National University of Computer and Emerging Sciences

Lahore Campus

Quiz 8

Total Marks: **12**
Time allowed: **5 Minutes**
Date: **April 24, 2025**
Section: **BCS-6A**

Q1: [2 Marks] In a typical MapReduce job what will be the value of the parameter ***R*** and how framework will calculate it?

Solution: R represents the number of buckets the output of the map will be hashed into.
The programmer that provided map and reduce functions will also provide the value of R.
A suitable value of R depends on that how much data we expect map to emit. If we expect a lot of data (as in sorting), then we should use a large value of R (usually some multiple of machines assigned to the job). If we expect map output to be short (as in grep when we somehow know a priori that not much results will be there in the input data), then we should use a small value of R.

Q2: [10 Marks] Let's assume you have crawled data of WWW saved in many large files in GFS. You want to provide a service that: given a CNIC (National ID Card) number, you want to search on which websites a specific given CNIC appears.

Write a MapReduce function for this purpose by clearly telling what the keys and values for Map and Reduce functions.

Solution:

Map(key = <WWW site URL, "your name"> , value = contents of the site)
{

- Read the site data from value above and search your name
- If name found: emit: key = <"your name"> , value = <URL of the website>

}

Reduce(key = "your name", values = List of URLs)
{

- Iterate over value and emit each <URL> as key = <URL>, value = <NULL>

}

By the way we should use R = 1 in this use case because there will be
Only one bucket that will have data. That is probably okay because
We don't expect occurrences of a specific CNIC much on the WWW.
(If that was not the case we would have done things bit differently so that
We could use larger value of R for better parallelism. Can you think how?)