HPC Big Data Certification Practice Test (Sample)

Study Guide

Everything you need from our exam experts!

Table of Contents

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.

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!

Questions

1. **What is the main function of Object Storage in Big Data Migration?**

   A. To transfer data via high bandwidth

   B. To provide an alternate storage solution during migration

   C. To host virtual machines

   D. To serve as a backup solution

2. **Which companies are known for heterogenous architecture involving CPU and GPU?**

   A. NVIDIA and AMD

   B. NVIDIA and Intel

   C. Intel and Apple

   D. AMD and Microsoft

3. **Which of the following has the highest price per core?**

   A. BM.Standard.2.52

   B. BM.GPU2.2

   C. BM.GPU3.8

   D. BM.HPC2.36

4. **What is Oracle Data Science primarily used for?**

   A. A platform for data scientists to create projects with real-time modeling

   B. A marketplace for data visualizations

   C. An API for developing mobile applications

   D. A software for database management

5. **What is a characteristic of Cloudera's enterprise version found within Oracle Big Data products?**

   A. Supports only SQL queries

   B. Integrates fully with traditional RDBMS

   C. Includes advanced data management capabilities

   D. Only usable in on-premises environments

6. What capacity is typically associated with archival storage?

A. Terabytes

B. Gigabytes

C. Petabytes

D. Exabytes

7. What is the recommended HDFS replication factor for DenseIO hosts to mitigate data loss?

A. 1

B. 2

C. 3

D. 4

8. Which of the following is considered the most important phase of the Terasort process?

A. TeraGen

B. TeraSort

C. TeraValidate

D. None of the above

9. How much latency is introduced when using TCP between two nodes?

A. 1.7 microseconds

B. 3 microseconds

C. 0.2 milliseconds

D. 5 microseconds

10. Where do customers store data and application code for Oracle Data Flow applications?

A. Object Storage

B. Local Disk

C. Database Instances

D. Data Lakes

Answers

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

Explanations

1. What is the main function of Object Storage in Big Data Migration?

A. To transfer data via high bandwidth

B. To provide an alternate storage solution during migration

C. To host virtual machines

D. To serve as a backup solution

The main function of Object Storage in Big Data Migration is to provide an alternate storage solution during migration. As organizations undertake the migration of large data sets to new environments, such as cloud infrastructures, they often need a reliable and flexible storage medium to accommodate this transition. Object Storage is designed to handle vast amounts of unstructured data, making it well-suited for big data workloads. It offers scalability by allowing data to be stored in a flat namespace, where each object can be accessed independently. During the migration process, Object Storage enables organizations to retain data availability and integrity while they move or transform data from one system to another. Its architecture is built for durability and can efficiently handle varied data formats. This flexibility makes it an ideal choice for storing data temporarily or permanently during the migration phase, ensuring that applications can continue functioning effectively without interruption. The alternatives, such as high bandwidth capability or backup solutions, do not specifically address the unique needs of transitioning data through a migration process, while hosting virtual machines focuses entirely on computational resources rather than data storage needs.

2. Which companies are known for heterogenous architecture involving CPU and GPU?

A. NVIDIA and AMD

B. NVIDIA and Intel

C. Intel and Apple

D. AMD and Microsoft

The selection of NVIDIA and Intel as companies known for heterogeneous architecture involving CPU and GPU is accurate because both companies actively develop technologies that utilize integrated CPU and GPU solutions to enhance computational performance. NVIDIA is renowned for its GPUs and has advanced technologies that enable the integration of GPUs with CPUs in systems designed for high-performance computing and graphics rendering. They have a strong focus on parallel processing and deep learning, areas where the synergy between CPU and GPU is crucial for optimizing performance. Intel, traditionally known for its CPUs, has been incorporating GPU capabilities into its architecture with products like Intel's integrated graphics found in many of their processors, as well as their discrete GPUs. They have also been investing in heterogeneous computing through their oneAPI initiative, which aims to provide a unified framework for programming across CPUs, GPUs, and other accelerators. In contrast, while other companies do contribute to heterogeneous computing, their primary products or designs do not focus as strongly on integrating CPU and GPU technologies as NVIDIA and Intel. For instance, AMD also plays a significant role in CPU and GPU development but primarily focuses on its own architectures rather than collaborations that mix these components as significantly as Intel. Companies like Apple are focused on their specific architectures and may not be as engaged in the broader context of heterogeneous architectures

3. Which of the following has the highest price per core?

A. BM.Standard.2.52

B. BM.GPU2.2

C. BM.GPU3.8

D. BM.HPC2.36

The choice of BM.GPU3.8 having the highest price per core is based on its configuration, which typically includes a powerful GPU architecture designed for high-performance computing tasks. This model is engineered for demanding applications that require not just raw processing power but also efficiency in handling complex computations. The GPU capabilities imply that it can perform parallel processing much more effectively than standard CPU-based configurations. In high-performance computing (HPC) environments, the inclusion of GPUs significantly influences pricing due to their advanced technology and capability to execute operations much faster than traditional CPU cores. As such, hardware with GPUs—especially more advanced versions like those found in the BM.GPU3.8—often costs more per core than standard processors or other types of node configurations. While other options might include powerful configurations as well, they don't combine the same level of advanced GPU integration that enhances performance and justifies a higher price point. This makes BM.GPU3.8 stand out in terms of price per core in the context of high-performance workloads.

4. What is Oracle Data Science primarily used for?

A. A platform for data scientists to create projects with real-time modeling

B. A marketplace for data visualizations

C. An API for developing mobile applications

D. A software for database management

Oracle Data Science is primarily designed as a platform that enables data scientists to collaborate and create projects focused on real-time modeling. This solution provides an integrated environment where data professionals can utilize machine learning algorithms, leverage extensive datasets, and apply advanced analytics to derive meaningful insights and predictions. The platform supports the entire data science lifecycle, from data preparation and model training to deployment and monitoring of models in production. This emphasis on real-time modeling allows organizations to make timely, informed decisions based on dynamic insights drawn from their data. While the other alternatives present important aspects related to data and technology, they do not capture the primary function of Oracle Data Science. It is not oriented towards visualizations, mobile application development, or traditional database management; rather, its core focus is on enabling data-driven decision-making through sophisticated analytical tools and methodologies in a collaborative setting.

5. What is a characteristic of Cloudera's enterprise version found within Oracle Big Data products?

A. Supports only SQL queries

B. Integrates fully with traditional RDBMS

C. Includes advanced data management capabilities

D. Only usable in on-premises environments

The characteristic of Cloudera's enterprise version found within Oracle Big Data products is its inclusion of advanced data management capabilities. This aspect highlights the comprehensive tools and functionalities that are designed to handle complex data environments, facilitating the efficient processing, storage, and retrieval of big data. Cloudera's offering brings advanced features such as data governance, security, and analytics, which are critical for enterprises looking to leverage large volumes of data for insights and decision-making. These capabilities are essential in environments where data variety and velocity create challenges that traditional systems may struggle to manage effectively. The other choices fail to capture the breadth of Cloudera's capabilities. For instance, supporting only SQL queries would imply a limitation in functionality, whereas Cloudera supports a variety of data processing paradigms beyond just SQL. The integration with traditional RDBMS is not a defining characteristic because Cloudera enhances rather than solely focuses on those systems. Additionally, the assertion that it is only usable in on-premises environments disregards the flexibility of deployment options that Cloudera offers, including cloud-based solutions. Thus, the focus on advanced data management capabilities correctly represents a key strength of Cloudera's enterprise version found in Oracle Big Data products.

6. What capacity is typically associated with archival storage?

A. Terabytes

B. Gigabytes

C. Petabytes

D. Exabytes

Archival storage is designed for the long-term retention of data that is not frequently accessed but must be preserved for regulatory, historical, or compliance reasons. This type of storage typically accommodates vast amounts of data, which makes the capacity of petabytes most appropriate. Petabytes are units of digital information that represent approximately 1,000 terabytes, or a million gigabytes. This level of capacity allows organizations to store extensive datasets effectively, which can include everything from scientific data and research findings to vast collections of digital images and videos. While terabytes and gigabytes are certainly relevant sizes in data storage, they do not represent the large-scale capacity commonly required for archival purposes. Exabytes, on the other hand, reflect an even larger magnitude than petabytes and are often used in discussions about the total capacity of massive systems or networks, rather than specific archival storage solutions. Thus, petabytes strikes the right balance, meeting the needs for extensive and efficient data archival.

7. What is the recommended HDFS replication factor for DenseIO hosts to mitigate data loss?

A. 1

B. 2

C. 3

D. 4

The recommended HDFS replication factor for DenseIO hosts being set to three is based on several considerations related to data reliability and fault tolerance within a Hadoop ecosystem. A replication factor of three means that each block of data is stored on three separate DataNodes. This redundancy plays a crucial role in ensuring that even if one or possibly two DataNodes become unavailable due to failures, the data remains accessible from the remaining node. This level of replication significantly enhances data durability and availability, which is critical for systems that handle large volumes of big data. DenseIO hosts are designed for high-performance computing and typically involve workloads that are I/O intensive. Given the use of fast storage and the need for high throughput in these environments, the extra data redundancy provided by a replication factor of three ensures that data can be accessed swiftly while still being protected against potential hardware failures. While lower replication factors may save on storage space, they do not provide the same level of fault tolerance. A replication factor of two is marginally better than one, but it may not offer sufficient protection if one of the nodes becomes unavailable. Therefore, choosing three strikes a balance between efficient storage usage and robust data safety, which makes it the standard recommendation for high-reliability environments like those utilizing DenseIO

8. Which of the following is considered the most important phase of the Terasort process?

A. TeraGen

B. TeraSort

C. TeraValidate

D. None of the above

The TeraSort process is a benchmark for sorting large datasets, particularly in Hadoop and Big Data environments. The most crucial phase of this process is TeraSort. During the TeraSort phase, the actual sorting of the massive data set occurs. This phase is where the core functionality of the Terasort benchmark is tested, as it demonstrates the system's ability to efficiently sort a dataset that can range into terabytes or even petabytes in size. It assesses the performance of the underlying system, including the efficiency of data shuffling, network utilization, and disk I/O operations during the sorting process. In contrast, TeraGen is primarily the stage where data is generated, while TeraValidate serves to ensure the integrity and correctness of the sorting outcome. While all three components are important to the overall Terasort process, TeraSort is the pivotal phase that provides direct insights into the performance and capabilities of the Big Data processing system.

9. How much latency is introduced when using TCP between two nodes?

A. 1.7 microseconds

B. 3 microseconds

C. 0.2 milliseconds

D. 5 microseconds

When discussing the latency introduced by TCP (Transmission Control Protocol) between two nodes, it is important to consider the nature of TCP as a reliable transport layer protocol. TCP establishes a full-duplex connection, ensures the reliable delivery of packets, and manages error checking, which can contribute to some degree of latency. The appropriate answer regarding latency for TCP connections typically falls in the low microsecond range under ideal conditions. Latency can vary based on several factors, including network conditions, the distance between nodes, and the specific configurations of the TCP stack, but the range usually does not extend to milliseconds in standard configurations. Considering the context of the other options provided, the introduction of latency as 0.2 milliseconds would imply significantly higher delay than what is typical for TCP communication in local area networks, where latency is usually expected to be measured in microseconds rather than milliseconds. Connecting two nodes via TCP tends to yield latencies closer to the single-digit microsecond range, depending on the speed and efficiency of the underlying network infrastructure. Therefore, the selection that reflects a realistic latency figure for TCP connections in high-performance environments would be in the microsecond range, aligning with what is typically observed in the industry. This understanding helps solidify the expectation of TCP performance in

10. Where do customers store data and application code for Oracle Data Flow applications?

A. Object Storage

B. Local Disk

C. Database Instances

D. Data Lakes

Oracle Data Flow applications utilize Object Storage to store both data and application code. This is primarily because Object Storage offers scalable, durable, and cost-effective storage solutions that are ideal for managing large volumes of unstructured data, which is commonly used in big data applications. By leveraging Object Storage, users can efficiently handle vast datasets, use various storage classes for optimization, and easily integrate with other Oracle Cloud services. In contrast, Local Disk is typically associated with individual machines and lacks the scalability and accessibility offered by Object Storage. Database Instances are designed for structured data and transactional processing, while they can also handle certain data storage needs, they are not the primary choice for big data applications which often deal with extensive datasets. Data Lakes, while they can serve as a storage solution for big data, are often built on top of Object Storage in cloud environments and aren't the primary choice for directly hosting application code associated with Oracle Data Flow. This makes Object Storage the most suitable option for such applications.

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://hpcbigdatacert.examzify.com

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