**Assignment 2**

**Big Data Analytics**

MET CS777

Faculty - Farshid Alizadeh-Shabdiz, PhD, MBA

## **Problem 1 (10 points)**

What are the differences between following RDD operations from functionality’s perspective, and also computation and cost of operation.

Assign a level of computation complexity from level 1 (less costly) to level 3 (most costly) to each one.

1. aggregateByKey()
2. reduceByKey()
3. groupByKey()
4. combineByKey()

You can find Spark documentation online at <https://spark.apache.org> .

# **Problem 2 (5 points)**

Name at least four differences between Spark and Hadoop MapReduce?

# **Problem 3 (5 points)**

How does Spark run an application and what are the driver layer functionalities? Explain from the Spark architecture’s perspective.

# **Problem 4 (5 points)**

What are the differences of running on multi-core computer versus running on multi worker/executor environment?

List pros and cons of each one.

# **Problem 5 (10 points)**

Why RDD is immutable? Was this a mistake in design of RDDs or has some advantage?

# **Problem 6 (10 points)**

Spark transformation is divided to narrow transformation and wide transformation. Referring to spark documentation, explain the differences.

# **Problem 7 (15 points)**

List 10 spark transformation operation with one line of example

List 5 spark action operation with one line example.

# **Problem 8 (20 points)**

We have a data which consists of following columns

* Row number
* First name of student
* Last name of student
* Course number
* Grade

Write an efficient Spark code to calculate

1. Min grade of each student
2. Max grade of each student
3. GPA
4. Number of courses taken

# **Problem 9 (20 points)**

Estimation area of a circle:

Use Spark to estimate area of the unit circle by "throwing darts" at the circle.

Assume you don’t know how to calculate area of a circle in a closed form, but you know how to calculate area of a square. You throw random darts/points in the 2 by 2 square ((-1, -1) to (1,1)) and count how many falls in the unit circle, a circle with radius of one. The fraction can be used to estimate of the area of the unit circle.

# **Academic Misconduct Regarding Programming**

In a programming class like our class, there is sometimes a very fine line between ”cheating” and acceptable and beneficial interaction between peers. Thus, it is very important that you fully understand what is and what is not allowed in terms of collaboration with your classmates. We want to be 100% precise, so that there can be no confusion.

The rule on collaboration and communication with your classmates is very simple: you cannot transmit or receive code from or to anyone in the class in any way—visually (by showing someone your code), electronically (by emailing, posting, or otherwise sending someone your code), verbally (by reading code to someone) or in any other way we have not yet imagined. Any other collaboration is acceptable.

The rule on collaboration and communication with people who are not your classmates (or your TAs or instructor) is also very simple: it is not allowed in any way, period. This disallows (for example) posting any questions of any nature to programming forums such as StackOverflow. As far as going to the web and using Google, we will apply the ”two line rule”. Go to any web page you like and do any search that you like. But you cannot take more than two lines of code from an external resource and actually include it in your assignment in any form. Note that changing variable names or otherwise transforming or obfuscating code you found on the web does not render the ”two line rule” inapplicable. It is still a violation to obtain more than two lines of code from an external resource and turn it in, whatever you do to those two lines after you first obtain them.

Furthermore, you should cite your sources. Add a comment to your code that includes the URL(s) that you consulted when constructing your solution. This turns out to be very helpful when you’re looking at something you wrote a while ago and you need to remind yourself what you were thinking.