COLLÈGE ROSEMONT

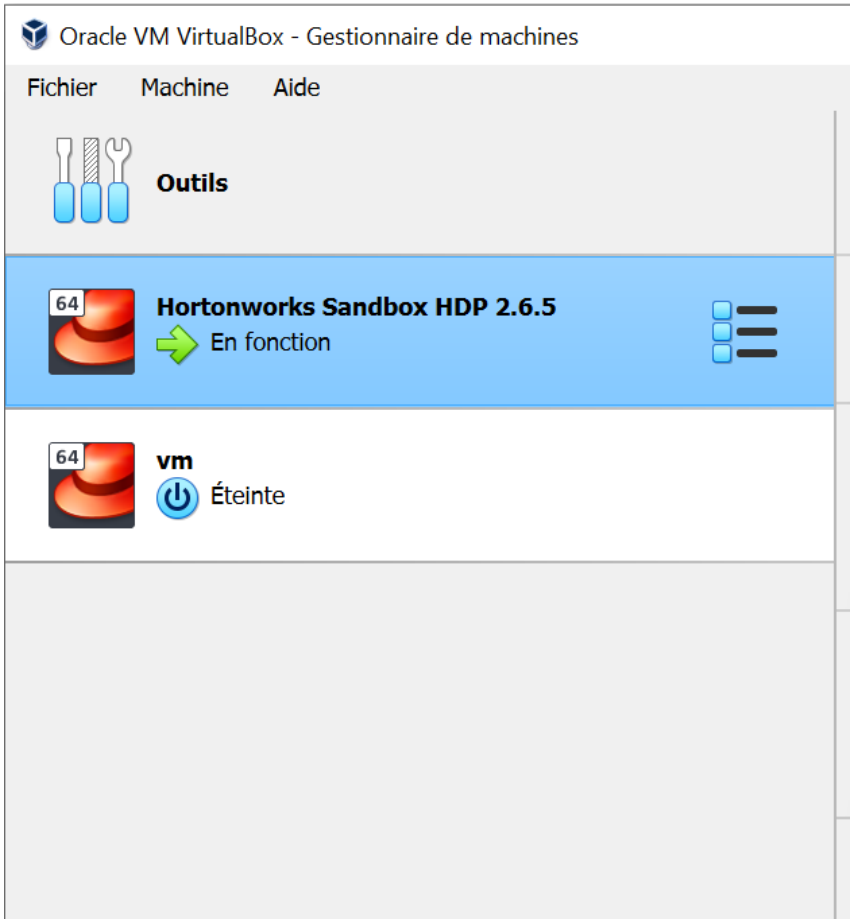
Valorisation de données

Par : Abderrazak Sahraoui

Sommaire

- Machine Virtuelle Hortonworks HDP 2.6.5
- Connexion à Ambari sur serveur local
- File View
- Hive View
- Requête Hive
- Tables et Bases de données sur Hive

Machine virtuelle avec HortonWorks HDP 2.6.5

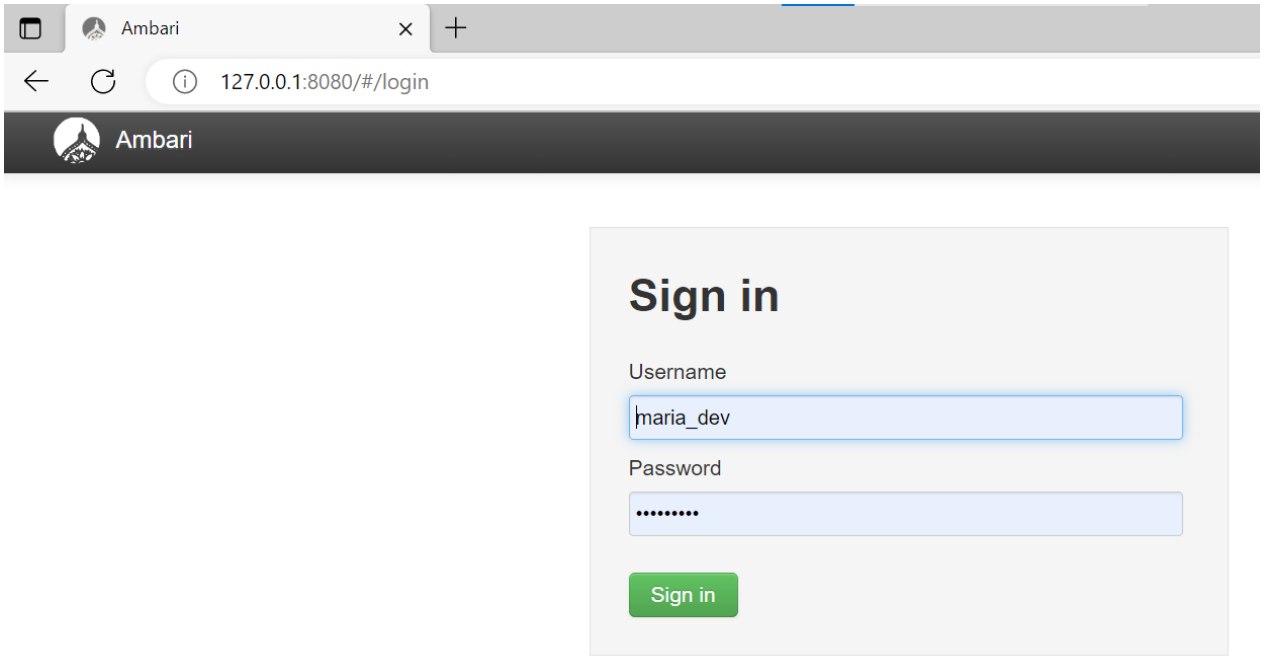


- Après installation de **Oracle VM VirtualBox,** télécharger et installer la plateforme **Hortonworks sandbox HDP 2.6.5.** sur la **VM VirtualBox.**

- Lancer **Hortonworks sandbox HDP 2.6.5. sur VirtualBox.**

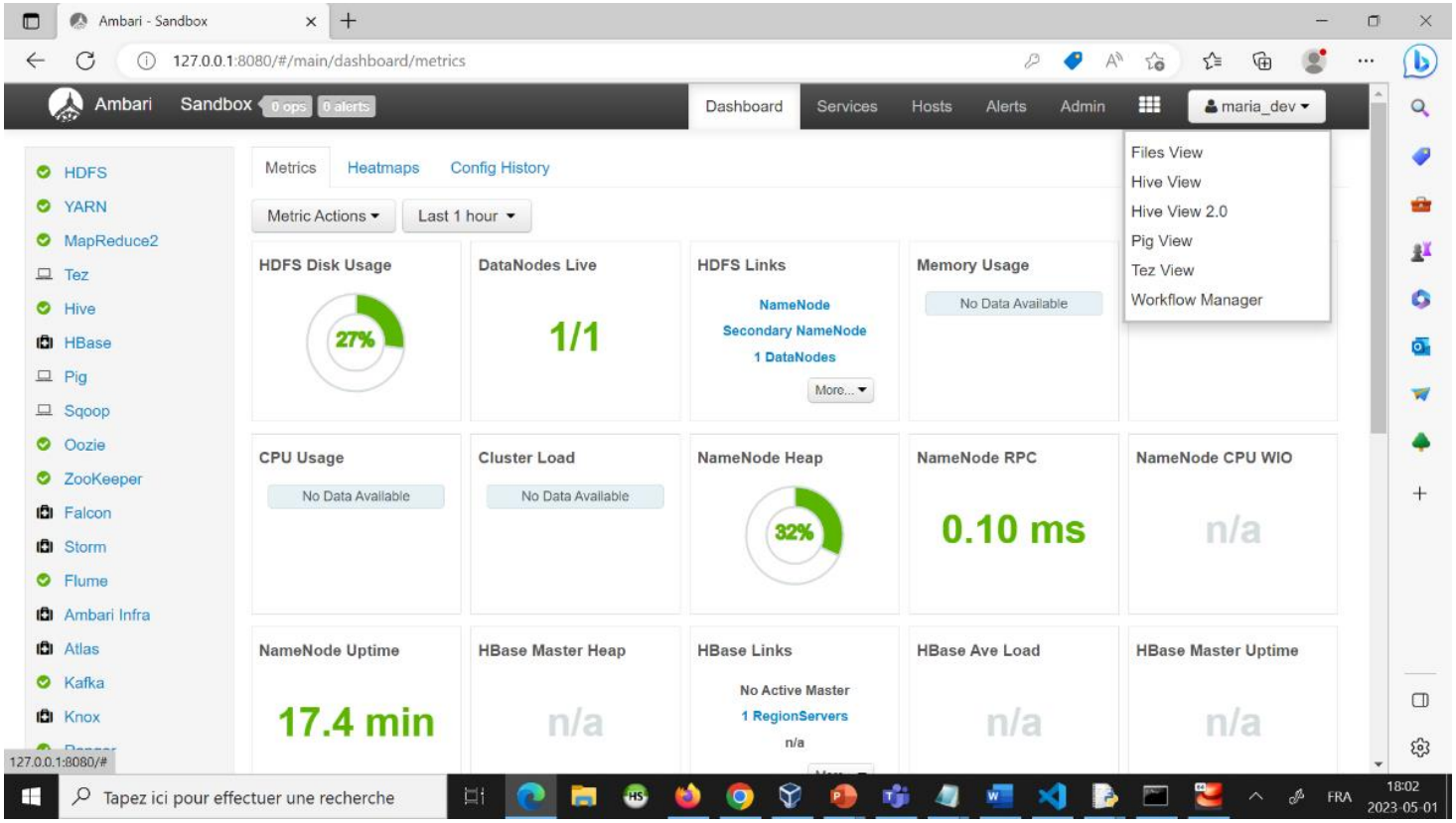
- Aller sur le serveur web local **127.0.0.1:8080**

Connection à Ambari par le serveur local



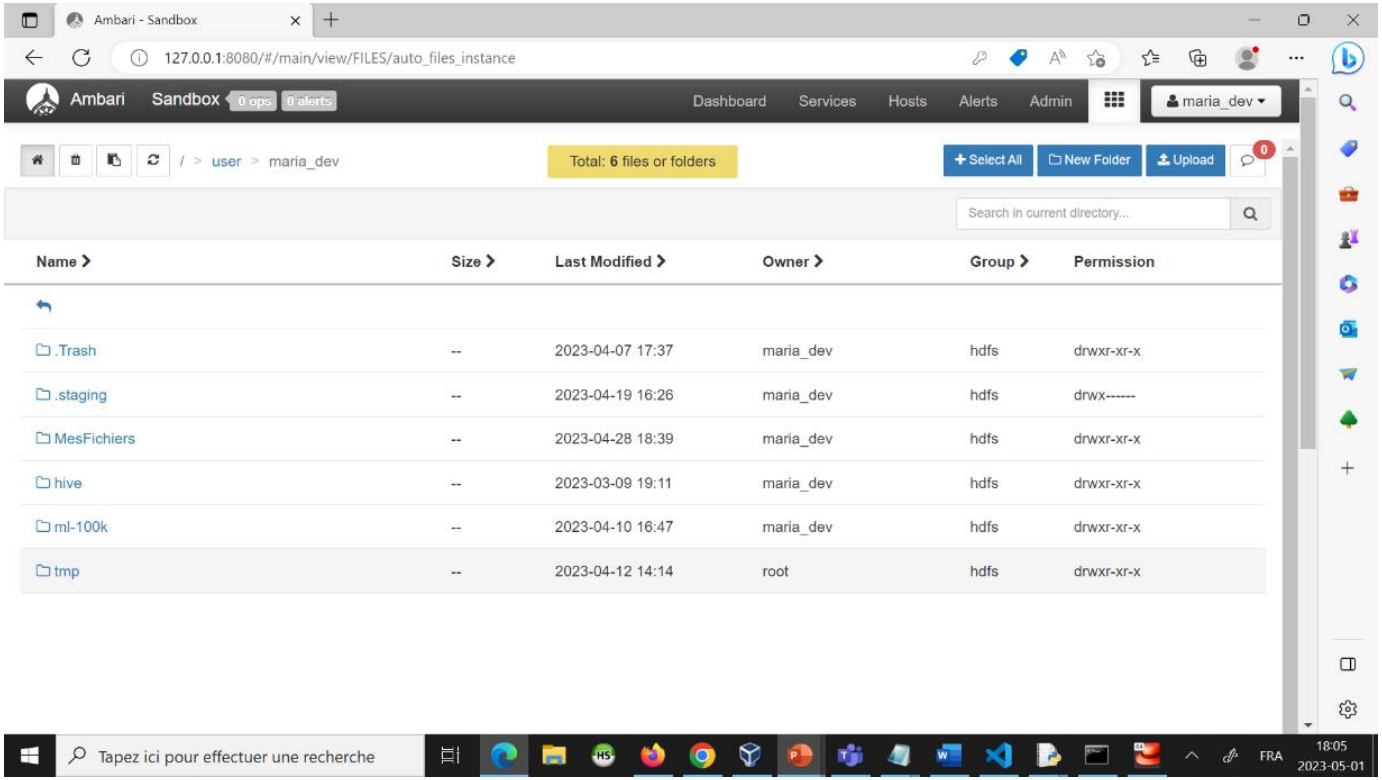
Se connecter à Ambari avec l'usager **maria_dev** (password **maria_dev**)

Ambari



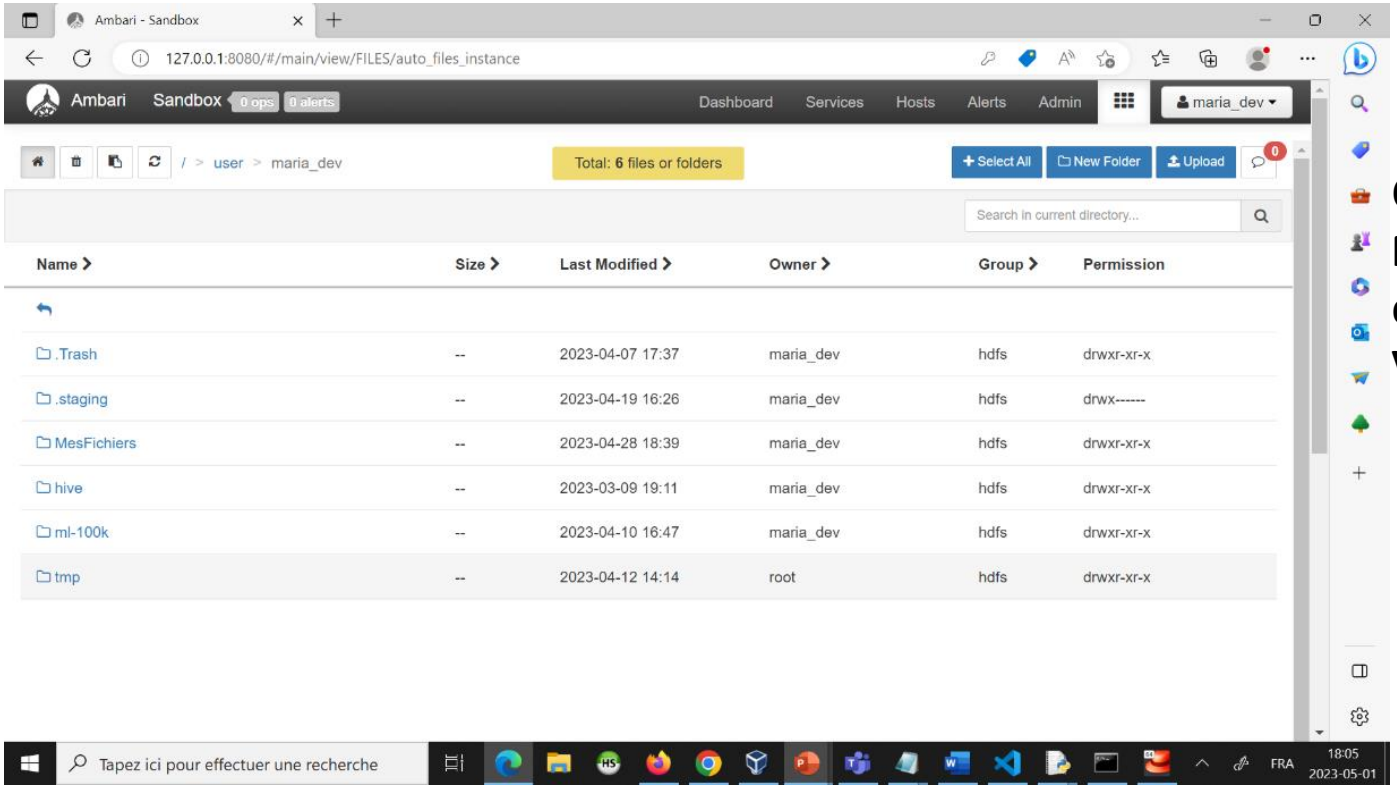
Lancer
Files View

Gestion de fichiers sur HDFS



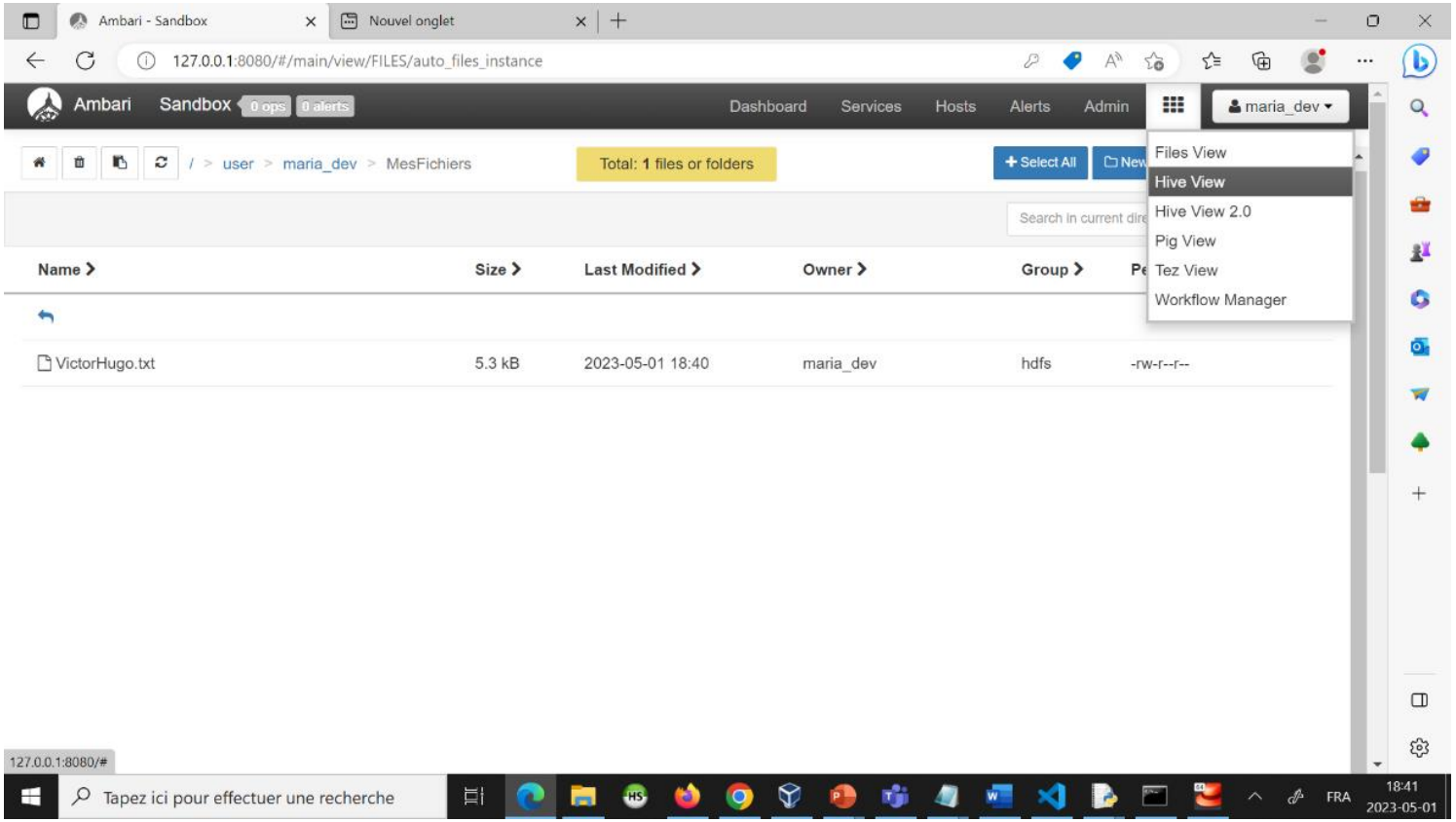
Ouvrer **user**
puis **maria_dev**

Gestion de fichiers sur HDFS



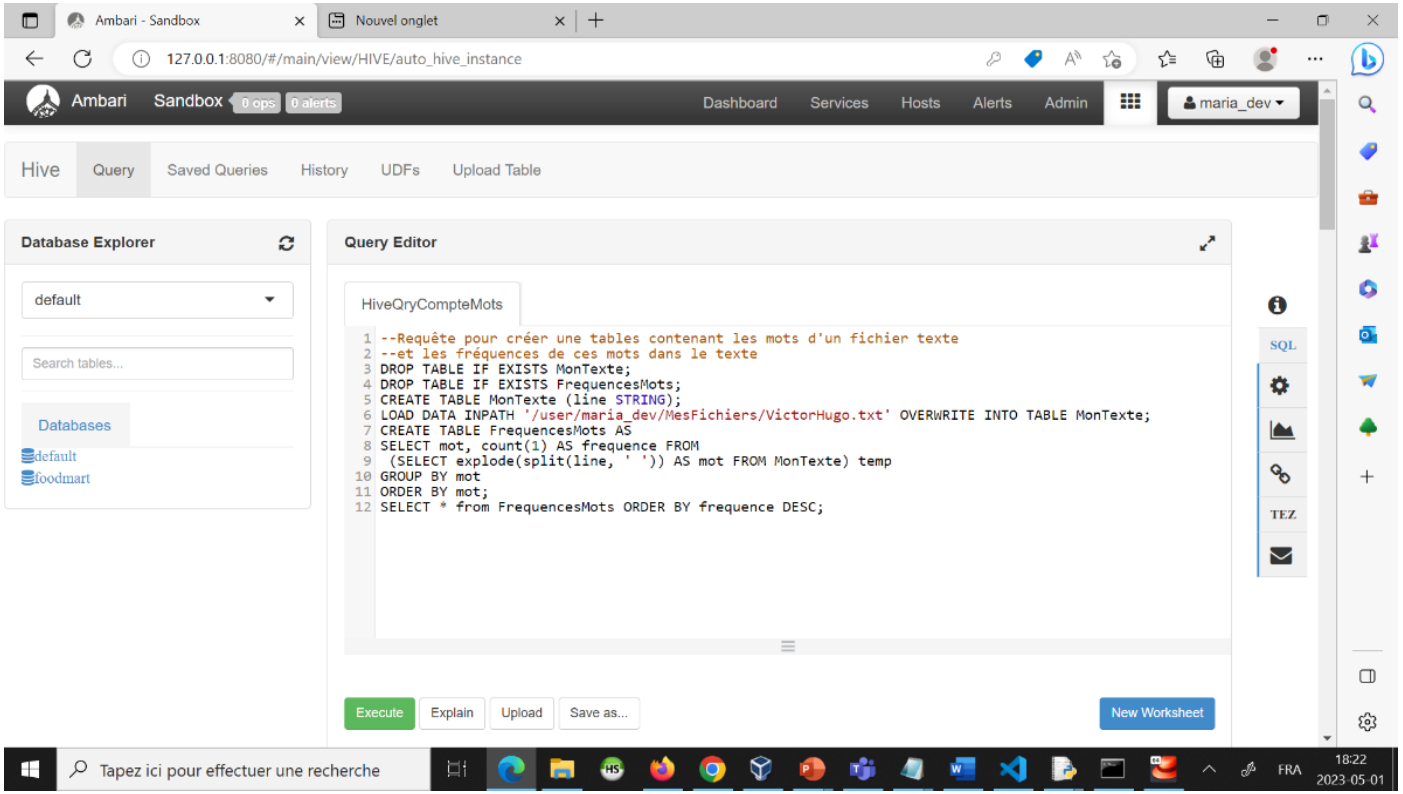
Créer le dossier **MesFichiers** et charger **VictorHugo.txt**

Hive View



Lancer
Hive View

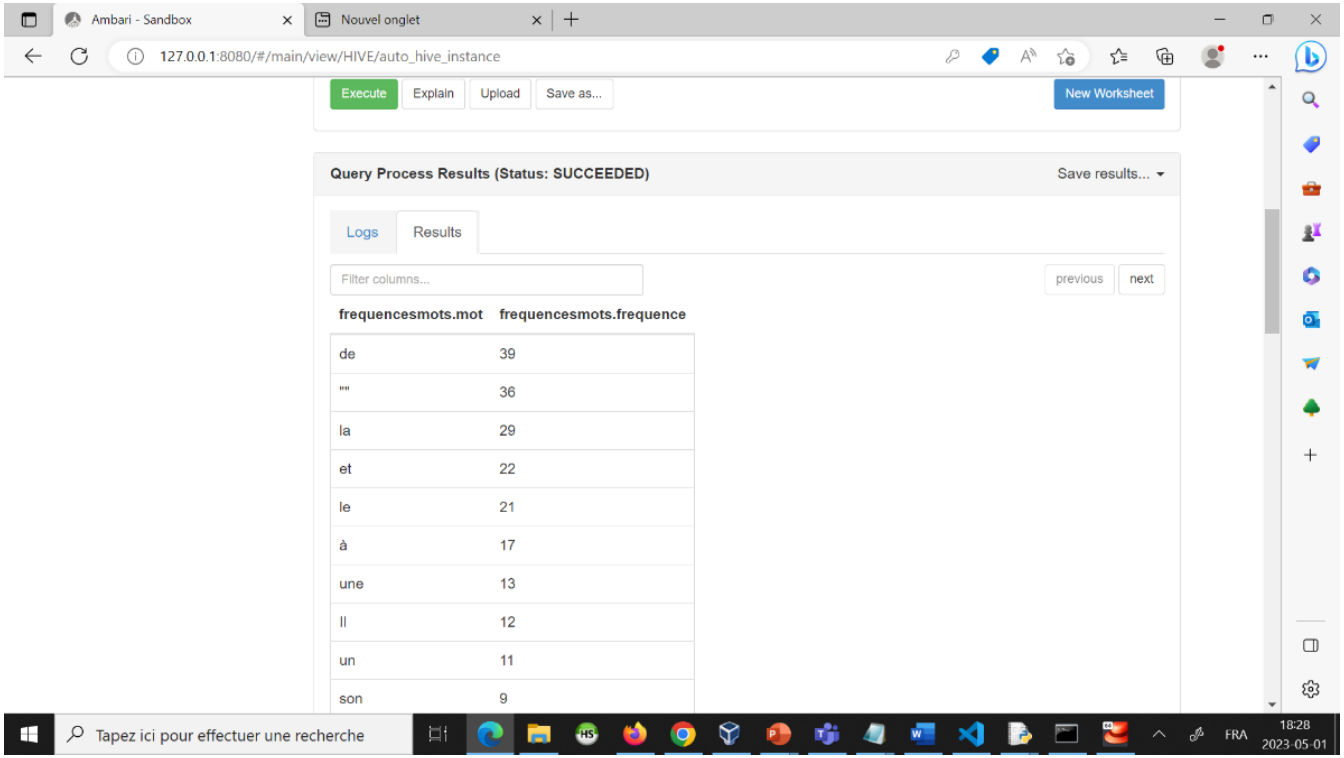
Requête Hive



Taper la **requête** ci-contre et l'enregistrer sous **HiveQryCompteMots**.

Exécuter la requête et observer le résultat.

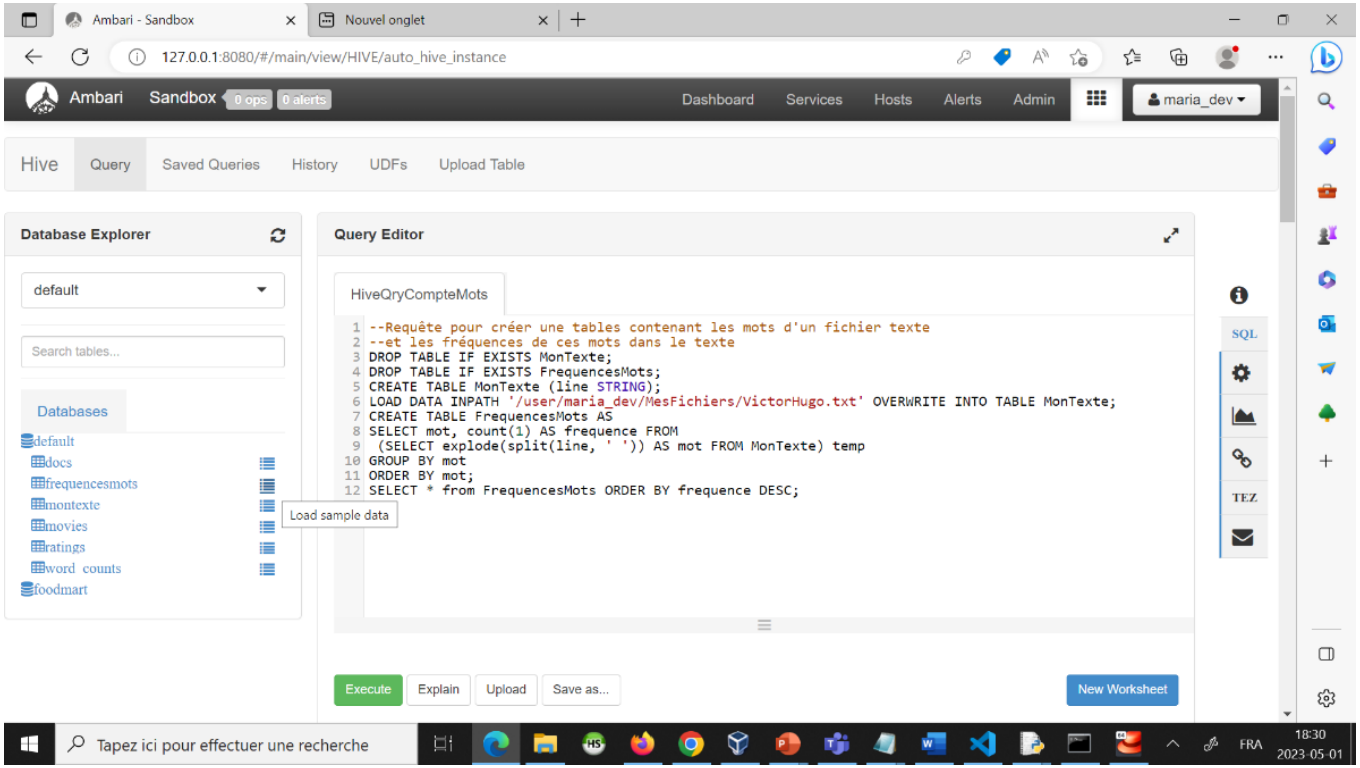
Résultats Requête



frequencesmots.mot	frequencesmots.frequence
de	39
""	36
la	29
et	22
le	21
à	17
une	13
Il	12
un	11
son	9

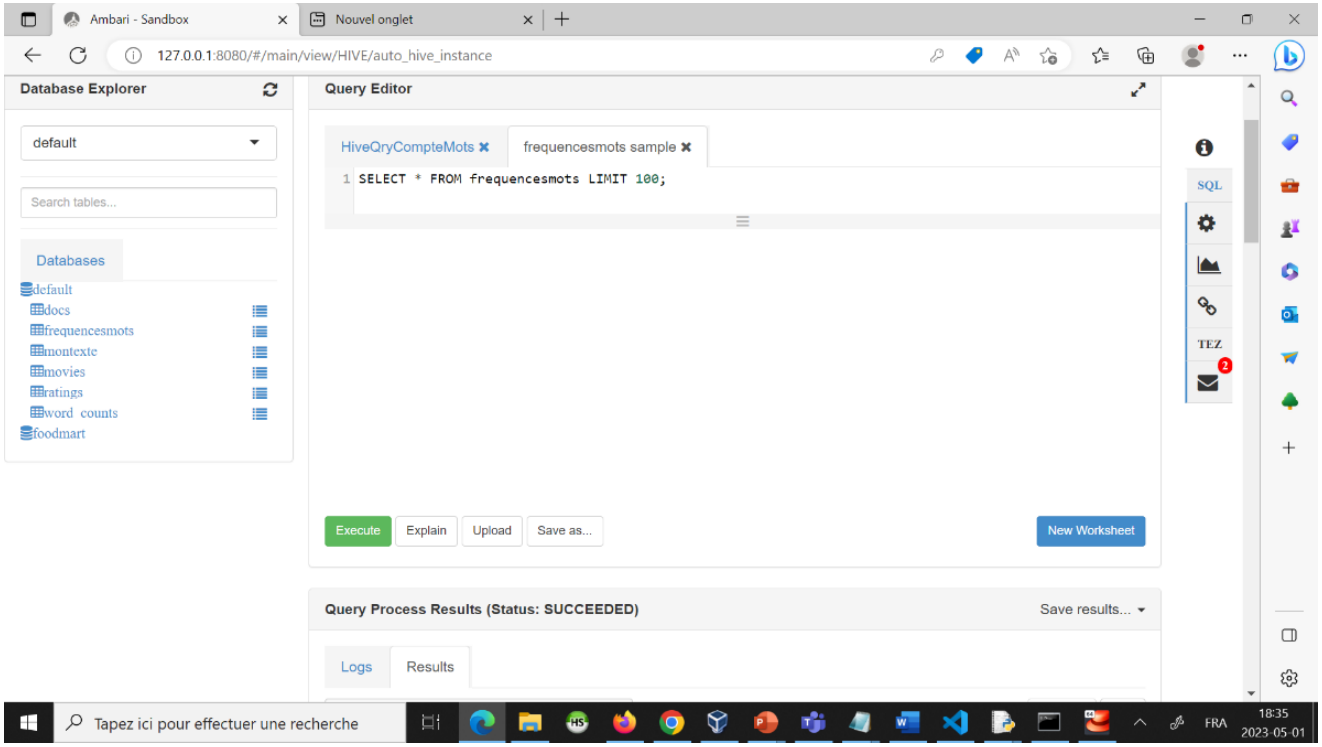
Fréquences des mots de VictorHugo.**txt triées par ordred décroissant** des fréquences

Table frequencemots



Dérouler le contenu de la base de données **default** et **charger** la table **frequencemots**

Table frequencemots



Exécuter la requête de l'affichage de la table **frequencemots** et **observer** son contenu**.**

Lien sur Cloudera

Getting Started with HDP Sandbox

OVERVIEW

1. Concepts
2. Loading Sensor Data into HDFS
3. Hive - Data ETL
4. Spark - Risk Factor
5. Data Reporting With Zeppelin

Outline

- Apache Hive Basics
- Become Familiar with Data Analytics Studio
- Create Hive Tables
- Explore Hive Settings on Ambari Dashboard
- Analyze the Trucks Data
- Summary
- Further Reading

Apache Hive Basics

Apache Hive provides SQL interface to query data stored in various databases and files systems that integrate with Hadoop. Hive enables analysts familiar with SQL to run queries on large volumes of data. Hive has three main functions: data summarization, query and analysis. Hive provides tools that enable easy data extraction, transformation and loading (ETL).

Become Familiar with Data Analytics Studio

Apache Hive presents a relational view of data in HDFS. Hive can represent data in a tabular format managed by Hive or just stored in HDFS irrespective in the file format the data is in. Hive can query data from RCFile format, text files, ORC, JSON, parquet, sequence files and many of other formats in a tabular view. Through the use of SQL you can view your data as a table

https://www.cloudera.com/tutorials/getting-started-with-hdp-sandbox/3.html