Exercice 2 : (12 points)

Un réseau social actif dans une communauté francophone souhaite évaluer la maîtrise de la langue française par ses abonnés. Il s'intéresse en particulier à la distinction entre le masculin et le féminin.
Pour cela, les textes des publications et des commentaires des abonnés sur plusieurs années sont sauvegardés dans un fichier « **comments.txt** » de taille **512Go** surHDFS.

1. Sachant que la taille d'un bloc est de **128Mo** et que le facteur de réplication est égal à **3**, calculer le nombre total de blocs appartenant à ce fichier sur HDFS?

2. Donner le nombre de tâches *Map* qui doivent se réaliser dans un job *MapReduce* pour traiter ce fichier. Comment expliquez-vous que le nombre de tâches *Map* réellement lancées sur le cluster peut être supérieur à la valeur calculée ?

3. On vous demande d'écrire un job *MapReduce* pour extraire tous les termes masculins et féminins existants dans ce document. En guise de simplification, on suppose que les mots masculins sont précédés par les termes « **le, un, mon, ce, cet** » et les mots féminins sont précédés par les termes« **la, une, ma, cette** ».

   a. Illustrer une solution utilisant le modèle de programmation *MapReduce*sur l'exemple suivant sans utiliser un *Combiner*.

Mon ami me donne son livre de lecture. La station est loin de ma maison. Cette fille est ma collègue.	⇒	**Masculin:** ami, livre **Féminin:** station, maison, fille, collègue

   b. Ecrire les pseudocodes des classes *Mapper* et *Reducer* correspondantes à ce problème.

   c. Illustrer la solution utilisant un *Combiner*.

   d. Dans Hadoop *MapReduce*, peut-on contrôler quelles clés vont à quel *reducer*? Expliquer comment et discuter le cas de cet exercice.

Bon courage

Université de Monastir

Institut Supérieur d'Informatique et Mathématiques Monastir

Devoir Surveillé:BIG DATA

Section :Mastère Professionnel - Génie Logiciel -Niveau 2

Signature des surveillants
Identifiant

Note/20	Devoir Surveillé-BIG DATA				Identifiant
	Documents :	**Non autorisés**	Section :	**MP2GL**	
	Date	**15/11/2021**	Durée :	**1h**	
	Enseignante :	**Asma KERKENI**	Nombre de pages :	**4**	

☒ *Le barème est donné à titre indicatif et peut subir éventuellement quelques modifications*
☒ *Répondre au QCM sur cette feuille et la joindre à votre feuille de réponse*

Exercice 1 : QCM (8 points)

Dans le Q.C.M. suivant, chaque question admet une ou plusieurs bonne(s) réponse(s). Encadrer la ou les bonne(s) réponse(s).
N.B : les réponses fausses sont pénalisées.

1. Les 3V dans le contexte Big data sont :
   A. Volume, vélocité et variété
   B. Volume, vélocité et vitesse
   C. Volume, valeur et variété
   D. Valeur, variété et vélocité

2. Les unités suivantes sont triées de la plus petite à la plus grande :
   A. *YottaByte ; PetaByte ; ExaByte ; ZettaByte*
   B. *PetaByte ; ExaByte ; ZettaByte ; YottaByte*
   C. *ExaByte ; PetaByte ; ZettaByte ; YottaByte*
   D. *ExaByte ; YottaByte ; PetaByte ; ZettaByte*

3. Le_____ permet de gérer les tâches d'une même application dans un *Cluster YARN*
   A. *NameNode*
   B. *JobTracker*
   C. *Ressource Manager*
   D. *Application Master*

NE RIEN ECRIRE ICI

4. _____ est un modèle de programmation utilisé pour développer des applications basées sur *Hadoop* pouvant traiter des quantités massives de données.
   A. *Map reduce*
   B. *Ressource Manager*
   C. *Yarn*
   D. *Mahout*

5. Lesquelles des phases suivantes se produisent simultanément ?
   A. *Map* et *Shuffle*
   B. *Map* et *Sort*
   C. *Combine* et *Reduce*
   D. *Shuffle* et *Sort*

6. Les composants clés de *Hadoop* sont :
   A. *Map redue*
   B. *HDFS*
   C. *Hive*
   D. *Hbase*

7. Comment fonctionne la distribution de fichiers sur *HDFS* ?
   A. Répartition en fonction de la taille des fichiers sur chaque nœud 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

8. 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* s'il devient indisponible.
   D. Administrer les transactions en autorisant ou non la lecture / écriture des fichiers.

9. L'outil suivant ne fait pas partie de l'écosystème de *Hadoop* :
   A. *MahoutOozie*
   B. *Yarn*
   C. *Mesos*
   D. *Pig*

10. Le principe de *« Rack-awarness »*dans *Hadoop* signifie que :
   A. Il sait combien de racks sont disponibles dans le cluster.
   B. Il est conscient du nombre de nœuds dans chaque rack.
   C. Il est conscient de la correspondance entre le nœud et le rack.
   D. Il sait quels nœuds de données ne sont pas disponibles dans le cluster.

11. Si l'adresse IP ou le nom d'hôte d'un *Datanode* change :
   A. Le*Namenode* n'a pas besoin de mettre à jour le *mapping* entre nom de fichier et nom de bloc.
   B. Le *namenode* met à jour le *mapping* entre nom de fichier et nom de bloc
   C. Les données de ce nœud de données sont définitivement perdues
   D. Aucune action n'est faite

12. La configuration High-availability de Hadoop implique :
   A. L'utilisation de plusieurs *Secondary NameNode*
   B. L'utilisation d'un ou plusieurs *StandBy NameNode*
   C. Deux *NameNode* au moins s'exécutent simultanément
   D. La résolution du problème du *Single Point Of Failure* de Hadoop.

13. Pour appliquer un *combiner*, quelle propriété doit être satisfaite par les valeurs émises par le *mapper*?
   A. La sortie du *mapper* et celle du *combiner* doivent être identiques.
   B. La sortie du *mapper* et celle du *reducer* doivent être identiques.
   C. La sortie du *mapper* et celle du *combiner* doivent avoir la même paire de valeurs de clé. Ce n'est possible que si les valeurs satisfont les propriétés associatives et commutatives.
   D. Le *combiner* peut toujours être appliqué à toutes les données.

14. La tolérance aux pannes dans Hadoop est assuré grâce à :
   A. La réplication des données
   B. L'utilisation du *StandBy NameNode*
   C. La décomposition des données en blocs.
   D. L'utilisation de plusieurs *racks* dans un même cluster.

15. La scalabilité horizontale :
   A. est l'augmentation des capacités des machines existantes.
   B. est l'ajout d'autres machines au cluster.
   C. est synonyme de *Scale up*.
   D. est l'approche adoptée par le Framework *Hadoop*.

Devoir Surveillés1 Semestre1 (15 - 20 Novembre 2021)