CSC8634 Cloud Computing Report

Angelos Nikolas (210444335)

Written report

Cloud analytics allow for enterprises to understand and process the data produced from cloud computing endeavors. Utilizing the information generated throughout all aspects of the tasks undertaken. Understanding the data can then produce beneficial techniques, reduce errors and pave the way of the creation of models that are able to be utilized alongside the development of the product. The outcomes of data science use cases on cloud computing vary, however some indicative elements are listed below: 1. Improved operational efficiency. 2. Generation of revenue in a sustainable procedure. 3. Pattern recognition (errors, negative impact of methods) 4. Utilization of data for product improvement. With these in mind this project focuses on the understanding of the data produced from Terascope terapixel rendering of the city of Newcastle. An Exploratory Data Analysis commences with six initial objectives which could provide an view in terms of performance and structure. The work undertaken is based on the Project Template framework in R Studio with a variety of libraries. In addition, Git version control is used to catalog all development activities. The data were in an excellent condition and almost ready to be utilized except the timestamp columns. These columns needed to be mutated to a new format using the dplyr library to be suitable for the work intended.

EDA Objectives

Objective 1 Which event types dominate task runtimes?

For answering this question, the application checkpoints was cleaned and a new subset was created containing the rows needed. Then each event type was filtered, for example the tilling event has a Start and a Stop timestamp each time it occurs. Each unique task ID contains 10 events with the time they are started and stopped. In order to find how much time, it took for tilling subtraction is used with the lubricate library. For the subtraction to work and be accurate the stop times and start times are left joined by taskId. This was replicated for all event types and the average time was computed and stored in a new data frame to be utilized for analysis.

Objective 2 What is the interplay between GPU temperature and performance?

For this question 2 assumptions/notes were made:

1. High temperature may cause performance issues, if the temperature is bouncing near the top threshold indicated by the GPU manufacturer (usually 55-60 the GPU may under perform)

2.The highest temperatures may indicate high levels of core usage percentages (the tasks undertaken require high computational performance).

The data set again is cleaned appropriately, the GPU dataset contains the temperatures and utilization values throughout all the processes having more than 1.5 million rows. Firstly, the temperature averages are extracted for each GPU by the unique serial number same for the utilization percentage. Both values range from 0-100 the results are merged resulted to 1024 rows for each GPU used in the project with their temperatures and utilization percentages stored as averages. This data set provides a general perfomance sense for each GPU, also by summarizing this new data frame a particular trend was spotted that can be seen and is explained in the analysis process.

Objective 3 What is the interplay between increased power draw and render time?

From objective 1 it was concluded that 96.7% of the whole process it's the rendering run time with that in mind it is safe to assume that power in Watt values are increasing over time while the rendering occurs. This can be seen by producing numerical summaries of the GPU data set. In addition the average power values for each GPU were extracted and examined to see if there were any patterns in the power distribution. Lastly, percentages were calculated to count how many instances the GPU's were operating above the 1st quartile. This could be an approximation about the power interplay on the rendering process.

Objective 4 Can we quantify the variation in computation requirements for particular tiles?

The rendered image contains 65,793 tiles each tile is assigned a task ID and each task ID relates to 5 events. A fixed number of tasks are distributed to each GPU to render several tiles. By tracing the task ID and matching timestamps assigned to particular tiles it is possible to extract the computational requirements. In addition, its possible to extract exactly what occurred for the creation of the tiles in hand an example was produced to depict this.

Objective 5 Can we identify GPU cards (based on their serial numbers) whose performance differs to other cards? (i.e., perpetually slow cards).

Each hostname cataloged refers to a specific gpu serial number, hostnames can be used in the application checkpoints dataset to trace the tasks this particular GPU completed. For this question the data were reshaped in order to extract the total time it took each GPU to complete every task assigned to it. To achieve this goal, subsets holding maximum and minimum time details for each hostname were created. This subsets helped calculate the time each GPU started it's first process and the last. After that some cleaning occurred and the results were merged with the corresponding GPU serial number. Further discussion can be found in the analysis document that contains the results.

Objective 6 What can we learn about the efficiency of the task scheduling process?

For this question some observations regarding the efficiency of the task scheduling process have been made throughout the EDA and will be discussed in the analysis document. Although, some further investigation was made regarding how the tiling process is scheduled. Specifically, if the tiles rendered by a single GPU are consistent on a specific area of the image. The whole process was discovered and showcased throughout the EDA based on the data available although how the tiling process is programmed and scheduled required deeper analysis. Some assumption have been made and discussed in the analysis file. A fast GPU was picked for this investigation from the Objective 5 results.

Success discussion

In terms of success the results were satisfactory the output of the data construction and transformation provided either a clear answer or enough evidence for discussion on the objectives proposed. Evidence of the results will be provided in the additional documentation that includes graphs and numerical summaries. The strength of the work conducted stems mainly from the data handling. The construction of the data is making use of pipes that enable the exploration to be engaged at several stages branching towards different approaches. Furthermore, the work conducted focused fully on EDA with a variety of questions, now that the foundation is set other analysis can be explored making use of the results and context provided. Lastly is important to mention that the project provides answers on an intermediate level in most questions and doesn't explore every possibility emerging regarding the objectives.

Personal Reflection

The work conducted can be used as a foundation of future investigations, for example a future project could be fully based on the GPU analysis and if another set of hardware could provide better performance. In the work conducted some performance summaries were produced to support the above statement. The possibilities for future work are numerous cloud computing analytics are significant in the domain of data science making analysis such as this very significant towards developing robust approaches.

This project was very interesting from a personal standpoint and as an inspiring to be professional. The process provided a practical understanding of hardware and cloud task scheduling. This was the second project using Project Template in R and a more confident and agile approach was made compared to the first one. The practice with time stamps and date types in general that this project required could be proven valuable for future projects. Having some predefined objectives can be very helpful to plan and execute the work. At this point setting clear objectives before starting working on the data is a challenging process that gets better with practice. In future work based on Cloud Computing collecting a standard set of questions from research on the domain cloud be a great asset.