Big Data Management Tutorial - MapReduce

Consider a *Student* table containing three columns (*studentID*, *courseID*, *semester*). Each tuple in the *Student* table records that a student registered for the course in the corresponding semester. Also consider a *Professor* table containing three columns (*professorID*, *courseID*, *semester*). Each tuple in the *Professor* table records that a professor teaches a course in the corresponding semester. A course may open in multiple semesters. If a student fails a course, he may retake the course in later semesters. **Example tuples** for the two tables are as follows.

studentID	courseID	semester
S001	C001	2021S1
S002	C001	2021S1
S002	C002	2021S2
S003	C001	2021S2

Table Q4.1: Example Student Table

professorID	courseID	semester
P001	C001	2021S1
P002	C002	2021S1
P001	C001	2021S2

Table Q4.2: Example Professor Table

The *Professor* table and the *Student* table are stored together in a file named *input_file*, with each tuple per line. There is an additional attribute to indicate whether this tuple is from the *Student* Table or the *Professor* Table. Based on the above example tuples, the file content is as follows.

Student-Table S001;C001;2021S1
Student-Table S002;C001;2021S1
Student-Table S002;C002;2021S2
Student-Table S003;C001;2021S2
Professor-Table P001;C001;2021S1
Professor-Table P002;C002;2021S1
Professor-Table P001;C001;2021S2

Please use MapReduce for the following scenarios and write down the pseudocode. Your pseudocode should start with a *Map* function that takes each line of the *input_file* as the input. The key in the *Map* function is the additional attribute (e.g., "*Student-Table*") in the line, and the value in the *Map* function is the remaining of the line (e.g.,

S001;C001;2021S1). You also need to design *Reduce* function if necessary. You can use multiple MapReduce jobs.

(a) Use MapReduce to collect for each student (represented by *studentID*) the number of distinct courses (represented by *courseID*s) he has registered.

(b) Use MapReduce to collect the courses (represented by *courseID*s) that have more than 50 registered students for at least two semesters. For example, a course will be output if there are 55 students for 2021S1 and 60 students for 2021S2.

(c) Use MapReduce to output every pair of (student, professor) (represented by *studentID* and *professorID*) that the student has attended at least one courses taught by the professor.