Hi All,

Below are answers to the questions posted about the project.

Note that I have extended the deadline for part I. The new deadline is April 17th.

Best,

Juliana

- For part I, you will only submit the *scripts*. You should not submit the output of the script. You will use the output to generate a summary of the data set (e.g., number of valid/invalid values for each column, number of null values, etc.)

  Your script should output for each value its base type, semantic type and whether the value is NULL, VALID, or INVALID.

  For example,

2015-01-15 19:23:42 DATETIME Pickup datetime VALID

1015-01-15 19:23:42 DATETIME Pickup datetime INVALID

(999)999-99999 TEXT phone NULL

(212)234-5678 TEXT phone VALID

- You can submit multiple scripts (one for each column) or a single script that handles all the columns. If you do the laltter, create one separate function per column.

- Your scripts must be *scalable* and should handle large data. The scripts should process the data in parallel, using Hadoop or Spark

- The project is supposed to give you hands-on experience with big data. Therefore it is not sufficient to work with a small sample (e.g., 1 month of data). You must use multiple years. In addition, since you have to explain patterns in the data, by looking at a small sample, you are unlikely to find many interesting patterns.