Université Mohamed Boudiaf M'sila
Faculté MI, Dépt Informatique

M1 Intelligence Artificielle
M'sila, Le 08/03/2021
Durée : 60 Minutes

Examen Final

Big Data et Science de Données

Exercice 1 (7 points)

Répondre brièvement à ces questions.
Question 1.1 : Décrivez les trois niveaux de structuration des données. Donnez un exemple pour chaque niveau. ***(1.5 Pts)***
Question 1.2 : Quels sont les trois caractéristiques des Big Data (aussi connus par les trois V). ***(1.5 Pts)***
Question 1.3 : Quels sont les domaines d'application de la science de données ? ***(1 Pt)***
Question 1.4 : Expliquer les expressions suivantes : Passage à l'échelle, Réplication, Sharding d'une BdD NoSQL ***(1.5 Pts)***
Question 1.5 : Dans HDFS, quel est rôle de chacun des nœuds suivants : NameNode active, Secondary NameNode, NameNode standby. ***(1.5 Pts)***

Exercice 2 (5 points)

Répondre par VRAI ou FAUX ***(1 Pt/réponse correcte)***

a- Le Big Data ne peut pas être manipulé par une seule machine.
b- Un algorithme distribué n'a pas besoin de circuler des données initiales entre les nœuds de calcul.
c- Le schéma de parallélisation MapReduce est : Map → Shuffle & Sort → Reduce
d- Dans une BdD NoSQL, une opération Map-Reduce permet de refaire un "Select...From...Where...Group by...".
e- YARN peut prendre en charge jusqu'à 4000 clusters.

Exercice 3 (08 points)

Soit un réseau social comportant des millions d'utilisateurs. Pour chaque utilisateur, on peut avoir dans la base de données la liste des utilisateurs qui sont ses amis sur le réseau. On souhaite afficher quand un utilisateur va sur la page d'un autre utilisateur une indication « *Vous avez N amis en commun* ».
Soient ici, les données d'entrée sous la forme Utilisateur => Amis:
A => B, C, D
B => A, C, D, E
C => A, B, D, E
D => A, B, C, E
E => B, C, D

On veut développer des programmes MAP et REDUCE pour cette opération.

Indications :

- Choisir pour clef la concaténation entre deux utilisateurs. Par exemple, la clef « A-B » désignera « les amis en communs des utilisateurs A et B ».
- Il faut que la clef soit toujours triée par ordre alphabétique (La clef « B-A » sera exprimée sous la forme « A-B ».
- MAP va prendre la liste des amis fournie en entrée, ensuite générer toutes les clefs distinctes possibles à partir de cette liste. La valeur sera simplement la liste d'amis, telle quelle. *Par exemple*, pour la première ligne: A => B,C,D On obtiendra les couples (clef;valeur): ("A-B"; "B C D") ("A-C"; "B C D") ("A-D"; "B C D").
- Après regroupement (shuffle & sort), on obtient pour chaque clef « USER1-USER2 », deux listes d'amis: les amis de USER1 et ceux de USER2. *Par exemple*, Pour la clef "A-B", la valeur sera composée de deux listes : Liste[0] = "A C D E" et Liste[1]="B C D".
- Le rôle de REDUCE sera très simple: déterminer quels sont les amis qui apparaissent dans les deux listes : Liste[0] et Liste[1].

BONNE REUSSITE

Dr. Tahar Mehenni

Université Mohamed Boudiaf M'sila **M1 Intelligence Artificielle**
Faculté MI, Dépt Informatique **M'sila, Le 08/03/2021**
Durée : 60 Minutes

Correction de l'Examen Final

Big Data et Science de Données

Exercice 1 (7 points)

Répondre brièvement à ces questions.

Question 1.1 : Décrivez les trois niveaux de structuration des données. Donnez un exemple pour chaque niveau. ***(1.5 Pts)***

Réponse :

- *Structuré : Une donnée structurée décrit une propiété d'une entité selon un modèle (ou template) fixé. Exemple ; Données stockées dans des feuilles (fichier excel)*
- *Une donnée non structurée décrit une entité qui ne poosède pas une structure à cause de ses propriétés qui ne peuvent pas être distinguées les unes des autres, Exemple : un texte*
- *Une donnée semi-structurée possède une structure où les entitées et leurs propriétés peuvent être facilement distinguées, MAIS l'organisation de la structure n'est pas rigoureuse comme celle de la table de la base de données. Exemples: documents XML, JSON, HTML.*

Question 1.2 : Quels sont les trois caractéristiques des Big Data (aussi connus par les trois V). ***(1.5 Pts)***

Réponse :

- *Volume: La taille du dataset.*
- *Vélocité: La nécessité de traitement des données à leur arrivée.*
- *Variété: La nature hétérogène des données (structurées, semi-structurées, non structurées).*

Question 1.3 : Quels sont les domaines d'application de la science de données ? ***(1 Pt)***

Réponse :

- *Entreprises commerciales: Aperçus de leurs clients, …*
- *Institutions financières: prédire les marchés boursiers, …*
- *Organisations gouvernementales: détection des fraudes, …*
- *Organisations non gouvernementales (ONG): collecter des fonds, …*

Question 1.4 : Expliquer les expressions suivantes : Passage à l'échelle, Réplication, Sharding d'une BdD NoSQL ***(1.5 Pts)***

Réponse :

- *Passage à l'échelle : Speed up + Size up + maîtrise des coût*
- *Réplication : stockage d'une ou plusieurs copies des blocs de données dans des noeuds différents.*
- *Sharding d'une BdD NoSQL : Le sharding est un découpage des différentes lignes d'une table par tranches dans différents chunks (au lieu de distribuer les colonnes)*

Question 1.5 : Dans HDFS, quel est rôle de chacun des nœuds suivants : NameNode active, Secondary NameNode, NameNode standby. ***(1.5 Pts)***

Réponse :

- *NameNode Actif: établit et conserve une cartographie de la répartition de tous les fichiers stockés dans HDFS.*
- *NomeNode Secondaire : Copie de sécurité de NameNode active en cas de panne.*
- *NameNode standby : Solution à haute disponibilité pour HDFS en cas de panne sur le NameNode*

Exercice 2 (5 points)

Répondre par VRAI ou FAUX ***(1 Pt/réponse correcte)***

a- Le Big Data ne peut pas être manipulé par une seule machine. ***VRAI***
b- Un algorithme distribué n'a pas besoin de circuler des données initiales entre les nœuds de calcul. ***FAUX***
c- Le schéma de parallélisation MapReduce est : Map → Shuffle & Sort → Reduce ***VRAI***
d- Dans une BdD NoSQL, une opération Map-Reduce permet de refaire un "Select...From...Where...Group by...". ***VRAI***
e- YARN peut prendre en charge jusqu'à 4000 clusters. ***FAUX***

Exercice 3 (08 points)

Puisqu'on est intéressé par l'information « amis en commun entre deux utilisateurs » et qu'on aura à terme une valeur par clef, on va choisir pour clef la concaténation entre deux utilisateurs. Par exemple, la clef « A-B » désignera « les amis en communs des utilisateurs A et B ».

Le pseudo code de la fonction MAP est:

```
utilisateur = [premiere partie de la ligne]
pour ami dans [reste de la ligne] faire:
        si utilisateur < ami:
                clef = utilisateur+"-"+ami
        sinon:
                clef = ami+"-"+utilisateur
        generer couple (clef; [reste de la ligne])
```

Une fois l'opération MAP effectuée, Hadoop va récupérer les couples (clef;valeur) de tous les fragments et les grouper par clef distincte. on obtient pour chaque clef « USER1-USER2 », deux listes d'amis: les amis de USER1 et ceux de USER2.

Le programme REDUCE va recevoir en entrée toutes les valeurs associées à une clef, ensuite déterminer quels sont les amis qui apparaissent dans les listes (les valeurs) qui nous sont fournies. Pseudo-code:

```
liste_amis_communs=[]
pour ami dans valeurs[0], faire:
        si ami egalement present dans valeurs[1], alors:
                liste_amis_communs+=ami
renvoyer liste_amis_communs
```

Rédigé par T. Mehenni