AWS Data Analytics Practice Test (Sample)

Study Guide

BY EXAMZIFY

Everything you need from our exam experts!

Sample study guide. For the full version with hundreds of questions, visit:
https://awsdataanalytics.examzify.com

Copyright © 2026 by Examzify - A Kaluba Technologies Inc. product.

ALL RIGHTS RESERVED.

No part of this book may be reproduced or transferred in any form or by any means, graphic, electronic, or mechanical, including photocopying, recording, web distribution, taping, or by any information storage retrieval system, without the written permission of the author.

Notice: Examzify makes every reasonable effort to obtain accurate, complete, and timely information about this product from reliable sources.

SAMPLE

Table of Contents

SAMPLE

Introduction

Preparing for a certification exam can feel overwhelming, but with the right tools, it becomes an opportunity to build confidence, sharpen your skills, and move one step closer to your goals. At Examzify, we believe that effective exam preparation isn’t just about memorization, it’s about understanding the material, identifying knowledge gaps, and building the test-taking strategies that lead to success.

This guide was designed to help you do exactly that.

Whether you’re preparing for a licensing exam, professional certification, or entry-level qualification, this book offers structured practice to reinforce key concepts. You’ll find a wide range of multiple-choice questions, each followed by clear explanations to help you understand not just the right answer, but why it’s correct.

The content in this guide is based on real-world exam objectives and aligned with the types of questions and topics commonly found on official tests. It’s ideal for learners who want to:

- **Practice answering questions under realistic conditions,**
- **Improve accuracy and speed,**
- **Review explanations to strengthen weak areas, and**
- **Approach the exam with greater confidence.**

We recommend using this book not as a stand-alone study tool, but alongside other resources like flashcards, textbooks, or hands-on training. For best results, we recommend working through each question, reflecting on the explanation provided, and revisiting the topics that challenge you most.

Remember: successful test preparation isn’t about getting every question right the first time, it’s about learning from your mistakes and improving over time. Stay focused, trust the process, and know that every page you turn brings you closer to success.

Let’s begin.

SAMPLE

How to Use This Guide

This guide is designed to help you study more effectively and approach your exam with confidence. Whether you're reviewing for the first time or doing a final refresh, here's how to get the most out of your Examzify study guide:

1. Start with a Diagnostic Review

Skim through the questions to get a sense of what you know and what you need to focus on. Your goal is to identify knowledge gaps early.

2. Study in Short, Focused Sessions

Break your study time into manageable blocks (e.g. 30 – 45 minutes). Review a handful of questions, reflect on the explanations.

3. Learn from the Explanations

After answering a question, always read the explanation, even if you got it right. It reinforces key points, corrects misunderstandings, and teaches subtle distinctions between similar answers.

4. Track Your Progress

Use bookmarks or notes (if reading digitally) to mark difficult questions. Revisit these regularly and track improvements over time.

5. Simulate the Real Exam

Once you're comfortable, try taking a full set of questions without pausing. Set a timer and simulate test-day conditions to build confidence and time management skills.

6. Repeat and Review

Don't just study once, repetition builds retention. Re-attempt questions after a few days and revisit explanations to reinforce learning. Pair this guide with other Examzify tools like flashcards, and digital practice tests to strengthen your preparation across formats.

There's no single right way to study, but consistent, thoughtful effort always wins. Use this guide flexibly, adapt the tips above to fit your pace and learning style. You've got this!

SAMPLE

Questions

SAMPLE

1. **Which formats can AWS Glue DataBrew output data in?**
   - A. Only XML and PDF
   - B. CSV, JSON, and Parquet
   - C. Text and HTML
   - D. Only Avro format

2. **What approach could improve query performance for a streaming application tied to Amazon Kinesis Data Streams?**
   - A. Increase the number of shards in Kinesis.
   - B. Run queries on a distributed cluster.
   - C. Merge the files in Amazon S3 to form larger files for more efficient querying.
   - D. Scale up the memory and CPU resources of the streaming application.

3. **What is the primary function of AWS Data Wrangler?**
   - A. To perform complex machine learning tasks
   - B. To manage network configurations for analytics
   - C. To simplify the process of data manipulation and ETL with Pandas integration
   - D. To enhance data visualization capabilities

4. **How can security for data in transit be ensured in AWS?**
   - A. By using API Gateway
   - B. By applying encryption protocols like TLS/SSL
   - C. By limiting network access to private IPs
   - D. By employing AWS Identity and Access Management (IAM)

5. **How does Amazon Redshift handle sudden increases in query traffic?**
   - A. By allocating additional user roles
   - B. By utilizing machine learning algorithms
   - C. Through concurrency scaling
   - D. By upgrading storage automatically

SAMPLE

6. **What method should a software company use to collect and analyze logs from EC2 instances after each deployment to ensure performance?**
   - A. Store logs directly in Amazon S3 for later analysis.
   - B. Use Amazon CloudWatch to monitor the logs closely.
   - C. Utilize the Amazon Kinesis Producer Library (KPL) agent to send data to Kinesis Data Firehose and visualize results.
   - D. Configure EC2 instances to send logs to AWS Lambda for real-time processing.

7. **What is a key benefit of using AWS Glue?**
   - A. Low-cost cloud storage
   - B. Fully managed ETL service
   - C. Real-time monitoring of database performance
   - D. On-premise database replication

8. **What is the first step a data analyst should take to load sensitive data from DynamoDB into Amazon Redshift securely?**
   - A. Create an AWS Lambda function to process the DynamoDB stream
   - B. Use IAM roles to control access to the data
   - C. Encrypt the data using AWS KMS managed keys
   - D. Save the output to an unprotected S3 bucket

9. **What is the primary purpose of Amazon Managed Streaming for Apache Kafka (MSK)?**
   - A. Building and running applications using Apache Kafka for real-time data processing
   - B. Managing ETL processes on large datasets
   - C. Providing relational database services for transactional data
   - D. Storing and analyzing log data in real-time

10. **Which tools does AWS provide for batch data processing?**
    - A. AWS Glue and Amazon S3
    - B. AWS Batch and Amazon EMR
    - C. Amazon Comprehend and Amazon OpenSearch
    - D. Amazon Redshift and AWS Lambda

SAMPLE

Answers

SAMPLE

1. B
2. C
3. C
4. B
5. C
6. C
7. B
8. A
9. A
10. B

SAMPLE

Explanations

SAMPLE

1. Which formats can AWS Glue DataBrew output data in?

A. Only XML and PDF

B. CSV, JSON, and Parquet

C. Text and HTML

D. Only Avro format

AWS Glue DataBrew supports outputting data in CSV, JSON, and Parquet formats. Each of these formats is commonly used for data storage and exchange. CSV (Comma-Separated Values) is widely utilized for its simplicity and ease of use in many applications, particularly for tabular data. JSON (JavaScript Object Notation) is popular for structured data storage and APIs, enabling easy integration with web applications and programming languages. Parquet, on the other hand, is an efficient columnar storage format that is optimized for use with big data processing frameworks like Apache Spark, allowing for better performance in analytical queries. The ability to output data in these formats means that users can easily work with their cleaned and transformed data in various downstream processes, such as analytics, machine learning, and integration with other AWS services. This versatility enhances the overall usability of AWS Glue DataBrew within data workflows. The other formats mentioned in the options do not match the capabilities of DataBrew, as the tool does not support XML, PDF, Avro, or plain text formats for output.

2. What approach could improve query performance for a streaming application tied to Amazon Kinesis Data Streams?

A. Increase the number of shards in Kinesis.

B. Run queries on a distributed cluster.

C. Merge the files in Amazon S3 to form larger files for more efficient querying.

D. Scale up the memory and CPU resources of the streaming application.

Increasing the number of shards in Kinesis can significantly enhance query performance for a streaming application. Each shard in a Kinesis Data Stream can handle a certain amount of read and write throughput, so by increasing the number of shards, you increase the overall capacity to read from the stream concurrently. This is especially useful for applications with high throughput requirements, allowing them to handle more data and to parallelize the processing. Running queries on a distributed cluster can also improve query performance, as it allows for parallel processing and can manage larger datasets more efficiently. However, this approach is typically more relevant for data stored in databases or data lakes rather than directly tied to the stream. Merging smaller files in Amazon S3 into larger ones can optimize query performance for analytics frameworks that read from S3, as larger files reduce the overhead of managing numerous objects. This is especially relevant when leveraging services like Amazon Athena or Amazon Redshift Spectrum, but it does not directly apply to query performance improvements within the Kinesis streaming context, as Kinesis handles data differently. Scaling up the memory and CPU resources of the streaming application can enhance the performance of specific processing tasks but may not address the concurrent read limitations of Kinesis Data Streams directly. Thus, increasing the number of shards allows

SAMPLE

3. What is the primary function of AWS Data Wrangler?

A. To perform complex machine learning tasks

B. To manage network configurations for analytics

C. To simplify the process of data manipulation and ETL with Pandas integration

D. To enhance data visualization capabilities

The primary function of AWS Data Wrangler is to simplify the process of data manipulation and extract, transform, load (ETL) tasks using integrated capabilities with Pandas, which is a powerful data analysis library in Python. This tool allows data engineers and data scientists to work seamlessly with AWS data services like Amazon S3, Amazon Redshift, and AWS Glue while performing data operations typically handled by Pandas. The integration with Pandas means that users can leverage familiar Python functions and methods for data manipulation, making it easier to preprocess and analyze data directly in the AWS cloud environment. This is particularly beneficial for tasks such as cleaning data, transforming datasets, and preparing data for further analytics or machine learning models without the need for extensive boilerplate code or complex configurations. While machine learning tasks are significant in the cloud ecosystem (as mentioned in one of the options), AWS Data Wrangler is not specifically designed for executing complex machine learning processes. Its purpose is more centered on data preparation and transformation. Similarly, managing network configurations for analytics and enhancing data visualization capabilities are beyond the scope of AWS Data Wrangler's functionalities. The tool is tailored towards data handling before analysis rather than networking or visualization.

4. How can security for data in transit be ensured in AWS?

A. By using API Gateway

B. By applying encryption protocols like TLS/SSL

C. By limiting network access to private IPs

D. By employing AWS Identity and Access Management (IAM)

Ensuring security for data in transit involves protecting data as it moves between locations, such as between a client and a server. The use of encryption protocols like TLS (Transport Layer Security) and SSL (Secure Sockets Layer) is a critical method for achieving this. These protocols encrypt the data, making it unreadable to anyone who might intercept it while it is in transit. Encryption ensures that even if the data packets are captured by unauthorized parties, they cannot decipher the contents without the appropriate cryptographic keys. Utilizing these protocols establishes a secure connection, commonly seen in HTTPS for web traffic, thereby protecting sensitive information such as personal data, payment information, and other confidential communications. This method is widely adopted across various AWS services and is fundamental for maintaining confidentiality and integrity of data during transmission. Other options may contribute to an overall security strategy but do not specifically address the need for encryption in transit. For instance, API Gateway is instrumental for managing APIs but does not inherently manage data encryption. Limiting network access to private IPs enhances the security of the network environment but does not safeguard data moving across those networks. Employing AWS Identity and Access Management (IAM) is vital for managing access rights and permissions but does not directly protect data during transmission.

SAMPLE

5. How does Amazon Redshift handle sudden increases in query traffic?

A. By allocating additional user roles

B. By utilizing machine learning algorithms

C. Through concurrency scaling

D. By upgrading storage automatically

Amazon Redshift handles sudden increases in query traffic through concurrency scaling, which is a feature designed to automatically add additional processing power to manage spikes in query demand without impacting performance. When the number of concurrent queries exceeds the capacity of the cluster, Redshift can provision additional clusters to handle the excess workload temporarily. This allows users to maintain fast query performance and ensures that users experience minimal wait times during peak periods. Concurrency scaling allows Redshift to dynamically adapt to varying workloads, which is crucial for businesses that may experience unpredictable traffic patterns. This feature operates transparently, so users would not have to make changes to their queries or configurations to benefit from the additional capacity. Instead, Redshift manages these resources automatically, reducing the complexity involved in scaling infrastructure. Other methods, while they may be related to Redshift's operation, do not specifically address handling sudden increases in query traffic in the same effective manner as concurrency scaling does.

6. What method should a software company use to collect and analyze logs from EC2 instances after each deployment to ensure performance?

A. Store logs directly in Amazon S3 for later analysis.

B. Use Amazon CloudWatch to monitor the logs closely.

C. Utilize the Amazon Kinesis Producer Library (KPL) agent to send data to Kinesis Data Firehose and visualize results.

D. Configure EC2 instances to send logs to AWS Lambda for real-time processing.

Utilizing the Amazon Kinesis Producer Library (KPL) agent to send data to Kinesis Data Firehose for visualization stands out as a highly effective method for collecting and analyzing logs from EC2 instances after each deployment. This approach is particularly advantageous because it allows for real-time data ingestion, enabling the software company to capture logs as they are generated and process them immediately. Kinesis Data Firehose integrates seamlessly with other AWS services, allowing for easy storage and analysis of logs. By visualizing results in real time, the company can monitor performance immediately after deployment, react to any issues swiftly, and gain insights into application behavior. This capability supports proactive management of application performance and enhances the decision-making process based on current data. While storing logs directly in Amazon S3 is a viable option for archival and later analysis, it may not provide the immediacy needed for active performance monitoring post-deployment. Monitoring logs closely with Amazon CloudWatch is beneficial, but it typically focuses on metrics and alarms rather than the deep analysis of log data. Configuring EC2 instances to send logs to AWS Lambda for real-time processing is another approach, but it may involve more overhead in terms of setup and management compared to Kinesis, especially for high throughput scenarios. Overall,

SAMPLE

7. What is a key benefit of using AWS Glue?

A. Low-cost cloud storage

B. Fully managed ETL service

C. Real-time monitoring of database performance

D. On-premise database replication

AWS Glue is a fully managed Extract, Transform, Load (ETL) service that automates the process of preparing and loading data for analytics. One of its key benefits is its ability to handle ETL tasks without requiring users to manage the underlying infrastructure. This means users can focus on data processing and analysis rather than worrying about server maintenance or scaling. By being fully managed, AWS Glue simplifies tasks such as data discovery, schema evolution, and providing a serverless environment where resources automatically scale based on demand. It integrates seamlessly with various AWS data services, such as Amazon S3, Amazon Redshift, and Amazon RDS, making it easy to transform and move data across different platforms. This integration capability enhances workflow efficiency, especially for data analytics and building data lakes. The other options, while they address features that may be part of the broader AWS service offering, do not specifically highlight the unique automation and management capabilities of AWS Glue that make it distinct as an ETL service. AWS Glue's primary value comes from streamlining the ETL process in a fully managed way.

8. What is the first step a data analyst should take to load sensitive data from DynamoDB into Amazon Redshift securely?

A. Create an AWS Lambda function to process the DynamoDB stream

B. Use IAM roles to control access to the data

C. Encrypt the data using AWS KMS managed keys

D. Save the output to an unprotected S3 bucket

The most effective first step for securely loading sensitive data from DynamoDB into Amazon Redshift is to use AWS Lambda to process DynamoDB streams. This approach allows for a serverless way to react to changes in your DynamoDB tables, ensuring that any data handling occurs in real time and securely. Using a Lambda function offers a controlled environment for accessing data. It can integrate tightly with both DynamoDB and Redshift, allowing you to extract, transform, and load (ETL) data efficiently. Furthermore, Lambda can be configured to access only the resources necessary for the task at hand, minimizing exposure and enhancing security. While options like using IAM roles and encrypting data with AWS KMS are important for overall security and access control, they are supplementary steps that enhance security rather than being the initial action for loading data. An unprotected S3 bucket should be avoided for sensitive data storage as it poses significant security risks.

SAMPLE

9. What is the primary purpose of Amazon Managed Streaming for Apache Kafka (MSK)?

A. Building and running applications using Apache Kafka for real-time data processing

B. Managing ETL processes on large datasets

C. Providing relational database services for transactional data

D. Storing and analyzing log data in real-time

The primary purpose of Amazon Managed Streaming for Apache Kafka (MSK) is indeed focused on building and running applications that utilize Apache Kafka for real-time data processing. Apache Kafka is a widely-used distributed event streaming platform that excels at handling real-time data feeds, which are essential for various applications such as real-time analytics, monitoring, and data integration. With Amazon MSK, users can easily set up, operate, and scale Apache Kafka clusters. It abstracts much of the administration and operational overhead involved in running Kafka, allowing developers to concentrate more on building and deploying their applications rather than on the intricacies of managing the underlying infrastructure. This service is particularly beneficial for companies looking to leverage real-time data streaming in their architectures without needing to deep-dive into the maintenance of Kafka clusters themselves. The other options pertain to different functionalities that are not the focus of Amazon MSK. For example, managing ETL processes typically involves tools designed for extracting, transforming, and loading data rather than real-time streaming. Similarly, providing relational database services is outside the realm of Kafka's purpose, as Kafka is not a relational database. Finally, while Kafka can indeed be used to process log data in real-time, its main utility is much broader than just handling logs;

10. Which tools does AWS provide for batch data processing?

A. AWS Glue and Amazon S3

B. AWS Batch and Amazon EMR

C. Amazon Comprehend and Amazon OpenSearch

D. Amazon Redshift and AWS Lambda

AWS provides specific tools that are designed to handle batch data processing effectively. The chosen answer highlights both AWS Batch and Amazon EMR as suitable solutions for this purpose. AWS Batch is a service that enables you to run batch computing workloads on AWS easily by dynamically provisioning the optimal quantity and type of compute resources based on the volume and resource requirements of the batch jobs you submit. This capability allows users to run batch jobs without having to manage the underlying infrastructure. Amazon EMR (Elastic MapReduce) is another powerful tool specifically designed for processing large amounts of data quickly and cost-effectively using tools like Apache Hadoop, Apache Spark, and others. It simplifies the setup and scaling of big data frameworks and allows for efficient processing of vast datasets. In contrast, other options either include tools that are not primarily designed for batch processing or mix components that focus on different aspects of data handling. For example, AWS Glue is a serverless data integration service that is often used for ETL (extract, transform, load) jobs rather than batch processing per se. Amazon S3 serves as a storage solution and isn't a processing tool. Similarly, while Amazon Comprehend and Amazon OpenSearch are powerful for natural language processing and search capabilities, they don't focus on batch data

SAMPLE

Next Steps

Congratulations on reaching the final section of this guide. You've taken a meaningful step toward passing your certification exam and advancing your career.

As you continue preparing, remember that consistent practice, review, and self-reflection are key to success. Make time to revisit difficult topics, simulate exam conditions, and track your progress along the way.

If you need help, have suggestions, or want to share feedback, we’d love to hear from you. Reach out to our team at hello@examzify.com.

Or visit your dedicated course page for more study tools and resources:

https://awsdataanalytics.examzify.com

We wish you the very best on your exam journey. You've got this!

SAMPLE