**Name: Nikhil Jadhav**

**Student ID: 801075504**

**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.

Map (key, record) //Key – A.ID, record – value of table

Output (value, null) // for each input row

Reduce (key, record)

Emit(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.

map (int key, int value)

//int key-intermediate key is generated after mapping

Output(rand(m), int value) //rand(m) to pick a random int between [1, m]

Reduce (key, value)

//Key: after random key generation, value: after shuffling and reducing

For value in value list

Emit(value)

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.

Let, x = 1000M, y = 400M, a = 1M

There will be 2 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)

= 2(600M)

= 1200M \* 1KB

= 1.12 TB (approx.)

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?

map (key, record)

Output [record (Supplier), record (Price)]

Reduce (Supplier, list of Price)

Emit (Supplier, Avg (Price))

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

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

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

**Answer:**

False, it is not necessary that the mapper and reducer must generate same number of output key/value pairs, as the reducer might reduce the generated key value pair.

1. **True or False:** The output type of keys/values of mappers/reducers must be of the same type as their input.

**Answer:**

False, mapper/reducer may produce output of any type.

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

**Answer:**

True, input to reducers are grouped by (key, value) pair.

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

**Answer:**

False, it is not possible to start reducers while some mappers are still running as the reducers need complete set of input from the mappers.