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

**Examen du 2ème semestre 2022/2023**

|  |  |  |
| --- | --- | --- |
| **La note :………….** |  | **Date : 02/06/2023** |
| **Module : BIG DATA** |  | **Durée : 4h** |
| **Nom & Prénom : Meggouri Ismail** |  | **Filière : BDCC** |

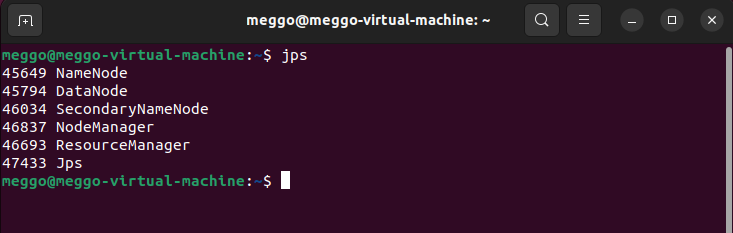
**Exercice 1: Manipuler le système de fichiers HDFS**  Tapez les commandes pour répondre aux questions suivantes :

1. Vérifiez la version Hadoop. :

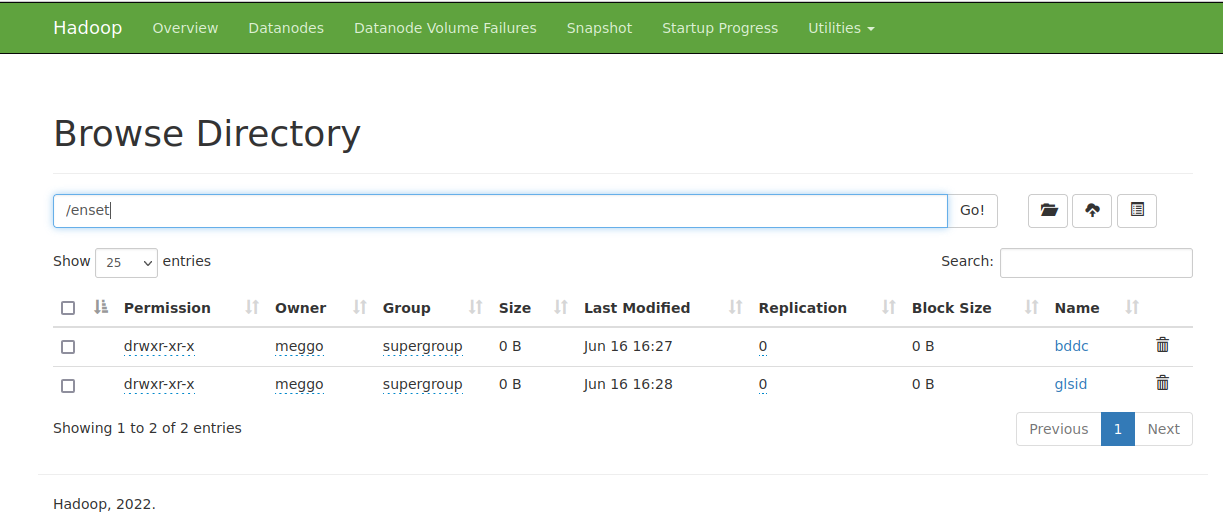
A screenshot of a computer

Description automatically generated

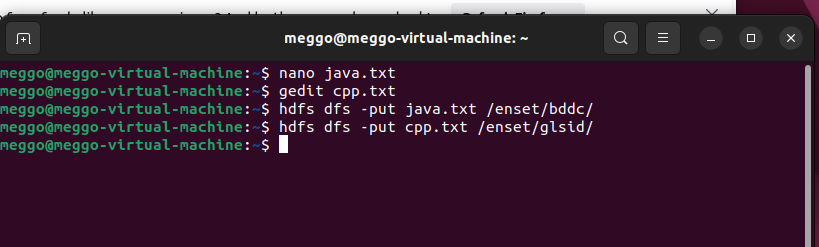
1. Démarrez HDFS et vérifiez qu’il est en cours d’exécution.



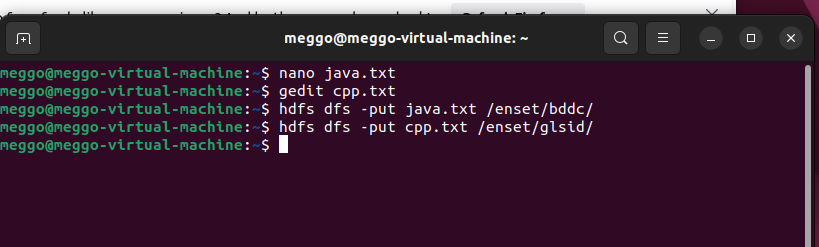
1. Créez deux nouveaux répertoires nommés **/enset/bddc** et **/enset/glsid** sur HDFS.



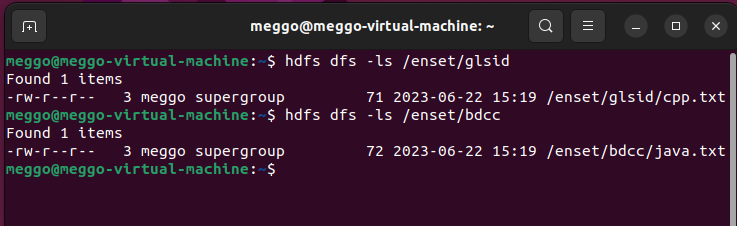
1. Créez un nouveau fichier **java.txt** contenant 10 lignes et **cpp.txt** contenant 10 lignes sur votre système local.



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



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

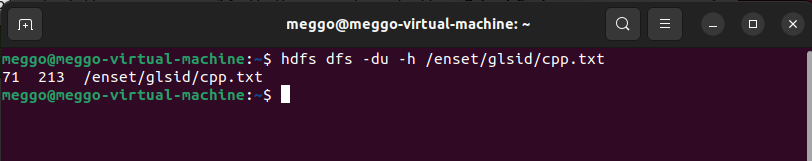


1. Affichez le contenu du fichier **java.txt** qui se trouve dans HDFS.

A screenshot of a computer

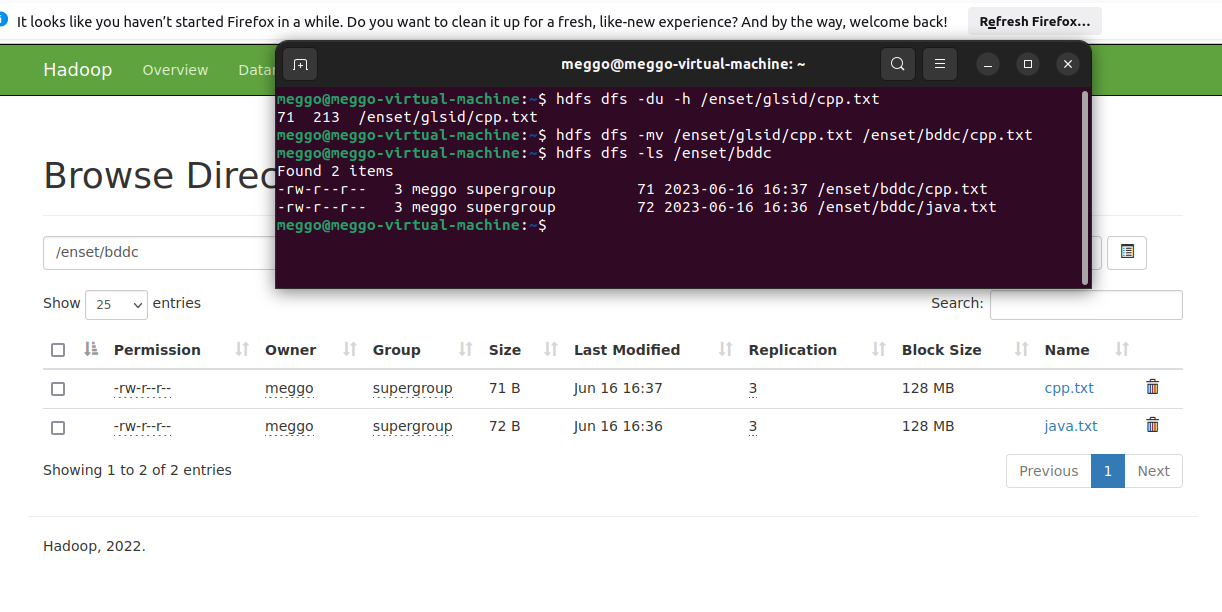
Description automatically generated

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

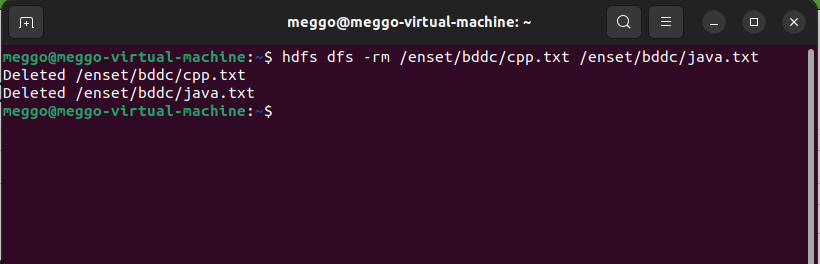
A picture containing text, screenshot, software

Description automatically generated

1. Déplacez le fichier **cpp.txt** vers **/enset/bddc** et vérifiersi le fichier est bien déplacé.



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



**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).

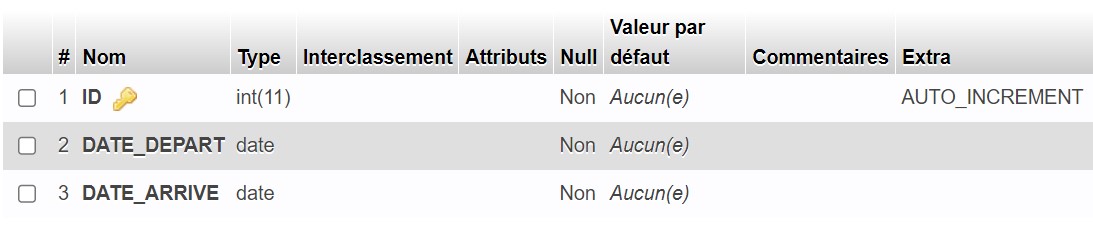
*Fig*

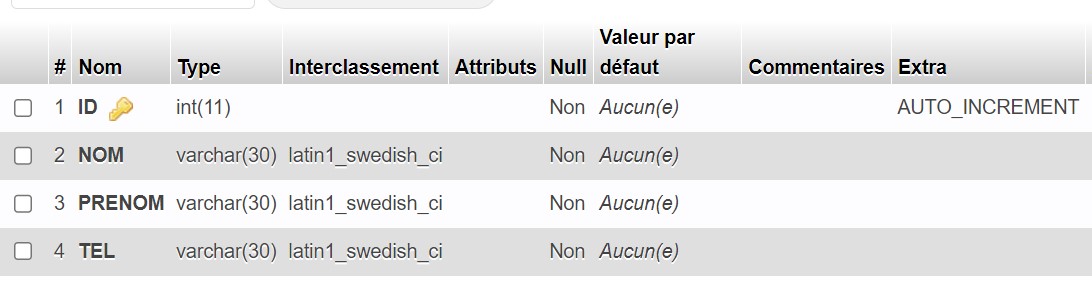
*ure*

*1*

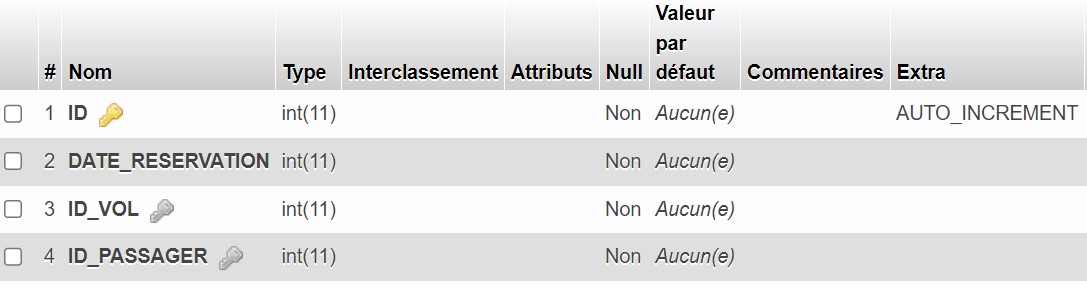
*:*

*Table Vols*





*Figure 2: Table Passagers*

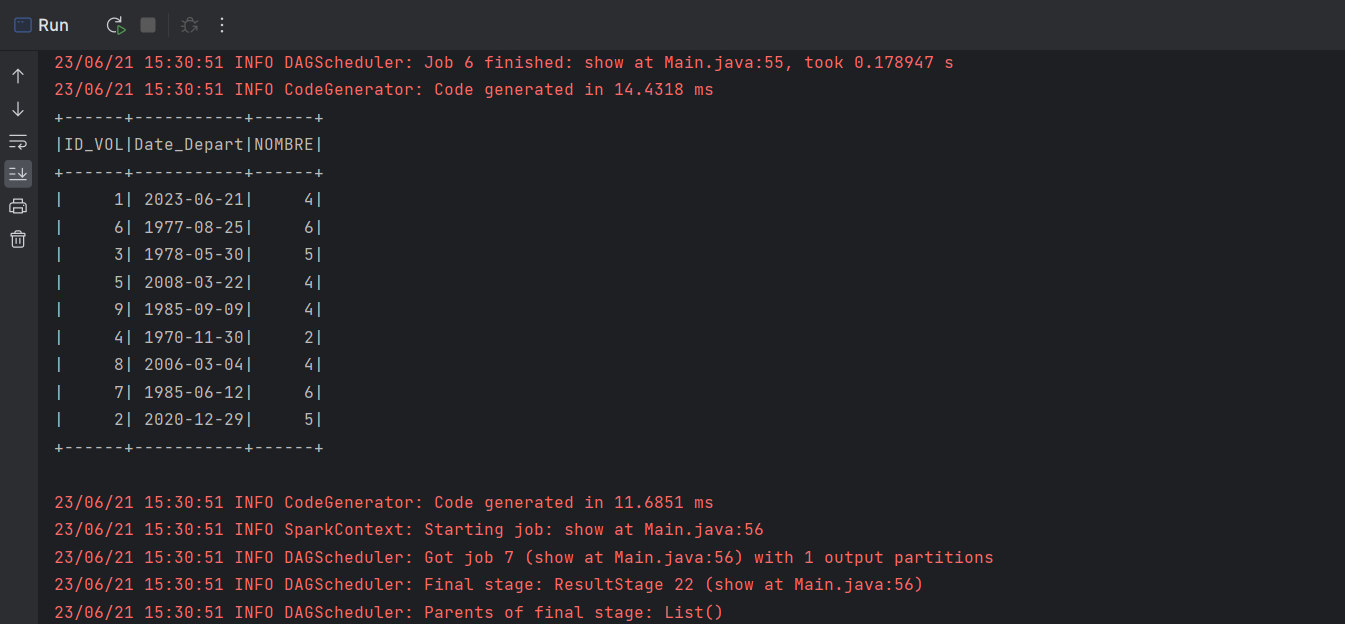
  *Figure 3: Table RESERVATIONS*

**Travail à faire :**

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

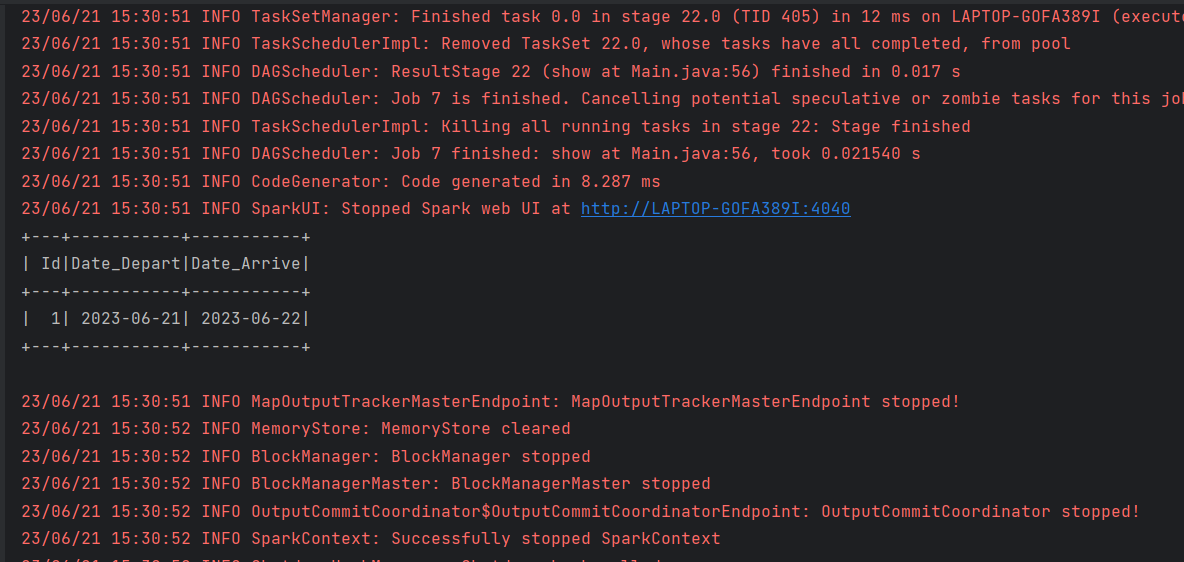
1. Afficher pour charque vol, le nombre de passagers selon le format d’affichage suivant :

ID\_VOL |DATE DEPART| NOMBRE



1. 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.

**sqoop import --connect "jdbc:mysql://localhost:3306/db\_aer" --username "root" --password "" --table vols --target-dir /vols**

* Créez un fichier nommé **vols.txt**, ajouter 3 vols, puis charger le fichier dans HDFS puis l’exportez vers la table VOLS avec scoop.

**sqoop export --connect "jdbc:mysql://localhost:3306/db\_aer" --username "root" --password "" --table vols --export-dir /enset/vols.txt --input-fields-terminated-by ',' --input-lines-terminated-by '\n'**

# Partie 3: Traitement de données en streaming

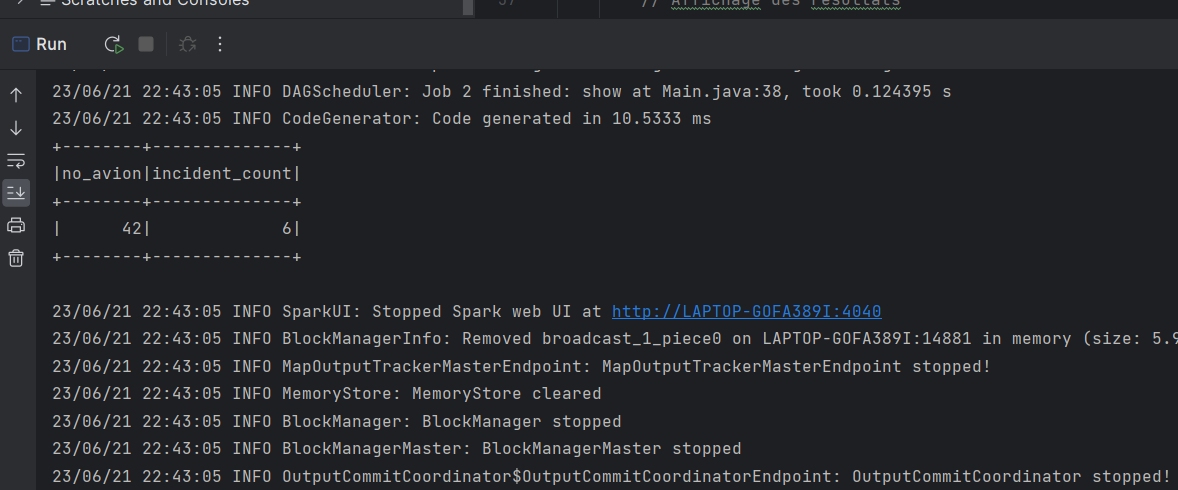
La société reçoit d’une manière contenu des fichiers CSV qui contient les incidents dans les avions, les fichiers sont stockés directement sur HDFS.

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

id, description, no\_avion, date

**Travail à faire :**

1. Afficher d’une manière continue l’avion ayant plus d’incidents.



1. Afficher d’une manière continue les deux mois de l’année en cours où il a y avait moins d’incidents.

A screen shot of a computer

Description automatically generated with medium confidence