWS DMS to ingest data into Amazon S3 for ML processing.

C. Ingest data into Amazon S3 using AWS DMS. Use AWS Glue to perform data curation and store the data in Amazon S3 for ML processing.

D. Take a full backup of the data store and ship the backup les using AWS Snowball. Upload Snowball data into Amazon S3 and schedule data curation jobs using AWS Batch to prepare the data for ML.

**Answer(s): C**

---

**14.** A US-based sneaker retail company launched its global website. All the transaction data is stored in Amazon RDS and curated historic transaction data is stored in Amazon Redshift in the us-east-1 Region. The business intelligence (BI) team wants to enhance the user experience by providing a dashboard for sneaker trends.

The BI team decides to use Amazon QuickSight to render the website dashboards. During development, a team in Japan provisioned Amazon QuickSight in ap-northeast-1. The team is having difficulty connecting Amazon QuickSight from ap-northeast-1 to Amazon Redshift in us-east-1.

Which solution will solve this issue and meet the requirements?

A. In the Amazon Redshift console, choose to configure cross-Region snapshots and set the destination Region as ap-northeast-1. Restore the Amazon Redshift Cluster from the snapshot and connect to Amazon QuickSight launched in ap-northeast-1.

B. Create a VPC endpoint from the Amazon QuickSight VPC to the Amazon Redshift VPC so Amazon QuickSight can access data from Amazon Redshift.

C. Create an Amazon Redshift endpoint connection string with Region information in the string and use this connection string in Amazon QuickSight to connect to Amazon Redshift.

D. Create a new security group for Amazon Redshift in us-east-1 with an inbound rule authorizing access from the appropriate IP address range for the Amazon QuickSight servers in ap-northeast-1.

**Answer(s): D**

---

**15.** An airline has .csv-formatted data stored in Amazon S3 with an AWS Glue Data Catalog. Data analysts want to join this data with call center data stored in Amazon Redshift as part of a daily batch process. The Amazon Redshift cluster is already under

a heavy load. The solution must be managed, serverless, well- functioning, and minimize the load on the existing Amazon Redshift cluster. The solution should also require minimal effort and development activity.

Which solution meets these requirements?

A. Unload the call center data from Amazon Redshift to Amazon S3 using an AWS Lambda function. Perform the join with AWS Glue ETL scripts.

B. Export the call center data from Amazon Redshift using a Python shell in AWS Glue. Perform the join with AWS Glue ETL scripts.

C. Create an external table using Amazon Redshift Spectrum for the call center data and perform the join with Amazon Redshift.

D. Export the call center data from Amazon Redshift to Amazon EMR using Apache Sqoop. Perform the join with Apache Hive.

**Answer(s): C**

---

**16.** A data analyst is using Amazon QuickSight for data visualization across multiple datasets generated by applications. Each application stores its data within a separate Amazon S3 bucket. AWS Glue Data Catalog is used as a central catalog across all application data in Amazon S3. A new application stores its data within a separate S3 bucket. After updating the catalog to include the new application data source, the data analyst created a new Amazon QuickSight data source from an Amazon Athena table, but the import into SPICE failed.

How should the data analyst resolve the issue?

A. Edit the permissions for the AWS Glue Data Catalog from within the Amazon QuickSight console.

B. Edit the permissions for the new S3 bucket from within the Amazon QuickSight console.

C. Edit the permissions for the AWS Glue Data Catalog from within the AWS Glue console.

D. Edit the permissions for the new S3 bucket from within the S3 console.

**Answer(s): B**

---

**17.** A team of data scientists plans to analyze market trend data for their company's new investment strategy. The trend data comes from several different data sources in large volumes. The

team wants to utilize Amazon Kinesis to support their use case. The team uses SQL-like queries to analyze trends and wants to send notifications based on certain significant patterns in the trends. Additionally, the data scientists want to save the data to Amazon S3 for archival and historical re-processing, and use AWS managed services wherever possible. The team wants to implement the lowest-cost solution.

Which solution meets these requirements?

A. Publish data to one Kinesis data stream. Deploy a custom application using the Kinesis Client Library (KCL) for analyzing trends, and send notifications using Amazon SNS. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.

B. Publish data to one Kinesis data stream. Deploy Kinesis Data Analytics to the stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SNS. Configure Kinesis Data Firehose on the Kinesis data stream to persist data to an S3 bucket.

C. Publish data to two Kinesis data streams. Deploy Kinesis Data Analytics to the first stream for analyzing trends, and configure an AWS Lambda function as an output to send notifications using Amazon SNS. Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.

D. Publish data to two Kinesis data streams. Deploy a custom application using the Kinesis Client Library (KCL) to the first stream for analyzing trends, and send notifications using Amazon SNS. Configure Kinesis Data Firehose on the second Kinesis data stream to persist data to an S3 bucket.

**Answer(s):** A

---

**18.** A company currently uses Amazon Athena to query its global datasets. The regional data is stored in Amazon S3 in the us-east-1 and us-west-2 Regions. The data is not encrypted. To simplify the query process and manage it centrally, the company wants to use Athena in us-west-2 to query data from Amazon S3 in both

Regions. The solution should be as low-cost as possible.

What should the company do to achieve this goal?

A. Use AWS DMS to migrate the AWS Glue Data Catalog from us-east-1 to us-west-2. Run Athena queries in us-west-2.

B. Run the AWS Glue crawler in us-west-2 to catalog datasets in all Regions. Once the data is crawled, run Athena queries in us-west-2.

C. Enable cross-Region replication for the S3 buckets in us-east-1 to replicate data in us-west-2. Once the data is replicated in us-west-2, run the AWS Glue crawler there to update the AWS Glue Data Catalog in us-west-2 and run Athena queries.

D. Update AWS Glue resource policies to provide us-east-1 AWS Glue Data Catalog access to us-west-2. Once the catalog in us-west-2 has access to the catalog in us-east-1, run Athena queries in us-west-2.

**Answer(s): B**

---

**19.** A large company receives files from external parties in Amazon EC2 throughout the day. At the end of the day, the files are combined into a single file, compressed into a gzip file, and uploaded to Amazon S3. The total size of all the files is close to 100 GB daily. Once the files are uploaded to Amazon S3, an AWS Batch program executes a COPY command to load the files into an Amazon Redshift cluster. Which program modification will accelerate the COPY process?

A. Upload the individual files to Amazon S3 and run the COPY command as soon as the files become available.

B. Split the number of files so they are equal to a multiple of the number of slices in the Amazon Redshift cluster. Gzip and upload the files to Amazon S3. Run the COPY command on the files.

C. Split the number of files so they are equal to a multiple of the number of compute nodes in the Amazon Redshift cluster. Gzip and upload the files to Amazon S3. Run the COPY command on the files.

D. Apply sharding by breaking up the files so the distkey columns with the same values go to the same file. Gzip and upload the sharded files to Amazon S3. Run the COPY command on the files.

**Answer(s): B**

---

**20.** A large ride-sharing company has thousands of drivers globally serving millions of unique customers every day. The company has decided to migrate an existing data mart to Amazon Redshift. The existing schema includes the following tables.

A trips fact table for information on completed rides.

A drivers dimension table for driver profiles.

A customers fact table holding customer profile information.

The company analyzes trip details by date and destination to examine profitability by region. The drivers data rarely changes. The customers data frequently changes.

What table design provides optimal query performance?

A. Use DISTSTYLE KEY (destination) for the trips table and sort by date. Use DISTSTYLE ALL for the drivers and customers tables.

B. Use DISTSTYLE EVEN for the trips table and sort by date. Use DISTSTYLE ALL for the drivers table. Use DISTSTYLE EVEN for the customers table.

C. Use DISTSTYLE KEY (destination) for the trips table and sort by date. Use DISTSTYLE ALL for the drivers table. Use DISTSTYLE EVEN for the customers table.

D. Use DISTSTYLE EVEN for the drivers table and sort by date. Use DISTSTYLE ALL for both fact tables.

**Answer(s): C**

---