Friday, April 13, 2018

CS455 A3 WC

---

1. How would you implement the complete set of tasks in HW3-PC using a single MapReduce job? Note that your program should emit outputs for each task into a separate file.

For this assignment, all of the seven questions I complete in a single MapReduce job.

Using question 1& 2 as an example, the goals of these two questions are to find the best time of the day(hour), day of week(Sunday to Saturday) and time of year(Month) to fly to minimize delays and to find the worst time of the day(hour), day of week(Sunday to Saturday) and time of year(Month) to fly to minimize delays.

According to the questions, I can know that the month, day of week, arrival time and arrive delay are the datas I need in the dataset. So I decided to make month, week, hour as my key and the arrive delay as my value.

So in the mapper part, it read the data set line by line, I split the line and store the valuable I need. To reduce the data, I set a if condition here. Only if the delay time is larger or equals to zero, the data will be passed. If we observe the data, we would find that there are a lot of negative delay time which means the plane arrived before the scheduled time. So this data is meaningless. So I blocked it so that the program will run faster.

Then I set a combiner. This is used to reduce a part of data before it passed to reducer. The map output key is “month, week, hour” and the map output value is “arrival delay in minute”. So it might happen one key with different values. So I just add them together and it will reduce a large amount of data. The program will be faster.

In the reducer, I initialized three maps. There are used to store <month, delay>, <day, delay>, <hour, delay>. Then, it is easy to find the max/min delay time and I can get the answer I want.

2. How would you extend your current solution to identify locations that are most similar to each other and also those that are most dissimilar to each other?

So for this question, I am going to identify locations that are the most similar to each other and the most dissimilar to each other.

According to the question, we need to define what is similar first. We can have a lot of similar things. So for this question, I am trying to say that two locations are the most similar because their flights are usually delayed because of the weather. Right now, we can simplify the questions like which locations experience the most weather-related delays and which location experience the less weather-related delays. The first question is find out the most similar and the other one is find the less similar.

Then, we can identify the data we need. Here location, I can say it is city and I also can say it is airport. Let's say it is city. In the main data set, there is no "city" so we need to use "itat" as the key to find out its city. Now I am trying to use two mapper which deal with two paths of data. In my mapper one, it read the main data set. I split the line and set the origin itat and dest itat as my output key. Then, I will set a if condition to check the weather delays are not "NA"(missing data) and they are larger than zero. This condition make sure the the flights delays because of weather and it reduce the data. So my output value is new Text("1") due to I only care about if the flight delayed. In my mapper two, it read the supplementary data, airports.csu. I split the line and set itat as my output key and city as my value.

In the reducer, it will deal with two kind of data. So I have to separate the values first. So, if the value is "1", it is the count, I will put the value in the map<itat, delayNum>. If not, it is city, I will put the value in another map<itat, city>. After that, I have two map now. For the first map, I can get the maximum value (delayNum), the second maximum value and the minimum value. I can get the itat according to the value and get the city according to the second map.

At last. I can get the answer like this. <city, delayNum>. The cities with two maximum delayNum is the most similar city, and the city with the maximum delayNum and the city with the minimum delayNum is the most dissimilar cities.

3. Assume that you are starting a new budget airline. What would be the best Cities to start services to? If new regulations are in place where delays are penalized at the rate of 20% of the ticket price, per hour of delay, where would you base your hubs if your goal is also to ensure greater coverage to cities? Please make sure that you consider Q3 without the datasets specified in Q4.

For this question, there are two things that we need two consider. The first concern is that delays are penalized at the rate of 20% of the ticket price, per hour of delay. The other consider is that my goal is to ensure greater coverage to cities.

According to these two considerations, we need to figure out the which city has less delay, also we need to find out the most busiest city.

To figure out which city has less delay, we need to calculate the average delay which is using the total number of delayed flights and also the total number of minutes that were lost to delays. Because delays are penalized at the rate of 20% of the ticket price, per hour of delay, so we need to find out the less value.

So in the mapper, it read the main data set. I split the line and set the origin itat and dest itat as my output key. Then, I will set a if condition to check the delays are not “NA”(missing data) and they are smaller than zero, if they are smaller than zero, they will be set to zero. Because negative value means they didn’t late. The out put value is the delay time.

In the reduce, we can calculate the the total number of delayed flights and also the total number of minutes that were lost to delays. The average delay time is (total number of minutes that were lost to delays) / (total number of delayed flights). And we can know the city by itat.

To find out the the most busiest city, we can calculator it by total flights numbers.

At last, we just need to find the most busiest city with the less average delay time. And this will be the city to start services to.

4. Consider the case where two additional (yearly) datasets have been made available to you. The first dataset includes information about migration patterns (alongside demographics such as age, gender, and educational levels) into and out of a state. The second dataset includes economic data such as number of new jobs that were created, the sectors that these jobs were created in, the average pay, educational levels requirements for jobs, etc. How would these datasets now influence your decision making for the questions you explored in Q3?

For this question, we have two extra datasets. These two datasets will influence my decision making for the questions you explored in Q3.

So let us start with the first datasets. The first data set includes information about migration patterns into and out of a state. It is the common knowledge that the higher migration population, the more favorable to us. I don't know if there is some data represent the migration population. But we can analyze the age, gender, and educational levels to speculate the migration population. People whose age between 20 to 40 are more likely migration. People with high educational levels are more likely migration. So we can consider the problem by age and educational levels.

For the second datasets, it includes economic data such as number of new jobs that were created, the sectors that these jobs were created in, the average pay, educational levels requirements for jobs, etc. The more number of new jobs that were created means the more job opportunities. So this may mean that there will be a large number of outsiders entering the city. Higher average pay mean that they have more economic power to travel. Higher average pay also mean the jobs will more attractive.

So to sum up these two data set we can know that, age from 20 to 40 are more likely to move. They are more likely find the job opportunities. The higher educational level are more likely get a high pay job. The more new jobs that were created, the more choices they have. They might pick those who have the higher educational level. For example, New York's migration population is always higher than Fort Collins.

After analyze these two data set, I will choose a city with more new jobs created, the higher pay, the higher educational level.

5. Your new airline decides to unveil a feature where likely delays are texted to clients 25 hours before the flight (i.e. before they check in online). You have access to weather forecasts from the National Weather Service. How would you design a scheme that allows you to generate accurate alerts if your cluster can only process ~30% of the weather forecast data in time for the alerts?

For this question, my new airline decides to unveil a feature where likely delays are texted to clients 25 hours before the flight. But my cluster can only process about 30% of the weather forecast data in time for the alerts?

To design a scheme that allows me to generate accurate alerts, I have to make the 30 percent procession more efficient. To do that, I need to reduce the data.

I don't know what the weather forecasts data set looks like. So I just use weather app on iso as the example. On the app, we can see the 8 days of the weather in the future. But do we really need these data? The answer is no. At first, we need to text to clients before 25 hours. That means, the flight will not depart in the 25 hours, we don't care the weather in these 25 hours. Also, we just need to text to clients the delays information. So we don't need to know the weather after the plane depart. Suppose the plane department needs one hour, we only use 1/(8*24) of the total data.

Furthermore, we just need to text to clients if their flight will be delayed. That means, we don't need to do anything if the weather is good in this hour. So if the weather is good in this hour, all of the flights that depart in this hour will be fine. In other words, if the weather is bad, the flights might be delayed, that means all of the flights that depart in this hour might be delayed. We have chosen the city which has lowest delays. So it is more likely that have a good weather.

To sum up, if we just use the data we need, it is totally fine if we can only process about 30% of the weather forecast data.