esprit Se former autrement	**EXAMEN** Semestre : 1 ■ 2 □ Session : Principale ■ Rattrapage □

Module: Big Data
Enseignantes : Asma Hamed, Henda Sfaxi, Ines Channoufi, Ines Slimene
Classes: 5 ARCTIC, 5 ERP-BI, 5 GL, 5 SIGMA, 5 TWIN
Documents autorisés : OUI □ NON ■ Nombre de pages : 4
Calculatrice autorisée : OUI □ NON ■ Internet autorisée : OUI □ NON ■
Date : 20/11/2017 Heure : 09h00 Durée : 1h30

Exercice 1 : (12 points)

1. Citer et expliquer les mécanismes qui permettent d'assurer la tolérance aux pannes dans HDFS. **(2pts)**
2. Citer et expliquer deux limites de Mapreduce 1 en proposant des solutions pour y remédier. **(2pts)**
3. Citer une bonne raison de choisir le langage de requête Hive plutôt que Mapreduce. **(0.5 pt)**
4. Soit la requête Hive suivante : **(2pts)**

```
hive> LOAD DATA LOCAL INPATH '/home/cloudera/hive/exemple.txt' OVERWRITE
  INTO TABLE exemple;
```

   a. Expliquer en détails la requête.
   b. Expliquer ce qui se produit si le terme **LOCAL** est omis.
   c. Expliquer ce qui se produit si le terme **OVERWRITE** est omis.

5. Soient les deux tables **CUSTOMERS** et **ORDERS** suivantes : **(1.5pts)**

```
+----+----------+-----+-----------+----------+
| ID | NAME     | AGE | ADDRESS   | SALARY   |
+----+----------+-----+-----------+----------+
| 1  | Ramesh   | 32  | Ahmedabad | 2000.00  |
| 2  | Khilan   | 25  | Delhi     | 1500.00  |
| 3  | kaushik  | 23  | Kota      | 2000.00  |
```

```
+-----+---------------------+-------------+--------+
|OID  | DATE                | CUSTOMER_ID | AMOUNT |
+-----+---------------------+-------------+--------+
| 102 | 2009-10-08 00:00:00 |           3 | 3000   |
| 100 | 2009-10-08 00:00:00 |           3 | 1500   |
| 101 | 2009-11-20 00:00:00 |           2 | 1560   |
```

Soit la table **CUST_ORDER** définie comme suit :

```
hive> create table cust_order( id int, name string, amount float,date0 timestamp );
OK
Time taken: 0.06 seconds
hive> insert overwrite table cust_order SELECT c.ID, c.NAME, o.AMOUNT, o.DATE0
    > FROM CUSTOMERS c JOIN ORDERS o ON (c.ID = o.CUSTOMER_ID);
```

Suite à l'exécution de la requête **select * from CUST_ORDER**, les données seront lues à partir de quel(s) fichier(s) ?

6. Citer et expliquer les étapes d'un programme Pig. **(1pt)**

7. Citer une bonne raison de choisir un système relationnel plutôt qu'un système NoSQL pour gérer les données. **(1pt)**

8. Soient les requêtes Hbase suivantes : **(1pt)**

```
hbase(main):009:0> scan 'emp'
ROW                     COLUMN+CELL
 1                      column=personal data:city, timestamp=1510680821804, value=hyderabad
 1                      column=personal data:name, timestamp=1510680797778, value=raju
 1                      column=professional data:designation, timestamp=1510680910238, value=manager
 1                      column=professional data:salary, timestamp=1510680942036, value=50000
1 row(s) in 0.0280 seconds

hbase(main):010:0> drop 'emp'

ERROR: [illegible]
```

La suppression de la table 'EMP' génère une erreur. Expliquer et proposer une solution.

9. Soient les requêtes Hbase suivantes : (**1pt)**

```
hbase(main):001:0> put 'Exemple','001','info:nom','Ali'
0 row(s) in 0.0920 seconds

hbase(main):002:0> put 'Exemple','001','info:nom','Salah'
0 row(s) in 0.0050 seconds

hbase(main):003:0> get 'Exemple', '001',{COLUMN=>'info:nom', VERSION=>2}
```

Quel est le résultat de la dernière requête ?

Exercice 2 : (2 points)

Le schéma ci-dessous correspond à l'architecture de Hadoop 1.

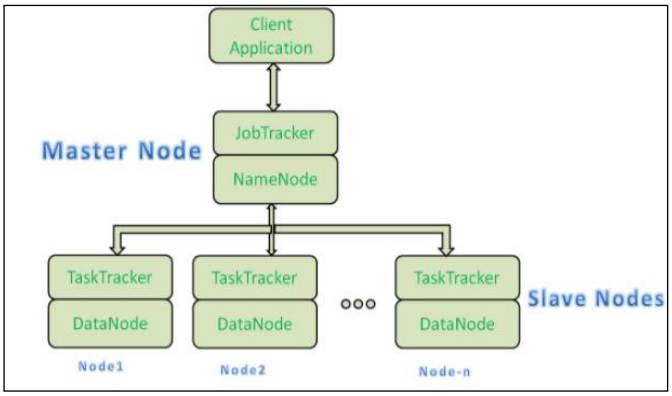

Figure 1 - Architecture Hadoop 1

Présenter l'architecture de Hadoop 2 en décrivant les différents composants.

Exercice 3 : QCM (6 points)

Choisir la bonne réponse.

Q1 : Comment changer le facteur de réplication par défaut pour les données stockées dans HDFS ?

- **a.** Changer la valeur du paramètre dfs.replication de hdfs-site.xml
- **b.** Changer la valeur du paramètre dfs.replication.max de hdfs-site.xml
- **c.** Changer la valeur du paramètre dfs.replication de core-site.xml
- **d.** Changer la valeur du paramètre dfs.replication.max de core-site.xml

Q2 : On souhaite exécuter un job MapReduce appelé « MRJob.jar » sur le fichier 'input.csv' dans un cluster hadoop. Laquelle des commandes suivantes est correcte :

- **a.** Hadoop job MRJob.jar /home/Cloudera/input.csv /home/Cloudera/OutputJob
- **b.** Hadoop jar MRJob.jar /user/Cloudera/input.csv OutputJob
- **c.** Hadoop job MRJob.jar /user/Cloudera/input.csv OutputJob
- **d.** Hadoop jar MRJob.jar /user/Cloudera/input.csv /home/Cloudera/OutputJob

Q3 : Pour créer une table HBASE :

- **a.** Il faut spécifier la liste des colonnes correspondants
- **b.** Il faut spécifier la liste des familles de colonnes correspondants
- **c.** Il faut spécifier les noms des Hfiles correspondants
- **d.** Il faut spécifier les noms des régions correspondants

Q4 : Puisque les données sont répliquées trois fois dans HDFS, cela veut dire que tout calcul effectué sur un nœud sera également reproduit sur les deux autres ?

- **a.** Vrai
- **b.** Faux

Q5 : Tout ce qui suit décrit Hadoop, à l'exception de la notion :

- **a.** Open source
- **b.** Real-time
- **c.** Java-based
- **d.** Distributed computing approach

Q6 : Apache HBase est :

- **a.** Un système de gestion de base de données en colonne qui fonctionne au-dessus de HDFS et qui remplit les propriétés cohérence (consistency) et disponibilité (availability) du théorème CAP
- **b.** Un système de gestion de base de données en colonne qui fonctionne au-dessus de HDFS et qui remplit les propriétés cohérence (consistency) et partitionnement du théorème CAP
- **c.** Un système de gestion de base de données en document qui fonctionne au-dessus de HDFS et qui remplit les propriétés cohérence (consistency) et disponibilité (availability) du théorème CAP
- **d.** Un système de gestion de base de données en document qui fonctionne au-dessus de HDFS et qui remplit les propriétés disponibilité (availability) et partitionnement du théorème CAP