Quiz 2: Storage Infrastructures

Tutorial on 08.05.2025

Performance summary

✓ Assessed

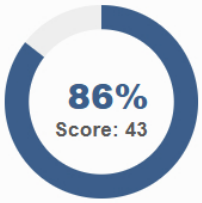
Success status	Score	Attempts
Undefined	43 of 50 points	1

Results

Course	Big Data SoSe 2025 ID: 4853531345 / 107639912247978
Test	Test 2 Big Data 2025 ID: 4893475104

This are your test results

Duration	0h 35m 47s 5/7/2025, 10:57 AM - 5/7/2025, 3:01 PM
Answered	13 of 13 questions (100%)
Your score	43 of 50 points (86%)



Recall

Parallel database architectures

Status	Answered	
Your score	6 / 6	100%

Response

	Shared nothing	Shared Disk	Shared Memory
Only scalable for relatively small number of the professor	☐	☑	☑
Extremely difficult to manage	☑	☑	☐
Efficient communication between processors	☐	☐	☑
Sending data requires the software interaction at both ends	☑	☐	☐
Can be easily scaled up to thousands of processors	☑	☐	☐
Might create a bottleneck at inter connection to the disk subsystem	☐	☑	☐

Single Master vs NameNode

Status	Answered	
Your score	5 / 6	83%

Response

Which of the following statements are true in context on Single Master vs NameNode systems?

Unanswered	Right	Wrong	
☐	☑	☐	GFS is a single master architecture
☐	☐	☑	Secondary NameNode is an extension to NameNode and hosts additional data
☐	☑	☐	Default block size of HDFS is 64MB
☐	☑	☐	Chunkservers in GFS and DataNodes in HDFS have similar role in the file system
☐	☐	☑	GFS is good for many small files
☐	☑	☐	In HDFS each slave machine hosts a DataNode daemon

☑ Degree of parallelism

Status	Answered		
Your score	4 / 4		100%

Response

Which of the following statements are true for Degree of parallelism?

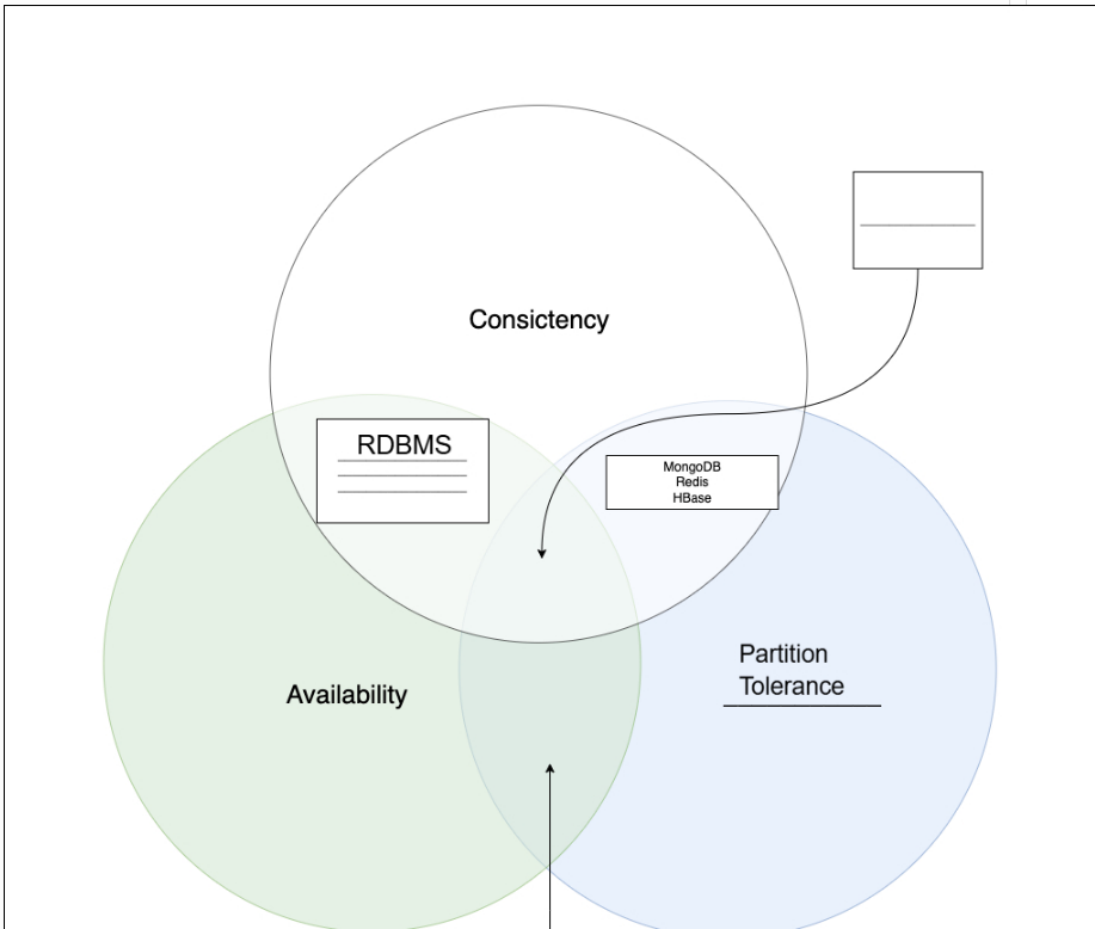
- ☑ It indicates the number of processors employed to run a single statment
- ☐ it indicates how many processors are in the system
- ☑ It indicates how many operations can be executed by the computer simultaniosly
- ☐ The maximum Degree of parallelism avalible is 32

CAP theorem

Status	Answered		
Your score	0 / 6		0%

Response

Fill the blanks on the CAP diagram



CouchDB
~~Cassandra~~
~~DynamoDB~~
Riak

System Component

Status	Answered	
Your score	4 / 4	100%

Response

	Manages execution across Spark cluster	Stores actual data blocks	Immutable distributed collection of elements	Manages metadata and file system namespace
DataNode	☐	☑	☐	☐
NameNode	☐	☐	☐	☑
RDD	☐	☐	☑	☐
SparkContext	☑	☐	☐	☐

☑ RDD

Status	Answered	
Your score	1 / 1	100%

Response

Which features are true about Apache Spark's Resilient Distributed Datasets (RDDs)?

- ☐ Eagerly evaluated
- ☐ Uses SQL as its core API
- ☑ Supports lazy evaluation
- ☑ Distributed
- ☑ Immutable

Knowledge questions

◉ Google File System (GFS)

Status	Answered	
Your score	2 / 2	100%

Response

What is the Google File System (GFS) primarily designed for?

- ⊘ c) Scalable distributed file system for large distributed data intensive applications
- ○ d) Personal file management
- ○ a) Handling small-scale data storage
- ○ b) Real-time data processing

☑ File system

Status	Answered	
Your score	2 / 2	100%

Response

What is the primary distributed storage system used by Hadoop applications, inspired by the Google File System (GFS)?

- ☐ b) Apache Kafka

☑ c) Hadoop Distributed File System (HDFS)

☐ a) HBase

☐ d) Apache Spark

HDFS

Status	Answered	
Your score	2 / 2	100%

Response

Why is replication pursued in Hadoop Distributed File System (HDFS) despite causing data redundancy?

- ☑ To achieve fault tolerance by storing data in different locations, ensuring data availability even in the event of node failures.
- ○ To improve data processing speed by distributing data across multiple nodes.
- ○ To increase storage capacity and reduce data redundancy by storing data in fewer locations.
- ○ To enhance data integrity by creating multiple copies of data on the same node.

Apache Spark Vs MapReduce

Status	Answered	
Your score	3 / 3	100%

Response

	Apache Spark	MapReduce
processes data in batches only	☐	☑
processes data in batches as well as in real-time	☑	☐
Suitable for batch processing & real-time processing	☑	☐

Components

Status	Answered	
Your score	8 / 8	100%

Response

What is the purpose for each of the components?

Spark SQL	Enables querying of structured data. → Spark SQL
Spark Streaming	Graph processing → GraphX
GraphX	Intended for real-time processing. → Spark Streaming
MLlib	Well-suited for machine learning tasks. → MLlib

Counting Juliet

Status	Answered

Your score	5 / 5	100%

Response

Unraveling Shakespeare's Classic with RDDs

Being RDDs composed of Java objects, they offer a familiar **object-oriented programing style**. RDDs' APIs allow low-level processing on the data structure: *map*, *filter*, *reduce*, are the most common operations to manipulate RDD objects using the typical object-oriented and functional programming features offered by Scala language.

Situation:

Suppose you have a text file containing Shakespeare's *Romeo and Juliet* where each line represents a line of the original tragedy. You want to count how many times the word "Juliet" appears in the whole text. For simplicity, consider you have the text without punctuation.

This parsing task fits well for RDDs. Basically, you have to place the below steps in the oder they would occur:

load the text file into an RDD, where each RDD element is a line of text. Then map all the lines to lower case letters;

take all those sentences, split them into separate words wherever you see a space, and put all those words together

filter (retain) all the words equal to "juliet"

then map all the elements to the number 1;

apply a reduce with a sum function to get the total result.

CAP Theorem

Status	Answered	
Your score	1 / 1	100%

Response

Which of the following best describes the trade-off emphasized by the CAP Theorem?

- ○ Disk Speed vs Memory Usage
- ⊘ Consistency vs Availability vs Partition Tolerance
- ○ Disk Speed vs Memory Usage
- ○ Scalability vs Speed

Test execution

Information

Availability: Expired at 5/8/2025, 4:00 PM

Max. attempts: Unlimited

Results of this test are visible to administrators and tutors of this course.

Start test

▸ Change log