UNIVERSITY OF MELBOURNE

Cluster and Cloud Computing Assignment 1 - Project Report

Matthias Bachfischer - Student ID 1133751

April 9, 2020

1 Introduction

This document serves as a project report for assignment 1 of the Cluster and Cloud Computing course at the University of Melbourne. It describes the general system architecture, commands and configuration parameters for invoking the HPC processing scripts as well as steps that were taken to speed up processing and leverage the available High Performance Computing (HPC) resources in an efficient manner.

1.1 Dataset

The task for this assignment was to process a dataset called *bigTwitter.json* consisting of multilingual Microposts that were extracted from the Twitter social networking platform[1]. The dataset has a total size of 20.7 Gigabyte (GB) and is structured in the JavaScript Object Notation (JSON) document format.

2 Program setup

The software is written in the Python programming language and makes use of the MPI for Python programming library **RN310** to ensure parallel execution of computing steps, e.g. in a multi-core / multi-node HPC environment. The software was designed to run on the Spartan HPC system operated by the University of Melbourne (Lafayette, Sauter, Vu, & Meade, 2016),

[1] Twitter social networking platform https://twitter.com

2.1 Instructions for processing on Spartan

In order to run the software on Spartan, copy or clone the contents of the folder containing the software into your home directory. Change into the `slurm` directory - it consists of three files that support the execution of the program in three configurations as stated in the assignment description:

Resource configuration	**SLURM script**
1 node and 1 core	tweetanalyzer_1node_1core.slurm
1 node and 8 cores	tweetanalyzer_1node_8cores.slurm
2 nodes and 8 cores	tweetanalyzer_2nodes_8cores.slurm

2.2 Steps taken to parallelize the code

3 Results

4 Discussion and potential improvement areas

References

Lafayette, L., Sauter, G., Vu, L., & Meade, B. (2016). Spartan performance and flexibility: An hpc-cloud chimera. In *Openstack summit*. doi:doi.org/10.4225/49/58ead90dceaaa