

École Supérieure de la Statistique et de l'Analyse de l'Information	Classe : 3ème Année
Année Universitaire : 2020-2021	Date : 27.11.2020
Examen : Big Data	Durée : 1h 30

Questions de cours

1. Définir les trois caractéristiques du Big Data
2. Décrire brièvement les étapes déclenchées lorsqu'un client Hadoop soumet un job pour exécution.
3. Quels sont les facteurs qui ont lancé l'ère du Big Data ?

Questions à choix multiples

1. Choisissez la bonne réponse :
 - a) Hadoop a besoin de matériel spécialisé pour traiter les données
 - b) Hadoop 2.0 permet le traitement en temps réel des données en temps réel
 - c) Dans le cadre de programmation Hadoop, les fichiers de sortie sont divisés en lignes ou enregistrements
 - d) Aucune des réponses précédentes
2. 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.
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.
4. Selon les analystes, en quoi les systèmes informatiques traditionnels peuvent-ils fournir un socle de base lorsqu'ils sont intégrés aux grandes technologies de données comme Hadoop ?
 - a) Big data et data mining
 - b) Data warehousing et Business Intelligence
 - c) Gestion des clusters Hadoop
 - d) Collecter et stocker des données non structurées
5. Choisissez la bonne réponse :
 - a) Hadoop est idéal pour la charge de travail analytique, post-opérationnelle, d'entrepôt de données
 - b) HDFS s'exécute sur un petit groupe de nœuds
 - c) Aucune des réponses précédentes
6. Hadoop est un framework qui fonctionne avec une variété d'outils connexes. Les cohortes communes incluent :
 - a) MapReduce, Hive and HBase
 - b) MapReduce, MySQL and Google Apps
 - c) Toutes les réponses précédentes

- [.5] 7. Tous les éléments suivants décrivent avec précision Hadoop, SAUF :
- a) Open source
 - b) Temps réel
 - c) Basé sur du Java
 - d) Approche de programmation distribué
- [.5] 8. NameNode est utilisé lorsque le NameNode primaire ne fonctionne plus.
- a) Rack
 - b) Data
 - c) Secondaire
 - d) Aucune des réponses précédentes
- [.5] 9. La machine est un point d'échec unique pour un cluster HDFS.
- a) DataNode
 - b) NameNode
 - c) ActionNode
 - d) Toutes les réponses précédentes
- [.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 besoin de réplication de données peut se produire dans divers scénarios comme :
- a) Le facteur de réplication est modifié
 - b) DataNode ne fonctionne plus
 - c) Les blocs de données sont corrompus
 - d) Tous les réponses précédentes
- [.5] 12. est le noeud esclave / travailleur et conserve les données utilisateur sous forme de blocs de données.
- a) DataNode
 - b) NameNode
 - c) Data block
 - d) Replication
- [.5] 13. est un modèle de calcul à usage général et un système d'exécution pour l'analyse de données distribuées.
- a) Hadoop
 - b) Sparks
 - c) Flume
 - d) Aucune des réponses précédentes
- [.5] 14. Un noeud sert d'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] 15. Choisissez la bonne réponse :
- a) MapReduce essaie de placer les données et le calcul le plus proche dans le temps
 - b) La tâche Map du MapReduce est exécutée à l'aide de la fonction Mapper()

- c) Réduire la tâche dans MapReduce est effectuée en utilisant la fonction Map()
 - d) Toutes les réponses précédentes
- [.5] 16. Le nombre de Maps est généralement déterminé par la taille totale des :
- a) Entrées
 - b) Sorties
 - c) Tâches
 - d) Aucune des réponses précédentes
- [.5] 17. L'entrée du est la sortie triée des Mappers.
- a) Reducer
 - b) Mapper
 - c) Shuffle
 - d) Toutes les réponses précédentes
- [.5] 18. Lesquelles des phases suivantes se produisent simultanément ?
- a) Shuffle & Sort
 - b) Reduce & Sort
 - c) Shuffle & Map
 - d) Toutes les réponses précédentes
- [.5] 19. L'interface réduit un ensemble de valeurs intermédiaires qui partagent une clé avec un ensemble plus petit de valeurs.
- a) Mapper
 - b) Reducer
 - c) Writable
 - d) Readable
- [.5] 20. YARN signifie :
- a) Yahoo's another resource name
 - b) Yet another resource negotiator
 - c) Yahoo's archived Resource names
 - d) Yet another resource need.

Exercice 2 : on administre un réseau social comportant des millions d'utilisateurs. Pour chaque utilisateur, on a dans notre base de données la liste des utilisateurs qui sont ses amis sur le réseau (via une requête SQL).

On souhaite afficher quand un utilisateur va sur la page d'un autre utilisateur une indication "Vous avez N amis en commun"

- [1.5] 1. Proposer des couples (clef ; valeur) avec lesquels le traitement soit parallélisable
- [1.5] 2. Ecrire les squelettes des programmes de Map et Reduce

Exercice 3 (2 points) : écrire les commandes sous Hadoop permettant de :

- a) stocker le fichier livre.txt sur HDFS dans le répertoire /data_input.
- b) obtenir le fichier /data_input/livre.txt de HDFS et le stocker dans le fichier local
- c) créer le répertoire /data_input
- d) supprimer le fichier /data_input/livre.txt

Correction de l'examen:

Questions de cours : Réponses

1. Définir les trois caractéristiques du Big Data :

Les trois caractéristiques principales du Big Data sont :

- **Volume** : Quantité massive de données générées à partir de diverses sources comme les réseaux sociaux, les capteurs, etc.
- **Vélocité** : Vitesse à laquelle les données sont générées, collectées et analysées.
- **Variété** : Diversité des types de données, incluant les données structurées, semi-structurées et non structurées.

2. Étapes déclenchées lorsqu'un client Hadoop soumet un job :

- Le **client** soumet le job au **JobTracker**.
- Le **JobTracker** divise le job en tâches et les assigne aux **TaskTrackers**.
- Les **TaskTrackers** exécutent les tâches (Map et Reduce).
- Les résultats des tâches sont combinés et renvoyés au client.

3. Facteurs ayant lancé l'ère du Big Data :

- Explosion des données générées par Internet, les réseaux sociaux et les appareils connectés.
- Avancées dans les technologies de stockage et de traitement distribués.
- Besoin croissant d'analyse en temps réel pour des décisions basées sur les données.

Questions à choix multiples : Réponses

1. Choisissez la bonne réponse :
d) Aucune des réponses précédentes.
2. Comment fonctionne la distribution de fichiers sur HDFS ?
b) Répartition en blocs répliqués sur les nœuds du cluster.
3. Quel est le rôle du NameNode ?
d) Administrer les transactions en autorisant ou non la lecture / écriture des fichiers.
4. En quoi les systèmes informatiques traditionnels peuvent-ils fournir un socle de base ?
b) Data warehousing et Business Intelligence.
5. Choisissez la bonne réponse :
a) Hadoop est idéal pour la charge de travail analytique, post-opérationnelle, d'entrepôt de données.
6. Les cohortes communes incluent :
a) MapReduce, Hive and HBase.
7. Tous les éléments suivants décrivent Hadoop, SAUF :
b) Temps réel.

8. Le ... NameNode est utilisé lorsque le NameNode primaire ne fonctionne plus :
c) Secondaire.
9. La machine ... est un point d'échec unique pour un cluster HDFS :
b) NameNode.
10. ... peut être décrit comme un modèle de programmation utilisé pour développer des applications basées sur Hadoop :
a) MapReduce.
11. Le besoin de réplication de données peut se produire dans divers scénarios comme :
d) Toutes les réponses précédentes.
12. ... est le nœud esclave/travailleur et conserve les données utilisateur sous forme de blocs de données :
a) DataNode.
13. ... est un modèle de calcul à usage général et un système d'exécution pour l'analyse de données distribuées :
b) Spark.
14. Un nœud ... sert d'esclave et est responsable de l'exécution d'une tâche qui lui est assignée par le JobTracker :
c) TaskTracker.
15. Choisissez la bonne réponse :
b) La tâche Map du MapReduce est exécutée à l'aide de la fonction Mapper().
16. Le nombre de Maps est généralement déterminé par la taille totale des :
a) Entrées.
17. L'entrée du ... est la sortie triée des Mappers :
a) Reducer.
18. Lesquelles des phases suivantes se produisent simultanément ?
a) Shuffle & Sort.
19. L'interface ... réduit un ensemble de valeurs intermédiaires qui partagent une clé avec un ensemble plus petit de valeurs :
b) Reducer.
20. YARN signifie :
b) Yet another resource negotiator.

Exercice 2 : Traitement des amis en commun

1. Couples (clé ; valeur) pour un traitement parallélisable :

Pour chaque utilisateur U , on peut générer des couples (clé, valeur) comme suit :

- **Clé** : Un couple d'utilisateurs triés (e.g., (U, V)).
- **Valeur** : La liste des amis de chaque utilisateur.

Exemple :

Si U_1 a pour amis $\{U_2, U_3\}$ et U_2 a pour amis $\{U_1, U_4\}$, les couples générés seront :


- Mapper output :
 - $(U_1, U_2) \rightarrow [U_2, U_3]$
 - $(U_2, U_1) \rightarrow [U_1, U_4]$

Le reducer combiner ces listes pour calculer les amis en commun.

2. Squelettes des programmes Python pour Map et Reduce :

Mapper :

python


 Copy code

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

for line in sys.stdin:
    # Chaque ligne contient un utilisateur et sa liste d'amis : "U\tF1,F2,F3,..."
    user, friends = line.strip().split("\t")
    friends = friends.split(",")
    for friend in friends:
        # Émettre une clé (user, friend) avec la liste des amis de user
        key = tuple(sorted([user, friend]))
        print(f"{key}\t{'.'.join(friends)}")
```

Reducer :

python

 Copy code

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

current_pair = None
friends_list = []

for line in sys.stdin:
    pair, friends = line.strip().split("\t")
    if pair != current_pair:
        if current_pair:
            # Trouver les amis en commun entre deux listes
            common_friends = set(friends_list[0]).intersection(set(friends_list[1]))
            print(f"{current_pair}\t{len(common_friends)}")
        current_pair = pair
        friends_list = [friends.split(",")]
```


```
        friends_list.append(friends.split(","))
    else:
        friends_list.append(friends.split(","))

# Dernier couple
if current_pair and len(friends_list) == 2:
    common_friends = set(friends_list[0]).intersection(set(friends_list[1]))
    print(f"{current_pair}\t{len(common_friends)}")
```

Exercice 3 : Commandes Hadoop

a) Stocker le fichier `livre.txt` sur HDFS dans le répertoire `/data_input` :


bash

 Copy code

```
hadoop fs -put livre.txt /data_input
```

b) Obtenir le fichier `/data_input/livre.txt` de HDFS et le stocker dans le fichier local :


bash

 Copy code

```
hadoop fs -get /data_input/livre.txt ./livre.txt
```

c) Créer le répertoire `/data_input` :


bash

 Copy code

```
hadoop fs -mkdir /data_input
```

d) Supprimer le fichier `/data_input/livre.txt` :

bash

 Copy code

```
hadoop fs -rm /data_input/livre.txt
```