Terapixel Project

Michael Luke Battle

30/12/2020

Project Requirements

Written Report Outline

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)

Discuss and implementation of CRISP-DM

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?

A brief reflection on your personal and professional learning in undertaking this project. Here, you can comment on how you found the process, what you learned about the technologies and methodologies you used, which aspects you found most difficult/straightforward, and any conclusions which will inform the way you undertake similar projects in the future

Structured Abstract

REMEMBER TO INCLUDE STRUCTURED ABSTRACT

EDA Findings.

You should also produce additional documentation detailing the findings from your exploratory analysis. You are encouraged to make use of a literate programming framework, e.g. R Markdown, to align analytic code with narrative tetx. You should submit the source file(s) for the notebooks as well as output saved in PDF Format. No limit on length for this document!

CRISP-DM Methodology:

Business Understanding

Determine Business Objectives

Background

Maybe include more background information on Newcastle University and data

With world-leading research in data, Newcastle University is always undertaking new challenges to advance the field. In order to keep up with ever-expanding amount of data produced by cities, one such challenge is to be able to effectively convey the information from this data to stakeholders. To address this issue, the University has created a terapixel image of the city of Newcastle upon Tyne, including environmental data such as temperature and humidity, from sensors across the city. This terapixel image contains over one trillion pixels and not only allows the entire city to be viewed, but is rendered at 12 different levels allowing users to zoom in whilst retaining full picture detail. However, rendering over one trillion pixels requires immense computing power, and so the University has made use of public cloud service *Microsoft Azure* to perform the rendering process. This facilitates a far quicker rendering process, which is more suitable for a primary requirement of the project - to ensure the terapixel image supports daily updates.

Business Objectives

Whilst Newcastle University has created this terapixel visualisation, it is important that the rendering process with cloud supercomputing is rigorously evaluated in order to identify any inefficiencies.

Business Success Criteria

Assess Situation

Determine Data Mining Goals

Produce Project Plan

Data Understanding

Describe Data

In order to evaluate the rendering process, Newcastle University has provided three files detailing various aspects of a single run of the rendering process. The file “application.checkpoints.csv” has dimensions and variable names:

```
dim(application.checkpoints)
```

```
## [1] 660400      6
```

```
colnames(application.checkpoints)
```

```
## [1] "timestamp" "hostname"  "eventName" "eventType" "jobId"     "taskId"
```

provides the times that each event as part of the rendering process started and stopped for each pixel, as well as the total render time. These events are “Tiling”, “Saving Config”, “Render” and “Uploading”. The file “gpu.csv” details

Three files, task.x.y, application.checkpoints and gpu.

Look for blanks

Look for duplicates

Data Preparation

Go through process of creating full merged dataset to combine all three

Modelling

Create several graphs to view different interactions of data set

Which event types dominate task runtimes? (app checkpoints)

What is the interplay between GPU temperature and performance? (gpu)

What is the interplay between increased power draw and render time? – Can we quantify the variation in computation requirements for particular tiles?

Can we identify particular GPU cards (based on their serial numbers) whose performance differs to other cards? (i.e. perpetually slow cards).

What can we learn about the efficiency of the task scheduling process?

Evaluation

Deployment