HPC BIG DATA VETERAN DECK PRACTICE TEST (SAMPLE)

STUDY GUIDE



SAMPLE STUDY GUIDE WITH LIMITED QUESTIONS

VISIT US ONLINE FOR THE FULL VERSION STUDY GUIDE

https://hpcbigdatavetdeck.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. If streaming is embarrassingly parallel, what shape do you choose?

A. CPU

B. GPU

C. MemoryOptimized

D. DenseIO

2. What does MPP stand for in database technologies?

A. Massively Parallel Processing

B. Managed Process Provisioning

C. Machine Performance Prediction

D. Multi-Protocol Processing

3. What is the role of a Data Lake in Big Data architecture?

A. To process structured data

B. To store vast amounts of raw data in its native format

C. To analyze real-time data

D. To provide data governance

4. In Big Data modeling, what does "schema on read" mean?

A. Data schema is applied during data ingestion

B. Data schema is defined before data is stored

C. Data schema is applied at the time of data retrieval rather than at ingestion

D. Data schema is fixed and cannot change

5. In Big Data, what does "volume" represent?

A. The complexity of data relationships

B. The amount of data being generated and stored

C. The variety of data formats available

D. The cost associated with data storage

6. What is the primary use of the VNCViewer?

A. File transfer

B. Remote desktop access

C. Network throughput monitoring

D. Security auditing

SAMPLE

7. Which cloud platform provides support for various Hadoop distributions, including Cloudera?

A. AWS

B. Google Cloud Platform

C. Oracle Cloud Infrastructure

D. Microsoft Azure

8. What role does a data engineer play in Big Data projects?

A. Designing and implementing machine learning algorithms

B. Creating user-facing data dashboards

C. Designing and maintaining the data infrastructure and pipelines

D. Conducting data analysis and reporting

9. Which are recommended storage options on OCI for a Persistent Filesystem of approximately 3TB?

A. Local NVMe and Object Storage

B. Block volume and FSS

C. Archive storage and Object Storage

D. File system and Cold Storage

10. Which programming language is often associated with Apache Spark?

A. Python

B. Java

C. Scala

D. Ruby

SAMPLE

Answers

SAMPLE

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

SAMPLE

Explanations

SAMPLE

1. If streaming is embarrassingly parallel, what shape do you choose?

A. CPU

B. GPU

C. MemoryOptimized

D. DenseIO

When streaming data is described as embarrassingly parallel, it means that the tasks can be performed independently and can be processed simultaneously without requiring significant interaction or communication between them. In this context, choosing a GPU (Graphics Processing Unit) is particularly advantageous due to its architecture, which is designed to handle multiple parallel tasks efficiently. GPUs are optimized for parallel processing, as they consist of hundreds or thousands of smaller cores that can execute many threads simultaneously. This makes them well-suited to handle workloads where the same operation is applied to large datasets, such as those encountered in streaming data applications. Tasks such as video processing, real-time data analysis, and complex computations can all benefit from the inherent parallelism that GPUs provide. In contrast, other shapes such as CPUs, Memory-Optimized, or DenseIO are not as specialized for the high level of parallelism found in streaming scenarios. While CPUs are capable of performing parallel operations, they are typically more suited for tasks that require tight coupling or significant inter-task communication. Memory-Optimized configurations focus on high memory bandwidth for data-intensive applications, and DenseIO provides enhanced storage performance which is not necessarily aligned with the needs of embarrassingly parallel workloads. Thus, leveraging a GPU for embarrassingly parallel

2. What does MPP stand for in database technologies?

A. Massively Parallel Processing

B. Managed Process Provisioning

C. Machine Performance Prediction

D. Multi-Protocol Processing

The term MPP in database technologies stands for Massively Parallel Processing. This architecture is designed to handle large volumes of data by dividing tasks among multiple processors that can run simultaneously. In an MPP system, data is distributed across many nodes, each of which processes its segment of the data independently. This approach allows for significant performance improvements when running complex queries or performing large-scale data analysis, as it can take advantage of parallel processing to speed up computations. The efficiency of MPP systems makes them suitable for big data applications where large datasets and intricate queries are common, allowing for quicker response times and better resource utilization. This concept is particularly relevant in data warehousing and analytics applications where the ability to process data rapidly is crucial for insights and business intelligence.

SAMPLE

3. What is the role of a Data Lake in Big Data architecture?

A. To process structured data

B. To store vast amounts of raw data in its native format

C. To analyze real-time data

D. To provide data governance

A Data Lake plays a crucial role in Big Data architecture by serving as a centralized repository that can store vast amounts of raw data in its native format. This characteristic is fundamental because it enables organizations to collect data from diverse sources, including structured, semi-structured, and unstructured formats, without the need for preprocessing or transformation before storage. By doing so, a Data Lake supports various data types and allows data scientists and analysts to access this raw data later for different analytical purposes or transformations, ensuring flexibility in how the data can be utilized. This capacity to handle raw data is particularly vital in environments where data needs to be retained for future analysis or compliance reasons. Unlike traditional data storage solutions that require data to be organized in a specific manner, a Data Lake's structure allows for the exploration and utilization of data without predefined schemas, making it significantly more versatile and scalable in addressing Big Data needs.

4. In Big Data modeling, what does "schema on read" mean?

A. Data schema is applied during data ingestion

B. Data schema is defined before data is stored

C. Data schema is applied at the time of data retrieval rather than at ingestion

D. Data schema is fixed and cannot change

"Schema on read" refers to the practice where the structure of the data is interpreted and defined at the time of data retrieval, rather than during the initial data ingestion process. This approach allows for greater flexibility, as it enables users to adjust how they want to view or analyze the data when they access it, accommodating various analytical needs or changes in requirements without having to alter the underlying data itself. In large datasets where the nature of the data can be diverse and unpredictable, this method of dynamically interpreting the schema allows for quick and efficient querying and analysis, particularly suited to environments like big data where traditional schema enforcement can be limiting. This concept contrasts with approaches like "schema on write," where the data must fit a predefined structure before it is stored, potentially leading to constraints and complexities in data management and modifications.

5. In Big Data, what does "volume" represent?

A. The complexity of data relationships

B. The amount of data being generated and stored

C. The variety of data formats available

D. The cost associated with data storage

Volume in the context of Big Data refers specifically to the sheer amount of data being generated and stored. It encompasses the scale of data, which can be measured in petabytes, exabytes, and beyond. As organizations gather information from various sources such as social media, sensors, transactions, and more, the volume of data can grow exponentially. This significant increase in data requires advanced storage solutions and processing power, emphasizing the importance of managing and analyzing large datasets effectively. While the other concepts related to data, such as complexity, variety, and cost, are important aspects of Big Data, they do not define "volume." Complexity refers to the intricate connections and relationships within data, variety pertains to the different formats and types of data available, and costs are associated with storing and processing data. The defining characteristic of volume is solely about the quantity of data present.

SAMPLE

6. What is the primary use of the VNCViewer?

A. File transfer

B. Remote desktop access

C. Network throughput monitoring

D. Security auditing

VNCViewer is primarily used for remote desktop access, allowing users to connect to and control a computer from a different location. This capability enables individuals to interact with the graphical user interface of a remote machine as if they were physically present at that computer. With VNCViewer, users can operate applications, troubleshoot issues, or perform system maintenance without needing to be on-site. This makes it an essential tool in environments where remote administration is crucial, such as in IT support, remote work situations, and server management. Although file transfer capabilities may accompany remote desktop applications, the core functionality of VNCViewer centers on providing a graphical interface to access and control a remote system effectively.

7. Which cloud platform provides support for various Hadoop distributions, including Cloudera?

A. AWS

B. Google Cloud Platform

C. Oracle Cloud Infrastructure

D. Microsoft Azure

The correct answer is Oracle Cloud Infrastructure. This cloud platform has been specifically designed to support various distributed computing technologies, including Hadoop ecosystems like Cloudera. Oracle has partnered with Cloudera to provide optimized services that leverage Cloudera's distribution, allowing users to take advantage of its capabilities for large-scale data processing and analytics. Oracle Cloud Infrastructure offers features such as flexible compute resources, high-performance storage options, and integrated management tools that enhance the performance of Hadoop distributions. This tailored support ensures that enterprises can efficiently deploy and manage big data applications using Cloudera on Oracle's infrastructure. In the context of big data workloads, having a cloud platform that directly supports specific distributions allows organizations to streamline operations and utilize familiar tools and frameworks efficiently. This capability typically results in better performance and simpler management for users working within the Hadoop ecosystem.

SAMPLE

8. What role does a data engineer play in Big Data projects?

A. Designing and implementing machine learning algorithms

B. Creating user-facing data dashboards

C. Designing and maintaining the data infrastructure and pipelines

D. Conducting data analysis and reporting

In Big Data projects, a data engineer plays a crucial role in designing and maintaining the data infrastructure and pipelines. This involves developing the architecture that allows for the efficient collection, storage, and processing of large datasets. Data engineers are responsible for building systems that handle the input and output of data, ensuring that data flows seamlessly between various sources and applications. Their expertise lies in selecting appropriate technologies and tools for data storage, such as databases and data warehouses, as well as implementing data pipelines that automate the movement and transformation of data. By doing so, data engineers enable data scientists and analysts to access clean, organized, and reliable data necessary for analysis and further processing. This foundational work is essential for the successful deployment of Big Data solutions and ensures that other team members can focus on analysis and application rather than data management issues. The tasks of designing machine learning algorithms, creating user-facing dashboards, and conducting data analysis and reporting belong more to the domains of data scientists and analysts. While these roles are important in the Big Data ecosystem, they rely heavily on the infrastructure and capabilities that data engineers provide.

9. Which are recommended storage options on OCI for a Persistent Filesystem of approximately 3TB?

A. Local NVMe and Object Storage

B. Block volume and FSS

C. Archive storage and Object Storage

D. File system and Cold Storage

The recommended storage options for a Persistent Filesystem of approximately 3TB on Oracle Cloud Infrastructure (OCI) are Block Volume and File Storage Service (FSS). Block Volume is ideal for applications that require high performance and consistent low-latency access to data, making it suitable for databases and enterprise applications. It allows attaching to compute instances and provides the flexibility of resizing volumes. File Storage Service (FSS), on the other hand, is designed for use cases that require scalable and shared file systems. It provides managed file shares that can be accessed by multiple instances and offers performance levels that make it appropriate for workloads that involve large data sets or require high throughput. The combination of Block Volume and FSS covers both high-performance needs and scalable storage requirements effectively. In contrast, options like Local NVMe, while offering high performance, are not suitable for persistent storage across instance lifecycles and may lead to data loss if the instance is stopped or terminated. Archive storage and Object Storage are designed for infrequently accessed data and are not optimized for performance-intensive applications. Cold Storage is also geared more towards archiving rather than providing the performance and access capabilities needed for a persistent filesystem. Thus, B is the most suitable choice, providing the necessary balance of performance,

SAMPLE

10. Which programming language is often associated with Apache Spark?

A. Python

B. Java

C. Scala

D. Ruby

Apache Spark is closely associated with Scala because Scala was the original language in which Spark was written. Spark is designed to take advantage of Scala's capabilities, including functional programming paradigms, which align well with Spark's design for distributed data processing. While other languages like Python and Java can interface with Spark through APIs (PySpark and Spark's Java API respectively), Scala remains the most native and idiomatic choice, allowing developers to utilize all features of Spark effectively. Additionally, many of Spark's advanced features and optimizations are often best accessed through Scala, making it a preferred language for Spark applications, especially in environments focused on performance and scalability. The association with Scala showcases not only the technical foundations of Spark but also the development community's preference for this language when leveraging the full potential of Spark's capabilities.

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://hpcbigdatavetdeck.examzify.com

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

SAMPLE