**Assignment 16.1**

**Problem Statement:**

Using Distributed cache join the two datasets and find out the number of temperature variations in each country using mapreduce.

**Input:**

Here, we have the temperatures collected every minute, from 20 top buildings all over the world. For this data analysis, you can download the necessary dataset from the below link:

<https://github.com/prateekATacadgild/DatasetsForCognizant/blob/master/Sensor%20data-20161205T052506Z.zip>

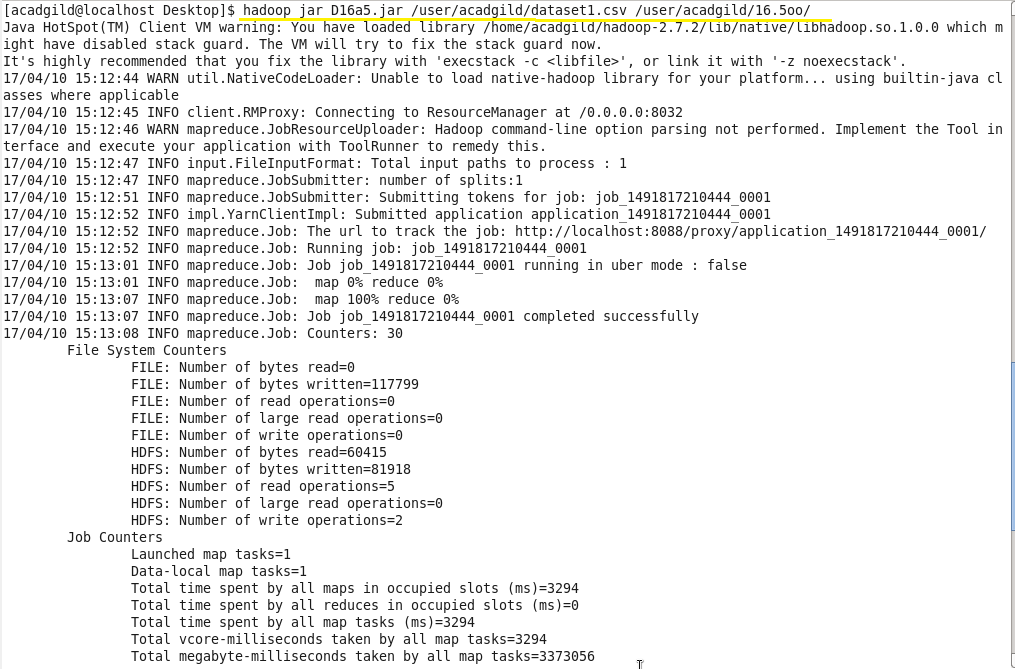
In the above link there are two datasets: building.csv contains the details of the top 20 buildings all over the world and HVAC.csv contains the target temperature and the actual temperature along with the building Id. HVAC (heating, ventilating/ventilation, and air conditioning) is the technology of indoor and vehicular environmental comfort. Its goal is to provide thermal comfort and acceptable indoor air quality. Through the HVAC sensors, we will get the temperature of the buildings.

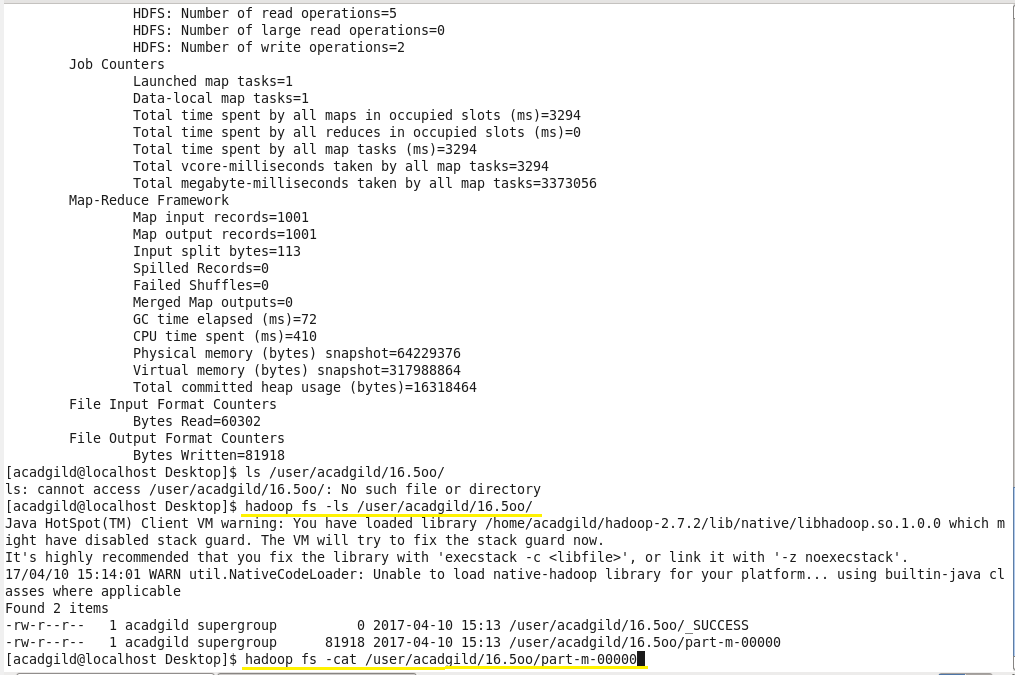
Here are the columns that are present in the datasets:

Building.csv – BuildingID, BuildingMgr, BuildingAge, HVACproduct,Country

HVAC.csv – Date, Time, TargetTemp, ActualTemp, System, SystemAge, BuildingID

**Program:**

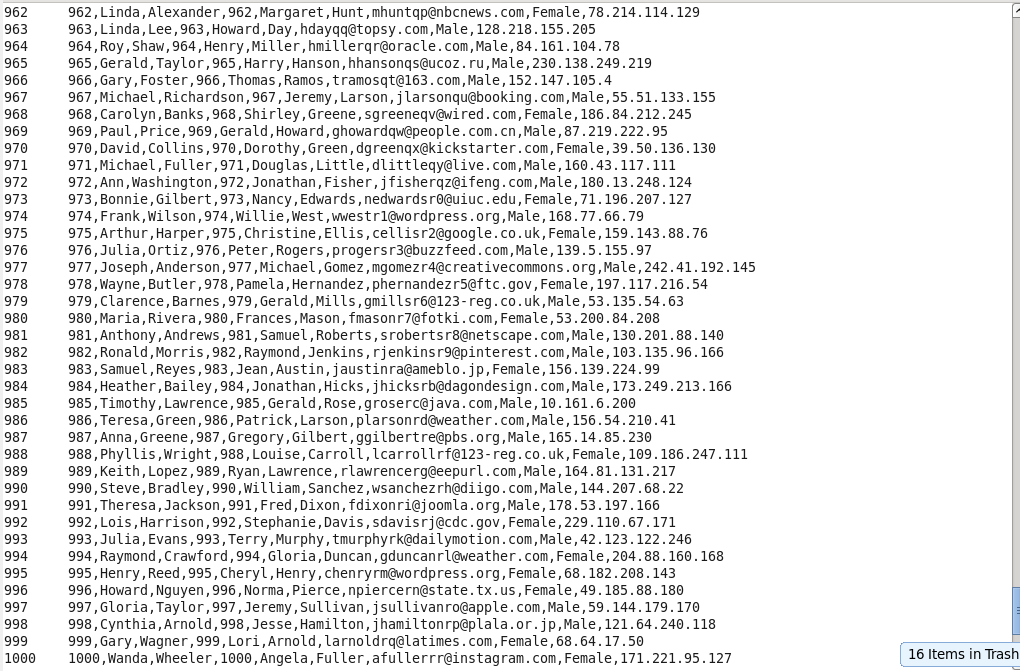
****

****

**Output:**

The records from dataset1.csv and dataset2.csv are joined using common field in both the datasets which is id.

As the output consists of large number of records screenshot of last few records is attached below:

****