aws training and certification

AWS Certified Data Analytics – Specialty (DAS-C01) Exam Guide

Introduction

The AWS Certified Data Analytics – Specialty (DAS-C01) exam is intended for individuals who perform a data analytics role. The exam validates a candidate's comprehensive understanding of how to use AWS services to design, build, secure, and maintain analytics solutions that provide insight from data.

The exam also validates a candidate's ability to complete the following tasks:

- Define AWS data analytics services and understand how they integrate with each other
- Explain how AWS data analytics services fit in the data lifecycle of collection, storage, processing, and visualization

Target candidate description

The target candidate should have a minimum of 5 years of experience with common data analytics technologies. The target candidate also should have at least 2 years of hands-on experience and expertise working with AWS services to design, build, secure, and maintain analytics solutions.

What is considered out of scope for the target candidate?

The following is a non-exhaustive list of related job tasks that the target candidate is not expected to be able to perform. These items are considered out of scope for the exam:

- Design and implement machine learning algorithms
- Implement container-based solutions
- Utilize high performance computing (HPC)
- Design online transactional processing (OLTP) database solutions

For a detailed list of specific tools and technologies that might be covered on the exam, as well as lists of in-scope and out-of-scope AWS services, refer to the Appendix.

Exam content

Response types

There are two types of questions on the exam:

- **Multiple choice:** Has one correct response and three incorrect responses (distractors)
- **Multiple response:** Has two or more correct responses out of five or more response options

Select one or more responses that best complete the statement or answer the question. Distractors, or incorrect answers, are response options that a candidate with incomplete knowledge or skill might choose. Distractors are generally plausible responses that match the content area.

Unanswered questions are scored as incorrect; there is no penalty for guessing. The exam includes 50 questions that will affect your score.

Unscored content

The exam includes 15 unscored questions that do not affect your score. AWS collects information about candidate performance on these unscored questions to evaluate these questions for future use as scored questions. These unscored questions are not identified on the exam.

Exam results

The AWS Certified Data Analytics – Specialty (DAS-C01) exam is a pass or fail exam. The exam is scored against a minimum standard established by AWS professionals who follow certification industry best practices and guidelines.

Your results for the exam are reported as a scaled score of 100–1,000. The minimum passing score is 750. Your score shows how you performed on the exam as a whole and whether or not you passed. Scaled scoring models help equate scores across multiple exam forms that might have slightly different difficulty levels.

Your score report could contain a table of classifications of your performance at each section level. This information is intended to provide general feedback about your exam performance. The exam uses a compensatory scoring model, which means that you do not need to achieve a passing score in each section. You need to pass only the overall exam.

Each section of the exam has a specific weighting, so some sections have more questions than other sections have. The table contains general information that highlights your strengths and weaknesses. Use caution when interpreting section-level feedback.

Content outline

This exam guide includes weightings, test domains, and objectives for the exam. It is not a comprehensive listing of the content on the exam. However, additional context for each of the objectives is available to help guide your preparation for the exam. The following table lists the main content domains and their weightings. The table precedes the complete exam content outline, which includes the additional context. The percentage in each domain represents only scored content.

Domain	% of Exam
Domain 1: Collection	18%
Domain 2: Storage and Data Management	22%
Domain 3: Processing	24%
Domain 4: Analysis and Visualization	18%
Domain 5: Security	18%
TOTAL	**100%**

Domain 1: Collection

1.1 Determine the operational characteristics of the collection system
- Evaluate that the data loss is within tolerance limits in the event of failures
- Evaluate costs associated with data acquisition, transfer, and provisioning from various sources into the collection system (e.g., networking, bandwidth, ETL/data migration costs)
- Assess the failure scenarios that the collection system may undergo, and take remediation actions based on impact
- Determine data persistence at various points of data capture
- Identify the latency characteristics of the collection system

1.2 Select a collection system that handles the frequency, volume, and the source of data
- Describe and characterize the volume and flow characteristics of incoming data (streaming, transactional, batch)
- Match flow characteristics of data to potential solutions
- Assess the tradeoffs between various ingestion services taking into account scalability, cost, fault tolerance, latency, etc.
- Explain the throughput capability of a variety of different types of data collection and identify bottlenecks
- Choose a collection solution that satisfies connectivity constraints of the source data system

1.3 Select a collection system that addresses the key properties of data, such as order, format, and compression
- Describe how to capture data changes at the source
- Discuss data structure and format, compression applied, and encryption requirements
- Distinguish the impact of out-of-order delivery of data, duplicate delivery of data, and the tradeoffs between at-most-once, exactly-once, and at-least-once processing
- Describe how to transform and filter data during the collection process

Domain 2: Storage and Data Management

2.1 Determine the operational characteristics of the storage solution for analytics
- Determine the appropriate storage service(s) on the basis of cost vs. performance
- Understand the durability, reliability, and latency characteristics of the storage solution based on requirements
- Determine the requirements of a system for strong vs. eventual consistency of the storage system
- Determine the appropriate storage solution to address data freshness requirements

2.2 Determine data access and retrieval patterns
- Determine the appropriate storage solution based on update patterns (e.g., bulk, transactional, micro batching)
- Determine the appropriate storage solution based on access patterns (e.g., sequential vs. random access, continuous usage vs.ad hoc)
- Determine the appropriate storage solution to address change characteristics of data (append-only changes vs. updates)
- Determine the appropriate storage solution for long-term storage vs. transient storage
- Determine the appropriate storage solution for structured vs. semi-structured data
- Determine the appropriate storage solution to address query latency requirements

2.3 Select appropriate data layout, schema, structure, and format
- Determine appropriate mechanisms to address schema evolution requirements
- Select the storage format for the task
- Select the compression/encoding strategies for the chosen storage format
- Select the data sorting and distribution strategies and the storage layout for efficient data access
- Explain the cost and performance implications of different data distributions, layouts, and formats (e.g., size and number of files)
- Implement data formatting and partitioning schemes for data-optimized analysis

2.4 Define data lifecycle based on usage patterns and business requirements
- Determine the strategy to address data lifecycle requirements
- Apply the lifecycle and data retention policies to different storage solutions

2.5 Determine the appropriate system for cataloging data and managing metadata
- Evaluate mechanisms for discovery of new and updated data sources
- Evaluate mechanisms for creating and updating data catalogs and metadata
- Explain mechanisms for searching and retrieving data catalogs and metadata
- Explain mechanisms for tagging and classifying data

Domain 3: Processing

3.1 Determine appropriate data processing solution requirements
- Understand data preparation and usage requirements
- Understand different types of data sources and targets
- Evaluate performance and orchestration needs
- Evaluate appropriate services for cost, scalability, and availability

3.2 Design a solution for transforming and preparing data for analysis
- Apply appropriate ETL/ELT techniques for batch and real-time workloads
- Implement failover, scaling, and replication mechanisms
- Implement techniques to address concurrency needs
- Implement techniques to improve cost-optimization efficiencies
- Apply orchestration workflows
- Aggregate and enrich data for downstream consumption

3.3 Automate and operationalize data processing solutions
- Implement automated techniques for repeatable workflows
- Apply methods to identify and recover from processing failures
- Deploy logging and monitoring solutions to enable auditing and traceability

Domain 4: Analysis and Visualization

4.1 Determine the operational characteristics of the analysis and visualization solution
- Determine costs associated with analysis and visualization
- Determine scalability associated with analysis
- Determine failover recovery and fault tolerance within the RPO/RTO
- Determine the availability characteristics of an analysis tool
- Evaluate dynamic, interactive, and static presentations of data
- Translate performance requirements to an appropriate visualization approach (pre-compute and consume static data vs. consume dynamic data)

4.2 Select the appropriate data analysis solution for a given scenario
- Evaluate and compare analysis solutions
- Select the right type of analysis based on the customer use case (streaming, interactive, collaborative, operational)

4.3 Select the appropriate data visualization solution for a given scenario
- Evaluate output capabilities for a given analysis solution (metrics, KPIs, tabular, API)
- Choose the appropriate method for data delivery (e.g., web, mobile, email, collaborative notebooks)
- Choose and define the appropriate data refresh schedule
- Choose appropriate tools for different data freshness requirements (e.g., Amazon Elasticsearch Service vs. Amazon QuickSight vs. Amazon EMR notebooks)
- Understand the capabilities of visualization tools for interactive use cases (e.g., drill down, drill through and pivot)
- Implement the appropriate data access mechanism (e.g., in memory vs. direct access)
- Implement an integrated solution from multiple heterogeneous data sources

Domain 5: Security

5.1 Select appropriate authentication and authorization mechanisms
- Implement appropriate authentication methods (e.g., federated access, SSO, IAM)
- Implement appropriate authorization methods (e.g., policies, ACL, table/column level permissions)
- Implement appropriate access control mechanisms (e.g., security groups, role-based control)

5.2 Apply data protection and encryption techniques
- Determine data encryption and masking needs
- Apply different encryption approaches (server-side encryption, client-side encryption, AWS KMS, AWS CloudHSM)
- Implement at-rest and in-transit encryption mechanisms
- Implement data obfuscation and masking techniques
- Apply basic principles of key rotation and secrets management

5.3 Apply data governance and compliance controls
- Determine data governance and compliance requirements
- Understand and configure access and audit logging across data analytics services
- Implement appropriate controls to meet compliance requirements

Appendix

Which key tools, technologies, and concepts might be covered on the exam?

The following is a non-exhaustive list of the tools and technologies that could appear on the exam. This list is subject to change and is provided to help you understand the general scope of services, features, or technologies on the exam. AWS services are grouped according to their primary functions. While some of these technologies will likely be covered more than others on the exam, the order and placement of them in this list is no indication of relative weight or importance:

AWS services and features

Analytics:

- Amazon Athena
- Amazon CloudSearch
- Amazon Elasticsearch Service (Amazon ES)
- Amazon EMR
- AWS Glue
- Amazon Kinesis (excluding Kinesis Video Streams)
- AWS Lake Formation
- Amazon Managed Streaming for Apache Kafka
- Amazon QuickSight
- Amazon Redshift

Application Integration:

- Amazon MQ
- Amazon Simple Notification Service (Amazon SNS)
- Amazon Simple Queue Service (Amazon SQS)
- AWS Step Functions

Compute:

- Amazon EC2
- Elastic Load Balancing
- AWS Lambda

Customer Engagement:

- Amazon Simple Email Service (Amazon SES)

Database:

- Amazon DocumentDB (with MongoDB compatibility)
- Amazon DynamoDB
- Amazon ElastiCache
- Amazon Neptune
- Amazon RDS
- Amazon Redshift
- Amazon Timestream

Management and Governance:

- AWS Auto Scaling
- AWS CloudFormation
- AWS CloudTrail
- Amazon CloudWatch
- AWS Trusted Advisor

Machine Learning:

- Amazon SageMaker

Migration and Transfer:

- AWS Database Migration Service (AWS DMS)
- AWS DataSync
- AWS Snowball
- AWS Transfer for SFTP

Networking and Content Delivery:

- Amazon API Gateway
- AWS Direct Connect
- Amazon VPC (and associated features)

Security, Identity, and Compliance:

- AWS AppSync
- AWS Artifact
- AWS Certificate Manager (ACM)
- AWS CloudHSM
- Amazon Cognito
- AWS Identity and Access Management (IAM)
- AWS Key Management Service (AWS KMS)
- Amazon Macie
- AWS Secrets Manager
- AWS Single Sign-On

Storage:

- Amazon Elastic Block Store (Amazon EBS)
- Amazon S3
- Amazon S3 Glacier

Out-of-scope AWS services and features

The following is a non-exhaustive list of AWS services and features that are not covered on the exam. These services and features do not represent every AWS offering that is excluded from the exam content. Services or features that are entirely unrelated to the target job roles for the exam are excluded from this list because they are assumed to be irrelevant.

Out-of-scope AWS services and features include the following:

- AWS IoT Core

1) A company ingests a large set of clickstream data in nested JSON format from different sources and stores it in Amazon S3. Data analysts need to analyze this data in combination with data stored in an Amazon Redshift cluster. Data analysts want to build a cost-effective and automated solution for this need.

Which solution meets these requirements?

A) Use Apache Spark SQL on Amazon EMR to convert the clickstream data to a tabular format. Use the Amazon Redshift COPY command to load the data into the Amazon Redshift cluster.
B) Use AWS Lambda to convert the data to a tabular format and write it to Amazon S3. Use the Amazon Redshift COPY command to load the data into the Amazon Redshift cluster.
C) Use the Relationalize class in an AWS Glue ETL job to transform the data and write the data back to Amazon S3. Use Amazon Redshift Spectrum to create external tables and join with the internal tables.
D) Use the Amazon Redshift COPY command to move the clickstream data directly into new tables in the Amazon Redshift cluster.

2) A publisher website captures user activity and sends clickstream data to Amazon Kinesis Data Streams. The publisher wants to design a cost-effective solution to process the data to create a timeline of user activity within a session. The solution must be able to scale depending on the number of active sessions.

Which solution meets these requirements?

A) Include a variable in the clickstream data from the publisher website to maintain a counter for the number of active user sessions. Use a timestamp for the partition key for the stream. Configure the consumer application to read the data from the stream and change the number of processor threads based upon the counter. Deploy the consumer application on Amazon EC2 instances in an EC2 Auto Scaling group.
B) Include a variable in the clickstream to maintain a counter for each user action during their session. Use the action type as the partition key for the stream. Use the Kinesis Client Library (KCL) in the consumer application to retrieve the data from the stream and perform the processing. Configure the consumer application to read the data from the stream and change the number of processor threads based upon the counter. Deploy the consumer application on AWS Lambda.
C) Include a session identifier in the clickstream data from the publisher website and use as the partition key for the stream. Use the Kinesis Client Library (KCL) in the consumer application to retrieve the data from the stream and perform the processing. Deploy the consumer application on Amazon EC2 instances in an EC2 Auto Scaling group. Use an AWS Lambda function to reshard the stream based upon Amazon CloudWatch alarms.
D) Include a variable in the clickstream data from the publisher website to maintain a counter for the number of active user sessions. Use a timestamp for the partition key for the stream. Configure the consumer application to read the data from the stream and change the number of processor threads based upon the counter. Deploy the consumer application on AWS Lambda.

3) A company is currently using Amazon DynamoDB as the database for a user support application. The company is developing a new version of the application that will store a PDF file for each support case ranging in size from 1–10 MB. The file should be retrievable whenever the case is accessed in the application.

How can the company store the file in the MOST cost-effective manner?

A) Store the file in Amazon DocumentDB and the document ID as an attribute in the DynamoDB table.
B) Store the file in Amazon S3 and the object key as an attribute in the DynamoDB table.
C) Split the file into smaller parts and store the parts as multiple items in a separate DynamoDB table.
D) Store the file as an attribute in the DynamoDB table using Base64 encoding.

4) A company needs to implement a near-real-time fraud prevention feature for its ecommerce site. User and order details need to be delivered to an Amazon SageMaker endpoint to flag suspected fraud. The amount of input data needed for the inference could be as much as 1.5 MB.

Which solution meets the requirements with the LOWEST overall latency?

A) Create an Amazon Managed Streaming for Kafka cluster and ingest the data for each order into a topic. Use a Kafka consumer running on Amazon EC2 instances to read these messages and invoke the Amazon SageMaker endpoint.
B) Create an Amazon Kinesis Data Streams stream and ingest the data for each order into the stream. Create an AWS Lambda function to read these messages and invoke the Amazon SageMaker endpoint.
C) Create an Amazon Kinesis Data Firehose delivery stream and ingest the data for each order into the stream. Configure Kinesis Data Firehose to deliver the data to an Amazon S3 bucket. Trigger an AWS Lambda function with an S3 event notification to read the data and invoke the Amazon SageMaker endpoint.
D) Create an Amazon SNS topic and publish the data for each order to the topic. Subscribe the Amazon SageMaker endpoint to the SNS topic.

5) A media company is migrating its on-premises legacy Hadoop cluster with its associated data processing scripts and workflow to an Amazon EMR environment running the latest Hadoop release. The developers want to reuse the Java code that was written for data processing jobs for the on-premises cluster.

Which approach meets these requirements?

A) Deploy the existing Oracle Java Archive as a custom bootstrap action and run the job on the EMR cluster.
B) Compile the Java program for the desired Hadoop version and run it using a CUSTOM_JAR step on the EMR cluster.
C) Submit the Java program as an Apache Hive or Apache Spark step for the EMR cluster.
D) Use SSH to connect the master node of the EMR cluster and submit the Java program using the AWS CLI.

6) An online retail company wants to perform analytics on data in large Amazon S3 objects using Amazon EMR. An Apache Spark job repeatedly queries the same data to populate an analytics dashboard. The analytics team wants to minimize the time to load the data and create the dashboard.

Which approaches could improve the performance? (Select TWO.)

A) Copy the source data into Amazon Redshift and rewrite the Apache Spark code to create analytical reports by querying Amazon Redshift.
B) Copy the source data from Amazon S3 into Hadoop Distributed File System (HDFS) using s3distcp.
C) Load the data into Spark DataFrames.
D) Stream the data into Amazon Kinesis and use the Kinesis Connector Library (KCL) in multiple Spark jobs to perform analytical jobs.
E) Use Amazon S3 Select to retrieve the data necessary for the dashboards from the S3 objects.

7) A data engineer needs to create a dashboard to display social media trends during the last hour of a large company event. The dashboard needs to display the associated metrics with a consistent latency of less than 2 minutes.

Which solution meets these requirements?

A) Publish the raw social media data to an Amazon Kinesis Data Firehose delivery stream. Use Kinesis Data Analytics for SQL Applications to perform a sliding window analysis to compute the metrics and output the results to a Kinesis Data Streams data stream. Configure an AWS Lambda function to save the stream data to an Amazon DynamoDB table. Deploy a real-time dashboard hosted in an Amazon S3 bucket to read and display the metrics data stored in the DynamoDB table.
B) Publish the raw social media data to an Amazon Kinesis Data Firehose delivery stream. Configure the stream to deliver the data to an Amazon Elasticsearch Service cluster with a buffer interval of 0 seconds. Use Kibana to perform the analysis and display the results.
C) Publish the raw social media data to an Amazon Kinesis Data Streams data stream. Configure an AWS Lambda function to compute the metrics on the stream data and save the results in an Amazon S3 bucket. Configure a dashboard in Amazon QuickSight to query the data using Amazon Athena and display the results.
D) Publish the raw social media data to an Amazon SNS topic. Subscribe an Amazon SQS queue to the topic. Configure Amazon EC2 instances as workers to poll the queue, compute the metrics, and save the results to an Amazon Aurora MySQL database. Configure a dashboard in Amazon QuickSight to query the data in Aurora and display the results.

8) A real estate company is receiving new property listing data from its agents through .csv files every day and storing these files in Amazon S3. The data analytics team created an Amazon QuickSight visualization report that uses a dataset imported from the S3 files. The data analytics team wants the visualization report to reflect the current data up to the previous day.

How can a data analyst meet these requirements?

A) Schedule an AWS Lambda function to drop and re-create the dataset daily.
B) Configure the visualization to query the data in Amazon S3 directly without loading the data into SPICE.
C) Schedule the dataset to refresh daily.
D) Close and open the Amazon QuickSight visualization.

9) A financial company uses Amazon EMR for its analytics workloads. During the company's annual security audit, the security team determined that none of the EMR clusters' root volumes are encrypted. The security team recommends the company encrypt its EMR clusters' root volume as soon as possible.

Which solution would meet these requirements?

A) Enable at-rest encryption for EMR File System (EMRFS) data in Amazon S3 in a security configuration. Re-create the cluster using the newly created security configuration.
B) Specify local disk encryption in a security configuration. Re-create the cluster using the newly created security configuration.
C) Detach the Amazon EBS volumes from the master node. Encrypt the EBS volume and attach it back to the master node.
D) Re-create the EMR cluster with LZO encryption enabled on all volumes.

10) A company is providing analytics services to its marketing and human resources (HR) departments. The departments can only access the data through their business intelligence (BI) tools, which run Presto queries on an Amazon EMR cluster that uses the EMR File System (EMRFS). The marketing data analyst must be granted access to the advertising table only. The HR data analyst must be granted access to the personnel table only.

Which approach will satisfy these requirements?

A) Create separate IAM roles for the marketing and HR users. Assign the roles with AWS Glue resource-based policies to access their corresponding tables in the AWS Glue Data Catalog. Configure Presto to use the AWS Glue Data Catalog as the Apache Hive metastore.
B) Create the marketing and HR users in Apache Ranger. Create separate policies that allow access to the user's corresponding table only. Configure Presto to use Apache Ranger and an external Apache Hive metastore running in Amazon RDS.
C) Create separate IAM roles for the marketing and HR users. Configure EMR to use IAM roles for EMRFS access. Create a separate bucket for the HR and marketing data. Assign appropriate permissions so the users will only see their corresponding datasets.
D) Create the marketing and HR users in Apache Ranger. Create separate policies that allows access to the user's corresponding table only. Configure Presto to use Apache Ranger and the AWS Glue Data Catalog as the Apache Hive metastore.

Answers

1) C – The Relationalize PySpark transform can be used to flatten the nested data into a structured format. Amazon Redshift Spectrum can join the external tables and query the transformed clickstream data in place rather than needing to scale the cluster to accommodate the large dataset.

2) C – Partitioning by the session ID will allow a single processor to process all the actions for a user session in order. An AWS Lambda function can call the UpdateShardCount API action to change the number of shards in the stream. The KCL will automatically manage the number of processors to match the number of shards. Amazon EC2 Auto Scaling will assure the correct number of instances are running to meet the processing load.

3) B – Use Amazon S3 to store large attribute values that cannot fit in an Amazon DynamoDB item. Store each file as an object in Amazon S3 and then store the object path in the DynamoDB item.

4) A – An Amazon Managed Streaming for Kafka cluster can be used to deliver the messages with very low latency. It has a configurable message size that can handle the 1.5 MB payload.

5) B – A CUSTOM JAR step can be configured to download a JAR file from an Amazon S3 bucket and execute it. Since the Hadoop versions are different, the Java application has to be recompiled.

6) C, E – One of the speed advantages of Apache Spark comes from loading data into immutable dataframes, which can be accessed repeatedly in memory. Spark DataFrames organizes distributed data into columns. This makes summaries and aggregates much quicker to calculate. Also, instead of loading an entire large Amazon S3 object, load only what is needed using Amazon S3 Select. Keeping the data in Amazon S3 avoids loading the large dataset into HDFS.

7) A – Amazon Kinesis Data Analytics can query data in a Kinesis Data Firehose delivery stream in near-real time using SQL. A sliding window analysis is appropriate for determining trends in the stream. Amazon S3 can host a static webpage that includes JavaScript that reads the data in Amazon DynamoDB and refreshes the dashboard.

8) C – Datasets created using Amazon S3 as the data source are automatically imported into SPICE. The Amazon QuickSight console allows for the refresh of SPICE data on a schedule.

9) B – Local disk encryption can be enabled as part of a security configuration to encrypt root and storage volumes.

10) A – AWS Glue resource policies can be used to control access to Data Catalog resources.