Assignment 2 – Map Reduce

KIT318 KIT418 Big Data and Cloud Computing

Cormac Collins 081136

Part 1 – unique word analysis

Task 1 aimed for all the unique words in all the text files to be entered as the 'InputFiles' to be mapped to the file containing the highest number of that unique word.

As shown in figure 1, this was achieved with a simple 'Hash-mapping' of the count for each word in a file and passing to the reducer to then loop through the list of filename and word counts to find the largest count.

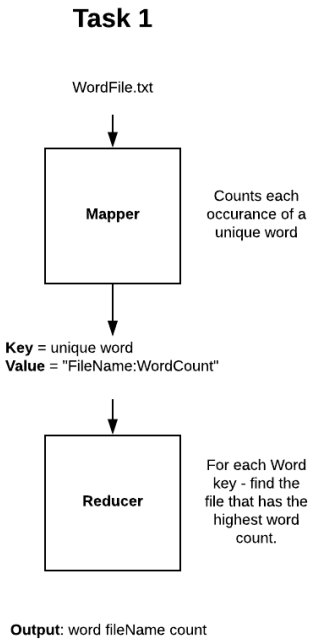

Figure 1: Mapping unique word count to file name.

Part 2 – Weather data average temperature (Min or Max)

Given provided weather data files, the program was expected to map this form of information (id, date, temperature type, temperature):

ITE00100554,17630122,TMIN,2,,,E,
ITE00100554,17630123,TMAX,27,,,E,

To a form similar to this:

ITE00100554 1763 10.065751
ITE00100554 1764 10.658197
*(Output form: ID Date AverageTemperature)

This works through mapping the id and temperature data to a unique year, before being reduced to an average temperature as shown in figure 2.

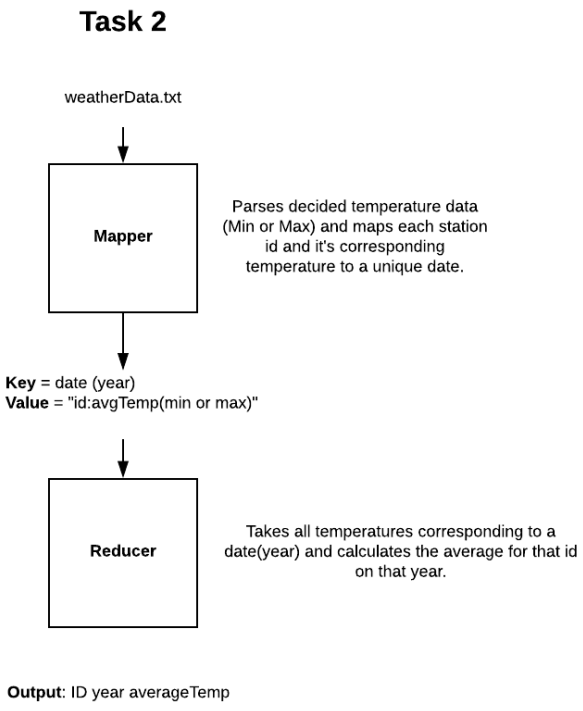

Figure 2: Calculating average temperature for each year for each station.

Task 2B – Station similarity scores

Taking the formula:

$$\text{Similarity}(p,q)=\frac{\sum_i |minT_i^p - minT_i^q|}{nY}$$

We can calculate the similarity for stations. This utilizes the output of the previous program in the form of (Comma delimiters added for easier parsing):

ITE00100554, 1763, 10.065751
ITE00100554, 1764, 10.658197

The Map-Reduce then calculates the similarity score for any id's that have matching dates:
ITE00100554 EZE00100082 4.2245913
(Output form: id1 id2 similarityScore).

This is all outlined in figure 3. We can see that this program utilizes a further combiner, simply because extra steps were needed, this included; Mapping the data to a unique date, Mapping the temperature scores to unique pairs of id's and finally calculating those scores for all unique pairs with scores. This was somewhat inspired from the inverted index model often used in Map-Reduce and having the understanding that 2 processes of creating a unique key were required. This was also felt to be analogous to utilizing joint primary-secondary key combinations in SQL for certain schema circumstances.

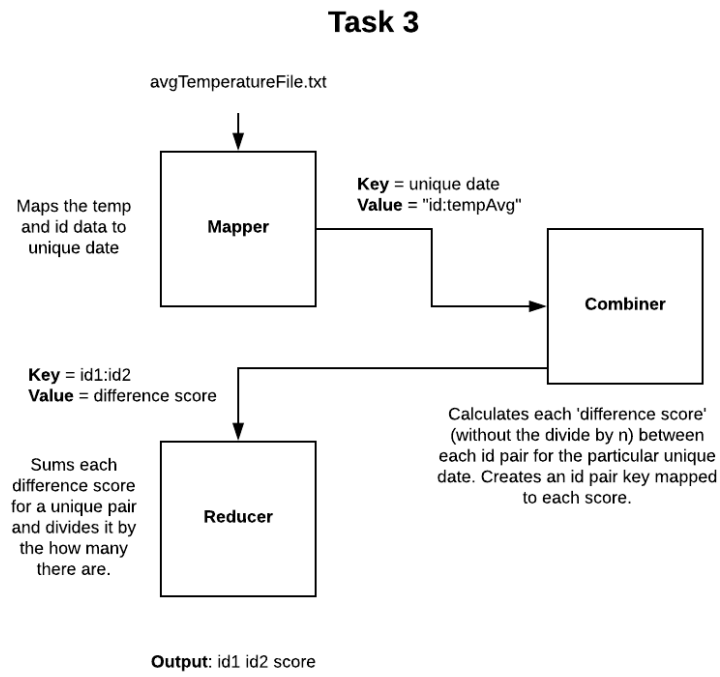

Figure 3: Difference score between weather stations for Min Temperature average.