

Ecole Normale Supérieure de l'Enseignement Technique
Département Mathématique et informatique

Examen du 2^{ème} semestre 2022/2023

La note :..... Date : 02/06/2023 Module : BIG DATA Durée : 4h Nom &

Prénom : Achaghour Zakaria Filière IIBDCC

Exercice 1: Manipuler le système de fichiers HDFS

Tapez les commandes pour répondre aux questions suivantes :

1. Vérifiez la version Hadoop.

hadoop version

2. Démarrez HDFS et vérifiez qu'il est en cours d'exécution.

start-dfs.sh

start-yarn.sh

jps

3. Créez deux nouveaux répertoires nommés **/enset/bddc** et **/enset/glsid** sur HDFS.

hdfs dfs - mkdir -p /enset/bddc /enset/glsid

4. Créez un nouveau fichier **java.txt** contenant 10 lignes et **cpp.txt** contenant 10 lignes sur votre système local.

5. Charger le fichier **java.txt** dans **/enset/bddc** et **cpp.txt** dans **/enset/glsid** sur HDFS.

hdfs dfs -moveFromLocal ./java.txt /enset/bddc

hdfs dfs -moveFromLocal ./cpp.txt /enset/glsid

6. Afficher le contenu du répertoire **/enset/bddc** et **/enset/glsid**.

hdfs dfs -ls /enset/bddc

hdfs dfs -ls /enset/glsid

7. Affichez le contenu du fichier **java.txt** qui se trouve dans HDFS.

```
hdfs dfs -cat /enset/bddc/java.txt
```

8. Déterminez la taille du fichier **cpp.txt** qui se trouve dans HDFS.

```
hdfs dfs -du -h /enset/glsid/cpp.txt
```

9. Déplacez le fichier **cpp.txt** vers **/enset/bddc** et vérifiez si le fichier est bien déplacé.

```
hdfs dfs -mv /enset/glsid/cpp.txt /enset/bddc/
```

```
hdfs dfs -ls /enset/bddc
```

10. Supprimez les fichiers **java.txt** et **cpp.txt** dans HDFS.

```
hdfs dfs -rm /enset/bddc/java.txt /enset/bddc/cpp.txt
```

Exercice 2 :

On souhaite traiter des données des vols d'une société aérienne au moyen d'une application Spark d'une manière parallèle est distribuée. L'entreprise possède des données stockées dans une base de données relationnel et des fichiers CSV. L'objectif est de traiter ces données en utilisant Spark SQL et SPARK Structured Streaming à travers les APIs DataFrame et Dataset pour extraire des informations utiles afin de prendre des décisions.

Partie 1 : Spark SQL

La société possède une application web pour gérer les réservations des vols, les données sont stockées dans une base de données MYSQL nommée **DB_AEROPORT**, qui contient trois tables **VOLS** et **PASSAGERS** et **RESERVATIONS** (Voir les figures 1, 2 et 3).


	#	Nom	Type	Interclassement	Attributs	Valeur par		Commentaires	Extra
						Null	défaut		
<input type="checkbox"/>	1	ID 	int(11)			Non	Aucun(e)		AUTO_INCREMENT
<input type="checkbox"/>	2	DATE_DEPART	date			Non	Aucun(e)		
<input type="checkbox"/>	3	DATE_ARRIVE	date			Non	Aucun(e)		

Figure 1: Table Vols

	#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra
<input type="checkbox"/>	1	ID 	int(11)			Non	Aucun(e)		AUTO_INCREMENT
<input type="checkbox"/>	2	NOM	varchar(30)	latin1_swedish_ci		Non	Aucun(e)		
<input type="checkbox"/>	3	PRENOM	varchar(30)	latin1_swedish_ci		Non	Aucun(e)		
<input type="checkbox"/>	4	TEL	varchar(30)	latin1_swedish_ci		Non	Aucun(e)		

Figure 2: Table Passagers




	#	Nom	Type	Interclassement	Attributs	Null	Valeur par défaut	Commentaires	Extra
<input type="checkbox"/>	1	ID 	int(11)			Non	Aucun(e)		AUTO_INCREMENT
<input type="checkbox"/>	2	DATE_RESERVATION	int(11)			Non	Aucun(e)		
<input type="checkbox"/>	3	ID_VOL 	int(11)			Non	Aucun(e)		
<input type="checkbox"/>	4	ID_PASSAGER 	int(11)			Non	Aucun(e)		

Figure 3: Table RESERVATIONS

Travail à faire :

Vous créez la base de données et les tables et vous répondez aux questions suivantes :

1. Afficher pour chaque vol, le nombre de passagers selon le format d’affichage suivant :

ID_VOL |DATE DEPART| NOMBRE

2. Afficher la liste des vols en cours selon le format d’affichage suivant :

ID_VOL |DATE DEPART| DATE ARRIVE

Partie 2 : Importer et exporter des données avec SQOOP

On souhaite à travers cet exercice d’importer et exporter des données entre une base de données sur MySQL et HDFS.

- On considère la base de données DB_AEROPORT dans MySQL contenant une table **VOLS**.
- Importez les données de la table VOLS dans HDFS en utilisant SQOOP.
- Créez un fichier nommé **vols.txt**, ajouter 3 vols, puis charger le fichier dans HDFS puis l’exportez vers la table VOLS avec scoop.

Partie 3: Traitement de données en streaming

La société reçoit d'une manière continue des fichiers CSV qui contiennent les incidents dans les avions, les fichiers sont stockés directement sur HDFS.

Le format de données dans les fichiers csv est la suivante :

id, description, no_avion, date

Travail à faire :

1. Afficher d'une manière continue l'avion ayant plus d'incidents.
2. Afficher d'une manière continue les deux mois de l'année en cours où il y avait moins d'incidents.