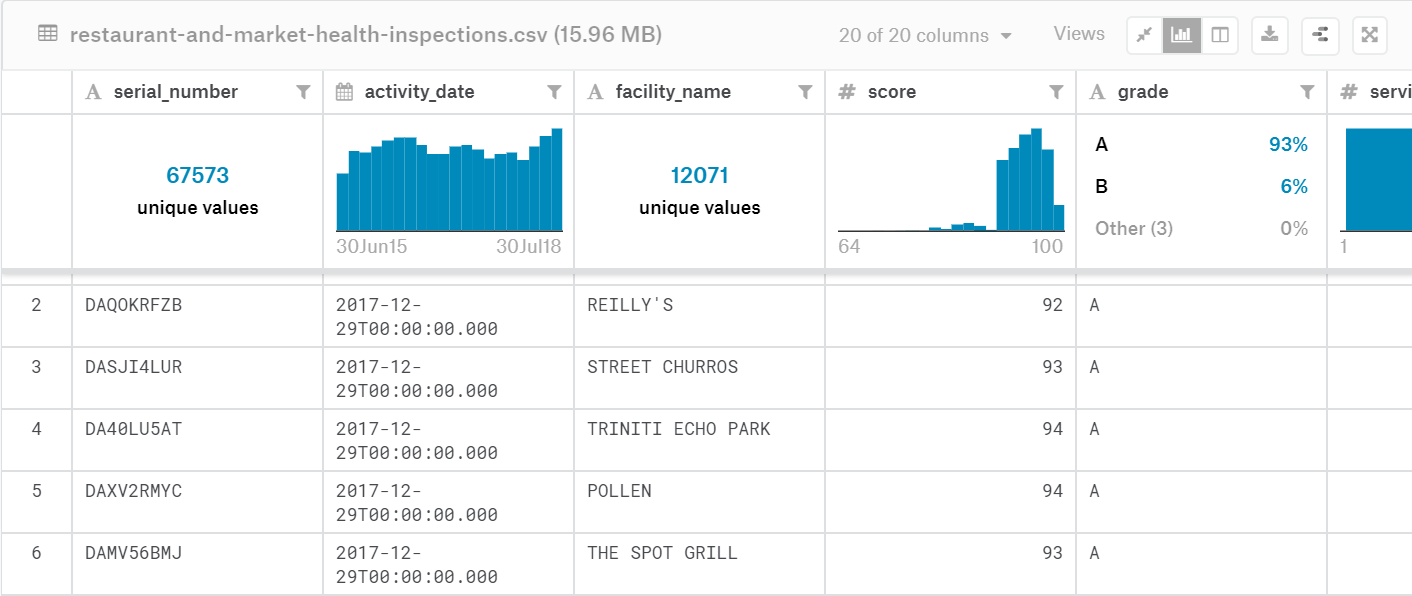
# Homework #1: Data Analysis via Spark & Hadoop

## Due: September 27, Friday

## 100 points

Consider the LA Restaurants & Market Health data set available at Kaggle: <https://www.kaggle.com/cityofLA/la-restaurant-market-health-data>. In particular, we consider the two CSV files: one for [inspection](https://www.kaggle.com/cityofLA/la-restaurant-market-health-data#restaurant-and-market-health-inspections.csv); the other for [violations](https://www.kaggle.com/cityofLA/la-restaurant-market-health-data#restaurant-and-market-health-violations.csv).

Kaggle displays useful statistics for each column, as shown below.



In this work, you are asked to compute several statistics about the two datasets using Spark (in Python) and Hadoop (in Java). To load CSV file into Spark, you may use spark.read.csv(), which will create a data frame (see <https://spark.apache.org/docs/latest/sql-getting-started.html>). You are required to turn data frame into RDD and use the operations covered in class for your homework.

* [60 points, Spark] Write a Spark script for each of the following tasks. Name the script Firstname\_Lastname\_q1.py, firstname\_lastname\_q2.py, and so on.

1. [10 points] Determine the number of unique values of facility names in the inspection dataset.
2. [20 points] Compute a histogram for the scores. The histogram shows the range by tens and the number of scores that fall into the range. For example, if the valid score ranges from 64 to 100, then the ranges include: [60, 69], …, [90,100]. Note that last range has 11 distinct values. Note also that Kaggle divides the scores into different ranges.
3. [15 points] Compute the percentages of different letter grades. For example, 93% are A.
4. [15 points] Determine the number of facilities (by their ids) that only appear in the inspection dataset, but no in the violation dataset.

For each script, please output the result in a text file named Firstname\_Lastname\_q1.txt, Firstname\_Lastname\_q2.txt and so on.

For q1, the submitted file should only contain a number.

For q2, please follow the format example shown below, where x is the number of scores that fall into the range:

[60,69]:x

[70,79]:x

…

[90,100]:x

For q3, please follow the format example shown below:

A:93

B:6

….

For q4, the submitted file should only contain a number.

* [40 points] Write a Hadoop MapReduce program “SQLCount.java” to find out the id, name, and the number of inspections done for the facility.

Execution format:

hadoop jar Firstname\_Lastname\_SQLCount.jar SQLCount <input-dir> <output-dir>

where <input-dir> stores the inspection CSV file with the header file removed.

The output file should be in the output dir and follow the format example shown below:

ID1,Name1:number1

ID2,Name2:number2

…..

Please upload a tar file, named Firstname\_Lastname\_hw1.tar. The tar file should contain all your programs. For Hadoop task, please upload both .jar and .java files.

Please double check your program could be run by the execution format successfully.

1. Your program will be manually terminated after a reasonable time. Programs that run for a long time, e.g., over 5 minutes, will not be graded.
2. If you name your files incorrectly, you may be subject to 10% deduction of points for the question.
3. If your program fails to compile, the program won’t be graded.
4. If the output file does not follow the correct format, e.g. wrong tag name, no points will be given for the question.

Please notice that your submission may be compared with programs from other students, previous students, and Github, for plagiarism checking purpose. All violations of academic honesty will be reported to academic office and you will receive zero for this assignment.