NAME: Sai Krishna Mannava

STUDENT ID: 801136361

**Cloud Computing for Data Analysis**

**Assignment – 1**

1. The DISTINCT(X) operator is used to return only distinct (unique) values for datatype (or column) X in the entire dataset.

As an example, for the following table A:

|  |  |  |
| --- | --- | --- |
| **A.ID** | **A.ZIPCODE** | **A.AGE** |
| 1 | 12345 | 30 |
| 2 | 12345 | 40 |
| 3 | 78910 | 10 |
| 4 | 78910 | 10 |
| 5 | 78910 | 20 |

DISTINCT(A.ID) = (1, 2, 3, 4, 5)

DISTINCT(A.ZIPCODE) = (12345, 78910)

DISTINCT(A.AGE) = (30, 40, 10, 20)

Implement the DISTINCT(X) operator using Map-Reduce. Provide the algo-

rithm pseudocode. You should use only one Map-Reduce stage, i.e. the algorithm should

make only one pass over the data.

**Ans:**

Here map-reduce is used to group together and remove duplicates. Use a mapper to emit X from each word as Key. Now reducer simply emits keys.

Pseudocode:

Map (key, record) //key is id, record is values

output (value, null) // implement for each and every row

reduce (key, record)

emit(key) // output of the key

1. The SHUFFLE operator takes a dataset as input and randomly re-orders it.

*Hint: Assume that we have a function rand(m) that is capable of outputting a random integer between [1, m].*

Implement the SHUFFLE operator using Map-Reduce. Provide the algorithm pseudocode.

**Ans:**

Mapper outputs the record as the value along with a random key. It is like, each record is sent to a random reducer and that reducer emits the values.

Pseudocode:

map(key,record):

rand(m)=picks a random integer in[1,m)

output(rand(m), record)

reduce(key, records):

for record in records:

output record

1. What is the communication cost (in terms of total data flow on the network between mappers and reducers) for following query using Map-Reduce:

**Get DISTINCT (A.ID from A WHERE A.AGE > 30)**

The dataset A has 1000M rows, and 400M of these rows have A.AGE <= 30. DISTINCT(A.ID) has 1M elements. A tuple emitted from any mapper is 1 KB in size.

**Ans:**

Assume, x=1000M, y=400M, z=1M elements

Here are two jobs and the output of WHERE is chained to DISTINCT:

WHERE emits(x-y) tuples from the mapper

DISTINCT emits(x-y) tuples from the mapper

Total=2(x-y) =1200M\*1KB=1.12TB

Total= (x-y) =600M\*1KB, if the values are filtered in the mapper

1. Consider the checkout counter at a large supermarket chain. For each item sold, it generates a record of the form [ProductId, Supplier, Price]. Here, ProductId is the unique identifier of a product, Supplier is the supplier name of the product and Price is the sales price for the item. Assume that the supermarket chain has accumulated many terabytes of data over a period of several months.

The CEO wants a list of suppliers, listing for each supplier the average sales price of items provided by the supplier. How would you organize the computation using the Map-Reduce computation model?

**Ans:**

Pseudo code:

map(key,reocrd):

output [record (SUPPLIER), record (PRICE)]

reduce (SUPPLIER, list of PRICE):

emit [supplier, AVG(PRICE)]

**\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\***

*For the following questions give short explanations of your Anss.*

**5) True or False:** Each mapper/reducer must generate the same number of output key/value pairs as it receives on the input.

Ans: **FALSE**, mappers/reducers may generate any number of output key/value pairs which may not be equal to number of inputs.

**6)** **True or False:** The output type of key/value of mappers/reducers must be of the same type as their input.

Ans: **FALSE**, mappers/reducers may generate key/value of any random type.

**7) True or False:** The input to reducers is grouped by key.

Ans: **TRUE**, the input key/value pair of the reducers are grouped by key.

**8) True or False:** It is possible to start reducers while some mappers are still running.

Ans: **FALSE**, As the inputs to the reducers is generated by the mappers.

Therefore, it is not possible to start reducers while some mappers are still running.