Lab 03 Map Reduce Programming

[IT494, Big Data Processing, Autumn'23]

Instructor: PM Jat (pm_jat@daiict.ac.in)

In this lab too we continue Map-Reduce Programming.

Exercise #1

Suppose you are given two files employee `"empc.csv"` and department `"depc.csv"` in `dataset` folder
https://drive.google.com/drive/folders/1Q0sy0NlD2nkjmzxuYURQoFt5XRZpcScs

These are from the company database of Elamasri/Navathe textbook and bit modified. Attributes of these files are as:

- `empc.csv: eno, name, dob, gender, salary, sup_eno, dno`
- `depc.csv: dno, name, mgr_eno, join_date`

Perform JOIN operation on these two files using the **map-reduce** approach. Let the joining condition be `"mgr_eno=eno"`

Let you use thefollowing alogirthm "Broadcast Join" and as discussed in the lectures.

```
//Original Algoritm A.4 in article [1]
void map_init() {
    HR <- build a hash table from referenced file R on Join Key

void map(K, V) {
    //V: value, a record from a split of referencing file L
    for each v from V {
        r <- HR.get( v.join_key )
        if r found
            emit(null, new record(r, v ))
}
```

Exercise #2

Let you yourself figure out a way a map-reduce based solution to compute moving average of time series data.

There is book titled **"Data Algorithms"** [6]. A copy of the book is placed in shared dataset folder itself. Chapter 6 of this book discusses the **computation of Moving average using map-reduce**. Refer related section for this purpose. Choose "Example 2: Time Series Data (URL Visits)" as data space.

Let you use dataset https://www.kaggle.com/datasets/bobnau/daily-website-visitors and compute monthly moving average of website visitors (first time and repeat)

Submission Required:

A document in pdf format (all in one pdf file) that contains following:

1. All source code is pasted here. Try having the code that is formatted and colored.
2. Link to “colab notebook” that contains source code and outputs of programs for all problems.

Using of markdown or latex for creating PDF would be better.

References

[1] Blanas, Spyros, et al. "A comparison of join algorithms for log processing in mapreduce." *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 2010.

[2] Parsian, Mahmoud. *Data algorithms: Recipes for scaling up with hadoop and spark*. " O'Reilly Media, Inc.", 2015.