 Se former autrement	<h1 style="margin: 0;">EXAMEN</h1> <p style="margin: 10px 0;">Semestre : 1 <input checked="" type="checkbox"/> 2 <input type="checkbox"/></p> <p style="margin: 10px 0;">Session : Principale <input checked="" type="checkbox"/> Rattrapage <input type="checkbox"/></p>
ETUDIANT(e) Nom et Prénom : Classe :	
Code :	
Module : BIG DATA Enseignantes : Asma Hamed, Ines Channoufi, Ines Slimene, Rayhan Ayadi Classes: 5ARCTIC, 5BI, 5GL, 5SIGMA, 5TWIN Documents autorisés : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Nombre de pages : 06 Calculatrice autorisée : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Internet autorisée : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Date : 22/11/2016 Heure : 09h00 Durée : 1h30	



Code	Note /20	Nom et Signature du Surveillant	Nom et Signature du Correcteur	Observations

Module :

Exercice 1 (4 pts)

Soit les deux fichiers custM.txt et custF.txt suivants :

	FirstName	LastName	Gender	Education	Occupation
1	Melvin	Lal	M	Partial Hi	Skilled Manual
2	Hunter	Green	M	Partial Co	Manual
3	Connor	Patterson	M	Graduate D	Management
4	Tyrone	Romero	M	Partial Hi	Clerical
5	Sebastian	Rivera	M	Partial Co	Professional

	FirstName	LastName	Gender	Education	Occupation
1	Stacy	Sanz	F	Partial Co	Professional
2	Claudia	Wu	F	Bachelors	Management
3	Haley	Green	F	Partial Co	Professional
4	Lindsey	She	F	High Schoo	Management
5	Amanda	Reed	F	Partial Co	Manual
6	Katherine	Murphy	F	Partial Hi	Clerical

1- Trouver une solution pour stocker les deux fichiers dans une seule table Hive.

.....

NE RIEN ECRIRE

2- Trouver une solution afin de garder le fichier sous HDFS après la suppression de la table.

.....

.....

.....

.....

3- Suite au chargement du fichier CustM.txt avec la requête Pig :

```
grunt> customers = LOAD 'exam/CustM.txt' USING PigStorage ('\t') AS (FirstName:chararray,LastName:chararray,Gender:chararray,Education:chararray,Occupation:chararray);
```

La commande suivante n'affiche pas de résultat. Expliquer pourquoi et proposer une solution.

```
grunt> perso_data = foreach customers generate $0,$1 ;  
grunt>
```

.....

.....

.....

.....

Exercice 2 : (2 points)

Expliquer et donner le résultat du script Pig suivant :

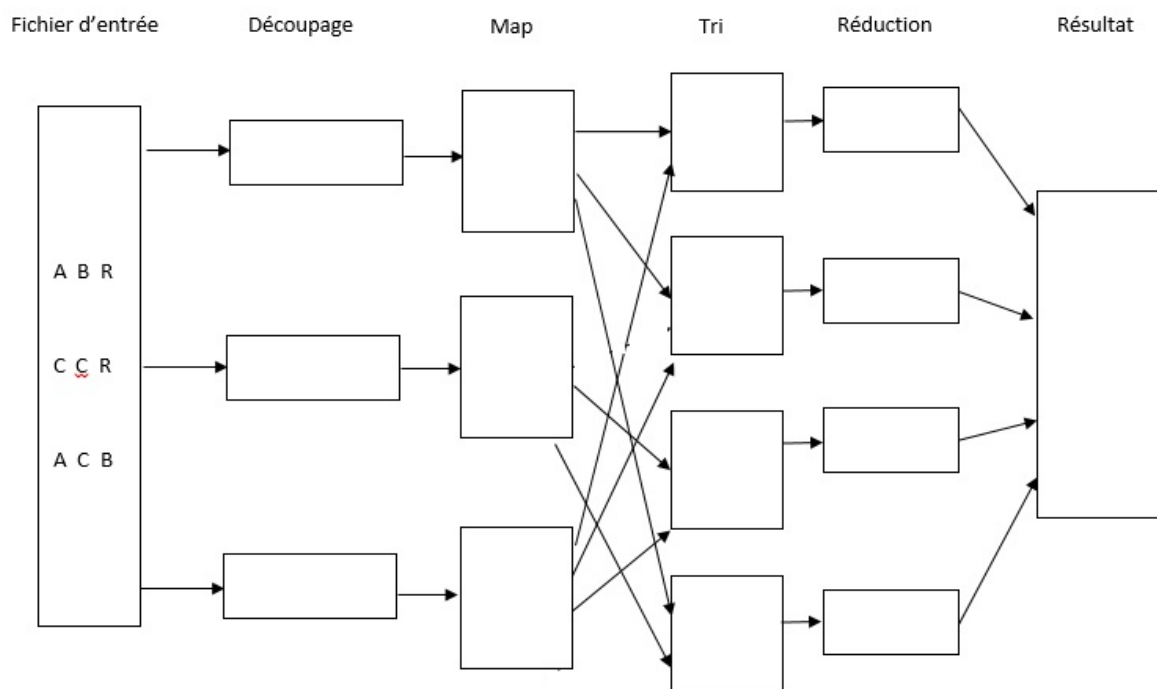
```
livres = LOAD 'exam/livres.txt' USING PigStorage(',') AS (id:int,titre:chararray,date_publication:chararray);  
ventes = LOAD 'exam/ventes.txt' USING PigStorage ('\t') AS (id:int,vendeur:chararray,date_vente:chararray);  
filter_livres = FILTER livres BY id > 1;  
livre_vente = JOIN ventes BY id, filter_livres BY id;  
group_by_vendeur = GROUP livre_vente by ventes::vendeur;  
nbr_livre_par_vendeur = FOREACH group_by_vendeur GENERATE group as vendeur_nom, COUNT(livre_vente) as count_livre ;  
store nbr_livre_par_vendeur into 'outputexam';
```

NE RIEN ECRIRE

.....
.....
.....
.....

Exercice 3 : (3 points)

Compléter le schéma ci-dessous afin d'expliquer les étapes d'un programme mapreduce permettant de compter le nombre d'occurrences de chaque lettre du fichier d'entrée.




NE RIEN ECRIRE

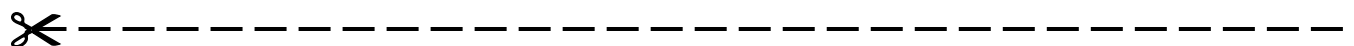
Exercice 4 : QCM (11 points)

Cocher la bonne réponse

N.B : Une seule réponse est correcte

1. Quand Hadoop est-il utile ?
 - ☐ Lorsque toutes les données sont non structurées
 - ☐ Lorsque les traitements peuvent être effectués en parallèle
 - ☐ Lorsque l'application requiert un accès aux données à faible latence
 - ☐ Lorsque l'application nécessite un accès aléatoire aux données
2. Qu'est ce qui est vrai à propos de Pig et Hive par rapport à l'écosystème Hadoop?
 - ☐ HiveQL exige la création d'un flux de données
 - ☐ Pig Latin exige que les données aient un schéma
 - ☐ HiveQL et Pig Latin nécessitent moins de lignes de code qu'un programme mapreduce
 - ☐ Tout ce qui précède
3. Où sont stockés les fichiers de sortie de la tâche Reduce?
 - ☐ Dans un entrepôt de données
 - ☐ En mémoire
 - ☐ Dans le DataNode
 - ☐ Dans le système de fichiers Linux

 Se former autrement	<h2 style="margin: 0;">EXAMEN</h2> <p>Semestre : 1 <input checked="" type="checkbox"/> 2 <input type="checkbox"/></p> <p>Session : Principale <input checked="" type="checkbox"/> Rattrapage <input type="checkbox"/></p>
ETUDIANT(e) Nom et Prénom : Classe:	Code :
Module : BIG DATA Enseignantes : Asma Hamed, Ines Channoufi, Ines Slimene, Rayhan Ayadi Classes: 5ARCTIC, 5BI, 5GL, 5SIGMA, 5TWIN Documents autorisés : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Nombre de pages : 06 Calculatrice autorisée : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Internet autorisée : OUI <input type="checkbox"/> NON <input checked="" type="checkbox"/> Date : 22/11/2016 Heure : 09h00 Durée : 1h30	



4. Le mécanisme qui permet d'éviter la perte de données sous HDFS est :
 - La réplication
 - La partition
 - La scalabilité
 - Yarn
5. Quelles sont les améliorations apportées par YARN par rapport au paradigme mapreduce v1 (MPR1)?
 - C'est complètement open source
 - Il divise JobTracker en deux processus : ResourceManager et ApplicationManager
 - Il divise le TaskTracker en deux processus : ResourceManager et ApplicationManager
6. Job Tracker fonctionne sur le :
 - NameNode
 - DataNode
 - NameNode secondaire
 - DataNode secondaire
7. Quel est le composant de stockage logique des lignes d'une table HBase ?
 - HDFS
 - Région
 - Master

NE RIEN ECRIRE

8. Lequel de ces éléments est responsable de la réplication des données dans Hadoop?
 - ☐ Task Tracker.
 - ☐ Job Tracker.
 - ☐ NameNode
 - ☐ DataNode
9. Quelle est la commande qui permet d'afficher la liste des bases de données Hive ?
 - ☐ DISPLAY ALL DB;
 - ☐ SHOW ME THE DATABASES;
 - ☐ DISPLAY DB;
 - ☐ SHOW DATABASES;
10. LOAD DATA LOCAL signifie que les données doivent être chargées à partir du HDFS ?
 - ☐ Vrai
 - ☐ Faux
11. Les Bags sont des groupes de tuples, les tuples sont des groupes de champs, les champs ont des types de données ?
 - ☐ Vrai
 - ☐ Faux