ENI-Carthage **January 2018**

Cloud Computing & Big Data Frameworks Exam

Instructor: Moussa R. **Groups :** IS Eng. 3 **Time Limit:** 1h30

Last Name :	First Name :	ID number :	Group:

✂---

Nota : Answers are in english or in french. Rèpondre en Français ou en Anglais.

Answer the following questions (7)

(1) Give 2 reasons not to move to the Cloud (2)

....

....

(2) Give 2 reasons to move to the Cloud (2)

....

....

(3) Give an example of a *continuous query* in a social net application such as twitter, facebook. (1.5)

....

....

(4) Why Spark is faster than Hadoop? (1.5)

....

....

Circle the correct answer(s) (3)

(5) The client reading the data from HDFS filesystem in Hadoop does which of the following?

a) Gets only the block locations from the namenode

b) Gets the data from the namenode

c) Gets both the data and block location from the namenode

d) Gets the block location from the datanode

(6) Is it a good practice to use HDFS for multiple small files? a) yes b) no

(7) Which type of data processing Spark offers?

a) Batch-based processing of data stream.

b) Real time processing of data stream.

c) Both.

MapReduce Exercise (6 pts)

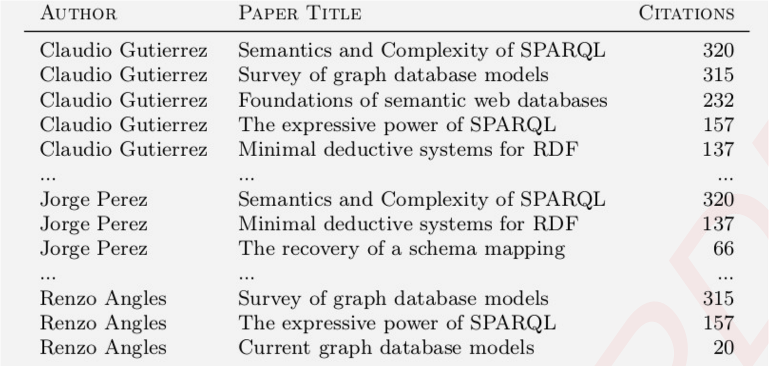
The table below is a large tab-separated values (TSV) file which contains millions of records about authors, their papers, and the citations of their papers. Multiple authors may write a single paper. Paper titles and author names can be assumed to be unique.

You are asked to compute a new file with pairs of co-authors and the sum of the number of citations of those papers they have co-authored together (Author1, Author2, sum of citations of co-authored papers). (indic: avoid duplicates by ensuring that *Author 1* is alphabetically lower than *Author 2*).

Created in Master PDF Editor

✂--

Given this input and desired output, design a series of MapReduce jobs to perform the required processing.



Author	Paper Title	Citations
Claudio Gutierrez	Semantics and Complexity of SPARQL	320
Claudio Gutierrez	Survey of graph database models	315
Claudio Gutierrez	Foundations of semantic web databases	232
Claudio Gutierrez	The expressive power of SPARQL	157
Claudio Gutierrez	Minimal deductive systems for RDF	137
...	...	...
Jorge Perez	Semantics and Complexity of SPARQL	320
Jorge Perez	Minimal deductive systems for RDF	137
Jorge Perez	The recovery of a schema mapping	66
...	...	...
Renzo Angles	Survey of graph database models	315
Renzo Angles	The expressive power of SPARQL	157
Renzo Angles	Current graph database models	20

Job #1

Mapper(s)

```
line in each input split is (K: offset of the line in
the file, V: line)
for each line in the input split do:
parse line: author, title, citations
emit(K: (title, citations), author)
```

Reducer(s)

```
input for each reducer
(K:(title, citations), L: list of authors)
 for each author a in L do:
    for each author b in L do:
        if (a < b) emit ((a,b), citations)
```

the framework sorts and groups
by (title, citations)

Job #2

Mapper(s)

input job 2 is the output of job 1

```
line in each input split (K: (author-a, author-b),
V: citations)

IDENTITY MAP: no transformation

emit (K: (author-a, author-b),
V: citations)
```

Reducer(s)

```
input for each reducer
K:(author-a, author-b), L:List of citations

for each pair
sum = 0
for each c in L do:
sum+= c

emit ((author-a,author-b), sum)
```

the framework sorts and groups
by (author-a, author-b)

Created in Master PDF Editor

✂--

Created in Master PDF Editor