ICT707 Data Science Practice

Semester 2, 2020

Task 2 Examination

First Name:	Gagandeep
Last Name:	Kaur
Student ID:	1121869

Examination Duration: 3 hours

Total marks: 40

Exam Conditions:

This is an open book exam.

Instructions to Students:

- Write your answers after each question. Please do not change the order of the questions.
- Your submission is to be made as a single Microsoft Word document – no other format is acceptable.
- Answer all questions and submit your Microsoft Word document on Blackboard in the Assessment area (Task 2).
- Please rename this file to "ICT707_T2 _***FirstName_LastName_ID***.docx" for submission.

Submission Declaration:

The submission will be checked by SafeAssign. By submitting this assessment item, you declare that your submission is your own work and is in accordance with the University's Student Academic Integrity Policy.

Section A – Short Answer Questions (total 20 marks with 5 marks each)

- What are the 4 categories of data analytics? In your own words discuss the difference between them. Give one example for each category.

Answer:

There are four categories of data analytics:

- **Descriptive analytics: It is used to give answers for event that has already occurred. 80%of generated analytics is descriptive in nature. It require basic skillset. It is often carried out via ad-hoc reporting. Reports generated are static in nature and display historic data. Operations are performed on stored data within an enterprice eg CRM OR ERP system.**

  Eg: What was the sales volume over the past 12 months?

- **Diagnosis analytics: it helps in determining the cause of phenomenon that occurred in past using question that focus on reason behind the event.it gives more value than descriptive and require more advanced skills. Huge amount of data from multiple sources is required that is stored in a structure that lends itself to performing drill-down and roll-up analysis. It require interactive visual tools that enable user to identify patterns and trends.**

  Eg: Why have there been more support calls originating from the Eastern region than from the Western region

- **Predictive analytics: it is carried out as an attempt to predict outcome of a future event. Prediction are made from patterns, events, current or historic events. It can be used to identify risk and opportunities. This technique require advance skillset than both descriptive and diagnostics. Information generated is improved to provide knowledge. Base of this model is association used to predict future from past events. Statistical intricacies are used as tools with user-friendly front-end.**

  Eg: If a customer has purchased Products A and B, what are the chances that they will also purchase Product C?

- **Prescriptive analytics: it provide best action to be taken based on the result of predictive analysis. It focus on best possible solution and it gives a reason for that also. Hence, it is used to gain advantage or mitigating risk. It use business rules and large amount of internal and external data to produce outcome and best course of action.**

  Eg: Among three drugs, which one provides the best results

- In your own words explain the features of RDD, and compare it with Python List.

Answer:

RDD

- Rdd is an immutable distributed collection of elements, it is read-only.
- It is partitioned and distributed for parallel execution
- It is easy retrieve a lost data block hence, it is fault-tolerant
- It is collection of various data set eg arrays, tuples, tables etc.
- Transformation and action methods are used in RDD.

List

- List are mutable that mean we can change it any time.
- List are ordered and have elements have fixed index.
- List consist of datatype like strings, integer or float.
- List can be nested.

- The Hadoop Distributed File System (HDFS) provides robust and reliable file management for the Hadoop ecosystem. Considering building an online video sharing platform which allows users to publish and watch 4K videos. Do you recommend using HDFS for storing users' files? Why?

Answer:

HDFS stores large files on distributed storage. It is only read-only that is once a file is uploaded it can not be modified, user can only read it. It is based on write-once-read-many access model.

As for the video sharing platform it is using 4k videos these file are of large size usually 100GB and with many-read model users can watch the videos many times.

Therefore, I will recommend using HDFS for video sharing platform.

- Amdahl's Law sets the upper limit the speed improvement that can be expected by doing parts of an algorithm in parallel.
    - Assume 60% of the program can be parallelised and the speedup for this part is 10. What is the speedup for the whole program? Please explain your calculation.
    - Assume 60% of the program can be parallelised and there are unlimited resources to boost the speed. What is theoretical

speedup for the whole program? Please explain your calculation.

Answer:

- S lateny(s)=1/(1-p)+p/s

  p is % time improved

  s is speed up part

S lateny(s)=1/(1-60)+60/10

= 2.1739

- S lateny(s)=1/(1-60)+60/100

= 2.4630

Section B – Coding Questions (total 20 marks)

- [**10 marks**] The following code implements the Word Count application.

```
# read input file
file_in is RDD after reading 'test.txt'
print('number of lines in file: %s' % file_in.count())

words = file_in.flatMap(lambda line: line.lower().split(" "))

words = words.filter(lambda x: len(x) > 4)

words = words.map(lambda w: (w,1))

words = words.reduceByKey(lambda x, y: x + y)

words = words.map(lambda x: (x[1], x[0])).sortByKey(False)

# take top 5 words by frequency
words.take(5)

```

Assume the content of "test.txt" has only 2 lines:

```
Write your answers after each question

Submit your answers on Blackboard
```

Please answer the following questions based on the given code (2 marks each).

- Write the content/value of "words" after executing Line 5.
- Explain the purpose of Line 7.
- Explain the purpose of Line 9.
- Explain the purpose of Line 11.
- Write the content/value of "words" after executing Line 13.

Answer:

- ["write","your","answers","after","each","question","submit","your","answer","on","blacboard"]

- Words with length more than 4 are filtered by the filter function. [“write”,”answers”,”after”,”question”,”submit”,”answer”,”blacboard”]
- Map function will produce (key,value) paired RDD

  [(“write”,1), (”answers” ,1), (”after” ,1), (”question” ,1), (”submit” ,1), (” answer” ,1),(”blackboard” ,1)]
- reduceByKey function will combine the values with same keys.it will add the values of similar keys.

  [(“write”,1), (”answers” ,2), (”after” ,1), (”question” ,1), (”submit” ,1),(” blackboard” ,1)]
- values are sorted in descending order.

  [(2,”answers”), (1,“write”), (1,”submit”), (1,”question”), (1,”blackboard”) (1,”after”)]

- [**10 marks**] We have a csv file “students.csv” consisting of the following data:

```
ID,FirstName,LastName,DateOfBirth,Gender,GPA

0001,Test,Exam,15/01/2000,Female,4.8

…many other records…
```

- [4 marks] Please provide two ways of reading this file to get a DataFrame. You do not need to write explicit code but explain your solutions in plain English.
- [3 marks] Write the code with DataFrame DSL to count the number of students for each gender.
- [3 marks] Write the code with DataFrame DSL to get the number of Male students with GPA greater than 4.

Answer:

- Two different methods of creating DataFrames in PySpark
  - From existing RDDs using SparkSession's createDataFrame() method.

    sqlContext.createDataFrame(rdd_var)

Each record is a Row object with schema

- From various data sources (CSV, JSON, TXT) using SparkSession's read method.

  user_df = sqlContext.*read.csv*(“students.csv", header = True, inferSchema = True)

  inferSchema = True: automatically detect the data type of each column; Otherwise all are string

- x=data_df.groupby('Gender')

  x.count().show()

- data_df.filter(data_df.Gender=='Male' && data_df.GPA>4).show()

END OF EXAMINATION