13th November, 2018

Week-9 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 ninth week of the work, I have accomplished following tasks:

1) Implemented and tested the Concurrent IO version of the MPI for minimum fare calculation by month and year.
2) Perform Concurrent IO MPI on the large data. The large data took around 8.91 second on SHARCNET with 4 processors.
3) Basic analysis of results of MPI results. Through modifications we found that in serial reading of data for MPI it took more than 19 seconds just to read the data, while concurrent IO version completed reading of file and processing one row at a time in less than 10 seconds.

Issues tackled in the current week:

1) Issues in concurrent I/O MPI version especially because of not being able to decide the split points based on no of lines. Analysis on few rows of data we found that there are about 180+ characters in each row on average and hence, used the overlap size of 250 characters.

Tasks for the upcoming week:

1) Visualize the Hadoop & MPI results further.
2) Report writing.

Expected Issues in the coming week:

1) Story telling from the results would be a challenging task.