COMP9313 - 2019T2
Assignment #2 (Spark) - 25 points total

1 Problem Statement

Given a log file that records HTTP requests (GET and POST) sent to a set of servers, you are asked to compute descriptive statistics on the amount of data (payload) communicated through such requests. These statistics need to be computed for each base URL (e.g., "`http://subdom0001.example.com`") in the log and the results must be reported in bytes. More specifically, you need to compute the following statistics per each sub-domain):

Minimum payload: The smallest payload communicated for each base URL.

Maximum payload: The largest payload communicated for each base URL.

Mean payload: Mean of payloads for each base URL. For computing the mean, consider the following formula (population mean):

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

where N is the size of the population being explored.

Variance of payload: The variance of payloads for each base URL. For computing the variance, consider the following formula (population variance):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

where N is the population size and μ is the population mean.

Notice that the statistics above need to be computed for each base URL in the log.

You are asked to develop a Spark solution to this problem (using Scala). Your solution must be runnable using the *spark-shell* interpreter. Your solution must rely on **RDDs** to solve this problem (the use Spark DataFrames and DataSets are ***not allowed*** for this assignment).

2 Input

The log file in input is a CSV file that contains one line HTTP per request. The file consists of three columns: Base URL of the HTTP request, endpoint, HTTP method, and size of payload. An excerpt of the log is shown below:

```
http://subdom0001.example.com,/endpoint0001,POST,3B
http://subdom0002.example.com,/endpoint0002,GET,431MB
http://subdom0003.example.com,/endpoint0003,POST,231KB
http://subdom0002.example.com,/endpoint0002,GET,29MB
http://subdom0001.example.com,/endpoint0001,POST,238B
http://subdom0002.example.com,/endpoint0001,GET,32MB
http://subdom0003.example.com,/endpoint0003,GET,21KB
```

Notice that the payload is given in different units of digital information (e.g. MB and KB). The log file for this assignment will contain the following units only: B (for bytes), KB (for kilobytes) and MB (for megabytes).

Assume that the log file is stored in a local filesystem (we will not use HDFS or similar for this assignment). We provide a sample log file in the link below:
https://webcms3.cse.unsw.edu.au/COMP9313/19T2/resources/28426

3 Output

The output consists in a CSV file that contains the list of base URLs along with the descriptive statistics (for each base URL) as presented in Section 1. We show a sample output below:

```
http://subdom0001.example.com,3B,238B,120B,13806B
http://subdom0002.example.com,30408704B,451936256B,171966464B,39193191483703296B
http://subdom0003.example.com,21504B,236544B,129024B,11560550400B
```

The columns in the output above are as follows: Base URL, minimum payload, maximum payload, mean payload, variance of payload. You do not need to provide a *header* for your CSV file.

Notice that the statistics must be expressed in bytes (B), as shown in the example above. Note also that you may get float / double numbers when computing means and variance. In such cases, *truncate* the numbers to keep just the whole number part. For example:

- 120938.32 -> 120938
- 9983.89 -> 9983

We provide a sample output in the URL below:
https://webcms3.cse.unsw.edu.au/COMP9313/19T2/resources/28428

4 Input file and output file specification

You will need to specify (in your Scala program) the path of your input file and output directory (where the results file will be stored) using two Scala values (*val*) with the following names:

```
val inputFilePath = "FULL_PATH_OF_YOUR_INPUT_FILE"

val outputDirPath = "FULL_PATH_OF_YOUR_OUTPUT_DIRECTORY"
```

Use this value names (`inputFilePath` and `outputDirPath`) whenever you need to read/write the input/output.

During assessment, we will change these values to the path of the input file and output directory used for assessing your work.

Please, locate the declaration of these values at the very beginning of your program. You can use the code template provided in the link below to write your solution:

https://webcms3.cse.unsw.edu.au/COMP9313/19T2/resources/28425

5 Running your Program

We will run your program using spark-shell. More specifically, we will use the following command to load and run your Scala program:

```
scala> :load assignment2.scala
```

6 Assignment Submission

Deadline: 21 July 2019 20:59:59

Log in to any CSE server (e.g. williams or wagner) and use the *give command* below to submit your solution:

```
$ give cs9313 assignment2 z9999999.zip
```

where you must replace z9999999 above with your own zID. The zip file above must contain the following:

- The file assignment2.scala containing your solution (Scala program)
- A PDF document, named assignment2_solution.pdf (maximum 1 page, 10 points font-size Arial), that explains your solution (use of figures is highly encouraged to explain your solution).

You can also submit your solution using WebCMS, or Give:

https://cgi.cse.unsw.edu.au/~give/Student/give.php

If you submit your assignment more than once, the last submission will replace the previous one. The late submission penalty (below) will be applied based on the timestamp of your last submission. To prove successful submission, please take a screenshot and keep it for your own record. If you face any problem while submitting your code, please e-mail the Course Admin (Maisie Badami, m.badami@student.unsw.edu.au)

7 Late submission penalty

10% reduction of your marks for the 1st day, 30% reduction/day for the following days.

8 Assessment

Your source code will be manually inspected and marked based on readability and ease of understanding. We will run your code to verify that it produces correct results. The code documentation (i.e. comments in your source code) and solution explanation (PDF document) are also important. Below, we provide an indicative assessment scheme (maximum mark: 25 points):

Result correctness	15 points
Documentation (PDF document)	5 points
Code structure and source code documentation (comments)	5 points

9 Plagiarism

This is an *individual assignment*. The work you submit must be your own work. Submission of work partially or completely derived from any other person or jointly written with any other person is not permitted. The penalties for such offence may include negative marks, automatic failure of the course and possibly other academic discipline. Assignment submissions will be examined manually.

Do not provide or show your assignment work to any other person - apart from the teaching staff of this course. If you knowingly provide or show your assignment work to another person for any reason, and work derived from it is submitted, you may be penalized, even if the work was submitted without your knowledge or consent.

Reminder: Plagiarism is defined as using the words or ideas of others and presenting them as your own. UNSW and CSE treat plagiarism as academic misconduct, which means that it carries penalties as severe as being excluded from further study at UNSW. There are several on-line sources to help you understand what plagiarism is and how it is dealt with at UNSW:

- Plagiarism and Academic Integrity
- UNSW Plagiarism Procedure

Make sure that you read and understand these. Ignorance is not accepted as an excuse for plagiarism. In particular, you are also responsible for ensuring that your assignment files are not accessible by anyone but you by setting the correct permissions in your CSE directory and code repository, if using one (e.g., Github and similar). Note also that plagiarism includes paying or asking another person to do a piece of work for you and then submitting it as your own work.

UNSW has an ongoing commitment to fostering a culture of learning informed by academic integrity. All UNSW staff and students have a responsibility to adhere to this principle of academic integrity. Plagiarism undermines academic integrity and is not tolerated at UNSW.