Quiz 4: Distributed Data Processing

Tutorial on 22.05.2025

Performance summary

✓ Assessed

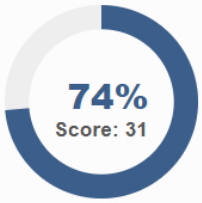
Success status	Score	Attempts
Undefined	31 of 42 points	1

▾ Results

Course	**Big Data SoSe 2025** ID: 4853531345 / 107761542276536
Test	**Test 4 Big Data 2025** ID: 4907958619

This are your test results

Duration	0h 20m 45s 5/16/2025, 11:59 AM - 5/16/2025, 12:20 PM
Answered	13 of 13 questions (100%)
Your score	31 of 42 points (74%)



Recall

☰ MapReduce

Status	Answered	
Your score	0 / 6	0%

Response

You are given the phases of MapReduce. Arrange them in the correct order.

Input
Split
Map
Shuffle
Sort
Reduce
Output

Features of Map-Reduce

Status	Answered	
Your score	1 / 1	100%

Response

Which of the following statements is true for Map-Reduce?

Unanswered	Right	Wrong	
☐	☑	☐	Map-Reduce re-runs the failed tasks
☐	☐	☑	Map tasks are scheduled close to the output when possible
☐	☑	☐	A Map-Reduce may specify how its input is to be read
☐	☑	☐	MapReduce is a programming model for distributed computing

☑ MLlib capabilities

Status	Answered	
Your score	0 / 1	0%

Response

Which of the following are part of Spark's MLlib capabilities?

- ☑ Distributed model training
- ☐ Model deployment in production
- ☑ Feature engineering
- ☐ High-speed transactional processing

True/false

Status	Answered	
Your score	2 / 2	100%

Response

Unanswered	Right	Wrong	
☐	☐	☑	Map tasks in Hadoop are always scheduled randomly across nodes.
☐	☐	☑	SparkR is used to write Spark applications in Python.
☐	☐	☑	MLlib cannot handle large-scale data due to memory constraints.
☐	☑	☐	Pig Latin scripts are automatically compiled into one or more MapReduce jobs.

Spark

Status	Answered	
Your score	4 / 4	100%

Response

Match the Spark Components and Features accordingly:

	Working with structured data using SQL/DataFrames
	Spark SQL
	Underlying general execution engine
	Spark Core
	Python API for Spark
	PySpark

Interface to use Spark with R language

SparkR

All about Pig

Pig & Pig Latin

Status	Answered	
Your score	4 / 5	80%

Response

Match the given statements with Pig or Pig Latin

	Pig	Pig Latin
handles erroneous/corrupt data entries gracefully	☑	☐
There is no need for a user to be aware of the algorithmic details in the map/reduce phases	☑	☐
It is a high-level language for expressing data flows	☐	☑
The script describes HOW to process the data	☐	☑

Pig Features

Status	Answered	
Your score	4 / 4	100%

Response

Which statements correctly describe features of Pig, a parallel data processing language implemented on Hadoop? (Select all that apply)

- ☐ Pig provides slow processing
- ☐ Pig is primarily designed for processing structured data
- ☑ Pig is easily controlled and modified
- ☑ Pig operates on any data.
- ☑ Pig is tightly integrated with Hadoop

PIG VS SQL

Status	Answered	
Your score	6 / 6	100%

Response

A procedural language.

Schema is mandatory

schema is optional. We can store data without designing a schema.

Pig

schema is optional. We can store data without designing a schema.

A procedural language.

SQL

Schema is mandatory

Pig Latin (Multiple choice)

Status	Answered

Your score	2 / 2		100%

Response

Which of the following best describes the motivation behind the development of Pig Latin and the Pig system?

(Refer Paper provided in the Lecture Material -> Further Reading -> Pig Latin: A Not-So-Foreign Language for Data Processing)

- [x] To bridge the gap between the declarative style of SQL and the procedural style of map-reduce, offering a more user-friendly alternative for data analysis tasks on large data sets.
- [ ] To provide a high-level, declarative language similar to SQL for analyzing large data sets, avoiding the cost and complexity of parallel database products like Teradata.
- [ ] To provide a debugging environment integrated with Hadoop, allowing for more efficient development and execution of data analysis tasks compared to traditional map-reduce implementations.
- [x] To address the limitations of the map-reduce programming model by offering a language that combines the ease of use of SQL with the flexibility of procedural programming.
- [ ] To introduce a new low-level, procedural programming language specifically tailored for analyzing terabytes of data collected daily by internet companies.

Further Reading

Apache Mahout -> Link Up ;)

Status	Answered		
Your score	6 / 6		100%

Response

Which of the following accurately describes Apache Mahout?

Match all the statements.

A library for machine learning on distributed dataflow systems, initially targeting MapReduce, now supporting Apache Spark and Apache Flink.	While Apache Mahout is widely used by leading web companies and is part of several commercial cloud offerings, it is not a platform for cloud computing itself.
A programming language for data preprocessing and model training, exclusively designed for linear algebraic computations.	A platform for cloud computing offered by leading web companies, specializing in large-scale machine learning since its inception in 2008.
A platform for cloud computing offered by leading web companies, specializing in large-scale machine learning since its inception in 2008.	Apache Mahout is primarily focused on machine learning tasks and is not specifically designed for data visualization and analysis.
A framework primarily used for data visualization and analysis, compatible with commercial cloud offerings and pioneering machine learning algorithms.	A framework primarily used for data visualization and analysis, compatible with commercial cloud offerings and pioneering machine learning algorithms.
A tool for managing distributed databases and streamlining data integration processes, widely utilized by industry professionals for recommendation algorithms.	Apache Mahout is not a programming language but rather a library/framework that provides implementations of machine learning algorithms.
	A programming language for data preprocessing and model training, exclusively designed for linear algebraic computations.
	Apache Mahout is not a tool for managing distributed databases or streamlining data integration processes; its main purpose is to provide machine learning algorithms for distributed data processing.
	A tool for managing distributed databases and streamlining data integration processes, widely utilized by industry professionals for recommendation algorithms.
	Correct answer
	A library for machine learning on distributed dataflow systems,

initially targeting MapReduce, now supporting Apache Spark and Apache Flink.

SparkSQL ✨

Status	Answered	
Your score	2 / 2	100%

Response

```
val sqlContext = new org.apache.spark.sql.SQLContext(sc)
val df = sqlContext.read.json("employee.json")
df.show()
df.select("name").show()
df.groupBy("age").count().show()
```

Which sequence of operations correctly describes the functionality of the provided Apache Spark SQL commands?

- ○ The commands initialize an SQL context, read data from a JSON file into a DataFrame, display all columns of the DataFrame, and then group the data by age and count the occurrences of each age.
- ○ The commands initialize a Spark session, load data from a CSV file, and then group the data by age and count the occurrences of each age.
- ○ The commands create an SQL context, read data from a JSON file into a DataFrame, display all columns of the DataFrame, and then filter the data based on age.
- ⊘ The commands create an SQL context, read data from a JSON file into a DataFrame, display the "name" column of the DataFrame, and then group the data by age and count the occurrences of each age.

MapReduce Research

Status	Answered	
Your score	0 / 2	0%

Response

Based on the detailed content of the paper **"MapReduce: an infrastructure review and research insights"** by Maleki, Rahmani, and Conti, please answer the following:

Unanswered	Right	Wrong	
☐	☐	☑	Data locality is ignored in modern scheduling approaches.
☐	☐	☑	Hadoop MapReduce supports task-level fault recovery via checkpointing.
☐	☑	☐	The paper presents a systematic review of MapReduce research from 2014 to 2017.
☐	☐	☑	Energy consumption in Hadoop clusters is proportional to system utilization.

☑ MapReduce Research

Status	Answered	
Your score	0 / 1	0%

Response

Which of the following are considered key challenges in Hadoop MapReduce infrastructure?

- ☑ Load balancing
- ☑ Fault tolerance
- ☑ Network latency
- ☑ Job scheduling

Test execution

Information

Availability: Expired at 5/22/2025, 4:00 PM

Max. attempts: Unlimited

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

Start test

Change log

Go to top