École Supérieure de la Statistique et de l'Analyse de l'Information	Classe : 3ème Année
Année Universitaire : 2024-2025	Date : 09.01.2025
Examen Big Data	**Durée : 1h 30**

Questions à choix multiples

.5 1. ..... est un utilitaire qui permet aux utilisateurs de créer et d'exécuter des tâches avec n'importe quel exécutable comme mapper et/ou reducer. :
   a) Hadoop Strdata
   b) Hadoop Streaming
   c) Hadoop Stream
   d) Aucune de ces réponses

.5 2. Un nœud ..... agit en tant qu'esclave et est responsable de l'exécution d'une tâche qui lui est assignée par le JobTracker.
   a) MapReduce
   b) Mapper
   c) TaskTracker
   d) JobTracker

.5 3. Quel est le rôle du NameNode ?
   a) Écrire ou lire les données sur les DataNodes.
   b) Vérifier la disponibilité des données sur les DataNodes.
   c) Remplacer un DataNode si un d'entre eux devient indisponible.
   d) Administrer les transactions en autorisant ou non la lecture / écriture des fichiers.

.5 4. La partie ..... de Map-Reduce est responsable du traitement d'un ou plusieurs morceaux de données et de la production des résultats de sortie.
   a) Mapper
   b) Reduce
   c) Map
   d) Aucune de ces réponses

.5 5. Quels sont les composants les plus critiques du Big Data ?
   a) MapReduce
   b) YARN
   c) HDFS
   d) Tous les composants ci-dessus

.5 6. Lesquels des éléments suivants sont des avantages du traitement des Big Data ?
   a) Réduction des coûts
   b) Réduction du temps
   c) Décisions commerciales plus intelligentes
   d) Aucune de ces réponses

.5 7. ..... implique l'exécution simultanée de plusieurs sous-tâches qui, ensemble, constituent une tâche plus importante.
   a) Traitement parallèle des données
   b) Traitement simple
   c) Traitement de données multiples
   d) Aucun des éléments mentionnés ci-dessus

.5 8. Comment fonctionne la distribution de fichiers sur HDFS ?
   a) Répartition en fonction de la taille des fichiers sur chaque nœuds du cluster.
   b) Répartition en blocs répliqués sur les nœuds du cluster.
   c) Répartition en nœuds répliqués sur les blocs du cluster.
   d) Répartition en fonction des choix de l'utilisateur au moment de l'upload.

.5 9. Quel est le rôle du NameNode ?
   a) Écrire ou lire les données sur les DataNodes.
   b) Vérifier la disponibilité des données sur les DataNodes.
   c) Remplacer un DataNode si un d'entre eux devient indisponible.
   d) Administrer les transactions en autorisant ou non la lecture / écriture des fichiers.

.5 10. ..... peut-être décrit comme un modèle de programmation utilisé pour développer des applications basées sur Hadoop qui peuvent traiter des quantités massives de données.
   a) MapReduce
   b) Mahout
   c) Oozie
   d) Toutes les réponses précédentes

.5 11. Le nombre de Maps est généralement déterminé par la taille totale des :
   a) Entrées
   b) Sorties
   c) Tâches
   d) Aucune des réponses précédentes

.5 12. L'entrée du ..... est la sortie triée des Mappers.
   a) Reducer
   b) Mapper
   c) Shuffle
   d) Toutes les réponses précédentes.

Exercice 1 (2 points) : Commenter et préciser les sorties des commandes suivantes :

a) hadoop jar /usr/lib/hadoop-0.20-mapreduce/contrib/streaming/hadoop-streaming-2.6.0-mr1-cdh5.13.0.jar -mapper code/mapper.py -reducer code/reducer.py -file code/mapper.py -file code/reducer.py -input myinput/file.txt -outputjoboutput

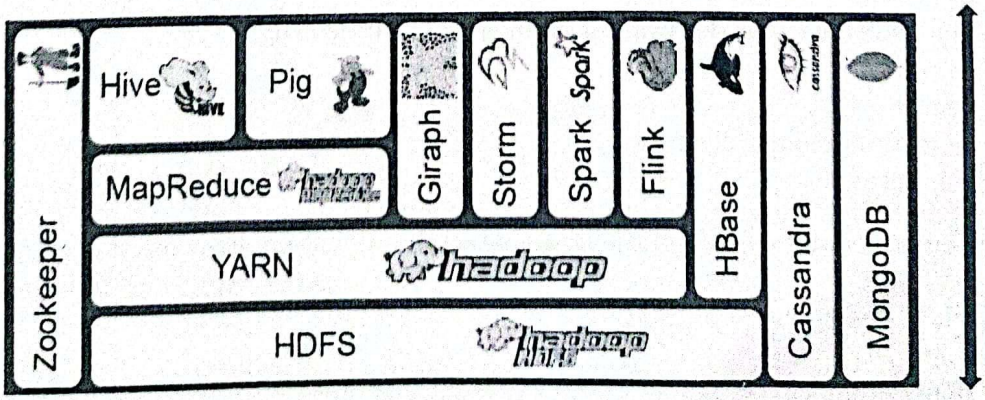
b) hadoop fs -cat joboutput/part-00000

c) head -50 Downloads/purchases.txt — code/mapper.py — sort —code/reducer.py

d) chmod +x code/reducer.py

Exercice 2 (3 points)

On considère le diagramme de couches d'Hadoop suivant :

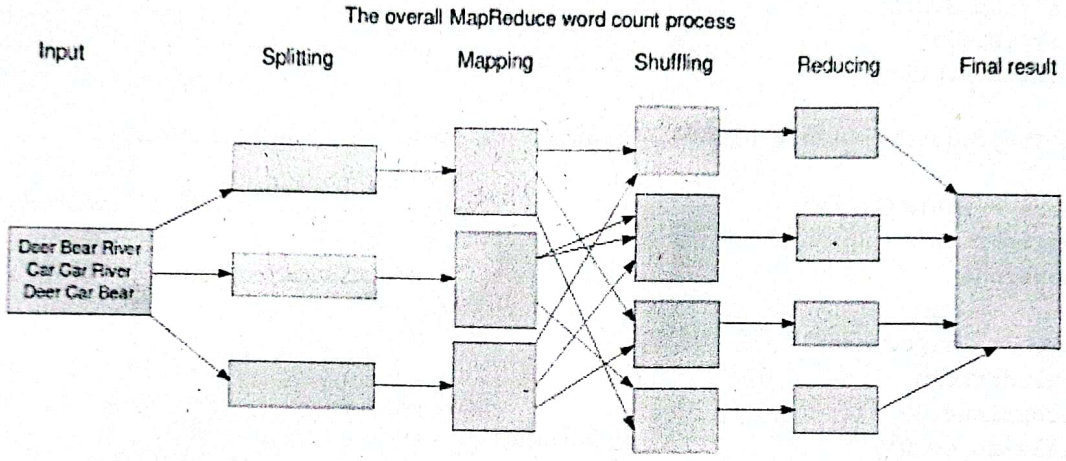


1.5 1. Commenter le diagramme de couches présenté ci-dessus.

1.5 2. Rappeler l'utilité des "frameworks" présents dans le diagramme

Exercice 3 (3 points) : Ecrire les commandes sous Hadoop permettant de :
a) stocker le fichier livre.txt sur HDFS dans le répertoire /data_input.
c) créer le répertoire /data_input
d) supprimer le fichier /data_input/livre.txt.

Exercice 4 (2 points) : Compléter les vides dans le diagramme suivant :



Exercice 5 (6 points) :

On dispose d'un autre fichier volumineux contenant les données (id,age,sexe,région,salaire,ancienneté) sur les employés d'une entreprise sous le format suivant :

```
Affichage de 6 lignes du fichier considéré :
0,25,homme,oran,28000,3
1,33,homme,oran,28000,2
2,46,homme,oran,54000,18
3,35,famme,alger,33000,10
4,23,famme,oran,25000,1
5,25,famme,mascara,25000,3
```

On vous donne les deux codes du programme Map-Reduce suivants :

Code 1 :

```
#!/usr/bin/env python

import sys
wordList = dict()
# input comes from STDIN (standard input)
for line in sys.stdin:
    line = line.strip()
    words = line.split(',')
    print '%s\t%s\t%s' % (words[1],words[4],1)
```

Code 2 :

```
#!/usr/bin/env python

from operator import itemgetter
import sys

current_age = None
max_salaire = 0
min_salaire = 0
current_counter = 0
age = None

wordList = dict()
# input comes from STDIN
for line in sys.stdin:
    line = line.strip()

    age, salaire,counter = line.split('\t', 3)
    try:
        salaire = int(salaire)
    except ValueError:
          continue
    try:
        counter = int(counter)
    except ValueError:
          continue
    if min_salaire == 0:
        min_salaire == salaire
    if current_age == age:
        if max_salaire < salaire:
            max_salaire = salaire
        if min_salaire > salaire:
            min_salaire = salaire
        current_counter += 1
    else:
        if current_age:
         print '%s\t%s\t%s\t%s' % (current_age, max_salaire , min_salaire , current_counter)
        current_age = age
        current_counter = counter
        current_salaire = salaire
        min_salaire =salaire
        max_salaire = salaire

if current_age == age:
    print '%s\t%s\t%s\t%s' % (current_age, max_salaire , min_salaire , current_counter)
```

[2] 1. Identifier le code Map et celui Reducer.

[2] 2. Expliquer les lignes de chaque code.

[2] 3. Préciser l'output de chaque code.