Clustering Job Accounting Data

CS5352 Course Project, Spring 2021

[This project accepts two students.]

**Contact**

Jie Li ([jie.li@ttu.edu](mailto:jie.li@ttu.edu))

**Description**

To productively using the HPC resources, researchers and domain scientists are expected to have the system-specific knowledge to develop codes and job submission scripts. To investigate the HPC platform performance and if HPC users have use the resources productively, we conduct the usage behavior analysis. In this project, we apply unsupervised machine learning algorithms on two months of job accounting data to identify usage behavior. You are expected to use two kinds of clustering algorithms (partitioning clustering and hierarchical clustering.) and determine the appropriate number of behavior categories (compute-intensive, data-intensive etc.). The following deliverables will be used for the evaluation:

1. Understand each metric in the job accounting information and extract features from it.
2. Select and apply appropriate clustering algorithms on the job accounting data.
3. Discuss findings from the cluster analysis (bonus).

**Requirements**

* Experience of Linux
* Experience of Python programming
* Knowledge and experience of clustering algorithms

**References**:

1. Univa Grid Engine accounting file format <https://docs.hpc.qmul.ac.uk/using/man/man5/accounting.5.txt>
2. HPC Usage Behavior Analysis and Performance Estimation with Machine Learning Techniques <https://search.proquest.com/docview/1416806444?pq-origsite=gscholar&fromopenview=true>
3. Towards understanding HPC users and systems: A NERSC case study <https://doi.org/10.1016/j.jpdc.2017.09.002>