Monash University

Bigdata Assignment-Programming

Marks: This assignment is worth 140 marks and it forms 40% of all marks for the unit

Type: Individual submission

Due Date: Week 10: **Mon 01-Oct-2018, 2pm**

Submission: The assignment must be uploaded to the unit's portal on Moodle as a single ZIP or GZIP file (using any other format will lead to your assessment being delayed); including a completed assignment submission cover sheet.

Lateness: A late submission penalty of 5% marks deduction per day will apply including the weekends and public holidays.

Extensions: If due to circumstances beyond your control, you are unable to complete the assignment by the due date, you should submit the incomplete assignment, presented in a professional manner, and complete a request for special consideration.

Authorship: This assignment is an individual assignment and the final submission must be identifiably your own work. You may be required to attend an interview with the marker to confirm that it is your own work (this submission will be submitted to Turnitin for screening by the marker).

Specifications

1. **Spark-Scala Programming Fundamentals [30 marks]**
   Provide spark-shell executable coding for the following tasks in a file named **q1.scala** (plain text). The program outputs must show clearly in spark-shell (failure to do so may lead to loss of marks). Your file must be appropriately commented to ensure that all significant programming steps have been clearly explained.

   a. Create a Spark data frame from a CSV file which has the headers in the first row (create a small CSV file or use ~/ /Documents/Datasets/simple.csv in the bigdata virtual machine) and verify. [4+1 = 5 marks]

   b. Print the data frame's schema. [1 marks]

   c. Convert the data frame to a RDD and display its contents. [1+1 =2 marks]

   d. Create a RDD by reading from a text file (create a text file or use $SPARK_HOME/README.md in the bigdata vm). [2 marks]

   e. Calculate the total length in characters, including white spaces, for all the lines in the $SPARK_HOME/README.md file. [5 marks]

   f. Count and display all the words as (String, Int) pairs, which occur in $SPARK_HOME/README.md file of the bigdata vm. [5 marks]

   g. Write a program which does word count of the $SPARK_HOME/README.md file using Spark. Explain the reduction operation. [2+3 = 5 marks]

   h. Factorial is an integer number calculated as the product of itself with all number below it e.g. Factorial of 3 or 3! = 3x2x1 = 6. Factorial of 0 is always 1. Using these rules write a compact program, which computes the factorials of an integer array X(1,2,3,4,5) and then sums these up into a single value. [5 marks]

Q1 Rubrics	
Evaluation Criteria	Marks %age
Correctness of the coded solutions	60%
Commenting	30%
Coding style and sophistication (e.g. use of fewer lines to express the program logic)	10%
Total	100%

2. **Data Exploration with Spark [20 + 10 = 30 marks]**
   Provide spark-shell executable code for the following task in a file named **q2-a.scala** (plain

text) and a PDF (or Word) file, **q2-b.pdf**, explaining the program design approach. The program outputs must show clearly in spark-shell (failure to do so may lead to loss of marks). Your code file must be appropriately commented to ensure that all significant programming steps have been clearly labeled.

a. Using a parquet-formatted dataset on flight data, flight_2008_pq.parquet/, available in bigvm's **`~/Documents/Datasets/flight_2008_pq.parquet`** (and also provided as flight_2008_pq.parquet.zip in Moodle), calculate and display the maximum flight departure delays (DepDelay) for up to 20 flights. Re-arrange and display the delays in descending order (listing the flight with the highest delays at the top).

b. Provide a written explanation, of no more than 500 words, of your Spark-Scala code for calculating these delays. The explanation should include your choice of Spark APIs and their brief explanation. Include a flowchart and explanatory figure/s where applicable.

Q2 Rubrics	
2-a Evaluation Criteria	Marks
Correctness of the code	12
Commenting	4
Coding style and sophistication (e.g. use of fewer lines to express the coding logic)	2
2-b Evaluation Criteria	
Write-up: clarity of explanation, well formatted, within word limit, referencing (if applicable)	6
Flow chart	2
Explanatory figure/s	2
Total	30

3. **Provide short answers (in one to two paragraphs) to the following questions [20+40 = 60 marks]**
Provide a PDF (or Word) file named **q3.pdf** for this question.
Programming Related:

a. Compare and contrast an Apache Spark data set with a data frame. (10 marks)

b. Compare and contrast reservoir sampling with bloom filter. (10 marks)

Framework Related:

c. Discuss the main differences between Apache HBase with Apache Spark. (10 marks)

d. List the main benefits of integrating Apache Spark with Hadoop HDFS. (5 marks)

e. Explain how Hadoop implements computational parallelism in terms of the parallel dwarf/s it employs and Flynn's taxonomy (5 + 5 = 10 marks).

f. Outline the main design features of the RDD abstraction for in-memory cluster computing? (15 marks)

Q3 Rubrics	
Evaluation Criteria	Marks as %age
Write-up: clarity of explanation, well formatted, referencing (if applicable)	80%
Table/s and Explanatory figure/s	20%
Total	100%

Lab Test (A-Prog Demos): Week11-12 [20 marks]

Your lab tutor will randomly select a small set of programming tasks and/or theory questions, in-total one to three*, from this assignment. These items must be re-worked and demonstrated during the lab. Time allocation will be upto **25* minutes** for completion of these tasks and **5 minutes/student** to demonstrate these tasks.

This test is, **Closed Book i.e. reference to your assignment submission, unit contents, personal notes/files, and web/Internet resources is not allowed with the exception of web resources [1] and [3] listed at the end, under resources, for this assignment.**

Lab Test Submission Instructions:

1. **Write your (i) full name, (ii) student id, (iii) lab day+time, and (iv) include the questions with your answers.**
2. **Provide plain text with .scala extension for codes and PDF/Word file for theory. Upload your outputs as a single ZIP/GZIP file to 'Moodle Lab Test' submission at the end of the lab test. Any issues with Moodle upload, you must email** the zip archive to your tutor before the end of your lab session. NO LATE SUBMISSION ACCEPTED.**

Lab Test Rubrics	
Evaluation Criteria	
All tasks demonstrated correctly	15
Tutor's questions answered satisfactorily	5
Total	20

*The number and duration of tasks may vary; these are indicative values only.
** You may use `sftp` or `scp` to transfer files between the virtual machine and your lab PC or personal notebook computer.

Interviews

During Week 11 and 12 each student in the lab will be interviewed as part of the in-lab test.

Interview duration: 5 minutes (max)
Interview procedure: Each student will be asked up to three questions to test their general understanding of practical terms in bigdata e.g. What is HBase? It is a non-relational database, which can be distributed over a cluster for scalability.

Pass criterion: At least one question must be answered correctly to pass this hurdle..

Final In-Lab Test marks (out of 20) will be allowed upon passing the interview.

Resources

[1] Spark 2.2.1 Scala API Docs

http://spark.apache.org/docs/2.2.1/api/scala/index.html#org.apache.spark.package

[2] Spark SQL, Dataframes and Datasets http://spark.apache.org/docs/2.2.1/sql-programming-guide.html

[3] Scala Docs http://scala-docs-sphinx.readthedocs.io/en/latest/style-guide/scaladoc.html