25th September, 2018

Week-2 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) Identify the best way to ingest the data. Use webhdfs or use hdfs command line. Most other tools connect through Restful web HDFS. Example of ingestion tools are Apache NIFI, Apache Kafka, Apache Gobblin, and others from [1].
2) Debug the Map Reduce code for a small number of input files. Map-Reduce failed for Snappy compression. Fault in library, need to build the compression library.
3) Programmed an MPI so that each slave reads one file as a proof of concept.

Issues tackled in the current week:

1) Building Snappy library in Visual studio. There were problem related to a library. So had to add a library in snappy header file. The error code was because _BitForward function is visual studio is in a header file not included in snappy header library.
2) Debugged MapReduce to get it working. Issues were in the Snappy library.

Tasks for the upcoming week:

1) Build Hadoop using newly compiled snappy library.
2) Prepare and give the sharcnet quiz.
3) Hands on basic MPI programs in SHARCNET.
4) Code documentation and synchronizing GITHUB for the rest of the project work.
5) Begin working on the end-of-september report.
6) Literature review of MPI for group by aggregate queries. (Monday & Tuesday).

Expected Issues in the coming week:

1) Sharcnet based debugging challenges.
2) Apache Hadoop path error for the snappy library.

References:

1) https://www.predictiveanalyticstoday.com/data-ingestion-tools/