Vendor: Amazon

Exam Code: DEA-C01

Exam Name: AWS Certified Data Engineer - Associate (DEA-C01) Exam

Version: DEMO

QUESTION 1
A company uses an Amazon S3 Standard bucket to maintain a self-managed transactional data lake that uses Apache Iceberg tables. The data lake ingests data both in real time and in batches.

Users report slow performance for real-time tables. A data engineer reviews the real-time tables and notices that the tables are made up of many small data files The data engineer must improve the performance of the real-time tables.

Which solution will meet this requirement?

A. Expire historic snapshots.
B. Archive historic snapshots.
C. Delete S3 objects that are not linked from the Iceberg table.
D. Apply compaction.

Answer: D
Explanation:
Compaction merges many small data files into fewer, larger files, which reduces file count and metadata overhead in Apache Iceberg tables, directly improving query performance for real-time workloads.

QUESTION 2
An ecommerce company collects daily customer transaction logs in CSV format and stores the logs in Amazon S3. The company uses Amazon Athena to scan a subset of attributes from the logs on the same day the company receives each log.

Query times are increasing because of increasing transaction volume. The company wants to improve query performance.

Which solution will meet these requirements with the SHORTEST query times?

A. Convert the CSV logs into multiple ORC files for better parallelism in Athena. Partition by date in Amazon S3. Use columnar pushdown filters.
B. Convert the CSV logs to JSON. Partition by date in Amazon S3. Use Athena with dynamic filtering to reduce data scans.
C. Convert the CSV logs to Avro. Partition by date in Amazon S3. Use Athena with projection-based partitioning.
D. Convert the CSV logs to a single Apache Parquet file for each day Partition the data by date in Amazon S3. Use Athena with predicate pushdown filters.

Answer: D
Explanation:
Partitioning by date limits each query to only the relevant day's data, and Parquet enables efficient columnar reads and predicate pushdown for scanning only the needed columns and row groups. Writing a single Parquet file per day minimizes file listing and small-file overhead, typically delivering the shortest Athena query times for same-day, subset-column analytics.

QUESTION 3
A company stores customer data in an Amazon S3 bucket. The company must permanently delete all customer data that is older than 7 years. Which solution will meet this requirement?

A. Configure an S3 Lifecycle policy to permanently delete objects that are older than 7 years.

B. Use Amazon Athena to query the S3 bucket for objects that are older than 7 years. Configure Athena to delete the results.
C. Configure an S3 Lifecycle policy to move objects that are older than 7 years to S3 Glacier Deep Archive.
D. Configure an S3 Lifecycle policy to enable S3 Object Lock on all objects that are older than 7 years.

Answer: A
Explanation:
An S3 Lifecycle policy can automatically and permanently expire (delete) objects after they reach an age of 7 years, ensuring the data is removed without ongoing manual processes.

QUESTION 4
Two data engineering teams use separate AWS accounts. Both teams request access to the same datashare in an Amazon Redshift cluster that is in a third AWS account. The datashare is named salesshare.

A data engineer must use the Amazon Redshift SQL interface to grant both data engineering teams' access to the datashare.

Which command or commands will meet this requirement?

A. GRANT USAGE ON DATASHARE salesshare TO ACCOUNTS '<account ID 1>' AND '<account ID 2>';
B. GRANT USAGE ON DATASHARE salesshare TO NAMESPACES '<account ID 1>' AND '<account ID 2>';
C. GRANT USAGE ON DATASHARE salesshare TO ACCOUNT '<account ID 1>';
GRANT USAGE ON DATASHARE salesshare TO ACCOUNT '<account ID 2>';
D. GRANT USAGE ON DATASHARE salesshare TO NAMESPACE '<account ID 1>';
GRANT USAGE ON DATASHARE salesshare TO NAMESPACE '<account ID 2>';

Answer: D
Explanation:
In Amazon Redshift data sharing, cross-account permissions for a datashare are granted to consumer cluster/serverless namespaces. Granting USAGE on the datashare to each consumer namespace enables both teams' Redshift environments in their respective accounts to create and query databases from the shared data.

QUESTION 5
A company regularly loads data into an Amazon DynamoDB table. The company must retain the data in DynamoDB for 1 year. After 1 year, the company must archive the data for 5 years. After 5 years, the company must delete the data. Which solution will meet these requirements in the MOST operationally efficient way?

A. Include a timestamp with the data that is loaded into DynamoDB. Create an AWS Lambda function to read the timestamps and move items that are 1 year old to Amazon S3 Glacier Flexible Retrieval. Create an S3 Lifecycle policy to delete the items after 5 years.
B. Enable Amazon DynamoDB Streams. Configure a TTL rule to delete items from the DynamoDB table after 1 year. Use an AWS Lambda function to consume the stream and send items to Amazon S3 Glacier Flexible Retrieval. Delete the items from S3 Glacier Flexible Retrieval after 5 years.
C. Include a timestamp with the data that is loaded into DynamoDB. Use an AWS Lambda function

to read timestamps and move items that are 1 year old directly to Amazon S3 Glacier Deep Archive. Configure the function to delete items after 5 years.

D. Enable Amazon DynamoDB Streams. Configure a TTL rule to delete items from the DynamoDB table after 1 year. Use an AWS Lambda function to consume the stream and send items directly to Amazon S3 Glacier Deep Archive. Create an S3 Lifecycle policy to delete the items after 5 years.

Answer: D
Explanation:
DynamoDB TTL provides a fully managed way to remove items after 1 year with minimal operational work. DynamoDB Streams can capture the item changes so a Lambda function can archive the expiring data to Amazon S3 before it is removed from the table. Using S3 Glacier Deep Archive minimizes archive storage cost for long retention, and an S3 Lifecycle policy can automatically delete the archived objects after 5 years, avoiding custom deletion logic.

QUESTION 6
A company must retain specific data for 1 year. A data engineer observes that one of the company's Amazon S3 buckets contains millions of objects that are older than 3 years. Versioning is enabled on the bucket.

To reduce costs, the data engineer implements an S3 Lifecycle rule to expire objects after 365 days. The new S3 Lifecycle rule causes the object count to double instead of decrease.

Which additional step must the data engineer take to permanently delete the old objects?

A. Disable versioning on the S3 bucket.
B. Use an AWS Lambda function to run a Python job to identify and delete objects that are older than 365 days.
C. Suspend versioning on the S3 bucket.
D. Add an additional S3 Lifecycle rule to delete the current and expired versions of objects that are older than 365 days.

Answer: D
Explanation:
With S3 versioning enabled, lifecycle expiration of an object typically adds a delete marker instead of removing prior versions, which can increase the total number of object versions. To permanently remove the older data, the lifecycle configuration must also delete noncurrent (previous) versions and expired delete markers based on age so that both current and prior versions are actually removed.

QUESTION 7
A manufacturing company uses AWS Glue jobs to process IoT sensor data to generate predictive maintenance models. A data engineer needs to implement automated data quality checks to identify temperature readings that are outside the expected range of -50°C to 150°C. The data quality checks must also identify records that are missing timestamp values.

The data engineer needs a solution that requires minimal coding and can automatically flag the specified issues.

Which solution will meet these requirements?

A. Create an AWS Glue DataBrew project to profile the sensor data Define completeness rules for timestamps. Set up numeric range validation for temperature values.

B. Use AWS Glue's Data Quality rules and machine learning (ML)-based anomaly detection to identify missing timestamps and to detect temperature anomalies.
C. Create an AWS Lambda function to scan the sensor data files to validate temperature ranges. Use AWS Glue Data Catalog tables to check timestamp completeness.
D. Create an AWS Glue DynamicFrame that uses a custom data quality operator to profile the sensor data. Use Amazon SageMaker Data Wrangler transforms to validate timestamps and temperature ranges.

Answer: A
Explanation:
AWS Glue DataBrew provides a low-code way to implement automated data quality checks, including completeness rules to flag missing timestamp values and numeric range validation to flag temperature readings outside -50°C to 150°C.

QUESTION 8
A media company wants to improve a system that recommends media content to customers 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 a Git repository.
D. Use Amazon Kinesis Data Streams to access and integrate third-party datasets from Amazon Elastic Container Registry (Amazon ECR).

Answer: A
Explanation:
AWS Data Exchange is designed to quickly subscribe to and consume third-party datasets through a managed service interface and APIs, minimizing the integration effort and operational overhead compared to building custom ingestion from other sources.

QUESTION 9
A company generates yearly financial statements for customers and stores the statements in an Amazon S3 bucket. Customers rarely access the documents after 1 week. The company must retain the statements for 7 years. The statements must remain readily accessible for customers. Which solution will meet these requirements in the MOST cost-effective way?

A. Create an S3 Lifecycle rule to transition objects to S3 Glacier Deep Archive after 7 days. Expire the objects after 7 years.
B. Set the S3 bucket to use S3 Intelligent-Tiering when new objects are uploaded. Set objects to expire after 7 years.
C. Create an S3 Lifecycle rule to transition objects to S3 Glacier Instant Retrieval after 7 days. Expire the objects after 7 years.
D. Set the S3 bucket to use S3 Glacier Instant Retrieval when new objects are uploaded. Create an AWS Lambda function that runs daily to delete any objects that are older than 7 years.

Answer: C
Explanation:
S3 Glacier Instant Retrieval is the most cost-effective option that still keeps the statements readily accessible because it is designed for archival data that is rarely accessed but must be retrieved immediately, with millisecond access. A lifecycle rule can keep the files in a standard S3 class for the first 7 days, then transition them to Glacier Instant Retrieval for the remainder of the 7-year retention period, and expire them automatically at the end. S3 Glacier Deep Archive is cheaper but does not provide immediate access, and S3 Intelligent-Tiering is less cost-effective here because the access pattern is already known.

Reference:
https://aws.amazon.com/ru/s3/storage-classes/glacier/instant-retrieval/

QUESTION 10
A company runs an extract, transform, and load (ETL) job in AWS Glue. The job processes personally identifiable information (PII) data and writes logs to an Amazon CloudWatch Logs log group. A data engineer needs to mask PII data in the CloudWatch logs group. Which solution will meet these requirements?

A. Attach an AWS Glue security configuration to the ETL job.
B. Configure a data protection policy. Attach the policy to the CloudWatch log group.
C. Run an Amazon Macie sensitive data discovery job.
D. Call AWS Glue sensitive data detection APIs in the ETL job.

Answer: B
Explanation:
A CloudWatch Logs data protection policy is specifically designed to detect, audit, and mask sensitive data such as PII in log events that are ingested into a log group. Attaching the policy directly to the CloudWatch log group meets the requirement to mask PII in the logs without changing the ETL logic itself.

Reference:
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/mask-sensitive-log-data.html
https://docs.aws.amazon.com/AmazonCloudWatch/latest/logs/cloudwatch-logs-data-protection-policies.html

QUESTION 11
A company creates a new non-production application that runs on an Amazon EC2 instance. The application needs to communicate with an Amazon RDS database instance using Java Database Connectivity (JDBC). The EC2 instances and the RDS database instance are in the same subnet. Which solution will meet this requirement?

A. Modify the IAM role that is assigned to the database instance to allow connections from the EC2 instances.
B. Modify the ec2_authorized_hosts parameter in the RDS parameter group to include the EC2 instances. Restart the database instance.
C. Update the database security group to allow connections from the EC2 instances.
D. Enable the Amazon RDS Data API and specify the Amazon Resource Name (ARN) of the database instance in the JDBC connection string.

Answer: C
Explanation:
Connectivity from Amazon EC2 to an Amazon RDS DB instance in the same subnet or VPC is

controlled by VPC security groups. Updating the database security group to allow inbound connections from the EC2 instances enables the application to connect to the database by using JDBC. IAM roles do not control direct JDBC network access, and the RDS Data API is not used through a JDBC connection string.

Reference:
https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Overview.RDSSecurityGroups.html
https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/USER_VPC.Scenarios.html

QUESTION 12
A retail company wants to implement real-time analytics for an ecommerce platform. The company needs to collect clickstream data from the company's website and mobile apps. The company needs to store the data for analytics. Which solution will meet these requirements with the LEAST ongoing maintenance?

A. Use Amazon Data Firehose to ingest the streaming data. Deliver the processed data directly to a provisioned Amazon Redshift cluster.
B. Deploy agents on Amazon EC2 instances to collect the streaming data. Use the AWS CLI to periodically batch upload the data to an Amazon S3 bucket.
C. Use Amazon Kinesis Data Streams to collect the streaming data. Use Amazon Data Firehose to deliver the data to an Amazon S3 bucket.
D. Use an Amazon Managed Streaming for Apache Kafka (Amazon MSK) broker to collect the data. Store the data in an Amazon RDS DB instance.

Answer: C
Explanation:
Amazon Kinesis Data Streams is well suited to collect high-volume real-time clickstream data from websites and mobile apps, and Amazon Data Firehose can then deliver that streaming data to Amazon S3 with very little operational effort. This approach is highly managed and avoids the ongoing administration required for EC2-based collectors, self-managed Kafka patterns, or maintaining a provisioned Redshift cluster for raw landing storage.

Reference:
https://docs.aws.amazon.com/firehose/latest/dev/what-is-this-service.html
https://docs.aws.amazon.com/firehose/latest/dev/basic-deliver.html

QUESTION 13
A company uses AWS Glue ETL pipelines to process data. The company uses Amazon Athena to analyze data in an Amazon S3 bucket.

To better understand shipping timelines, the company decides to collect and store shipping and delivery dates in addition to order data. The company adds a data quality check to ensure that shipping date is greater than order date and that delivery date is greater than shipping date. Orders that fail the quality check must be stored in a second S3 bucket.

Which solution will meet these requirements MOST cost-effectively?

A. Use the AWS Glue DataBrew DATEDIFF function to create two additional columns. Check the new columns.
B. Use Athena to query all three date columns, and compare the columns.
C. Use AWS Glue Data Quality to create a custom rule that uses the three date columns.
D. Use an AWS Glue crawler to populate an AWS Glue Data Catalog. Use the three date columns to create a filter.

Answer: C
Explanation:
AWS Glue Data Quality is the most cost-effective choice because it is built for ETL data validation and supports custom rules that compare multiple columns, such as ensuring shipping date is after order date and delivery date is after shipping date. It can also identify the specific records that fail validation so those failed orders can be written to a separate Amazon S3 bucket as part of the ETL flow.

Reference:
https://docs.aws.amazon.com/glue/latest/dg/glue-data-quality.html
https://docs.aws.amazon.com/glue/latest/dg/tutorial-data-quality.html
https://docs.aws.amazon.com/glue/latest/dg/dqdl-rule-types-CustomSql.html

QUESTION 14
A data engineer is designing a log table for an application that requires continuous ingestion. The application must provide dependable API-based access to specific records from other applications. The application must handle more than 4,000 concurrent write operations and 6,500 read operations every second. Which solution will meet these requirements?

A. Create an Amazon Redshift table with the KEY distribution style. Use the Amazon Redshift Data API to perform all read and write operations.
B. Store the log files in an Amazon S3 Standard bucket. Register the schema in AWS Glue Data Catalog. Create an external Redshift table that points to the AWS Glue schema. Use the table to perform Amazon Redshift Spectrum read operations.
C. Create an Amazon Redshift table with the EVEN distribution style. Use the Amazon Redshift Java Database Connectivity (JDBC) connector to establish a database connection. Use the database connection to perform all read and write operations.
D. Create an Amazon DynamoDB table that has provisioned capacity to meet the application's capacity needs. Use the DynamoDB table to perform all read and write operations by using DynamoDB APIs.

Answer: D
Explanation:
Amazon DynamoDB is the best fit because it is designed for dependable API-based access with very high request rates and can be provisioned to support the required thousands of concurrent write and read operations per second. It also supports direct item retrieval and query operations through DynamoDB APIs, which matches the requirement for other applications to access specific records reliably. Amazon Redshift is designed for analytics workloads rather than high-throughput transactional API reads and writes.

Reference:
https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/read-write-operations.html
https://docs.aws.amazon.com/amazondynamodb/latest/developerguide/Query.html

QUESTION 15
A company needs to store and analyze a large amount of IoT sensor data. The company needs to retain the data indefinitely. The company analyzes the data in an Amazon Redshift cluster. Which solution will meet these requirements MOST cost-effectively?

A. Store the data in an Amazon S3 bucket in JSON format. Configure auto-copy data ingestion from the S3 bucket to the Redshift cluster.

B. Store the data in an Amazon S3 bucket in Apache Parquet format. Configure query access through Amazon Redshift Spectrum.
C. Store the data in an Amazon S3 bucket in JSON format. Configure query access through Amazon Redshift Spectrum.
D. Store the data in an Amazon S3 bucket in Apache Parquet format. Configure auto-copy data ingestion from the S3 bucket to the Redshift cluster.

Answer: B
Explanation:
Storing the IoT data in Amazon S3 in Apache Parquet format and querying it through Amazon Redshift Spectrum is the most cost-effective approach because the data can be retained indefinitely in low-cost S3 storage while still being analyzed from Redshift without loading all of it into the cluster. Parquet further reduces cost because it is a columnar format that lets Redshift Spectrum scan only the needed columns, which lowers data scanned and improves query efficiency.

Reference:
https://docs.aws.amazon.com/redshift/latest/dg/c-spectrum-external-tables.html
https://docs.aws.amazon.com/redshift/latest/dg/c-spectrum-external-performance.html

QUESTION 16
A company has an Amazon S3 based data lake. The data lake contains datasets that belong to multiple departments. The data lake ingests millions of customer records each day.

A data engineer needs to design an access and storage solution that allows departments to access only the subset of the company's dataset that each department requires. The solution must follow the principle of least privilege.

Which solution will meet these requirements with the LEAST operational effort?

A. Define IAM policies and IAM roles for each department. Specify the S3 access paths from the data lake that each team can access.
B. Set up Amazon Redshift and Amazon Redshift Spectrum as the primary entry points for the data lake. Define an IAM role that Amazon Redshift can assume. Configure the IAM role to grant access to the data that is in Amazon S3.
C. Set up AWS Lake Formation. Assign LF-Tags to AWS Glue Data Catalog resources. Enable Lake Formation tag-based access control (LF-TBAC).
D. Deploy an Amazon RDS for PostgreSQL database that has the aws_s3 extension installed. Configure AWS Step Functions events to invoke an AWS Lambda function to sync the data lake with the database.

Answer: C
Explanation:
AWS Lake Formation with LF-Tags and LF-TBAC is designed for fine-grained, scalable access control across shared data lake resources. It lets the company grant each department access only to the datasets it needs while following least privilege, and it reduces ongoing administrative work compared with managing separate IAM policies and S3 paths for every team.

QUESTION 17
A healthcare company stores patient records in an on-premises MySQL database. The company creates an application to access the MySQL database. The company must enforce security protocols to protect the patient records. The company currently rotates database credentials every 30 days to minimize the risk of unauthorized access.

The company wants a solution that does require the company to modify the application code for each credential rotation.

Which solution will meet this requirement with the LEAST operational overhead?

A. Assign an IAM role access permissions to the database. Configure the application to obtain temporary credentials through the IAM role.
B. Use AWS Key Management Service (AWS KMS) to generate encryption keys. Configure automatic key rotation. Store the encrypted credentials in an Amazon DynamoDB table.
C. Use AWS Secrets Manager to automatically rotate credentials. Allow the application to retrieve the credentials by using API calls.
D. Store credentials in an encrypted Amazon S3 bucket. Rotate the credentials every month by using an S3 Lifecycle policy. Use bucket policies to control access.

Answer: C
Explanation:
AWS Secrets Manager is purpose-built to store sensitive credentials and rotate them automatically without requiring the application code to be updated for each password change. The application retrieves the current secret dynamically through API calls, so credential rotation happens transparently while minimizing operational overhead.

QUESTION 18
A retail company needs to implement a solution to capture data updates from multiple Amazon Aurora MySQL databases. The company needs to make the updates available for analytics in near real time. The solution must be serverless and require minimal maintenance. Which solution will meet these requirements with the LEAST operational overhead?

A. Set up AWS Database Migration Service (AWS DMS) tasks that perform schema conversions for each database. Load the changes into Amazon Redshift Serverless.
B. Use Amazon Managed Streaming for Apache Kafka (Amazon MSK) Connect with Debezium connectors to load data into Amazon Redshift Serverless.
C. Use AWS Database Migration Service (AWS DMS) to set up binary log replication to Amazon Kinesis Data Streams. Load the data into Amazon Redshift Serverless after schema conversion.
D. Use Aurora zero-ETL integrations with Amazon Redshift Serverless for each database to load Aurora MySQL changes in Amazon Redshift Serverless.

Answer: D
Explanation:
Aurora zero-ETL integration with Amazon Redshift Serverless is the most direct fit because it delivers near real-time data availability for analytics with a fully managed approach and minimal maintenance. It avoids building and operating custom CDC pipelines, streaming infrastructure, or connector frameworks, which gives it the least operational overhead for multiple Aurora MySQL databases.

QUESTION 19
A data engineer is writing a query to join two tables in Amazon Athena. The data engineer needs to choose the correct join order for the tables to optimize query performance. Which solution will meet these requirements?

A. Specify the smaller table on the left side of the join and the larger table on the right side of the join.

B. Specify the larger table on the left side of the join and the smaller table on the right side of the join.
C. Use AWS Glue to pre-process the tables before performing the join.
D. Use table statistics to automatically determine the join order.

Answer: A
Explanation:
Amazon Athena (based on Presto/Trino) performs joins more efficiently when the smaller table is placed on the left side because it is used as the build side of the join and can be loaded into memory. This minimizes data shuffling and improves performance when joining with a larger table on the right side.

QUESTION 20
A company generates reports from 30 tables in an Amazon Redshift data warehouse. The data source is an operational Amazon Aurora MySQL database that contains 100 tables. Currently, the company refreshes all data from Aurora to Amazon Redshift every hour, which causes delays in report generation. Which combination of steps will meet these requirements with the LEAST operational overhead? (Choose two.)

A. Use AWS Database Migration Service (AWS DMS) to create a replication task. Select only the required tables.
B. Create a database in Amazon Redshift that uses the integration.
C. Create a zero-ETL integration in Amazon Aurora. Select only the required tables.
D. Use query editor v2 in Amazon Redshift to access the data in Aurora.
E. Create an AWS Glue job to transfer each required table. Run an AWS Glue workflow to initiate the jobs every 5 minutes.

Answer: BC
Explanation:
Aurora zero-ETL integration replicates operational data from Aurora into Amazon Redshift with low latency and minimal management overhead, and AWS supports filtering so only the required tables are replicated instead of all 100 tables. After the integration is created, Amazon Redshift requires a destination database to be created from that integration so the replicated data can be queried for reporting.

Reference:
https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/zero-etl.filtering.html
https://docs.aws.amazon.com/redshift/latest/mgmt/zero-etl-using.creating-db.html
https://docs.aws.amazon.com/AmazonRDS/latest/AuroraUserGuide/zero-etl.html

Thank You for Trying Our Product

Lead2pass Certification Exam Features:

★ More than **99,900** Satisfied Customers Worldwide.

★ Average **99.9%** Success Rate.

★ **Free Update** to match latest and real exam scenarios.

★ **Instant Download** Access! No Setup required.

★ Questions & Answers are downloadable in **PDF** format and **VCE** test engine format.

★ Multi-Platform capabilities - **Windows, Laptop, Mac, Android, iPhone, iPod, iPad**.

★ **100%** Guaranteed Success or **100%** Money Back Guarantee.

★ **Fast**, helpful support **24x7**.

View list of all certification exams: http://www.lead2pass.com/all-products.html

10% Discount Coupon Code: **ASTR14**