6th November, 2018

Week-8 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 eigth week of the work, I have accomplished following tasks:

1) Upload data on Cedar cluster for MPI performance analysis. The size of data is 933 MB.
2) Perform MPI on the large data. The large data took around 70 second on local computer while it took 25.1 second on Sharcnet with 4 processors.
3) Basic analysis of results for HADOOP results
4) Implementation of MPI using concurrent I/O. This allows me to process the data using constant memory and O(n) time (i.e. one pass algorithms)

Issues tackled in the current week:

1) Changed the usage of MPI_Scatter to MPI_Scatterv to accommodate the need of varying size of data points when no of data points is not a multiple of the no of processors.
2) Updates to the MPI program to keep all the functions in C++ as it has better exception handling. This was useful for converting string to float or integer because the functions throw invalid_exceptions which can be caught and handles and stops breaking the scripts.
3) To review all the previous logs of execution of Map Reduce tasks shifted to Job-History sever which still uses YARN for tracking metrics.
4) Issues in concurrent I/O MPI version especially because of not being able to decide the split points based on no of lines.

Tasks for the upcoming week:

1) Improvise a plan for concurrent I/O version of MPI. Only way forward seems to be to allow duplicate rows. And this seems reasonable since we are doing min operation which has no effect of duplicate values. Had it been another operation like sum or average it would have been a completely different scenario.
2) Visualize the Hadoop results further.

Expected Issues in the coming week:

1) Debugging and implementation challenges.