CSC 4760/6760 Big Data Programming

All Students

Exam 2

13:30 pm - 16:00 pm, 05/04/2020

Name: _____

Panther ID: _____

Note: You may use additional white papers if the space is not enough.

This table is to be filled by the grader.

Problem	Total Score	Score
1	10	
2	10	
3	16	
4	12	
5	14	
6	6	
7	12	
8	20	
Total	100	

This page is intentionally left blank.

But there are contents on the backside of the remaining pages.

Problem 1. Rob is installing Spark in Ubuntu 16.04 OS. Please help him with the installation.

Step 1: Since Spark needs Scala, Rob needs to install Scala first. He downloads the Scala (scala-2.12.4.tgz). He unzips the file in the "/home/rob/" directory and renames the folder as "scala". Therefore "/home/rob/scala" is the root directory for Scala. After that, please tell Rob how to update the environment variables "SCALA_HOME" and "PATH".

Answer: Open the ".bashrc" file by using the following command:

$

Add the following two lines to the end of the file to update the environment variables "SCALA_HOME" and "PATH".

1:

2:

Step 2: He downloads the Spark (spark-2.2.1-bin-hadoop2.7.tgz). He unzips the file in the "/home/rob/" directory and renames the folder as "spark". Thus "/home/rob/spark" is the root directory for Spark. Now Rob needs to update the environment variables "SPARK_HOME" and "PATH".

Answer: Open the ".bashrc" file again and add the following two lines to the end of the file.

1:

2:

Step 3: After Spark installation, we can use the following commands to verify the Spark installation. To start the Python Spark shell, we should type:

$

To start the Scala Spark shell, we should type:

$

Step 4: Rob wants to run the WordCount example in the batch mode. Suppose that the Python source code is in the file "WordCount.py", please give the command for running this Python Spark source code file. Suppose the input file name and output file directory are hard coded in the source code, so you do not need to pass those parameters in the command line.

$

Problem 2. Rob has successfully installed Spark in Ubuntu. He also runs the WordCount example successfully. He now can start understanding each line in the WordCount.py file. He opens the pyspark shell. Suppose that the Spark context "sc" is already defined when he opens the shell.

He wants to load the data "/home/rob/data/peterpan.txt". Please tell him how to load the data in pyspark (please provide the command).

Answer: $

What is the type of the return variable?

Answer:

What is the command to split each line into words with the whitespace symbol?

Answer: $

What is the type of the return variable?

Answer:

What is the type of the function you called? transformation or action?

Answer:

Please write the command to emit the (key = word, value = 1) pairs.

Answer: $

Please write the command to reduce the (key = word, value = 1) pairs.

Answer: $

The command to display the contents in the return RDD into the terminal is

Answer: $

Problem 3. Suppose we have the following Fisher's Iris dataset "/home/rob/data/fisher.txt".

0	2	14	33	50
1	24	56	31	67
1	23	51	31	69
0	2	10	36	46
1	20	52	30	65

The numers in each line are separated by the tab symbol "\t". The meaning of the five columns are "type of iris flowers", "petal width", "petal length", "sepal width", and "sepal length" respectively. Please provide Python Spark commands to achieve the following functions.

Please read this "fisher.txt" from the disk by using Pyspark command.

Answer: $

Please split each line into 5 tokens:

Answer: $

Please keep the second to the fifth columns and convert the value from string to float. Please use inline function when you pass a function to Spark.

Answer: $

Please define a Python function and pass it to Spark to achieve the same goal.

Answer:

$ def convert_to_float(x)
$
$
$
$

Please write the command to pass the above newly defined function to Spark

Answer: $

Problem 4. Rob designs two algorithms for solving the Word Counting problem. The two algorithms are shown in the following table.

Algorithm A	Algorithm B
book = sc.textFile("/home/rob/data/peterpan.txt") book.count() book.first() wordCount = book.flatMap(lamba line : line.split(" ")) \ .map(lambda word : (word, 1)) \ .reduceByKey(lambda x, y : x + y) wordcount.collect()	book = sc.textFile("/home/rob/data/peterpan.txt")**.persist()** book.count() book.first() wordCount = book.flatMap(lamba line : line.split(" ")) \ .map(lambda word : (word, 1)) \ .reduceByKey(lambda x, y : x + y) wordcount.collect()

The only difference between Algorithm A and B is that we add ".persist()" at the end of the first line in Algorithm B. Which one (Algorithm A or B) runs faster and why?

Answer:

Instead of persist(), we can also use cache(). What is the difference between persist() and cache()?

Answer:

In the Algorithm A, how many RDDs are there? Please tell the type of the RDD for each. Standard string RDD or key-value pair RDD? Please also explain the meaning of the elements in each RDD.

Answer:

Problem 5. Hadoop and Spark are both share-nothing paradigms. But they do support sharing immutable data structures among all workers (computer nodes) in the computer cluster. Please enumerate two techniques to achieve this goal and fill them in the following table.

Hadoop	Spark

In Spark, Rob needs to create an **accumulator** with initial value "3.14" of double type. Please tell him how to do that:

Answer:

$

Rob wants to increase the value of this accumulator by "1.1" of double type in each executor. Please him how to do that.

Answer: In the source code for each executor, add the following two lines:

$

$

When Rob implements the join algorithm for joining the phone book and country code lookup table, he needs that each executor can access the code lookup table. Suppose that the lookup table is stored in an RDD named "LookupTable1", please tell Rob how to replicate this RDD to all the executors in the driver (master) program.

Answer:

$

How to get the broadcasted data in each executor?

Answer:

$

For each executor, the Accumulators are _____-only variables; the broadcast variables are _____-only variables. Please choose from "read" or "write" when you fill in the previous two blanks.

Problem 6. Basics for Spark Streaming and GraphX

(Spark Streaming) Suppose we create a Spark Streaming Context "streamingContext" in Python Spark.

streamingContext = StreamingContext(sc, 3)

where "sc" is the SparkContext instance. What does the second parameter "3" represent here?

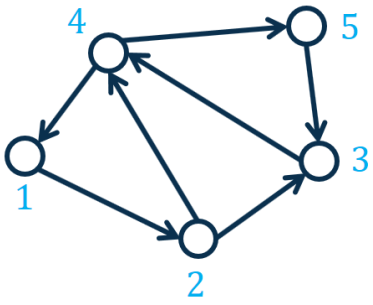
Answer:

We want to stream the data from TCP source with IP Address “123.45.67.89” and port number “4321”. Please provide the function for streaming data from that TCP source.

Answer:

The return variable is a _____.

(Spark GraphX) Suppose we have a graph in the following.



1	2
2	3
2	4
3	4
4	1
4	5
5	3

The left figure shows the graph, and the right column shows the edge list file. Suppose the full directory of the file is "/home/rob/data/ToyGraph01.txt". “sc” is the Spark Context. Please provide the command to load the graph edge list file in Scala.

Answer:

1:

If we want to get the number of vertices, we should use the following command in Scala.

Answer:

1:

Problem 7. Basics for Pig, Hive, and HBase

(Pig) The command to open Pig shell, Grunt, in a local mode is:

Answer:

$

Suppose we design a Pig program to count the number of workers. Suppose the script file name is “CountWorkers.pig”. How to submit the script file in the terminal such that it will run in local mode? Please provide the command.

Answer:

$

(Hive) In Hive, what is the command to start a Hive shell?

Answer:

$

In Hive, what is the command to run a script in non-interactive mode? Suppose the script file is "example.sql"

Answer:

$

In Hive, Please provide the command to list all the databases:

Answer:

hive>

(HBase) Previously, in Hadoop, we used "start-dfs.sh" and "start-yarn.sh" commands to start HDFS and YARN services. In HBase, what is the command to start HBase service?

Answer:

$

This file is in which directory?

Answer:

After we start the HBase service, we can start the HBase shell. Please provide the command to start the HBase shell?

Answer:

$

Suppose we already have a table "workers". The workers' information has been put into this table. Rob wants to view all the information. Which command should Rob type in the shell in order to see all the workers' information in the table "workers"?

Answer:

$

If we program a "CountWorkers.java" source code file for using HBase to count the workers, how should we run this HBase program? Rob figures out we need to first compile it into a runnable "CountWorkers.jar" file. Then what is the command to run this "CountWorkers.jar" file?

Answer:

$

Problem 8. Suppose we have the following workers' information.

name	age	gender	occupation
Mike	19	Male	Computer Scientist
Paul	26	Male	Computer Scientist
Bob	25	Male	Computer Scientist
Olivia	30	Female	Accountant
Rob	32	Male	Computer Scientist
Susan	36	Female	Computer Scientist
David	35	Male	Accountant
Emma	44	Female	Accountant
Lisa	32	Female	Accountant

The data is stored in a json file "/home/rob/exam2/workers_spark.json". The data file is uploaded into iCollege.

We want to compute the number of workers above age 20 in each gender and each occupation. That is, we want to get the following table from the above one.

gender	occupation	count
Male	Computer Scientist	3
Female	Computer Scientist	1
Male	Accountant	1
Female	Accountant	3

Note that Mike is 19, which is smaller than 20. Therefore, he should be filtered out and not counted.

From the above result table, we can see that there are more male workers in Computer Science and more female workers in Accounting. This is what we learnt from the original workers' information table.

Now please design Python Spark algorithm to implement this function. You are required to use Spark Dataframe APIs. The data file is uploaded into iCollege. You may want to program with the data and debug and make sure that your answers are correct. The last line should show the results in the terminal.

Answer: (Only show the key lines of the source code. Do not need the preparation code for Spark Context and Spark SQL Context)

1:

2:

3:

4:

We can also use Spark SQL to implement the same function. Spark SQL allows you to use the SQL-like sentences like “SELECT * FROM …” to operate on the dataset. In this method, you need to call “createOrReplaceTempView()” and “spark.sql(…)” functions to achieve filter and group by functions. Please provide the source code. The last line should show the results in the terminal.

Answer: (Only show the key lines of the source code. Do not need the preparation code for Spark Context and Spark SQL Context)

1:

2:

3:

4:

5:

6:

If we want to use Pig to achieve the same goal, what are the source code for doing that?

The data is stored in a different json file “/home/rob/exam2/workers_pig.json”. The format in this file is slightly different than that in the previous “workers_spark.json” file because Spark and Pig have different parsers for Json files. The json data file is also uploaded into iCollege (in the folder of “Exam 2”). You may want to program with the data and make sure that your answers have no bugs. The last line of the code should write the results into the folder “PigOutput” on the disk.

Answer: (Please provide the entire source code for Pig)

1:

2:

3:

4:

5:

(This is the end of Exam 2.)