ENSEEIHT – Valdom - Année 2020-2021 Prénom/Nom : …….

Big Data

Examen

Daniel Hagimont

Durée: 1h, documents autorisés

Questions de cours (10 points)

Chaque réponse doit être concise et rédigée dans le cadre.

Quelle différence principale y a t'il entre un cluster HPC (High Performance Computing) pour le calcul parallèle et un cluster Big Data (devant exécuter des applications Hadoop) ? (1 point)

…….
…….
……
…….

Donnez deux raisons pour lesquelles les blocs sont répliqués dans HDFS ? (1 point)

……..
……..
……..

A quoi correspondent des partitions dans l'exécution d'une application Spark ? (2 points)

…….
……..
……..
……..

Dans Spark, quelle différence essentielle y a t'il entre un map() et un reduceByKey() du point de vue des communications réseau ? (2 points)

……..
……..
……..
……
…….

Expliquer la différence fondamentale entre Spark et Spark-streaming (2 points)

…….
……..
……..

Résumez en une phrase ce qu'on entend par scalabilité dans Spark (2 points)

……..
……..
……..
……..

Problème (10 points)

On considère un grand fichier contenant les enregistrements des ventes (appelées transactions) d'une chaîne de magasins.
Pour chaque transaction, le fichier inclut une ligne prenant la forme :

storeid,productid,number,totalprice

- storeid : l'identifiant du magasin
- productid : l'identifiant du produit
- number : le nombre de produits vendus dans la transaction
- price : le prix total de la transaction (un produit peut être vendu à différents prix dans différents magasins)

Tous ces champs sont des entiers.
On peut extraire d'une ligne L ces champs respectivement avec L.split(" ")[0], L.split(" ")[1], etc.
Vous disposez dans un programme Spark du RDD suivant (qui a été initialisé avec un fichier disponible dans HDFS) :

```
JavaRDD<String> data = sc.textFile(inputFile);
```

Donnez le programme Spark qui calcule pour chaque produit le nombre de produits vendus globalement (2 points)

……
……
……
……
……
……
……
……

Complétez ce programme Spark pour afficher le produit le plus vendus globalement (4 points)
Indication : vous pouvez utiliser les méthodes SortByKey() et take(n) décrites dans le cours

……
……
……
……
……
……
……

Donnez le programme Spark qui calcule pour chaque produit le nombre de magasins où il est vendu (4 points)
Indication : tout RDD permet d'utiliser la méthode distinct() qui en retire les doublons

……
……
……
……
……
……
……
……