Lab 10 STATS 769 Due 4:00pm Tuesday 22 October

The main point of this lab is to try out **high performance computing** frameworks. We will perform some simple analyses using both standard R packages and two high performance computing environments.

The Data

The data set for this lab is a single large CSV file containing just the vehicle type, trip distance, and trip duration variables for electric vehicle trips. This file is available on the VMs at the following location:

`/course/data.austintexas.gov/type-durn-dist.csv`

The Task

1. Read the entire CSV file into R using `data.table::fread()`.

   Define an "input format" for the CSV file using `rmr2::make.input.format()`.

   Load the CSV file into Spark storage using `sparklyr::spark_read_csv()`.

   Compare the sizes of the three R objects that are created as a result of these function calls. Why are they so different?

2. For trips with non-positive durations and distances, generate two new variables, log duration and log distance, using **data.table**.

   Do the same operation on the data in Spark storage using **dplyr** functions (`filter()` and `mutate()`).

   Compare the sizes of the R objects that are created.

3. Calculate average trip distances and durations (for trips with both positive distance and positive duration) for both bicycle trips and scooter trips, using **data.table**.

   Do the same calculation using `rmr2::mapreduce()`. You might want to start by just trying to get the average trip distance, then break that into average trip distances for different types of trips, then expand to averages for both trip distances and trip durations for different types of trips.

   Do the same calculation using the data in Spark storage, using **dplyr** functions (`group_by()` and `summarise()`).

   Check that the results from all three calculations are the same.

4. Fit a simple linear regression model that predicts log duration from log distance using `stats::lm()`.

   Fit the same model using the data in Spark storage.

   Measure and compare the memory usage in R for both approaches.

   Check that the coefficients from the two model fits are the same.

The Report

Your submission should consist of a *tar ball* (as generated by the `tar` shell tool) that contains an R Markdown document *and* a Makefile *and* a processed version of your R Markdown document, submitted via Canvas.

You should write your document and your Makefile so that the tar ball can be extracted into a directory anywhere on one of the virtual machines provided for this course (`sc-cer00014-04.its.auckland.ac.nz` or `sc-cer00014-05.its.auckland.ac.nz`) and the markdown document can be processed just by typing `make`.

Your report should include:

- A description of the data format.
- A description of each of "The Task"s.
- A conclusion summarising the analysis.

Your report should NOT be longer than **10 pages**.

Your report should NOT take longer than **2 minutes** to process.

Your tar ball should NOT be larger than **1 Mb**.

Marks will be lost for:

- Submission is not a tar ball.
- More than 10 pages in the report.
- R Markdown file does not run.
- R Mardown file takes too long to process.
- Tar ball is too large.
- Section of the report is missing.
- R Markdown file is missing.
- Processed file (pdf or docx or html) is missing.
- Makefile is missing.
- Significantly poor R (or other) code.