**Code for data preparation:**

*#pig*

*#records = LOAD 'input/ncdc/19011910.txt' AS (usaf:chararray, wban:int, year: int, temp: int);*

*#station = LOAD 'input/ncdc/stationlistshort.txt' AS (usaf:chararray, wban:int, name:chararray);*

**Analysis 1:**

The following code join the observed temperature data with the station name data so that the location name will be included within each observation in the relation:

*# recordnstation = JOIN records BY $0, station BY $0;*

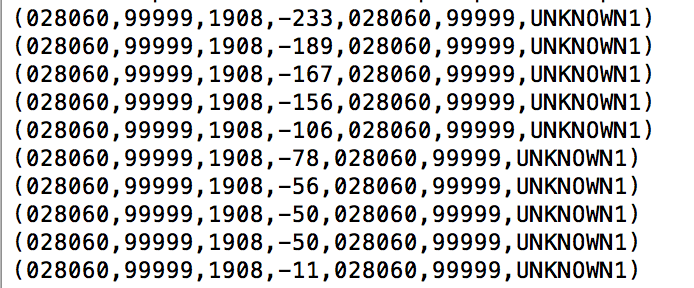
*#lim\_result = LIMIT recordnstation 10;*

*#DESCRIBE lim\_result;*

*#DUMP lim\_result;*

The result of first 10 entries is shown below:

Column Names: 1. usaf, 2. wban, 3. year, 4. temperature, 5. usaf, 6. wban 7. location name



From the above result above, we can see that the location name has been added to each observation accordingly.

**Analysis 2:**

*Code:*

*#C = GROUP recordnstation BY name;*

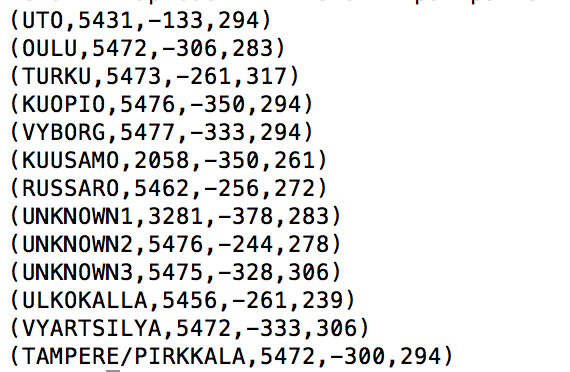
*#MaxMin\_TEMP = FOREACH C GENERATE group,COUNT($1),MIN(recordnstation.temp), #MAX(recordnstation.temp) ;*

*#DESCRIBE MaxMin\_TEMP;*

*#DUMP MaxMin\_TEMP;*

The number of trusted temperature observations, the minimum and maximum temperatures by station are shown as below:

Column Names : 1. Station, 2. Number of temperature observations, 3. Min Temp, 4. Max Temp



**Analysis 3:**

*Code*

*#ord = ORDER MaxMin\_TEMP by $3 DESC;*

*#max\_record = LIMIT ord 1;*

*#DUMP max\_record;*

*Output:*

*Column names: 1. Location name, 2. Count of observations, 3. Min Temp, 4. Max Temp*

**

*Hence, the location with highest max temp is TURKU.*

*#filtered = FILTER recordnstation BY name == max\_record.$0;*

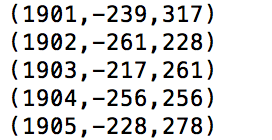
*#grp\_records = GROUP filtered BY year;*

*#maxmintemp = FOREACH grp\_records GENERATE group,MIN(filtered.temp), #MAX(filtered.temp) ;*

*#DESCRIBE maxmintemp;*

*#DUMP maxmintemp;*

Column Names: 1. Year, 2. Min Temp, 3. Max Temp (for the station with highest maximum temperature)



**Analysis 4:**

*code*

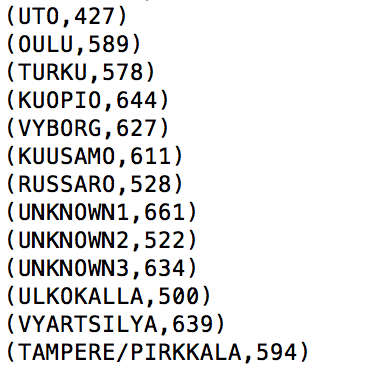
*#range = FOREACH MaxMin\_TEMP GENERATE group, $3-$2;*

*#DESCRIBE range;*

*#DUMP range;*

The temperature range for each location is shown below:

Col names: 1. Station, 2. temperature range



The following code find the station name and temperature range for the station with smallest temperature range for the time period:

*#ord\_range = ORDER range by $1;*

*#min\_range = LIMIT ord\_range 1;*

*#DESCRIBE min\_range;*

*#DUMP min\_range;*

Column names:

1. Station with minimum temperature range, 2. range of temperature



Hence, the station with minimum temperature range is UTO with a range of 427.

To obtain that station’s temperature ranges by year:

Code:

*#filtered\_mr = FILTER recordnstation BY name == min\_range.$0;*

*#grp\_mr = GROUP filtered\_mr BY year;*

*#range\_mr = FOREACH grp\_mr GENERATE group, MAX(filtered\_mr.temp) - #MIN(filtered\_mr.temp);*

*#DESCRIBE range\_mr;*

*#DUMP range\_mr;*

Hence, the station’s temperature ranges by year is as below:

Column Names: 1. year, 2. Range of Temperature

