# 202122 - Analyzing Big Data - S1



| | |
|---|---|
| Started on | Tuesday, 11 January 2021, 6:25 PM |
| State | Finished |
| Completed on | Tuesday, 11 January 2022, 7:45 PM |
| Time taken | 1 hour 19 minutes |

**Exam instructions**

1. This is an open book individual exam with a duration of 60 minutes
2. Camera must be on
3. Use the microphone to use your mobile phone not to communicate with anyone but the professor
4. The exam has two sections
5. Section I consists of multiple-choice questions. Each correct answer is worth 0.5 points. Wrong answers have a penalty of 0.15
6. Section II consists of open questions that may be answered either in English or Portuguese. Each question is worth 3 points

Quiz navigation

ABD2021 1st Exam 11-Jan

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Show one page at a time

Finish review

Question 1 Complete Marked out of 0.50 Flag question

Information

What is the meaning of the data processing model: "Scale Out"?

Select one:
- a. Means adding bigger, more powerful machines
- b. Means adding more, smaller machines
- c. Means processing data with scalable machines
- d. Means processing data outside the cluster

Question 2 Complete Marked out of 0.50 Flag question

In Databricks notebooks you can?

Select one:
- a. Program only in Python and Spark
- b. Program only in the language defined in the Notebook creation
- c. Select a language at the cell level
- d. Select a language for a set of cells

Question 3 Complete Marked out of 0.50 Flag question

In a Databricks notebook, to access the cluster driver node console, what magic command is used?

Select one:
- a. %ls
- b. %dbfse
- c. dbutils.fs.mounts()
- d. %sh

Question 4 Complete Marked out of 0.50 Flag question

What is a DataFrame?

Select one:
- a. A data processing method in Spark
- b. A distributed dataset in-memory
- c. A distributed dataset on-disk
- d. A distributed dataset on-disk and in-memory

Question 5 Complete Marked out of 0.50 Flag question

What is MapReduce?

Select one:
- a. A programming language
- b. A programming model
- c. A set of functions for processing big data
- d. The original Apache Software Foundation query engine for big data

Question 6 Complete Marked out of 0.50 Flag question

What is Avro in Hadoop?

Select one:
- a. A program to load data with high parallelization
- b. A column-based data format
- c. A row-based data format
- d. A text-based data format for compatibility and portability

Question 7 Complete Marked out of 0.50 Flag question

What is a Pair RDD?

Select one:
- a. Two RDDs in sequence in a transformation statement
- b. An RDD with only two rows
- c. An RDD with a pair operation
- d. An RDD with key-value pairs

Question 8 Complete Marked out of 0.50 Flag question

What is the result of the Spark statement bellow?

```
sc.parallelize(mytuples,3)
```

Select one:
- a. Creates an RDD with a minimum of 3 partitions
- b. Creates an RDDs named mytuple and the value 3
- c. Generates an error cf too many parameters
- d. Creates 3 RDDs with mytuple

Question 9 Not answered Marked out of 0.50 Flag question

What Spark function has, as an output, a Pair RDD?

Select one:
- a. map
- b. flatmap
- c. textFile
- d. wholeTextFiles

Question 10 Complete Marked out of 0.50 Flag question

Select the right statement to create a DataFrame:

Select one:
- a. DF = spark.createDataFrame(["Line 1", "Line 2"])
- b. DF = spark.table("mytable")
- c. DF = spark.load("mytable")
- d. DF = sc.parallelize(["Line1", "Line2"])

Question 11 Complete Marked out of 0.50 Flag question

Select the right statement regarding reduceByKey():

Select one:
- a. reduceByKey() is a wide transformation
- b. reduceByKey() is a narrow transformation
- c. reduceByKey() is a key transformation
- d. reduceByKey() is an action

Question 12 Complete Marked out of 0.50 Flag question

Select the **false** statement regarding Spark terminology:

Select one:
- a. A Job is a set of tasks executed as a result of an action
- b. A Stage is a set of tasks in a job that can be executed in parallel
- c. A task is an individual unit of work sent to an executor
- d. An Application can only contain one job

Question 13 Complete Marked out of 0.50 Flag question

What is a lambda function?

Select one:
- a. It's a function defined without a name and with only one parameter
- b. It's a function defined without a name and with only one expression
- c. It's a function that can be reused with many parameters
- d. It's a function that can be reused with many expressions

Question 14 Complete Marked out of 0.50 Flag question

collect() is a Spark function that?

Select one:
- a. You can extensively use to display DataFrames content
- b. It's not available for DataFrames
- c. You can extensively use to display RDDs content
- d. You should use with caution to display RDDs content

Question 15 Complete Marked out of 0.50 Flag question

With the instruction sc.textFile("file:/data") you are?

Select one:
- a. Reading a file from your hdfs file system
- b. Reading a file called "file:/data"
- c. Reading a file from your local non-Hadoop file system
- d. Reading a file called "data" stored in a folder called "file"

Question 16 Complete Marked out of 0.50 Flag question

Window operations in Spark are used for?

Select one:
- a. Define a window to display data
- b. Select a window of data to display
- c. Freeze data in memory from a window of data
- d. Apply transformations over a window of data

Question 17 Complete Marked out of 0.50 Flag question

What is the data type that results from the following spark instruction: spark.range(10)?

Select one:
- a. A list
- b. A tuple
- c. An RDD
- d. A DataFrame

Question 18 Complete Marked out of 0.50 Flag question

Select the right statement regarding Spark transformations:

Select one:
- a. Wide transformations are very efficient because they don't move data from the node
- b. Narrow transformations are very efficient because they don't move data from the node
- c. Both wide and narrow transformations move data from the node
- d. None of the narrow or wide transformations move data from the node

Question 19 Complete Marked out of 0.50 Flag question

In Spark, lazy execution means that:

Select one:
- a. Execution will take some time because it needs to be sent to the worker nodes
- b. Execution will take some time because the code is interpreted
- c. Execution is done one line at the time
- d. Execution is triggered only when an action is found

Question 20 Complete Marked out of 0.50 Flag question

What is the difference between Spark Streaming and Structured Streaming?

Select one:
- a. Structured Streaming is for structured streaming data processing and Spark Streaming is for unstructured streaming data processing
- b. Spark Streaming is the new ASF library for Streaming Data and Structured Streaming the old one
- c. Structured Streaming is a stream processing engine and Spark Streaming is an extension to the core Spark API to streaming data processing
- d. Structured Streaming relies on micro batch and RDDs while Spark Streaming relies on DataFrames and Datasets

Question 21 Complete Marked out of 0.50 Flag question

What is a Tumbling Window in Spark Streaming?

Select one:
- a. A fixed-sized, non-overlapping and contiguous window of data
- b. An overlapping and contiguous window of data
- c. A non-contiguous window of data
- d. A dynamic size window of data

Question 22 Complete Marked out of 0.50 Flag question

Spark ML library can be classified as:

Select one:
- a. A mature ML library with a very wide range of predictive and descriptive models to choose from
- b. A strong Deep Learning library
- c. A complete ML framework for data analysis
- d. A ML library with a reasonable set of models but still with work in progress

Question 23 Complete Marked out of 0.50 Flag question

What is the main difference between Spark MLlib and ML?

Select one:
- a. There is no difference apart from the bigger set of algorithms available on Spark ML
- b. Spark ML works with Streaming Data
- c. Spark MLlib is faster
- d. Spark ML is based on DataFrames

Question 24 Complete Marked out of 0.50 Flag question

In a Spark ML program, what is the purpose of the code bellow?

```
model.transform(mydata)
```

Select one:
- a. Create a machine learning model based on the data of 'mydata'
- b. Apply the model in 'model' to the data in 'mydata'
- c. Create a new model based on 'mydata'
- d. Adjust the model based on the data in 'mydata'

Question 25 Complete Marked out of 0.50 Flag question

Is a Spark ML program, what is the purpose of the code bellow?

```
model.fit(mydata)
```

Select one:
- a. Train a machine learning model based on the data of 'mydata'
- b. Apply the model in 'model' to the data in 'mydata'
- c. Create a new model based on 'mydata'
- d. Adjust the model based on the data in 'mydata'

Question 26 Complete Marked out of 0.50 Flag question

What is the result of the Spark ML instruction bellow?

```
lr = LogisticRegression(maxIter=10)
```

Select one:
- a. A logistic regression object is declared with a maximum of 10 interactions
- b. A logistic regression is executed with a maximum of 10 interactions
- c. A logistic regression is trained with a maximum of 10 interactions
- d. A logistic regression is estimated with a maximum of 10 interactions

Question 27 Complete Marked out of 0.50 Flag question

The vertex DataFrame in a GraphFrame is:

Select one:
- a. A tree form DataFrame
- b. A DataFrame that must contain a column named 'id'
- c. A DataFrame that must contain a column named 'src' and 'dst'
- d. A DataFrame that must contain a column named 'id', 'src' and 'dst'

Question 28 Complete Marked out of 0.50 Flag question

What is DSL used for in GraphFrames?

Select one:
- a. Formatting the output of a GraphFrame query
- b. Declare a GraphFrame object
- c. Search for patterns in a graph
- d. Define properties in a GraphFrame

Question 29 Complete Marked out of 0.50 Flag question

Based on the figure bellow, explain the shown algorithm line by line and the expected outcome with an example (of the outcome). Complementary, explain what is supposed to be input content for the algorithm and where is it (the name of the the file object) identified in the code.

```
count = sc.textFile("mydata").map(lambda line: line.split()) \
          .map(lambda word: (word, 1)) \
          .reduceByKey(lambda a,b: a+b)
```

The algorithm should most likely be used in the processing of a text file (e.g. txt) with multiple lines of works or sentences.

1 -rdd.flatMap : creates an RDD with flattened transformed element (from an eventual nested list), using a flatmap a flat list with all words is returned
2- for each element of the previous operation apply a lambda function that returns a tuple with each element - word, and 1 e.g. ('word1', 1), ('word2', 1), ('word1', 1)
3 - group the the similar words (keys) and sum the number of occurrences (values = 1)

The result should be a list of tuples with a key for each unique word and the sum of occurrences of that work in the file.

The input of the code is in "mdp" , e.g. a previous command would be similar to rdd = spark.read.text('capitv').....

Question 30 Complete Marked out of 0.50 Flag question

What are the general steps of a Spark ML program? Give an example.

The general steps in a SparkML program would be:
1. Data ingestion (load the data) , e.g. dataset = spark.read.format("csv").schema(schema).load("/datasets/ricks-datasets/ufo/ufo4.data")
2. Data preparation (analyze, clean and transform data), e.g. Eliminate duplicates with distinct() or dropDuplicates(); dropDuplicates allows columns selection. Ex: dropDuplicates(["name", "age"]), identify outliers, train-test split
3. Model creation and training, e.g. model = model_type().model.fit(training_data)
4. Model analysis and test , e.g. predict(testSetX), printScore(predictedY)
5. Make predictions, e.g. prediction = model.transform(test_data)

Question 31 Complete Marked out of 0.50 Flag question

Explain the main differences between real time data processing and batch data processing.

- Data Scope: batch processing can process all the data in the dataset. Stream processing typically only has access to the most recent data or within a rolling time window.
- Data Size: batch processing is suitable for handling large datasets efficiently. Stream processing is intended for individual records or micro-batches consisting of few records
- Performance: the latency for batch processing is typically a few hours. Stream processing typically occurs immediately, with latency in the order of seconds or milliseconds. Latency is the time taken for the data to be received and processed
- Analysis: you typically use batch processing for performing complex analytics. Stream processing is used for simple response functions, aggregates or calculations such as rolling averages.

Finish review



NOVA IMS

Useful Links

Support
Serviços de Informática / IT Services
sdi@novaims.unl.pt

NOVA IMS Accreditations and Certifications

Collaborating