COMP4434 Big Data Analytics

Assignment 3 Solutions

PolyU, Hong Kong

Problem 1 (3 points) Assume that a large table is distributed across multiple files, each containing partial rows of the table. Each row is composed of the following data:
(student_name, department, salary).
For example, (Bob, Computing, 30,000) means Bob graduated from the department of Computing and the salary of his first job is 30,000.
The objective is to determine, in each department, the total number of graduated students whose salary is more than 25,000 in their first job.

(a) What are the relationships between MapReduce and Apache Hadoop?

(b) Provide a concise pseudo-code for the Map workers, specifying the input and output (key, value) pairs.

(c) Provide a concise pseudo-code for the Reduce workers, specifying the input and output (key, value) pairs.

Answer:

(a) MapReduce is a programming framework of Hadoop that is used for parallel processing of data. Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model.

(b)
```
Map(key, value) {
// key: file name
// value: file content
for each row in value {
     if (row.salary > 25000)
                emit(row.department, 1); // intermediate key value pairs
}
}
```

(c)
```
Reduce(key, values) {
// key: department
// values: a list of 1
emit(key, size(values));
}
```

Problem 2 (2 points) We have four text files as follows, storing the student grades of four subjects.

math.txt	physics.txt	chemistry.txt	art.txt
James, 81	James, 57	James, 78	James, 67
John, 83	John, 78	John, 92	John, 89
Robert, 75	Robert, 68	Robert, 68	Robert, 88
Michael, 71	Michael, 71	Michael, 91	Michael, 87
David, 79	David, 79	David, 77	David, 87
Mary, 73	Mary, 69	Mary, 74	Mary, 79
Linda, 83	Linda, 79	Linda, 89	Linda, 94
Susan, 67	Susan, 76	Susan, 87	Susan, 78
Lisa, 76	Lisa, 74	Lisa, 92	Lisa, 91

Our goal is to calculate the total scores of students in all four subjects.

(a) Write a concise pseudo-code for the Map workers, specifying the input and output (key, value) pairs.

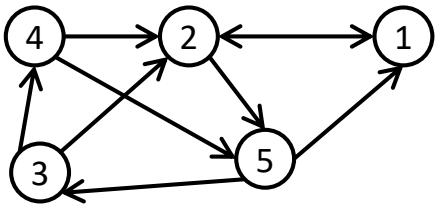
(b) Write a concise pseudo-code for the Reduce workers, specifying the input and output (key, value) pairs.

Answer:

(a)
```
Map(key, value) {
// key: file name
// value: all (name, grade) pairs in the file
Open File F.txt;
while has lines left{
     L = Read a line
     for each tuple in L{
          emit(name, grade)
     }
}
}
```

(b)
```
Reduce(key, values) {
// key: name
// values: a list of grades
// (key, value) with same key will shuffle to the same machine
// this Reduce function is invoked once per unique name
total = 0
for each grade v in values{
     total +=v
     }
emit(name, total)
}
```

Problem 3 (3 points) Assume that the connections among 5 webpages are represented as a graph as follows. We use the PageRank equation (with random teleports) to update the rank value of each webpage. Assume that the initial (iteration 0) rank value of each webpage is 1/5, and the damping factor β is 0.85.



(a) Formulate the PageRank equation for each webpage, ensuring the inclusion of specific weights.

(b) Compute the PageRank values for all five webpages during iterations 1 and 2. Subsequently, arrange the webpages in descending order based on their PageRank values. Note that the ranking after the second iteration remains consistent with the final ranking.

Answer:

(a) $r(1) = 0.2 * (1 - 0.85) + 0.85(r(2)/2 + r(5)/2)$.
$r(2) = 0.2 * (1 - 0.85) + 0.85(r(1) + r(3)/2 + r(4)/2)$.
$r(3) = 0.2 * (1 - 0.85) + 0.85(r(5)/2)$.
$r(4) = 0.2 * (1 - 0.85) + 0.85(r(3)/2)$.
$r(5) = 0.2 * (1 - 0.85) + 0.85(r(2)/2 + r(4)/2)$.

(b)

1	2	3	4	5
0.2	0.37	0.115	0.115	0.2
0.27225	0.29775	0.115	0.078875	0.236125

Final ranking: **2, 1, 5, 3, 4**