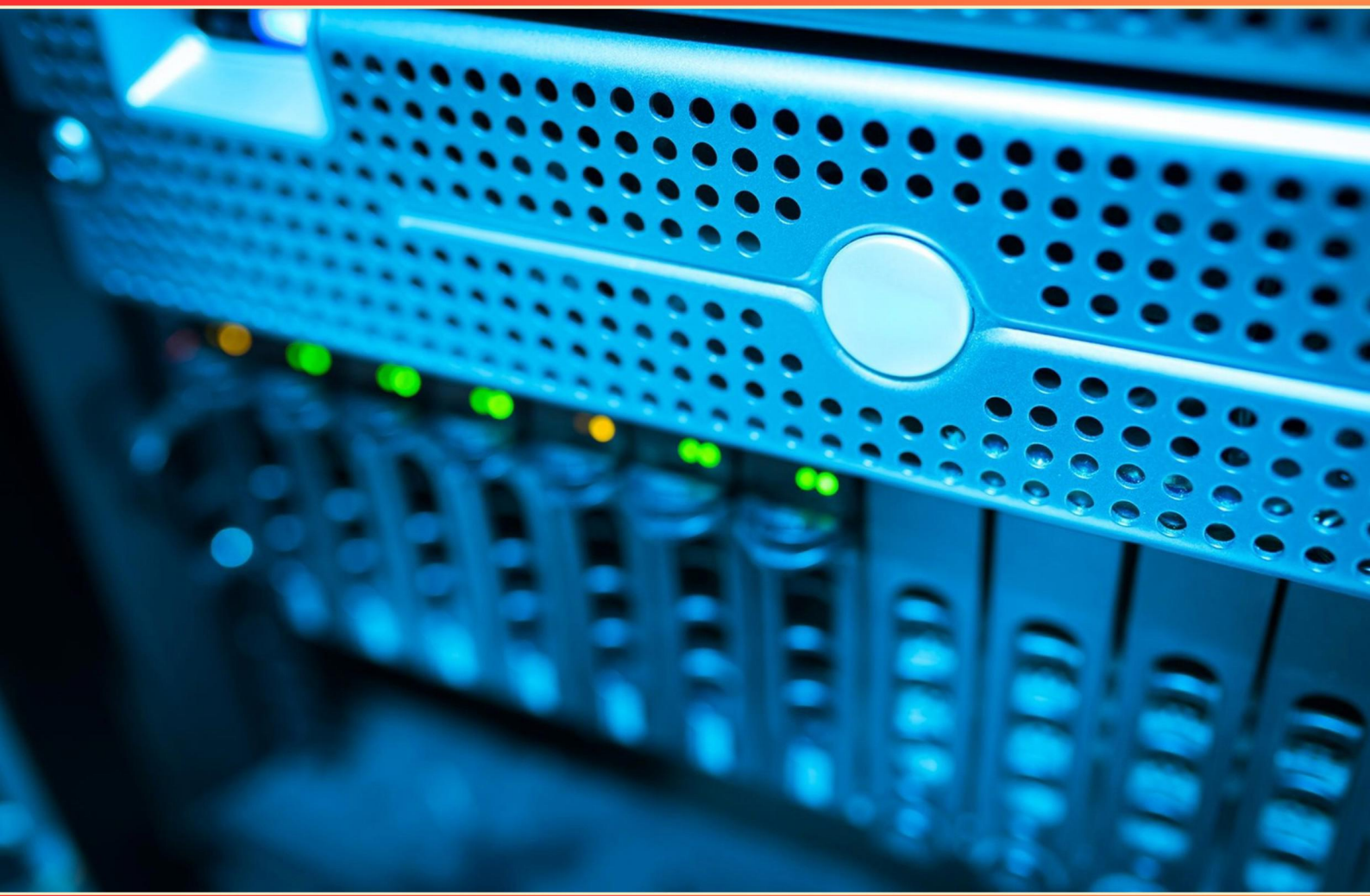
AWS ACADEMY DATA ENGINEERING PRACTICE TEST (SAMPLE)

STUDY GUIDE



SAMPLE STUDY GUIDE WITH LIMITED QUESTIONS

VISIT US ONLINE FOR THE FULL VERSION STUDY GUIDE

https://awsacademydataengr.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

Copyright 1
Table of Contents 2
Introduction 3
How to Use This Guide 4
Questions 5
Answers 8
Explanations 10
Next Steps 16

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 tool is described as an open-source, in-memory structured query language (SQL) query engine?**

   A. Hadoop

   B. Apache Presto

   C. Apache Hive

   D. Amazon Athena

2. **What feature of AWS Lambda allows it to be truly serverless?**

   A. Automatic scaling based on demand

   B. Integration with EC2 instances

   C. Pre-provisioned resources

   D. Static resource allocation

3. **What is the primary purpose of AWS Snowball?**

   A. Transferring large amounts of data securely

   B. Backup and recovery of data

   C. Data synchronization across regions

   D. Real-time data streaming

4. **Which of the following best describes Amazon DynamoDB?**

   A. A relational database service

   B. A fully-managed NoSQL database service

   C. A data warehouse solution

   D. An analytics service

5. **Which AWS service is best suited for ingesting data from a software as a service (SaaS) application?**

   A. Amazon S3

   B. Amazon AppFlow

   C. Amazon Kinesis

   D. Amazon RDS

SAMPLE

6. Which stages are part of every modern data pipeline?

A. Storage and processing

B. Analysis and visualization

C. Rechecking and auditing

D. Storage and analysis

7. What is the main advantage of using Apache Spark for machine learning?

A. Cost-effectiveness

B. Built-in workflow management

C. Better support for batch processing

D. High performance for iterative algorithms

8. Which tool helps in data transformation jobs in AWS?

A. AWS Glue

B. AWS Lambda

C. Amazon EC2

D. Amazon RDS

9. Which database would be BEST suited for a high-traffic computer game's leaderboard?

A. Amazon RDS

B. Amazon DynamoDB

C. Amazon Aurora

D. Amazon S3

10. What is NOT a valid data preprocessing strategy?

A. Normalization

B. Scaling

C. Compressing

D. Encoding

SAMPLE

Answers

SAMPLE

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

SAMPLE

Explanations

SAMPLE

1. Which tool is described as an open-source, in-memory structured query language (SQL) query engine?

A. Hadoop

B. Apache Presto

C. Apache Hive

D. Amazon Athena

Apache Presto is accurately described as an open-source, in-memory structured query language (SQL) query engine. It was designed specifically to handle large amounts of data and to execute fast analytical queries across various data sources. Presto allows users to run SQL queries over large datasets that reside in different locations, including Hadoop, AWS S3, and other data types, and it does so with the ability to query data quickly in real-time. Being an in-memory engine, Presto optimizes query execution performance by keeping data in memory, which reduces the need for disk I/O operations and allows for faster data processing. This capability is particularly beneficial for interactive and ad-hoc analysis, making it a preferred choice for analytics workloads that require low latency. Other options like Hadoop and Apache Hive are more focused on distributed data storage and processing, with Hive specifically using a different approach to SQL-like querying by converting SQL queries into MapReduce jobs, which are typically slower than Presto's in-memory execution. Amazon Athena, while it utilizes Presto, is a serverless service focused on querying data stored in S3 rather than being a standalone query engine itself. Thus, Presto stands out as the tool that meets the criteria laid out in the question.

2. What feature of AWS Lambda allows it to be truly serverless?

A. Automatic scaling based on demand

B. Integration with EC2 instances

C. Pre-provisioned resources

D. Static resource allocation

The characteristic of AWS Lambda that allows it to be truly serverless is automatic scaling based on demand. This feature means that AWS Lambda can automatically handle the number of requests and execute functions in response to these requests without requiring manual intervention or the need to provision or manage servers. As usage increases or decreases, Lambda can adapt by automatically scaling the resources needed to handle incoming requests, ensuring that performance remains efficient and cost-effective. This serverless architecture abstracts the underlying server infrastructure away from the user. Users do not need to worry about capacity planning, scaling, or maintenance of the servers, allowing them to focus on writing code and building applications. The operational burden of managing the infrastructure is offloaded to AWS, enabling developers to build and deploy applications more rapidly and efficiently. Other options, while they may have relevance in different contexts, do not capture the essence of what makes Lambda serverless. For instance, integration with EC2 instances does not inherently confer serverless capabilities, as EC2 involves managing instances. Pre-provisioned resources and static resource allocation also suggest a level of management and operational overhead that is contrary to the core serverless principle of automatically managing scaling and capacity based on demand.

SAMPLE

3. What is the primary purpose of AWS Snowball?

A. Transferring large amounts of data securely

B. Backup and recovery of data

C. Data synchronization across regions

D. Real-time data streaming

The primary purpose of AWS Snowball is indeed to facilitate the secure transfer of large amounts of data. AWS Snowball is a physical device provided by Amazon Web Services that enables organizations to move massive data sets into and out of the AWS cloud efficiently. When organizations face limitations with network bandwidth or require moving significant data quickly, Snowball becomes an invaluable solution by providing a secure and swift method of transporting data. The device is designed to be rugged and secure, often equipped with encryption to ensure data safety during transit. Users can load their data onto the Snowball device, and then ship it back to AWS, where the data is uploaded directly into services like Amazon S3. This process is particularly useful for data collection from remote locations or for organizations with compliance requirements that necessitate controlled physical transport of sensitive data. In contrast to other options, Snowball is not primarily intended for backup and recovery, which typically involves different strategies and services tailored for data redundancy and restoration. Similarly, data synchronization across regions is a different use case, often handled by services like AWS DataSync or S3 replication. Real-time data streaming, on the other hand, involves continuous data flow and is facilitated by services such as Amazon Kinesis, not Snowball.

4. Which of the following best describes Amazon DynamoDB?

A. A relational database service

B. A fully-managed NoSQL database service

C. A data warehouse solution

D. An analytics service

Amazon DynamoDB is best described as a fully-managed NoSQL database service. This designation highlights several key attributes of DynamoDB. As a NoSQL database, it is designed to handle unstructured and semi-structured data, providing a flexible schema that is particularly suitable for large-scale applications with varying data types. The fully-managed aspect means that AWS handles the operational overhead, including hardware provisioning, software patching, setup, configuration, and backups, allowing users to focus more on application development rather than database management. Moreover, DynamoDB offers high availability and scalability, automatically adjusting to accommodate varying workloads and ensuring quick and consistent performance regardless of the amount of data or the number of users accessing it. The service also provides features such as automatic scaling, built-in security, and in-memory caching, making it a versatile choice for applications that require low-latency data access and can benefit from a distributed NoSQL architecture.

SAMPLE

5. Which AWS service is best suited for ingesting data from a software as a service (SaaS) application?

A. Amazon S3

B. Amazon AppFlow

C. Amazon Kinesis

D. Amazon RDS

Amazon AppFlow is the most appropriate service for ingesting data from a software as a service (SaaS) application due to its specialized design for seamlessly connecting SaaS applications with AWS services. This managed integration service allows users to securely transfer data between various SaaS applications and AWS services with minimal coding and setup. It supports the connection to a wide range of popular SaaS applications, making data ingestion straightforward and efficient. By providing built-in connectors, AppFlow can handle data transformations and facilitate features like filtering and mapping before it moves the data to its destination, such as Amazon S3 or other AWS services. This integration capability simplifies the process of automating the flow of data, which can be particularly useful for businesses looking to analyze and utilize data from multiple sources without heavy lifting or complicated ETL (Extract, Transform, Load) processes. In contrast, while Amazon S3 is primarily used for object storage and can receive data, it does not have the specialized capabilities needed for direct integration with SaaS applications like AppFlow. Similarly, Amazon Kinesis is more suited for real-time data streaming and processing and is not specifically designed for SaaS data ingestion. Amazon RDS is a relational database service that can store data but does not provide the direct,

6. Which stages are part of every modern data pipeline?

A. Storage and processing

B. Analysis and visualization

C. Rechecking and auditing

D. Storage and analysis

In a modern data pipeline, the stages of storage and processing are fundamental components that enable the effective handling and transformation of data. Storage refers to the methods and technologies used to retain data in a way that it can be easily accessed and used later. This is essential for ensuring that data is available for processing, analysis, and eventual consumption. Various storage options are utilized in modern pipelines, including cloud storage solutions like Amazon S3 and databases, which ensure that data is both scalable and reliable. Processing involves the transformation and manipulation of the stored data. This step can include data cleaning, aggregation, and enrichment, allowing raw data to be converted into a format that is useful for analysis and other applications. Tools and frameworks, such as Apache Spark and AWS Glue, are often employed to efficiently handle and process large volumes of data. Together, these stages form the backbone of any modern data pipeline, enabling the pathway from raw data collection to meaningful insights, which are critical for informed decision-making in organizations. Other stages, while important, are not universally required in every pipeline; therefore, storage and processing stand as essential to all modern data workflows.

SAMPLE

7. What is the main advantage of using Apache Spark for machine learning?

A. Cost-effectiveness

B. Built-in workflow management

C. Better support for batch processing

D. High performance for iterative algorithms

The primary advantage of using Apache Spark for machine learning lies in its high performance for iterative algorithms. Apache Spark is designed to handle large-scale data processing effectively, and its architecture allows for execution of tasks in memory. This capability significantly enhances the efficiency of iterative algorithms, which are common in many machine learning applications. For instance, when training models, especially those that require multiple passes over the dataset, such as gradient descent methods, Spark's ability to keep data in memory reduces the overhead of reading and writing data to disk repeatedly, resulting in faster computations and smoother experience when dealing with iterative processes. Additionally, Spark's distributed computing framework allows it to scale out across multiple nodes, further improving the performance and speed of machine learning tasks. This makes Spark particularly useful for handling large datasets and complex computations inherent in machine learning workflows, leading to faster model training and evaluation cycles.

8. Which tool helps in data transformation jobs in AWS?

A. AWS Glue

B. AWS Lambda

C. Amazon EC2

D. Amazon RDS

AWS Glue is specifically designed for data transformation tasks within the AWS ecosystem. It is a fully managed extract, transform, load (ETL) service that enables users to prepare and transform data for analytics. With AWS Glue, you can create data transformation jobs that allow you to manipulate your data from various sources, cleaning and enriching it as needed, and then loading it into a target data store such as Amazon S3 or a data warehouse like Amazon Redshift. The service includes features such as automatic schema discovery and a visual interface that simplifies the process of designing ETL workflows. Additionally, AWS Glue supports various programming languages like Python and Scala, giving you flexibility in how you implement your transformations. This makes it particularly suitable for tasks that require ongoing data preparation as part of larger analytics pipelines. Other options mentioned serve different purposes: AWS Lambda is a serverless computing service for running code in response to events, Amazon EC2 provides resizable compute capacity in the cloud, and Amazon RDS is a managed relational database service. While these services can play a role in data workflows, they are not primarily focused on data transformation jobs in the same way that AWS Glue is.

SAMPLE

9. Which database would be BEST suited for a high-traffic computer game's leaderboard?

A. Amazon RDS

B. Amazon DynamoDB

C. Amazon Aurora

D. Amazon S3

Choosing Amazon DynamoDB as the best database for a high-traffic computer game's leaderboard makes sense due to its ability to handle large volumes of read and write operations with low latency. DynamoDB is a fully managed NoSQL database service that is designed to provide fast and predictable performance, even under heavy workloads. In the context of a leaderboard, where there are frequent updates (e.g., player scores) and real-time reads (e.g., fetching the current top scores), DynamoDB's performance capabilities shine. It can scale automatically to meet the needs of fluctuating traffic, crucial for games that may experience sudden spikes in active users, particularly during events or launches. Moreover, DynamoDB offers features such as provisioned throughput and on-demand scaling, which are beneficial for workloads that can vary widely in usage. Its ability to automatically distribute data across multiple nodes ensures high availability and durability, making it a reliable choice for applications demanding high availability, such as gaming leaderboards. In contrast, while Amazon RDS and Amazon Aurora are relational database services that can handle structured data effectively, they may face limitations in terms of scaling with high traffic scenarios compared to a NoSQL solution like DynamoDB. Additionally, Amazon S3 is primarily an object storage service, unsuitable for

10. What is NOT a valid data preprocessing strategy?

A. Normalization

B. Scaling

C. Compressing

D. Encoding

Compression is not considered a valid data preprocessing strategy in the context of preparing data for machine learning or data analysis. Preprocessing typically involves techniques that enhance the quality and usability of the data for modeling purposes. Normalization adjusts the values into a specific range, scaling alters the range of features, and encoding converts categorical data into a numerical format suitable for algorithms. While compression can reduce the overall size of the dataset, it doesn't directly relate to modifying the data for improved performance in machine learning processes. Instead, it focuses on storage efficiency rather than preparing the data for analysis or learning. Thus, the other strategies listed—normalization, scaling, and encoding—are integral to ensuring the data is appropriately conditioned for training models and drawing meaningful insights.

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://awsacademydataengr.examzify.com

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

SAMPLE