CSC 8634 - Cloud Computing

Marc Birkett

17/12/2021

Introduction

This project will be in exploratory data analysis project. To bring structure to this project the CRISP-DM methodology will be followed as “it is soundly based on the practical, real-world experience of how people conduct data mining projects.” (Chapman et al, 2000). To aid organisation and repeatability of the project various packages from the Tidyverse will be used, particularly ReadR, DPlyr, GGPlot2 and ProjectTemplate. The methodology splits a data mining project into 5 stages which will provide structure to this document, these are Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation.

Business Understanding

(Business objectives, assess situation, goals, project plan)

What is the need for the project? - Justify your choice of response (i.e. the nature of, and your plan for, your project). To give strength to your argument, you should reference to practice elsewhere (e.g. in academic literature, or industry practices.

Hoefler, Torsten and Belli (2015) state that the “measuring and reporting performance of parallel computers constitute the basis of scientific advancement of high performance computing … and that the state of practice is lacking”. This paper conducts a performance evaluation of terapixel rendering in cloud super computing. The solution was rendering using an Infrastructure as a Service (IaaS) cloud environment and up to 1024 graphical process unit (GPU) nodes which was used to compute a realistic visualisation of Newcastle Upon Tyne and its environmental data as captured by the Newcastle Urban Observatory. The data was subsequently provided for analysis via comma separated value files. There will also subsequently be a dashboard created to allow investigation of the data set. The completion of this paper will contribute to the some of knowledge regarding the measurement and assessment of metrics on cloud based supercomputers.

Data Understanding

(collect initial data, describe, explore, data quality) What, concisely, did you do?

Data Preparation

(select, clean, construct, integrate, format) What, concisely, did you do?

Modelling

(technique, test design, build, assess) R-Shiny What, concisely, did you do?

Evaluation

(evaluate, review, next steps) How successful has it been? Provide evidence, using appropriate evaluation methodologies, and comment on the strengths/weaknesses of your evidence in answering this question. What are the future implications for work in this area? If applicable, which areas of extension work are now possible due to the foundational work you have performed in this project?

References

Hoefler, Torsten, and Roberto Belli, 2015, “Scientific benchmarking of parallel computing systems: twelve ways to tell the masses when reporting performance results.” In Proceedings of the international conference for high performance computing, networking, storage and analysis, p. 73. ACM, 2015.

Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas

Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler), 2000, CRISP-DM Step-by-step data mining guide

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

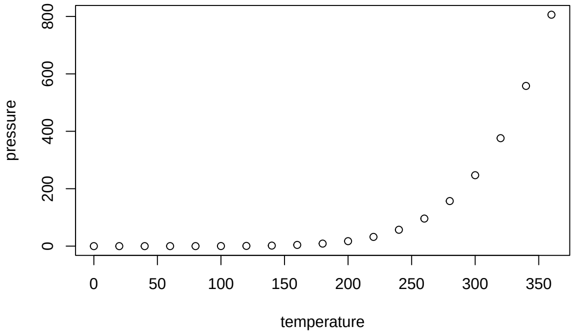
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.