**Cloud Computing for Data Analysis**

**Assignment – 1**

**STUDENT NAME: BALASUNDARAM AVUDAI NAYAGAM**

**STUDENT ID: 801104903**

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.

A)

**ALGORITHM:**

Map(Key, Value):

output(Value, null) from X in each record

Reduce (Key, Values):

output Key

**NOTE: X denotes the column**

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.

A)

**ALGORITHM:**

Map(Key, Value):

rand(m)= a random integer between [1,m]

output(rand(m),Value)

Reduce(Key, Values):

for each Value in Values:

output 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.

1. There are totally two jobs and since there is a ‘WHERE’ and ‘DISTINCT’, the result of WHERE is combined with the output of DISTINCT.

**Given:**

Total records (a) = 1000M

Records with Age<=30 (b) = 400M

Record with distinct ID (c) = 1M

Sum = 2\*(a-b) = 2\*(1000-400) = 2\*(600M) = 1200M \* 1KB = 1.12 TB

When the Values are filtered, Sum = (a-b) = 600M \* 1KB = 0.56 TB

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?

1. We can use the algorithm below to get the list of suppliers

Map(Key, Value):

output [Value(Supplier), Value(Price)]

Reduce(Supplier, Price List):

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.
2. **False**

**Because, it is possible that each mapper/reducer can generate different number of output key/value pairs as it receives on the input.**

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

**In general, the mappers can contain key-values of any type. It is not necessary that they must contain same type as input.**

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

**All the reducers are given with key-vale pairs grouped by key.**

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

**No it is not possible since the mappers that are running may produce key-value pairs which are already processed by the reducer.**