MONASH University

Semester Two 2019

Examination Period

Faculty of Information Technology

EXAM CODES: **FIT5202**

TITLE OF PAPER: **DATA PROCESSING FOR BIG DATA**

EXAM DURATION: 2 hours 10 minutes

THIS PAPER IS FOR STUDENTS STUDYING AT: (tick where applicable)

□ Caulfield □ Clayton □ Parkville □ Peninsula

□ Monash Extension □ Off Campus Learning □ Malaysia □ Sth Africa

□ Other (specify)

During an exam, you must not have in your possession any item/material that has not been authorised for your exam. This includes books, notes, paper, electronic device/s, mobile phone, smart watch/device, calculator, pencil case, or writing on any part of your body. Any authorised items are listed below. Items/materials on your desk, chair, in your clothing or otherwise on your person will be deemed to be in your possession.

No examination materials are to be removed from the room. This includes retaining, copying, memorising or noting down content of exam material for personal use or to share with any other person by any means following your exam.
Failure to comply with the above instructions, or attempting to cheat or cheating in an exam is a discipline offence under Part 7 of the Monash University (Council) Regulations, or a breach of instructions under Part 3 of the Monash University (Academic Board) Regulations.

AUTHORISED MATERIALS

OPEN BOOK **NO**

CALCULATORS **YES**
SPECIFICALLY PERMITTED ITEMS **NO**
if yes, items permitted are:

Candidates must complete this section if required to write answers within this paper
STUDENT ID: __ __ __ __ __ __ __ __ DESK NUMBER: __ __ __ __ __

Question 1

a. What is Apache Spark? What are the two advantages of unified stack in Spark? **(1 + 2 = 3 Marks)**
Write your answer below

b. Define RDD. RDDs are lazily evaluated. What does it mean? Can RDDs be shared between SparkContexts? Please explain. Give two examples of how RDD can be created using SparkContext (assume 'sc' as a SparkContext object). **(1 + 1 + 1 + 2 = 5 Marks)**
Write your answer below

Continue your answer below

c. Write a program that does word count of the words ERROR and WARN found in the web server logs called “logs.txt”. The contents of “logs.txt” file is shown below. Please complete the program in the space provided. **(2 Marks)**

```
INFO This is a message with content
INFO This is some other content
WARN This is a warning
ERROR Something bad happened
WARN More details on the bad thing
```

```
from pyspark import SparkContext

# Start your code here

sc = SparkContext(master=“local[2]”, appName=“Errors and warnings Count”)
lines = sc.textFile(“logs.txt”)

# End your code here
```


Office use only

Question 2

a. What are broadcast variables? Why do we need broadcast variables when working with Apache Spark? **(1 + 1 = 2 Marks)**
Write your answer below

b. We want to perform a log analysis. The input data consists of log messages of varying degrees of severity, along with some blank lines. We want to compute how many log messages appear at each level of severity. The contents of "input.txt" file is shown below.

```
INFO This is a message with content
INFO This is some other content
(empty line)
INFO Here are more messages
WARN This is a warning
(empty line)
ERROR Something bad happened
WARN More details on the bad thing
INFO back to normal messages
```

The expected output of the operations is as below.

```
[('INFO', 4), ('WARN', 2), ('ERROR', 1)]
```

Write the code that will produce the expected output Assume spark context object **'sc'** has already been initialised. **(5 Marks)**

Write your answer below

```
from pyspark import SparkContext

# Start your code here

sc = SparkContext(master="local[2]", appName="Errors and warnings Count")
Input = ssc.textFile("input.txt")

# End your code here
```

c. List four different sections of Spark Web UI and briefly explain all of them. **(1 + 2 = 3 Marks)**
Write your answer below

Office use only

Question 3:

a. What is data visualisation? List two importance of data visualisation? What are the factors you need to be aware of before visualising the data? Please explain. **(1 + 1 + 2 = 4 Marks)**
Write your answer below

b. What are the benefits of using Apache Spark and MongoDB together? Assume you have a database named “FIT5202” and a collection named “zips” in MongoDB database. The information on the attributes are as follows

- The **_id** field holds the zip code as a string.
- The **city** field holds the city name. A city can have more than one zip code associated with it as different sections of the city can each have a different zip code.
- The **state** field holds the two-letter state abbreviation.
- The **pop** field holds the population.
- The **loc** field holds the location as a latitude longitude pair.

Assume that spark session object (i.e. spark) has been initialised. The analysis required is **“Find the states with populations above 10 Million”**. Alice, our data analyst, is only familiar with SQL queries so she provided you with the following SQL query:

```
SELECT state, SUM(pop) AS totalPop
FROM zips
GROUP BY state
HAVING totalPop >= (10*1000*1000)
```

You read the data from the MongoDB using the command below:

```
zips_df = spark.read.format("com.mongodb.spark.sql.DefaultSource").load()
```

Use the **functions provided by dataframe** to find the states with populations above 10 million. (1 + 5 = 6 Marks)
Write your answer below

Continue your answer below

Office use only

Question 4

a. What is Machine Learning and why should you use machine learning with Spark? In Apache Spark, machine learning pipelines provide a uniform set of high-level APIs built on top of DataFrames. It makes easier to combine multiple algorithms into a single pipeline, or workflow. The key concepts introduced by the Pipelines API are DataFrame, Transformer, Estimator, Pipeline, and Parameter. **What is a Transformer and an Estimator? (2 + 2 = 4 Marks)**
Write your answer below

b. Suppose we have a set of data comprising; height, weight and shoe size of some customers. The aim is to predict the shoe size of a new customer given only height and weight information.

Height (in cm)	Weight (in kg)	Shoe Size
158	58	36
158	59	36
158	63	36
160	59	38
160	60	38
163	60	38
163	61	38
163	64	40
165	64	40
165	61	40
165	62	40
168	65	40
168	62	40

Write the formula to calculate the Euclidean distance? A new customer named "Matthew" has height 161 cm and weight 61 kg. Using kNN, for k = 5, what is the (most) unlikely shoes size of Matthew? **(2 + 4 = 10 Marks)**

Write your answer below

Office use only

Question 5

a. What is the difference between Supervised Learning and Unsupervised Learning? Mention any two differences. **(2 Marks)**

b. Consider the following data set consisting of the scores of two variables on each of seven individuals:

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Use the k-means algorithm to cluster the data in two clusters. The distance of each data point and the centroid (or mean) is calculated using Euclidean distance. The formula to calculate the euclidean distance is given below. **(8 Marks)**

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Write your answer below

Continue your answer below

Office use only

Question 6

a. 'People who bought this also bought…' recommendations seen on Amazon is based on which algorithm? What is the difference between Association Rules and Collaborative Filtering? **(1 + 1 = 2 Marks)**

b. A simplistic implementation using ALS (Alternating Least Squares) is given below. The goal is to recommend movies to users. Details of the dataset is given below. The full data set ("u.data"), 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. This is a tab separated list of

user id | item id | rating | timestamp

The time stamps are unix seconds since 1/1/1970 UTC. The sample data contents of "u.data" file is shown below.

```
196     242     3       881250949
186     302     3       891717742
22      377     1       878887116
244     51      2       880606923
166     346     1       886397596
```

Write the necessary code below to develop an ALS based recommendation model for movie recommendations. First examine the dataset and perform necessary steps to convert the dataset into DataFrame to make it ready for the algorithm. **(8 Marks)**
Write your answer below

```
from pyspark import SparkContext
from pyspark.sql import SparkSession, Row
from pyspark.ml.recommendation import _____
from pyspark.ml.evaluation import_____

appName="Collaborative Filtering with PySpark"
# initialize the spark session
spark = SparkSession.builder.appName(appName).getOrCreate()

# get sparkcontext from the sparksession
sc = spark.sparkContext

# Step1: the data is loaded to an RDD

movielens_rdd =


# Step 2: process the data into appropriate structure




# Step 3: convert the rdd to dataframes




# Step 4: split the dataset into training and test data (70% training and 30% test)

(trainingData, testData) =

# Step 5: build the recommendation model using ALS on the training data
# Use maxIter = 5, regParam = 0.01, coldStartStrategy = "drop",
# implicitPrefs = False
```

```
# Step 6: predict the top movies for some selected users
predictions = model.transform(testData)

# Step 7: find and print the accuracy of the model
```


Office use only

This page can be utilised to perform rough calculations and will not be marked unless clearly indicated.

This page can be utilised to perform rough calculations and will not be marked unless clearly indicated.