Institut Supérieur d'Informatique et de Mathématiques de Monastir

ISIMM

Devoir surveillé – S1 – 2024/2025

Filière : *Ing2_Info*	**Matière : Bases de données NOSQL et Big Data**		**Enseignante : Asma KERKENI**
Date : 22 / 11 / 2024	Nbr de Crédits : **3**	Coefficient : **3**	**Calculatrice : autorisée** **Documents autorisés : Non**
Durée de l'épreuve : 1h	Régime d'évaluation : **Mixte** EX (45%) + DS (22%) + TP (33%)		**Nombre de pages : 5**
Nom & Prénom :			**Matricule :** .....
Signature :	**Code confidentiel :**		**Classe :** ..... **N° Place :** .....

NOTE : Répondre directement sur les feuilles de l'examen /

Note ... / 20

Exercice 1 : (10 points)

Cet exercice nécessite des réponses concises et précises aux questions posées. Répondez uniquement dans l'espace prévu, en fournissant la réponse attendue. Évitez les explications ou définitions longues.

1. Que signifie le terme "véracité" dans le contexte du Big Data ?

.....

.....

2. Qu'est-ce qu'un Data Lake?

.....

.....

3. Quelle cause majeure a entraîné l'explosion de données à l'origine de l'émergence du Big Data ?

.....

.....

.....

4. Comment le NameNode détecte-t-il les nœuds en panne dans Hadoop ?

.....

.....

5. Quelle est la principale limitation qui freine la scalabilité de Hadoop dans sa version 1.

.....

.....

.....

6. Quel est le rôle du Job History Server dans Hadoop 2 ?

.....

.....

7. Quel démon gère les ressources sur un nœud worker dans YARN ?

.....

.....

8. Un fichier de 256 Go doit être stocké dans un cluster Hadoop avec une politique de réplication de 2. Si chaque DataNode dispose de 50 Go de capacité libre, combien de DataNodes sont nécessaires (taille de bloc de 128 Mo) pour stocker le fichier ?

.....

.....

.....

.....

9. Le taux de parallélisme est défini comme le ratio entre les tâches simultanées et le total de tâches. Un job MapReduce contient 20 tâches Map, et le cluster dispose de 5 nœuds, chacun pouvant exécuter 3 tâches simultanément. Quel est le taux de parallélisme pour ce job ?

.....

.....

.....

.....

10. Pour traiter 512 Go de données dans un cluster Hadoop avec une taille de bloc de 128 Mo et 10 slots par nœud, combien de tâches Map seront générées, et combien de nœuds sont nécessaires pour un parallélisme optimal ?

.....

.....

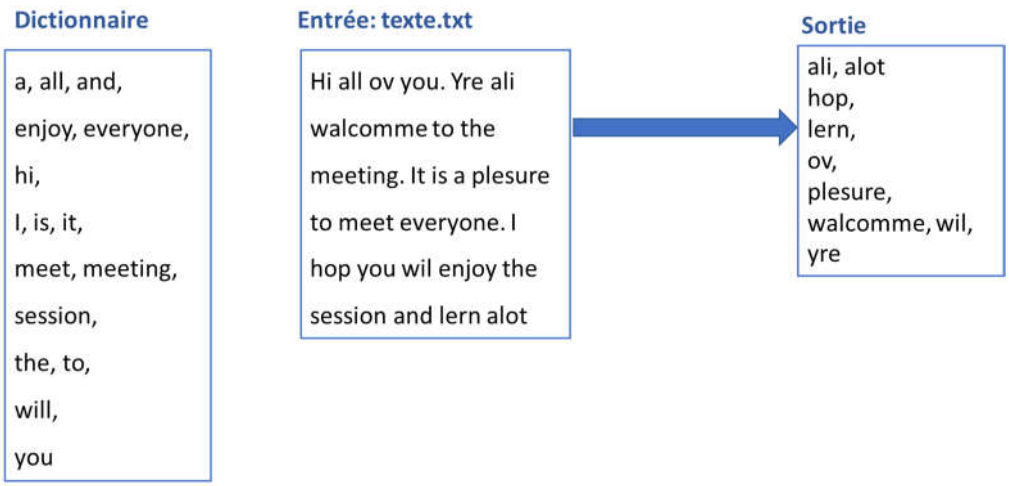
.....

.....

Exercice 2 : (10 points)

L'objectif de cet exercice est de développer un programme utilisant Hadoop MapReduce pour traiter un ensemble de fichiers contenant des textes écrits en anglais. Le but est d'extraire tous les mots qui ne figurent pas dans un dictionnaire de référence de la langue anglaise et de créer un nouveau dictionnaire, trié par ordre alphabétique, contenant uniquement les mots extraits.

Exemple :



1. Illustrer une solution utilisant le modèle de programmation *MapReduce* sur l'exemple.

2. Erire les pseudocodes des fonctions *Map* et *Reduce* correspondantes à ce problème.