ROLL NO : 33281

Camlin | Page No. | Date / /

ASSIGNMENT 3

Title : Design and develop a distributed application using MapReduce

Problem Statement : Design and develop a distributed application to find the hottest / coolest year from the available weather data. Use weather data from the Internet and process it using MapReduce.

Objectives :

- To understand use of various aggregate operation in MapReduce.
- To understand use of various Key & value passing techniques in MapReduce.
- To understand preprocessing of input data file.

Theory :

Dataset Description —

The data we had used is from National Climatic Data Center or NCDC. The data is stored in line-oriented ASCII format in which each line is a record.

The format supports a rich set of meteorological elements, many of which are optional or with variable data length for

simplicity, we focus on the basic elements, such as temperature which are always present and are of fixed width.

MapReduce program execution explanation

MapReduce works by breaking the processing into two phases, the map phase and the reduce phase. Each phase has key-value as input and output. The input to our map phase is raw NCDC data. We choose text input format that gives us each line in the dataset as a text-value. The key is the offset of the beginning of the line from the beginning of the file. We pull out year and temperature in our map function. The output from the map function is processed by MapReduce framework before being sent to the reduce function. This processing sorts and groups the key-value pair by key so that each year will come with an list of its temperature. All reduce function is to now do is iterate through the list and add it to sum and then

divide by count of list to get average temperature of the year and then we will compare average with min and max if and according set values of min and max and atleast value of min and max year will be given as output.

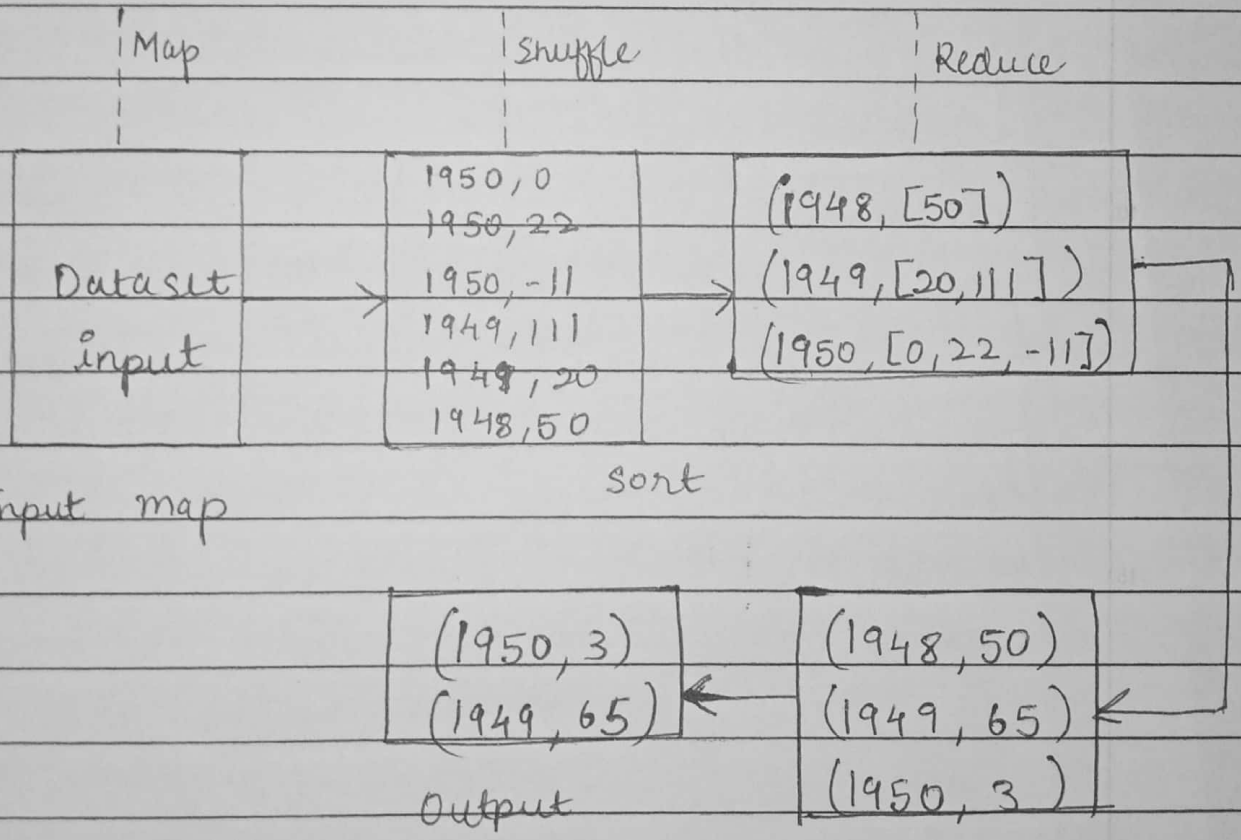

Fig: MapReduce logical data flow

Camlin | Page No.
Date / /

Conclusion :

In this assignment, I have understood the use of various aggregate functions using mapReduce and successfully designed and developed a distributed application to find hottest / coolest year from available weather data using hadoop.