Nom :
Prénom :

BDLE – Seconde Partie

Exemen réparti du 24 Novembre 2017
Durée : 2 Heures (au total)
Documents autorisés

Les réponses doivent être inscrites dans les cadres réservés puis sur intercalaires s'il manque de la place.

1 Algèbre RDD de Spark (4 pts)

Soit le jeux de données suivant qui décrit des informations sur des personnes. Chaque ligne décrit un nuplet qui consiste en un ensemble de paires clé-valeur (ex. `nom:mcgill`) séparées par des virgules.

```
nom:mcgill,prenom:ben,age:22
nom:smith,prenom:lara,niveau:4
nom:snod,prenom:rick,age:27,niveau:5
nom:kirch,prenom:lars,pays:russia
```

users.txt

Le but est d'extraire l'ensemble des attributs en indiquant s'ils sont obligatoires (présents dans tous les nuplets) ou optionnels. On utilisera une classe `attribut` ayant deux champs : le nom de l'attribut et un booléen qui est vrai si l'attribut est obligatoire et faux sinon.

Question 1 (4 points)

Compléter les instructions suivantes.

Réponse :

```
val data = sc.textFile(path+"users.txt")

case class attribut(cle:String,pres:Boolean)


/* prend en entrée un String de la forme cle:val et produit
un objet attribut(cle, true)*/
def parseElem(in: String): attribut = {

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

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

     }
```

```
/* prend en entrée  une liste de  String de la forme cle:val
et produit une liste d'objets attribut(cle, true) en utilisant parseElem*/
def parseTuple(in:List[String]): List[attribut] = {

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

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

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

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

     }

/*transforme la RDD data en une RDD de listes d'attributs.
Chaque liste est triée par les clés de ses attributs*/
val parsed =   data.map ......................

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

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


/*voici un extrait de parsed*/
scala> parsed.collect.foreach(println)
List(attribut(age,true), attribut(nom,true), attribut(prenom,true))
List(attribut(niveau,true), attribut(nom,true), attribut(prenom,true))
...

/* combine deux listes d'attributs en une seule liste
où le champ pres de chaque objet attribut est mis à false si
cet objet est absent dans l1 ou l2, sinon il est inchangé.
 Ne pas écrire le crops de cette méthode!!*/
def magic(l1: List[attribut], l2: List[attribut]): List[attribut]

/*produit le résultat final*/
val synthese = parsed. ......................

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


scala> synthese.collect
List(attribut(age,false), attribut(nom,true), attribut(prenom,true),
attribut(niveau,false), attribut(pays,false) )
```

2 Algèbre Dataset de Spark (6 pts)

On s'intéresse à la formulation de requêtes graphe en utilisant l'algèbre Dataset de Spark. On considère des triplets de la forme n, p, m, décrivent qu'un arc p démarre du noeud n et arrive au noeud m. Exprimer en Dataset les requêtes suivantes.

Question 1 (3 points)

Retourner le nombre de circuits (cycles) de longueur 2. (Rappel : un circuit est une suite d'arcs consécutifs dont les deux sommets extrémités sont identiques. La longueur d'un circuit et le nombre des noeuds qu'il traverse.)

Réponse :

On étend l'ensemble de triplets avec des triplets de la forme n, isA, t indiquant que le noeud n a le type t. Pour simplifier, chaque n a un seul type t.

Question 2 (3 points)

Retourner les arcs p communs à au moins deux triplets n, p, n' et m, p, m' tels que n et m ont le même type tandis que n' et m' ont deux types distincts.

Réponse :

3 SQL en MapReduce (2 pts)

On voudrait appliquer la technique de transformation de requêtes SQL vers des jobs MapReduce les plus compacts possible (cf. cours 4). Considérer le schéma et l'arbre algébrique ci-dessous. Toutes les jointures sont naturelles.

$$R(A, B) \quad S(B, C, D) \quad T(C, D, E)$$

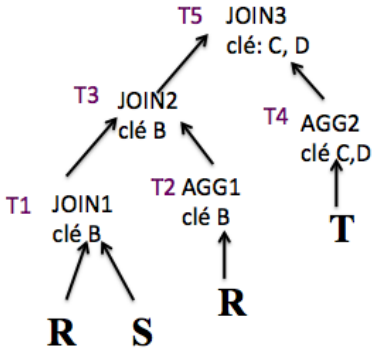

FIGURE 1 – Arbre

Question 1 (1 point)

Extraire les différents types de corrélation, i.e Input, Transit et JobFlow.

Réponse :

Question 2 (1 point)

En appliquant l'optimisation, combien de jobs sont nécessaires pour exécuter cette requête ? Préciser pour chaque job les taches qui y sont exécutées en indiquant les/les règle(s) utilisée(s).

Réponse :