ENI-Carthage **February 2019**

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 (10)

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

.....

.....

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

.....

.....

(3) What are the different types of time-windows for handling streams(3)

.....

.....

.....

.....

(4) Let's take the following file:

$0 $1 $2 $3 $4 $5

```
Frank,19,44,1st_year,12
John,23,,2nd_year,-1
Tom,21,,,0
Frank,9,44,2st_year,10
```

and the following Pig Latin script with $i (i : 0..4) is the ith attribute

```
A = load 'dataset.txt' using PigStorage(',');
B = filter A by $1>20;
C = group B by $2;
dump C;
```

john and Tom lines selected

(null, ((john,23,,), (Tom,21,,..)))

How many records will be generated as output when running this script? (2)

a)0 b) 1 c) 2 d) 3

5) Let's consider the file above once more. You are tasked with writing a Pig Latin script that outputs the unique names (first column) occurring in this file. Which Pig Latin operators do you use (choose the minimum number)? (1)

a) foreach, distinct b) filter, distinct c) foreach, filter d) foreach

MapReduce Exercise (10 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 triplets of co-authors and the sum of the

✂----------

number of citations of those papers they have co-authored together (Author1, Author2, Author3, sum of citations of co-authored papers > 10). (indic: avoid duplicates by ensuring that *Author 1* is alphabetically lower than *Author 2 and Author2* is alphabetically lower than *Author3*).

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)	*Reducer(s)*
TO DO	
Job #2	
Mapper(s)	*Reducer(s)*