

Group B
Assignment No. 3

AIM: Locate dataset (e.g., sample_weather.txt) for working on weather data which reads the text input files and finds average for temperature, dew point and wind speed.

THEORY:

Introduction to Map- Reduce:

MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner. MapReduce is a processing technique and a program model for distributed computing based on java.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes

The Algorithm

MapReduce program executes in three stages, namely:

1. Map stage
2. Shuffle stage
3. Reduce stage.

Inserting Data into HDFS:

- The MapReduce framework operates on pairs, that is, the framework views the input to the job as a set of pairs and produces a set of pairs as the output of the job, conceivably of different types.

- The key and the value classes should be in serialized manner by the framework and hence, need to implement the Writable interface. Additionally, the key classes have to implement the Writable-Comparable interface to facilitate sorting by the framework.

INPUT:

OUTPUT:

CONCLUSION: Thus, we have learnt how to process input data using map reduce framework.