Université Tunis El Manar — Année universitaire 2019/2020
Institut Supérieur d'Informatique

DEVOIR SURVEILLÉ
MATIÈRE : BIG DATA

Classe / Filière	**:**	3ème IDL	le 31 octobre 2019 à 8h30
Documents	**:**	**Non autorisés**	**Nom&Prénom :…….**

EXERCICE 1 :

On considère le code pig ci-dessous, précisez ce que affiche chaque « dump » donné en a), b) et c) :

```
-- Chargement des documents de « journal-small.txt » (voir annexe)
articles = load 'journal-small.txt'
    as (year: chararray, journal:chararray, title: chararray) ;
sr_articles = filter articles BY journal=='SIGMOD Record';
year_groups = group sr_articles by year;
count_by_year = foreach year_groups generate group, COUNT(sr_arti-
cles.title);
```

a) `dump sr_articles;`
b) `dump year_groups;`
c) `dump count_by_year;`

Le contenu du fichier « journal-small.txt »

```
2005    VLDB neural networks.
1997    VLDB Big Data.
2003    SR   Management
2001    VLDB E-Services.
2003    SR   Optimization.
1998    VLDB Memory in Databases.
1996    VLDB Query in Oracle
1996    VLDB Relational Algebra.
1994    SR   DDL.
2002    SR   Data Mining.
```

EXERCICE 2 :

Cocher une ou plusieurs réponse(s) correcte(s) :

1) HDFS stocke les données sur
   a) plusieurs « Nodes »
   b) Plusieurs « Blocks »
   c) Plusieurs « Ranks »

2) Si HDFS détecte un Data Node défectueux, alors il réalise la fiabilité par
   a) La substitution des données sur un autre Node.
   b) La réplication des données sur plusieurs autres Nodes.
   c) L'isolation de ce Node défectueux.

3) Un "Node" est le rassemblement de :
   a) De « CPU », « Memory » et plusieurs « Disks »
   b) De « CPU » et plusieurs « Disks »
   c) De « CPU », plusieurs « Disks » et plusieurs « Memory »

4) En HDFS, de nouveaux "Nodes" peuvent être ajoutés sans changement de:
   a) Comment les données sont supprimées
   b) Comment les données sont stockées
   c) Comment les jobs sont écrits et lus

5) En 2001, Google a proposé « Google File System » et
   a) HDFS
   b) Hadoop Common
   c) MapReduce

6) Hadoop est fait pour les travaux « OLTP » :
   a) Vrai b)Faux

7) Seules les pannes des « Tasktracker » et « Child » sont tolérées :
   a) Vrai b)Faux

8) Hive est un environnement pour la gestion et l'interrogation de données

   a) Structurées
   b) Semi Structurées
   c) Non Structurées

9) Clé/Valeur est le résultat de l'étape :
   a) split
   b) map
   c) Shuffle

10) Les MPP (Massive Parallel Processing) ne permettent pas d'absorber les pics de distribution, de charge et de temps de réponse

   a) Vrai b)Faux