Technical Report of Terascope Project

The description of this project is about performance evaluation of the Terapixel image that has been rendered based on the observations made by the Newcastle Urban Observatory. At the initial analysis, the main objective or need of the project cannot be determined with just the project description as the analysis may vary based on the obtained datafile. In-order to decide the need for the project a methodology is followed which allows us to focus on the analysis with the given datafiles and then finalize the approach of the findings.

To proceed with the findings, the CRISP-DM methodology is used, which allows us to initially do the Business understanding followed by the Data understanding. As far as concerned the Business description of the project is to evaluate the performance of the Supercomputer. Based on the CRISP-DM methodology, the data is collected, and it is viewed to find whether the datasets can be related to each other. Once the data is obtained, we could find a mere correlation between all the three datasets particularly with the following three variables **taskId, hostname and timestamp.**

The data pre-processing phase comes into account for proceeding further with the analysis part. Based on the three variables which was mentioned previously, the datasets are merged and joined together which gives us a clear-cut idea of how to proceed with the analysis. Once the datasets are completely merged there where "NA" values in the subsequent variables and these values where then replaced by using a Last carry forward function. Further to support our analysis, the GPU performance variables where then grouped together based on both hostname and taskId from which the total power consumption and average of GPU temperature and utilizations where calculated. Initially the analysis is made based on the rendering time of each tasks where we were able to identify some of the key outliers of the uploading event type. To specifically continue the analysis of finding the observations which has some high runtime for this event the given information on dataset is not sufficient. Hence this analysis can be implicated for the future analysis purpose.

The summarization done on the GPU performance variables favoured the path for the analysis to take place based on the performance of each GPU's with the respect to the rendering tasks. With some furthermore graphical analysis, it was evident that some of the GPU nodes were taking higher power for completing the whole rendering process. At this point, it is decided that the Business solution for this project is to analyse the performance of each GPU nodes and to make suggestion if there needs to be any change in the GPU's in order to increase the Performance of the Supercomputer while rendering the Terapixel image. The final findings of this analysis determines that there are five GPU nodes which consume higher power for rendering the Terapixel image.

From the findings of the analysis made, it is clear that by replacing those high-power consuming GPU's with other good GPU's the performance would significantly increase. Further to this analysis, a graphical representation of time series based on the GPU performance can be implemented by creating the Animation plot which would be easy to represent in front of the stakeholders about the GPU node changes. As explained earlier, there are some outliers which still needs some more information to proceed with the analysis. Such analysis of rendering time of each event types can be made if furthermore information is observed and used.

The preprocessing step plays a major role in this analysis. The setting up of data in order to proceed with the analysis took a lot amount of time. Many tries of generating the complete dataset was performed before fixing up with the final dataset. The whole analysis looked easy and simple after getting the perfect dataset for proceeding it. These preprocessing, analysis and all those steps were

made easy by using the project template right from the start and it was made even more easy by following the CRISP-DM methodology. Instead of having n number of lines of code, each pre-processing step is stored separately in the munge folder of the project template and each analysis files are stored separately in the src folder of the project template. The project template plays a key role in simplifying the analysis process which is easy to understand if there occurs any change in code for future analysis purpose. As the datafiles are of huge size, the Cache file of the datasets are stored in the Cache folder which helps the project code to be reproducible and can be run through any environment.

The usage of Project template, CRISP-DM methodology pays off well as their usage has predominantly saved the time and made a clear information of how the analysis would be. In the future, with this approach and specific focus on data pre-processing similar projects could be undertaken in a better version as this adds on to the experience.