3rd October, 2018

Week-3 Report
D Swami

Project Title: On studying the performance of Hadoop MapReduce vs MPI for Aggregation Operations: A Big Data Challenge

The following project aimed at benchmarking various parameters of Map Reduce & MPI for parallel I/O. In the first week of the work, I have accomplished following tasks:

1) Build Hadoop using newly compiled snappy library.
2) Hands on basic MPI programs in SHARCNET.
3) Code documentation and synchronizing GITHUB for the rest of the project work.
4) Literature review of MPI for group by aggregate queries.
5) Map Reduce benchmarking with 5 GB data and completed the task under 12 minutes.
6) Map Reduce overload without YARN (No Cluster Manager) is less than 2 minutes and with YARN its 3-5 minutes.

Issues tackled in the current week:

1) Debugged MapReduce to get it working. Issues were in the YARN Configurations. Problems was Virtual memory to main memory ratio. Currently disabled the check for virtual memory to main memory.

Tasks for the upcoming week:

1) Debug Map Reduce to get rid of logical errors.
2) Hard test Map Reduce for the complete dataset.

Expected Issues in the coming week:

1) Disk space related issues if the amount of disk space is not enough for Map Reduce to process 215 GB data. We have a current capacity of more than 300 GB. Hence, we don’t expect the same but also should not be ignored.
2) Map Reduce Main memory allocation issues: The current setup in YARN allocated main memory in multiples of 256 MB for a maximum of 2 GB in total for all the mappers spawned concurrently.