Faculté des Sciences
Aix-Marseille Université

Année universitaire 2022/2023

Site : ☐ Luminy ☐ St-Charles ■ St-Jérôme ☐ Cht-Gombert ☐ Aix-Montperrin ☐ Aubagne-SATIS

Sujet session de : ■ 1re semestre - ☐ 2ème semestre - ☐ Session 2 Durée de l'épreuve : 2 heures

Examen de : ☐L1/☐L2/☐L3 - ☐M1/■M2 - ☐ LP - ☐ DU Nom diplôme : Master Informatique – Parcours SID

Code Apogée du module : SINCU89 Libellé du module : Big Data

Documents autorisés : ☐ OUI - ■ NON Calculatrices autorisées : ☐ OUI - ■NON

N.B : l'examen comporte 2 parties à rédiger sur des copies séparées.

Pour le QCM, une réponse fausse vaut -0,25.

Partie A (13 points)

1 - Concepts Big Data (7 points)

1.1. Donner une définition du Big Data (4 lignes max)

1.2. Donner une définition de Spark (2 lignes max)

1.3. Représenter une architecture Big Data avec n couches en utilisant le tableau ci-dessous :

a. Choisir une valeur appropriée de n,

b. Renseigner dans chaque couche les éléments Xi se rapportant à cette architecture.

c. Pour chaque élément Xi, dire brièvement (2 lignes maximum) quel est son rôle.

X1
.....
Xn

1.4. Cocher la bonne réponse

2 ○ a. Hadoop 2.0 permet le traitement en temps réel des données en temps réel

○ b. Hadoop a besoin de matériel spécialisé pour traiter les données.

○ c. Hadoop 2.0 permet le traitement en temps réel des données en temps réel
Dans le cadre de programmation Hadoop, les fichiers de sortie sont divisés en lignes ou enregistrements

○ d. Aucune des réponses précédentes

1.5. Choisir l'affirmation vraie :

○ a. Hadoop est idéal pour la charge de travail analytique, post-opérationnelle, d'entrepôt de données

2 ○ b. HDFS s'exécute sur un petit groupe de noeuds

○ c. NewSQL est souvent le point de collecte pour le big data

○ d. Aucune affirmation

1.6. Tous les éléments suivants décrivent avec précision Hadoop, SAUF :

○ a. Approche de programmation distribuée

○ b. Temps réel

2 ○ c. Open source

1.7. Parmi les affirmations suivantes, laquelle est fausse ?

○ a. Map s'applique à chaque élément du RDD et retourne un résultat dans un autre RDD.

2 ○ b. Map peut retourner 0, 1, ou plusieurs éléments.

○ c. Map transforme un RDD de taille N en un autre RDD de taille N

○ d. Un développeur peut implementer sa propre logique métier dans un Map.

1.8. Cocher l'affirmation fausse

- a. Map s'applique à chaque élément du RDD et retourne un résultat dans un autre RDD.
- b. Map peut retourner 0, 1, ou plusieurs éléments.
- c. Map transforme un RDD de taille N en un autre RDD de taille N
- d. Un développeur peut implementer sa propre logique métier dans un Map.

1.9. Quel est le rôle d'un moteur Spark

- a. Planifier les tâches
- b. Distribuer les données sur un Cluster
- c. Monitorer les données sur un cluster
- d. Toutes les réponses

1.10. Dans un RDD, la tolérance aux pannes est assurée grâce à :

- a. Le propriété d'immutabilité des RDD
- b. Le graphe acyclique direct
- c. L'évaluation paresseuse
- d. Aucune des réponses ci-dessus

1.11. Quel est le point d'entrée d'une application Spark ?

- a. Une SparkSession
- b. Un SparkContext
- c. Aucune des réponses

1.12. Dans Spark, une action RDD :

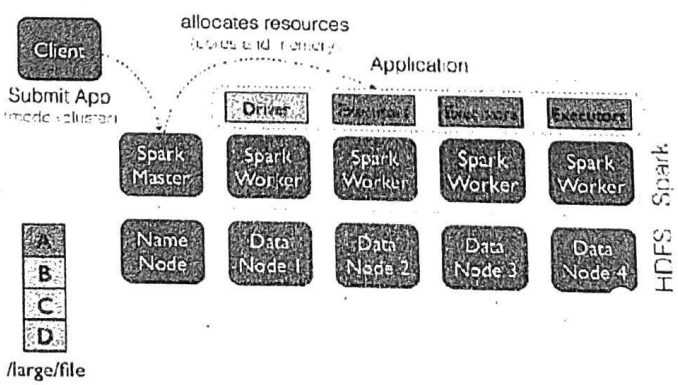
- a. Prend en entrée un RDD et retourne en sortie un ou plusieurs RDD
- b. Crée un ou plusieurs nouveaux RDD
- c. Envoie des résultats des exécuteurs vers le driver
- d. Toutes les réponses sont correctes

1.13. Spark RDD améliore Hadoop/MapReduce grâce :

- a. L'évaluation paresseuse
- b. à la structure de Graphe Acyclique Direct
- c. au traitement en mémoire
- d. Toutes les réponses

2 – Spark-Architecture (3 points)

On considère le schéma ci-dessous qui illustre le traitement par Spark d'un fichier (/large/file).



2.1. Expliquer le fonctionnement de Spark sur la base du traitement sur ce fichier.
2.2. Décrire brièvement le rôle des briques principales représentées sur ce schéma.
2.3. Quelles parties du fichier sont affectées aux 4 Data Node ? Justifier votre réponse

3 – Spark-Programmation (3 points)

On considère le fichier fruits.txt qui contient les lignes suivantes :

```
Pomme Raisin Orange
Raisin Datte Banane
Pomme Raisin Datte
```

3.1. On considère le programme ci-dessous, et on suppose qu'on est en mode console.
Que fait ce programme ?
Quel résultat apparaît à la suite de l'exécution des 3 premières lignes ?

```
> val mydata = sc.textFile("fruits.txt")
> val mydata_uc = mydata.map(line => line.toUppercase())
> val mydata_filt = mydata_uc.filter(line => line.startsWith(''POMME'')
>
```

3.2. On considère le noyau de programme ci-dessous. Compléter ce programme afin qu'il réalise le comptage de mots du fichier `fruits.txt`.

```
> val lines = sc.textFile("fruits.txt")
> val counts = lines.flatMap(lambda line : line.split(-----))
               .map(mot lambda : -----)
               .reduceByKey(lambda x, y : -----))
```

Partie B (7 points)

Spark MLIB

1. Expliquer les différentes étapes d'une procédure d'apprentissage supervisé et lister le nom des fonctions principales à utiliser en Spark MLIB (3 points).
2. Expliquer les concepts d'estimator, transformer et pipeline en Spark MLIB (2 points).
3. Expliquer les fonctions existantes en Spark MLIB pour effectuer le réglage des hyperparamètres et comment les utiliser (2 points).