COLLÈGE ROSEMONT

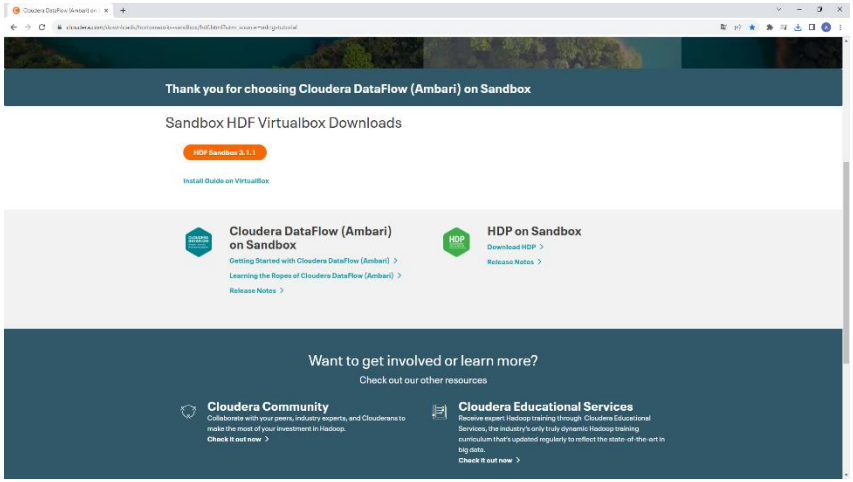
Valorisation de données

Par : Abderrazak Sahraoui

Sommaire

- Machine Virtuelle Hortonworks HDP 2.6.5
- Connexion à Ambari sur serveur local
- Connexion au web shell
- Hive 2.0 View
- Requêtes Hive

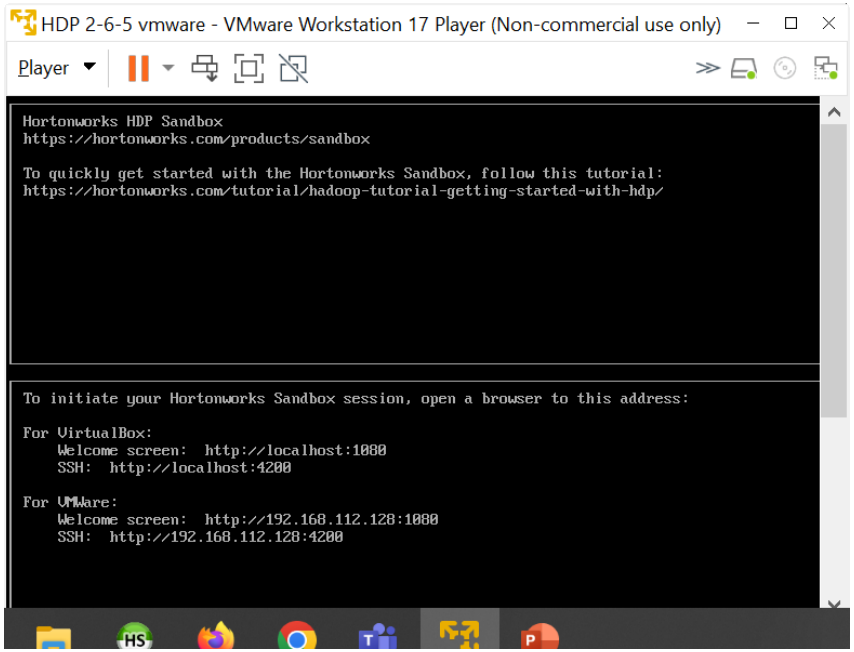
Machine virtuelle HortonWorks HDP 2.6.5



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

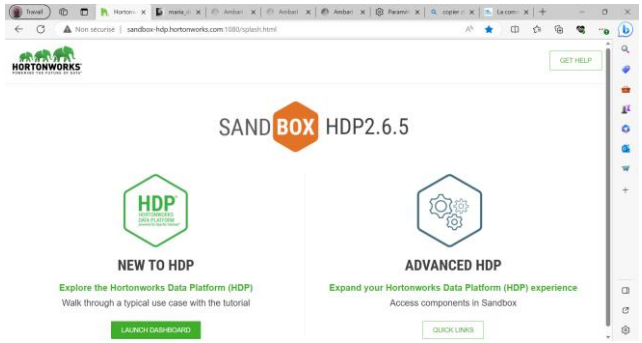
https://www.cloudera.com/downloads/hortonworks-sandbox/hdf.html?utm_source=mktg-tutorialm

Machine virtuelle avec HortonWorks HDP 2.6.5



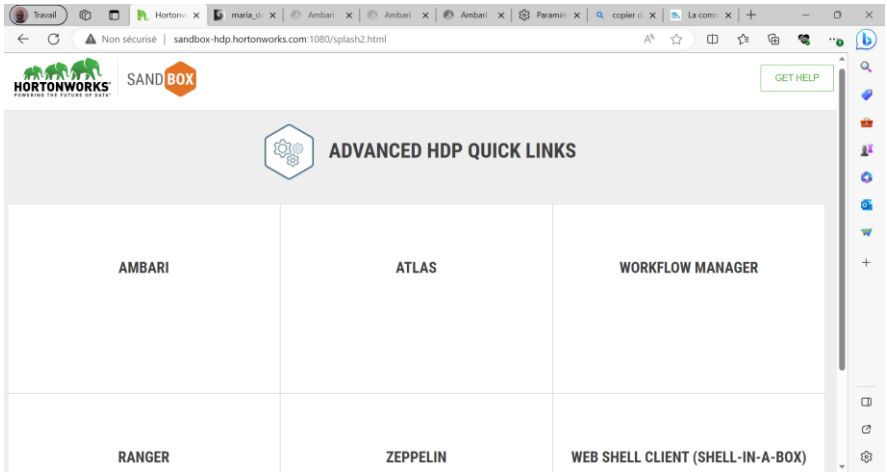
- Aller sur le serveur web dont l'ip a été généré par VMWare.

Connection aux pages web du serveur



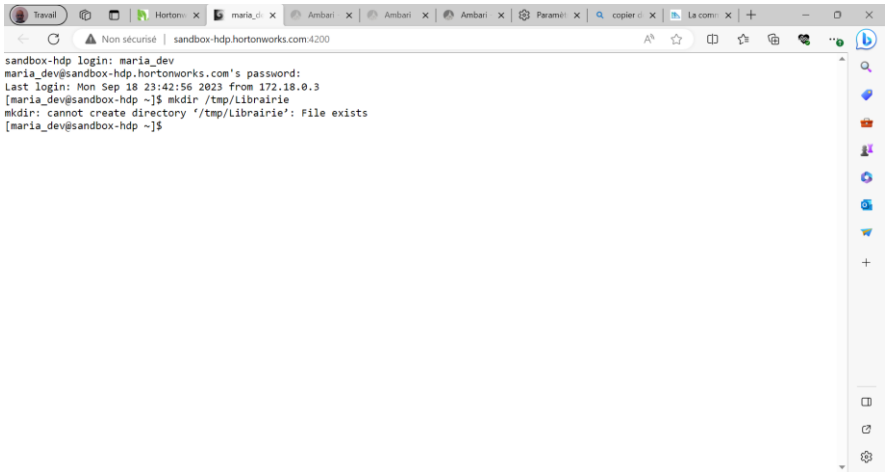
- Se connecter à http://sandbox-hdp.hortonworks.com:1080/
- Choisir l'option Advanced HDP

Connection à Ambari / shell



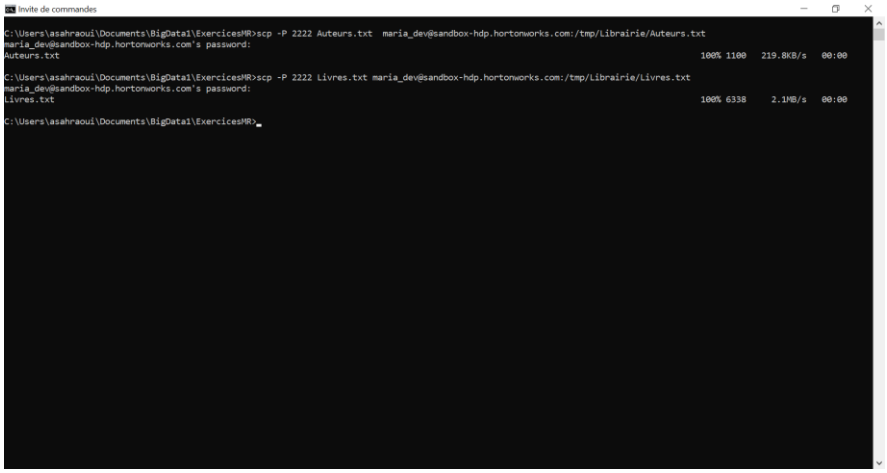
- Lancer le web shell

Connection au shell



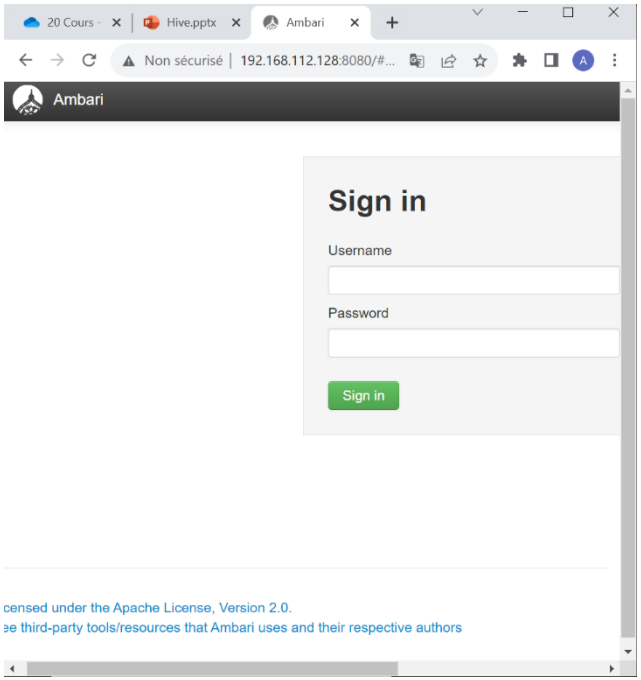
- Créer un dossier Librairie dans le dossier /tmp

Copie de fichiers vers le serveur



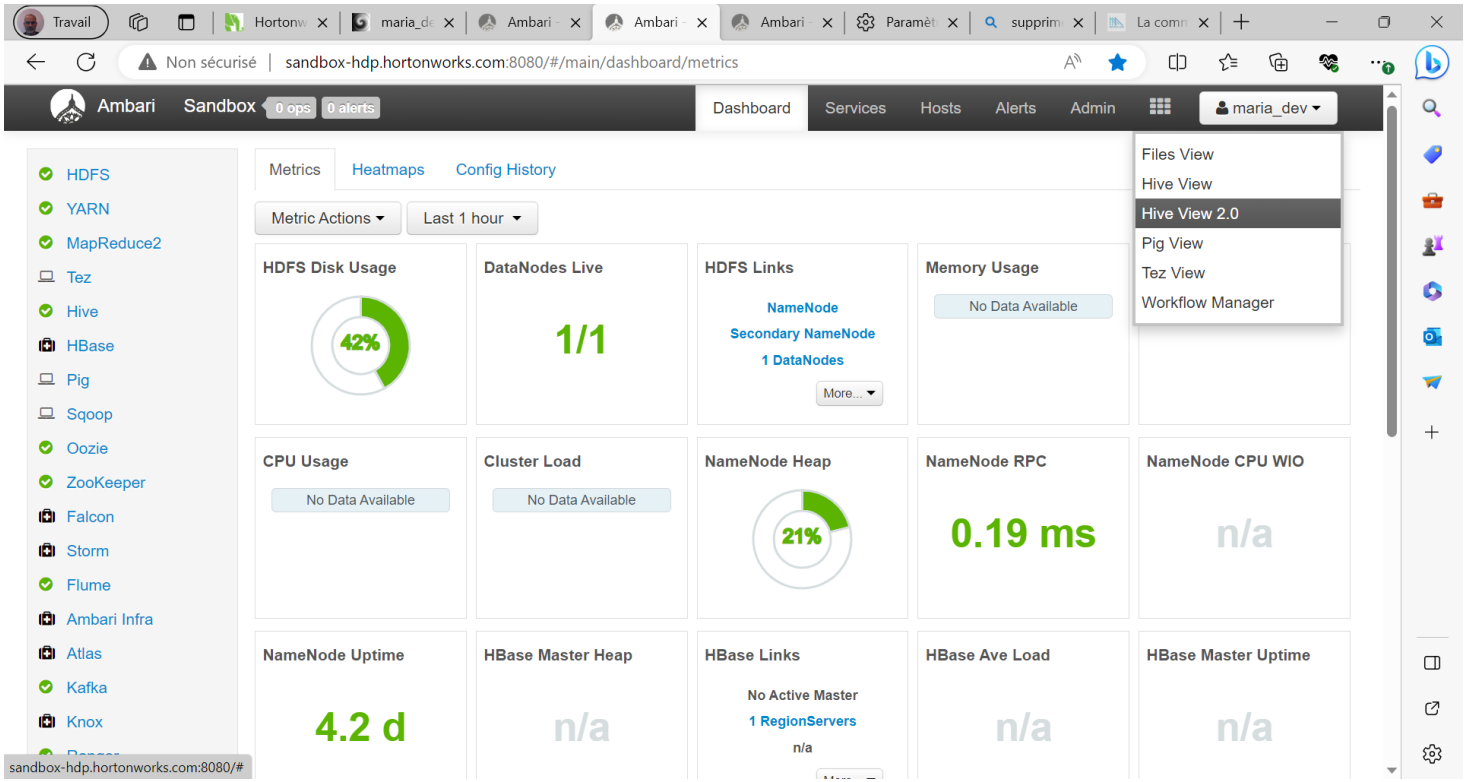
- Copier les fichiers Auteurs.txt et Livres.txt de Windows vers le serveur sandbox-hdp.hortonworks.com dans le dossier /tmp/Librairie.

Connection à Ambari



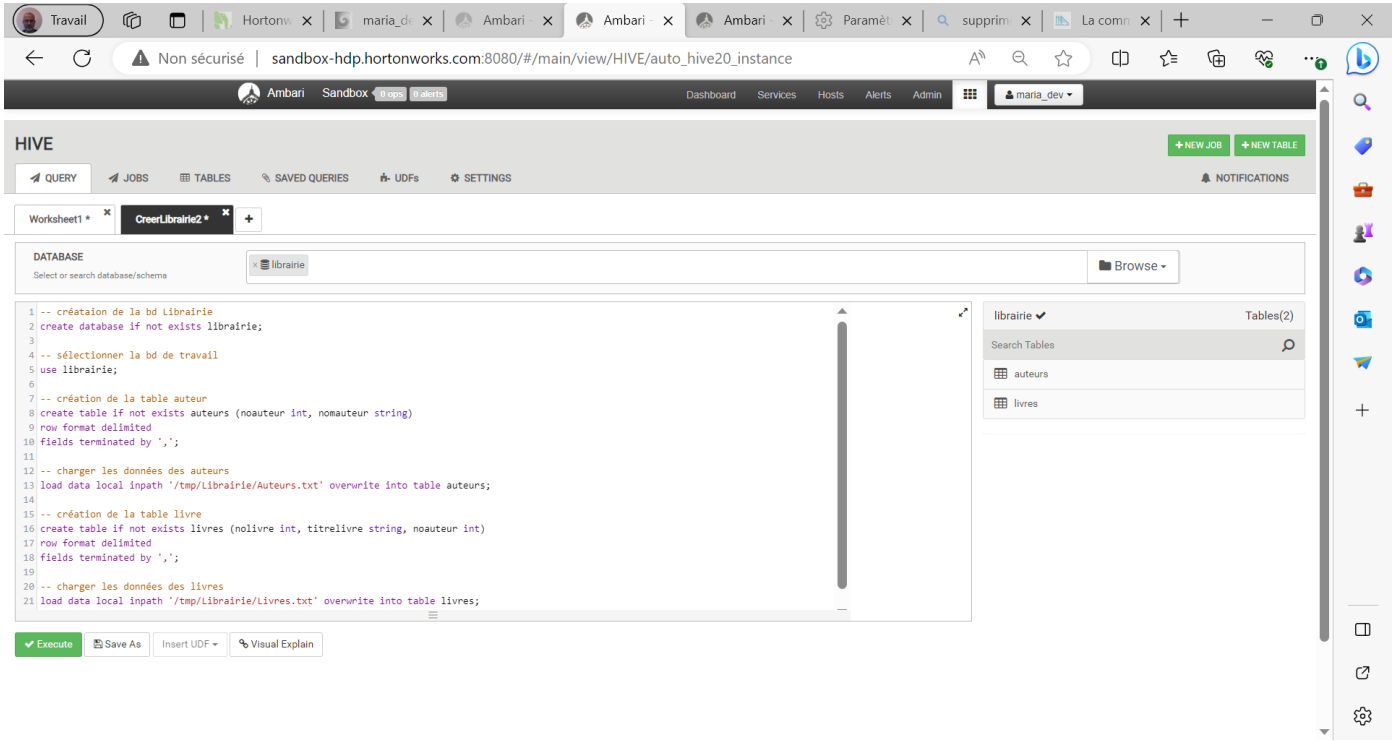
- Se connecter à Ambari avec l'usager **maria_dev** (password **maria_dev**)

Hive View 2.0



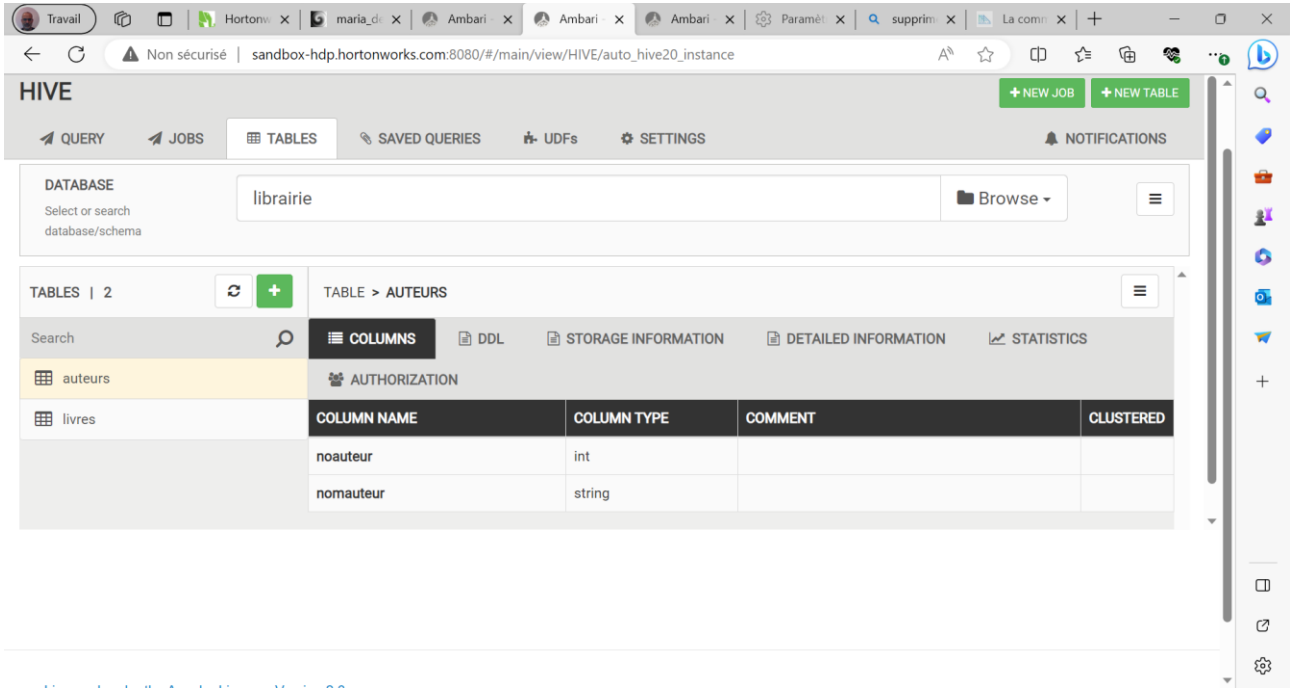
Lancer **Hive View 2.0**

Requête (DDL) pour créer une base



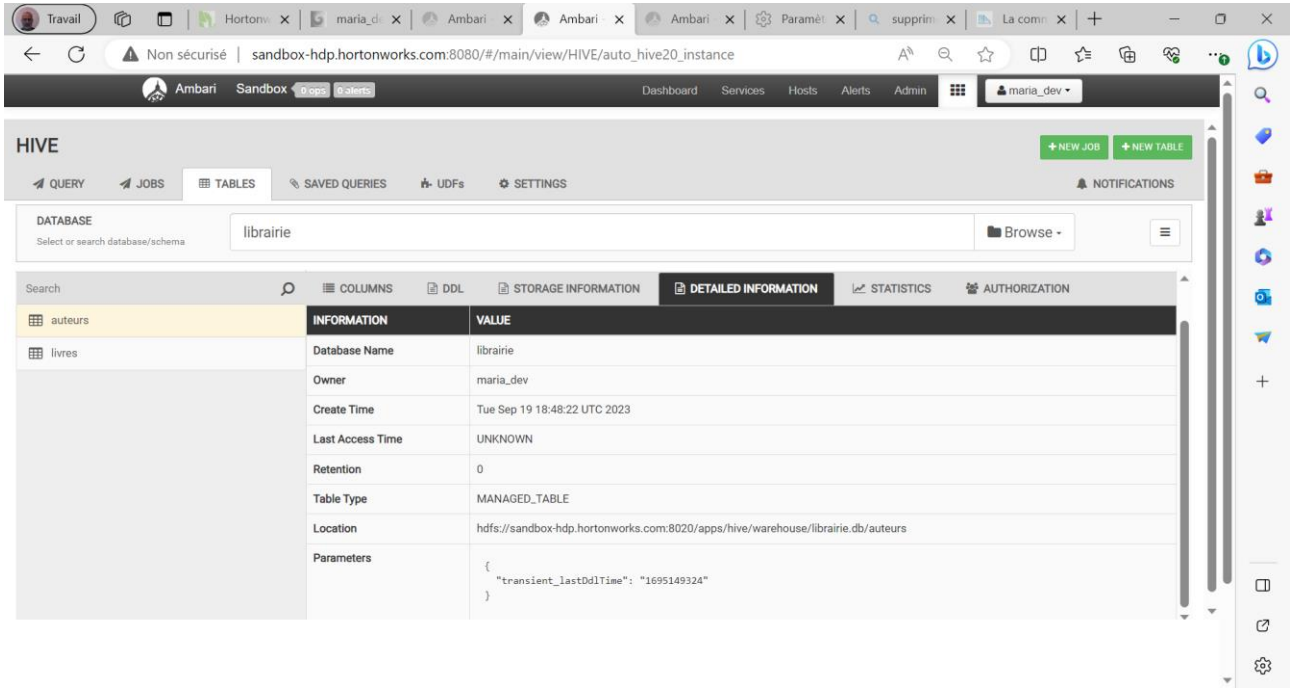
- Taper la requête permettant de créer la BD Librairie et ses tables Auteurs et Livres à partir des fichiers stockées dans le dossier local /tmp/Librairie
- Enregistrer requête par SaveAS.
- Puis exécuter la requête.

Visualisation de tables



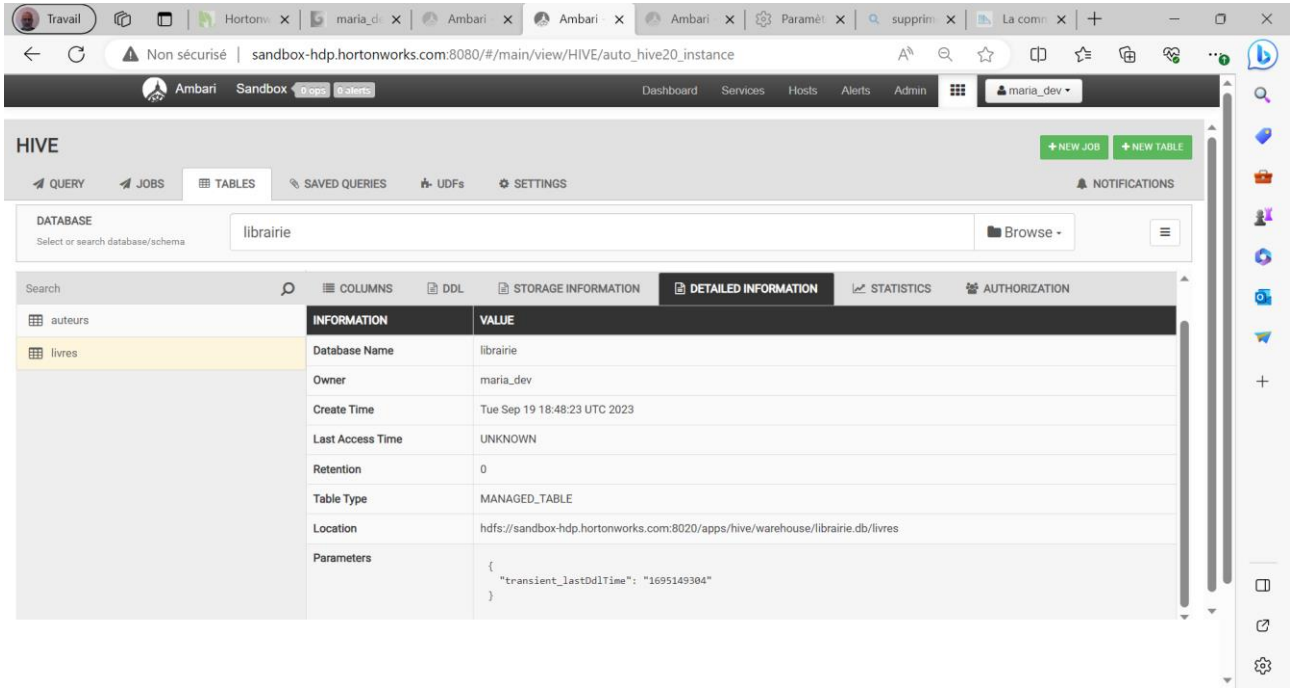
- Visualiser la structure des tables de la BD Librairie.
- ChoisirLibrairie par Browse et choisir l'onglet Table.
- Sélectionner auteurs dans la section de gauche et visualiser.
- Puis visualiser livres

Visualisation de tables



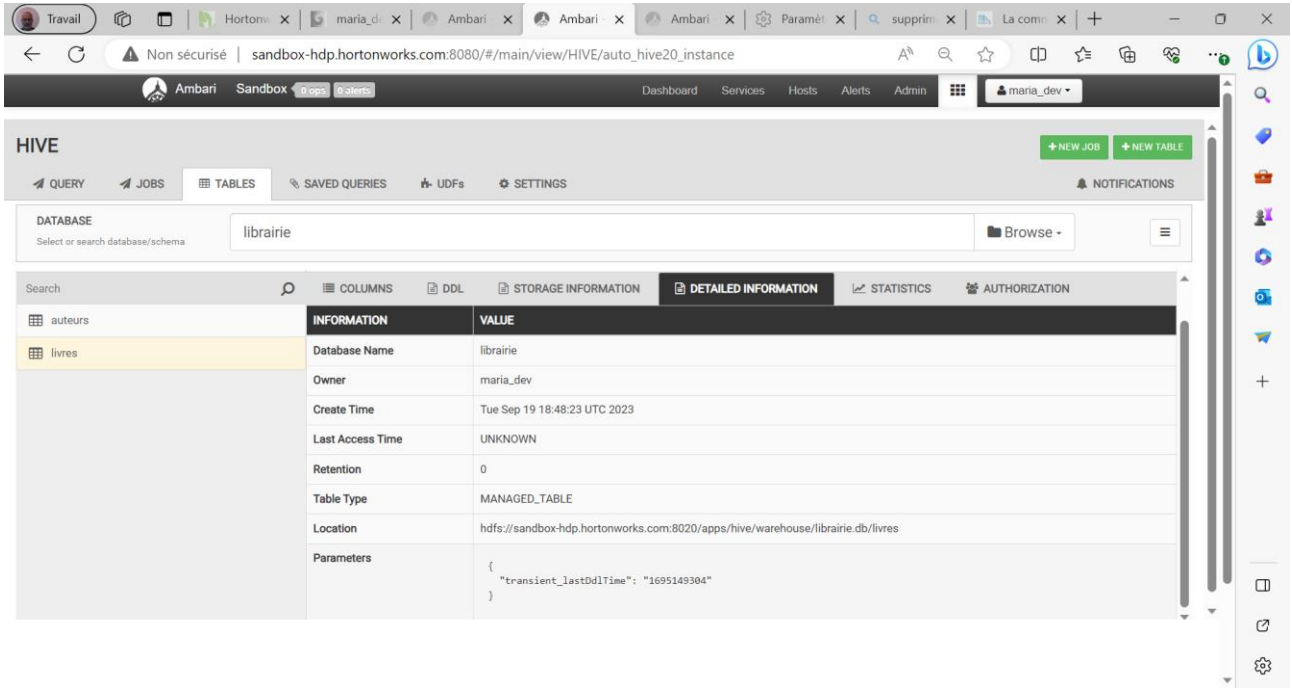
- Localisation de la BD et ses tables sur HDFS.
- La BD librairie se trouve dans le dossier warehouse
- La table auteurs se trouve dans le dossier auteurs dans le dossier librairie dans le dossier warehouse.

Visualisation de tables



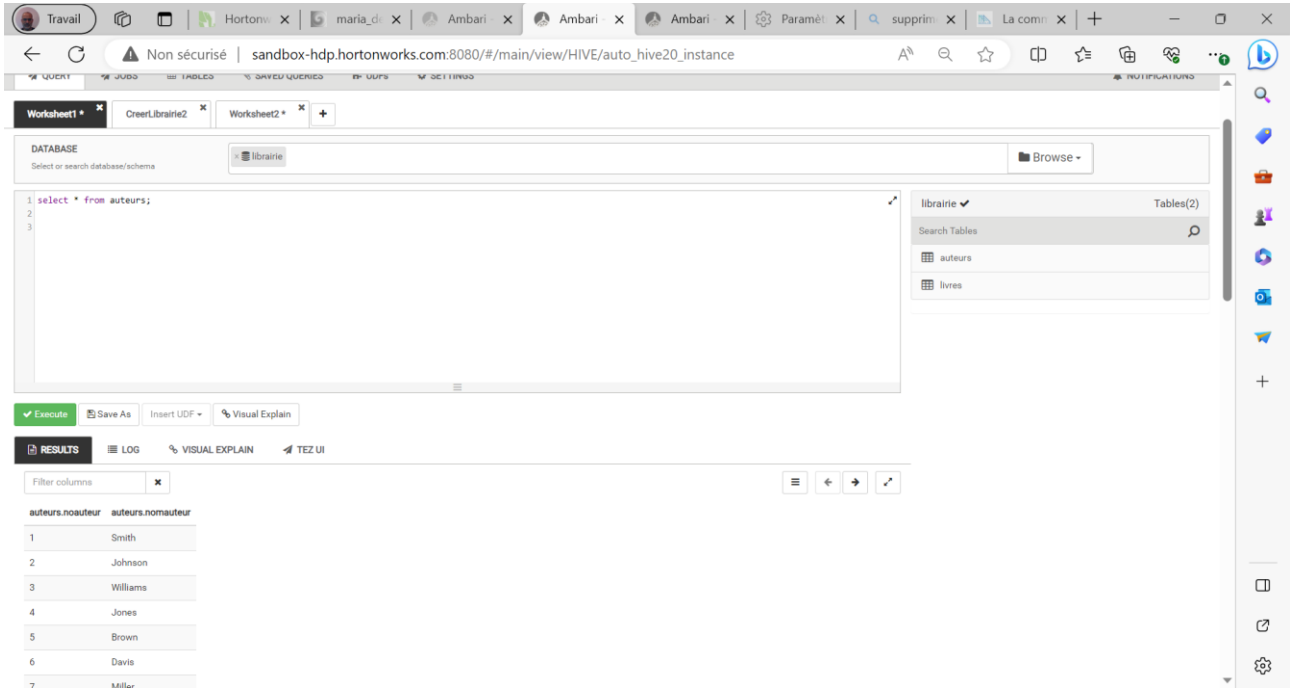
- Localisation de la BD et ses tables sur HDFS.
- La BD librairie se trouve dans le dossier warehouse
- La table livres se trouve dans le dossier livres dans le dossier librairie dans le dossier warehouse.

Visualisation de tables



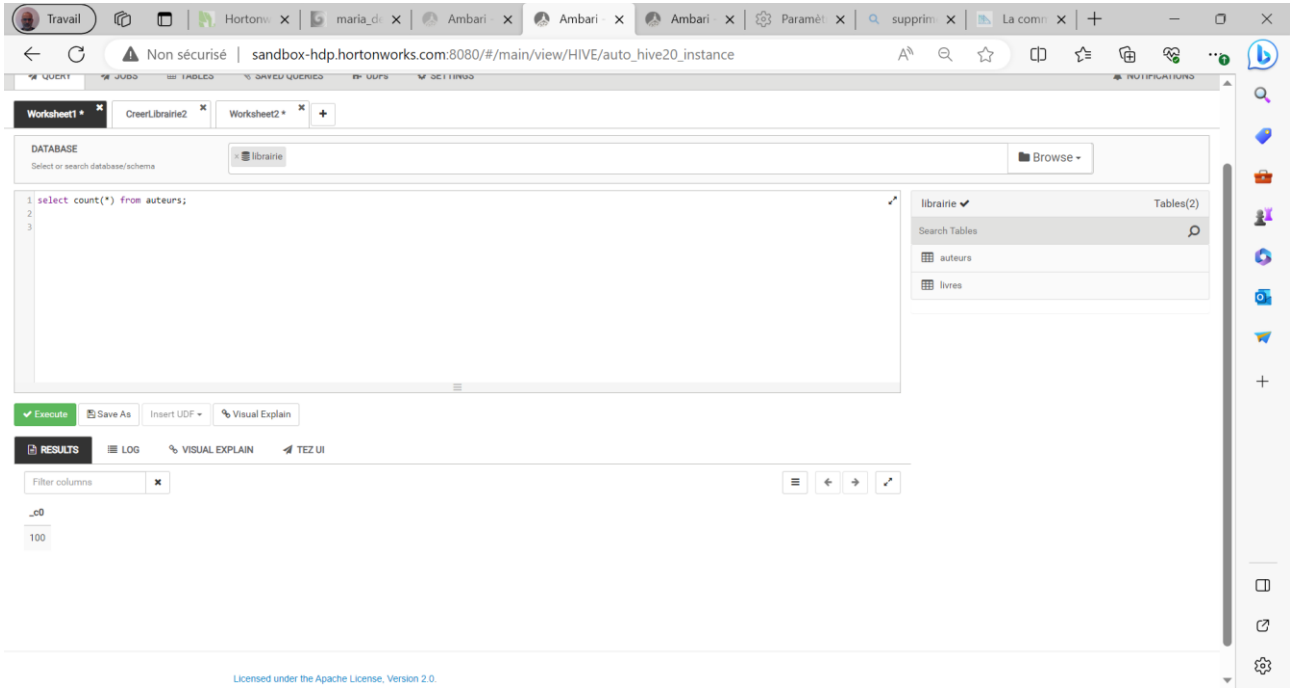
- Localisation de la BD et ses tables sur HDFS.
- La BD librairie se trouve dans le dossier warehouse
- La table livres se trouve dans le dossier livres dans le dossier librairie dans le dossier warehouse.

Requête d'interrogation



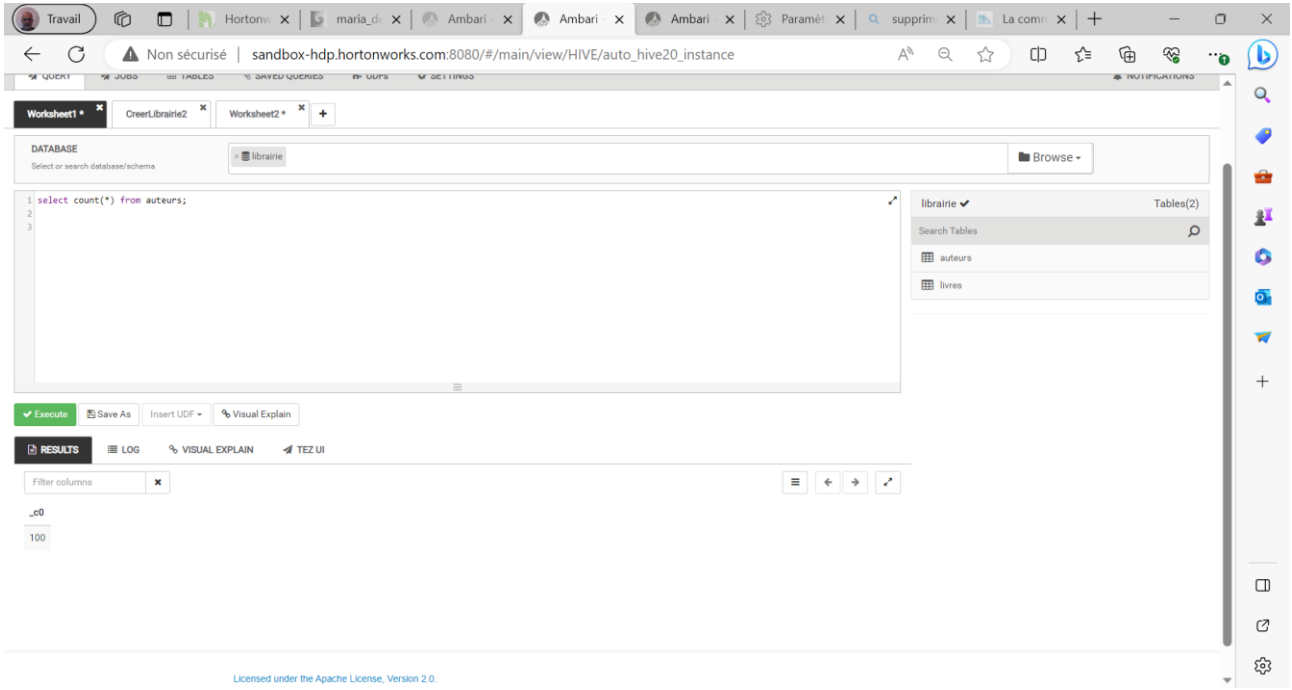
- Afficher les données de la table auteurs.

Requête d'interrogation



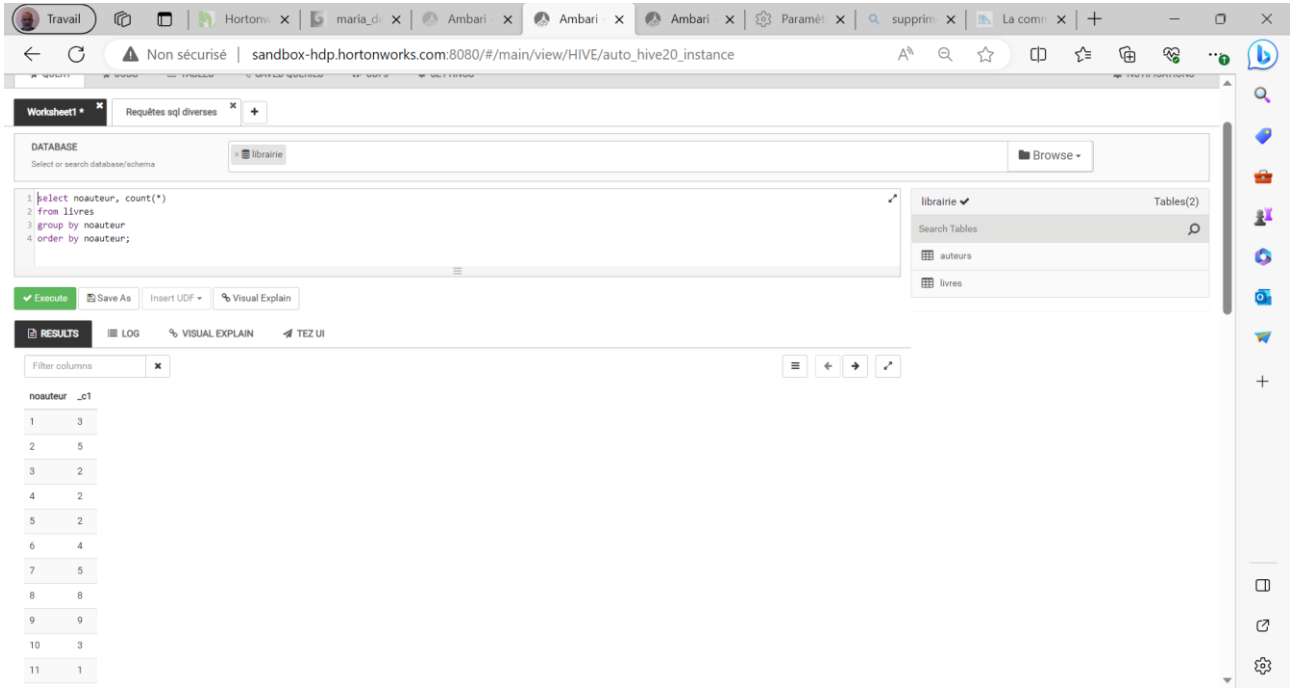
- Compter les auteurs.

Requête d'interrogation



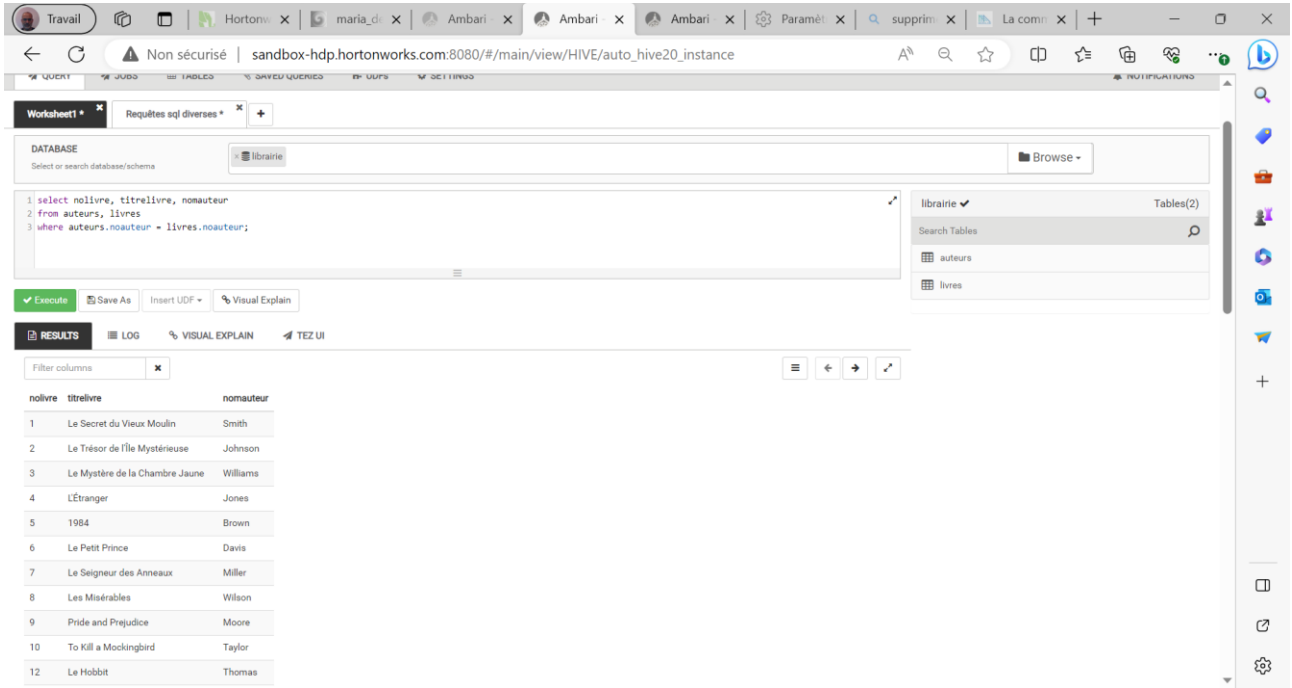
- Ordonner les auteurs selon l'ordre descendant du noAuteur.

Requête d'interrogation



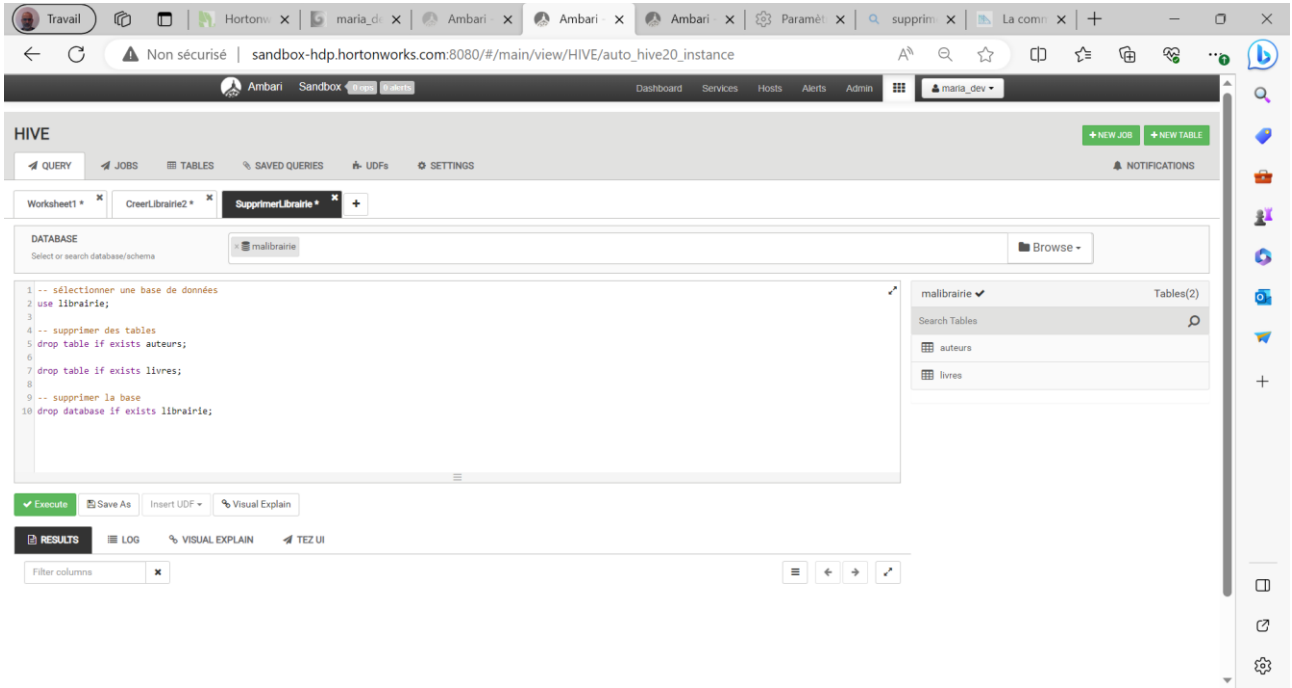
- Grouper les livres par noauteur et les compter.
- Ordonner les résultats selon l'ordre ascendant de noauteur.

Requête de jointure



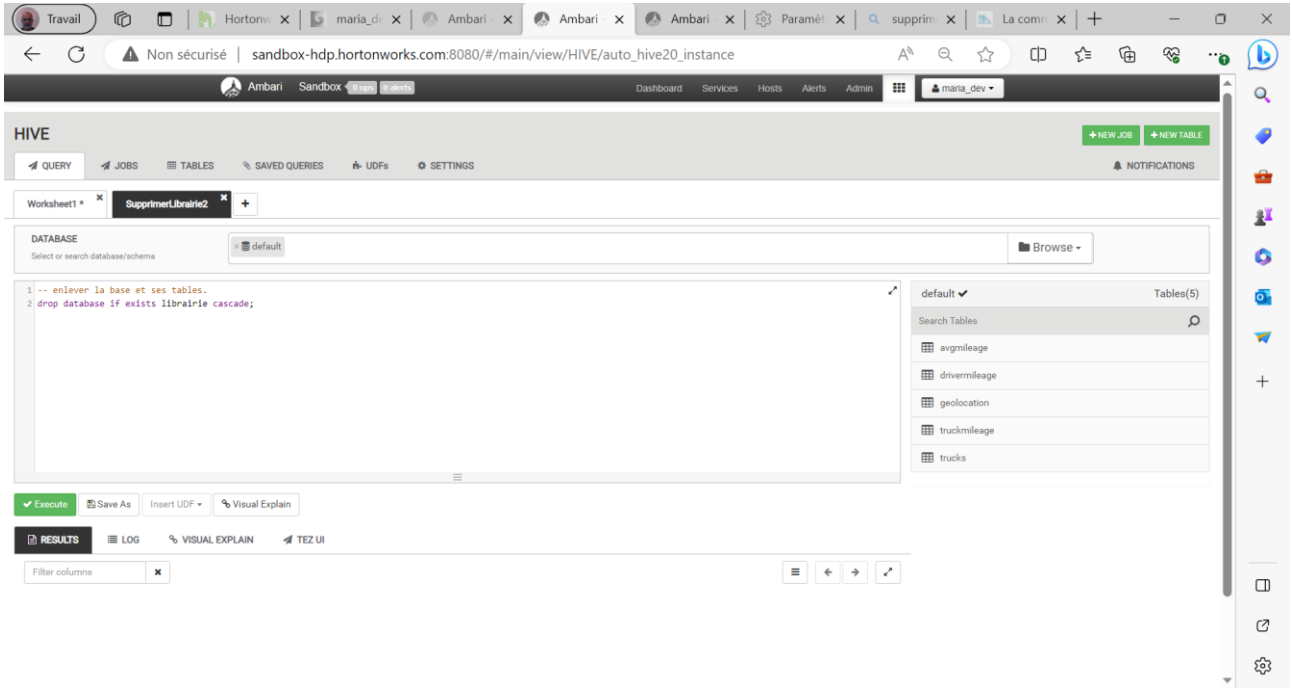
- Grouper les livres par noauteur et les compter.
- Ordonner les résultats selon l'ordre ascendant de noauteur.

Requêtes (DML) de suppression



- Supprimer la base librairie et ses tables.

Requêtes (DML) de suppression



- Supprimer la base librairie et ses tables.

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

Tutorial sur Apache Hive

https://cwiki.apache.org/confluence/display/hive/tutorial

Apache Hive

- Hive est un outil open source pour créer des entrepôts de méga données structurées et semi structurées.
- Hive utilise l'écosystème Hadoop ; il permet de stocker des fichiers de méga données (Big Data) sur HDFS et utilise le paradigme MapReduce pour implémenter des requêtes ETL sur ces méga données.
- Hive prend en compte différents types et formats de fichier : Texte, Sequence, ORC, RC(Record Columnar)
- Avec Hive, l'utilisateur n'a pas besoin d'écrire ses requêtes ETL en MapReduce, Hive lui offre un langage de requête flexible similaire au langage SQL. Ce langage est connu sous le nom HiveQl.
- L'utilisateur peut définir un schéma de base de données relationnelle sur les données stockées par Hive et peut opérer toutes sortes de requête sur ces données en utilisant le langage HiveQL.
- Le langage HiveQL permet d'effectuer des opérations classiques de type LDD, LMD et requête. Il permet également de gérer le stockage des données sur HDFS selon des partitions et des buckets (compartiments). Il offre à l'utilisateur un ensemble de fonctions utiles prédéfinies et il lui permet de définir et utiliser ses propres fonctions UDF (fonctions définies par l'utilisateur) spécifiques et personnalisées pour le nettoyage, le filtrage et la transformation des données.

https://www.guru99.com/introduction-hive.html

Apache Hive

- Hive possède un composant appelé Metastore qui utilise une base de donées relationnelle MySQL/PostgreSQL pour stocker les métadonnées (noms de colonnes, types de données, commentaires, etc.)
- Hive est schéma en lecture uniquement. Il n'est donc pas possible de mettre à jour et de modifier les données.

https://www.guru99.com/introduction-hive.html

Types

- Types numériques
  - Tiny int, small int, int, big int, float, double, decimal
- Types String
  - Char, Varchar, String
- Types Date
  - Timestamp, Date
- Types complexes
  - Arrays, Maps, Structs, Union

https://www.guru99.com/introduction-hive.html

Types complexes

- array : souvent appelé tableau est une collection de une ou plusieurs valeurs de même type.
  L'accès à une valeur se fait selon son indice dans la collection

  auteurs array<string>

  ["Margaret Weis","Tracy Hickman"]

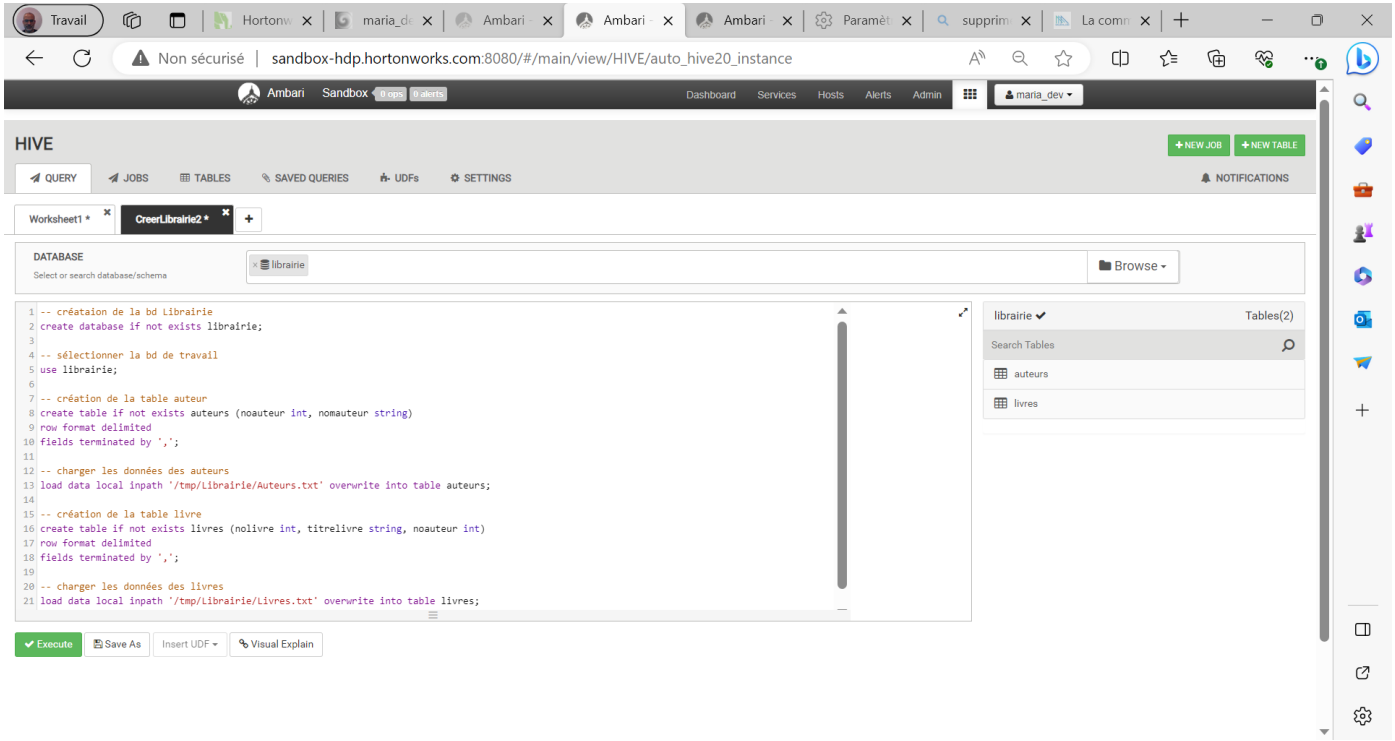
  auteurs[0] permet d'accéder à la première valeur du tableau auteurs

  auteurs[i] permet d'accéder à la (i+1)ème valeur

  size(auteurs) permet d'obtenir la taille du tableau

https://www.guru99.com/introduction-hive.html

Requête DDL



- Taper la requête permettant de créer la BD Librairie3 et sa table Livres2 à partir des fichiers stockées dans le dossier local /tmp/Librairie
- Enregistrer requête par SaveAS.

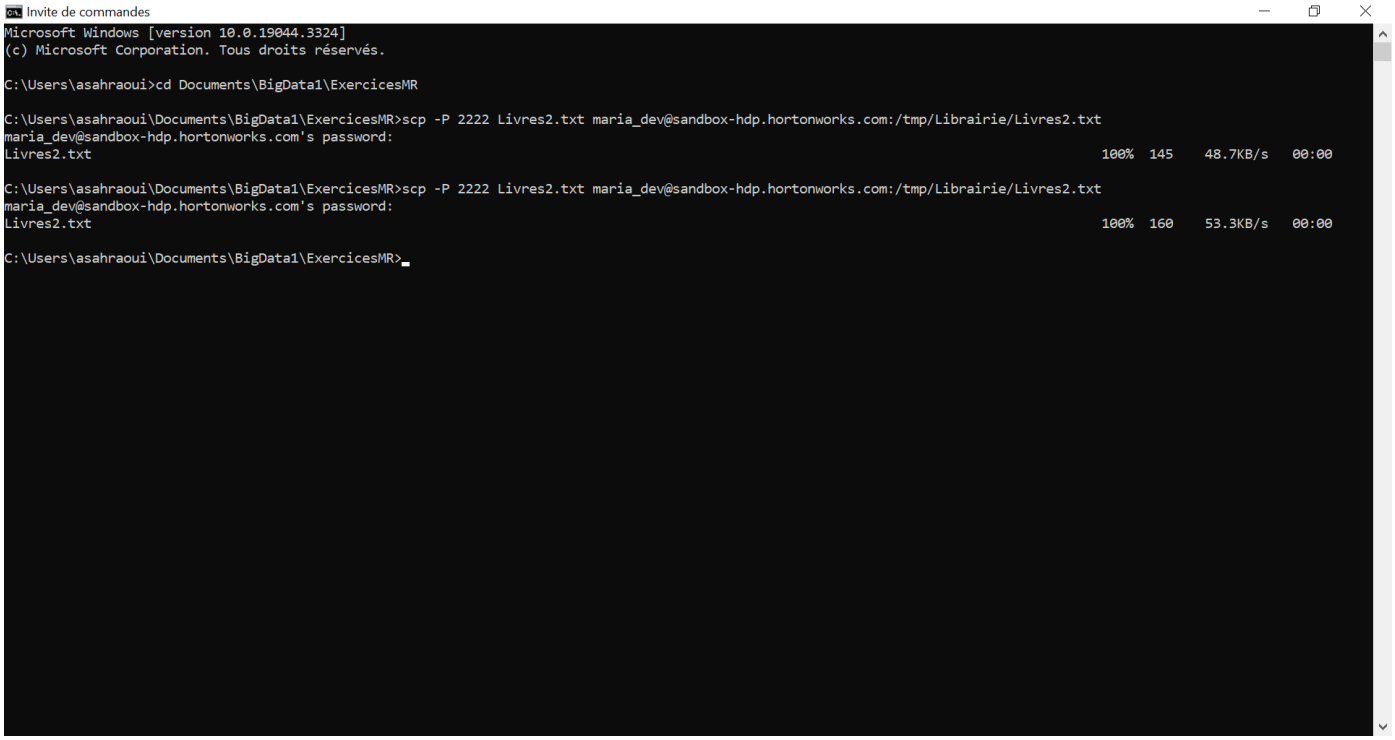
Fichier à charger

Livres2.txt

```
74,La Trilogie de l'Empire,Raymond E. Feist#Janny Wurts
76,Le Cycle de l'Âge de la Mort,Margaret Weis#Tracy Hickman
78,La Trilogie de l'Éveil,Pauline Alphen
```

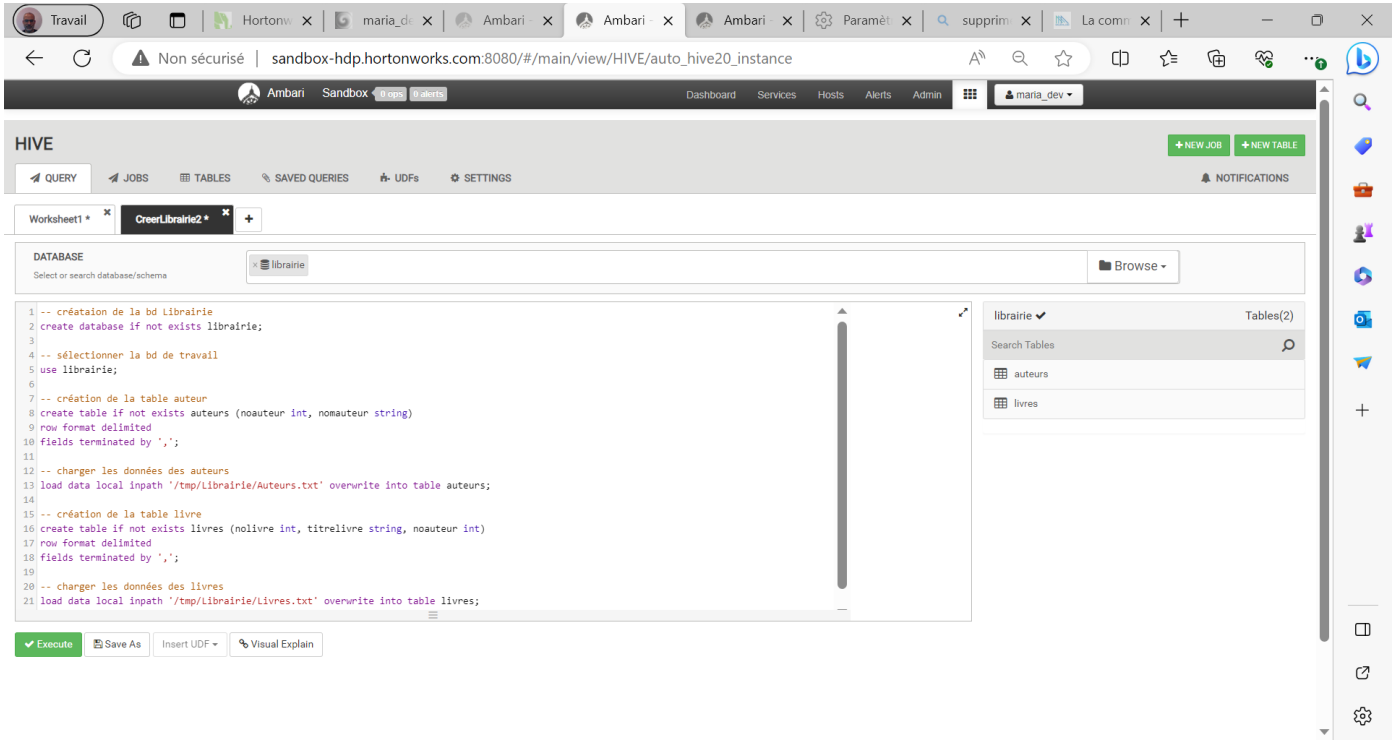
- Créer le fichier livres2.txt

Copie de fichier



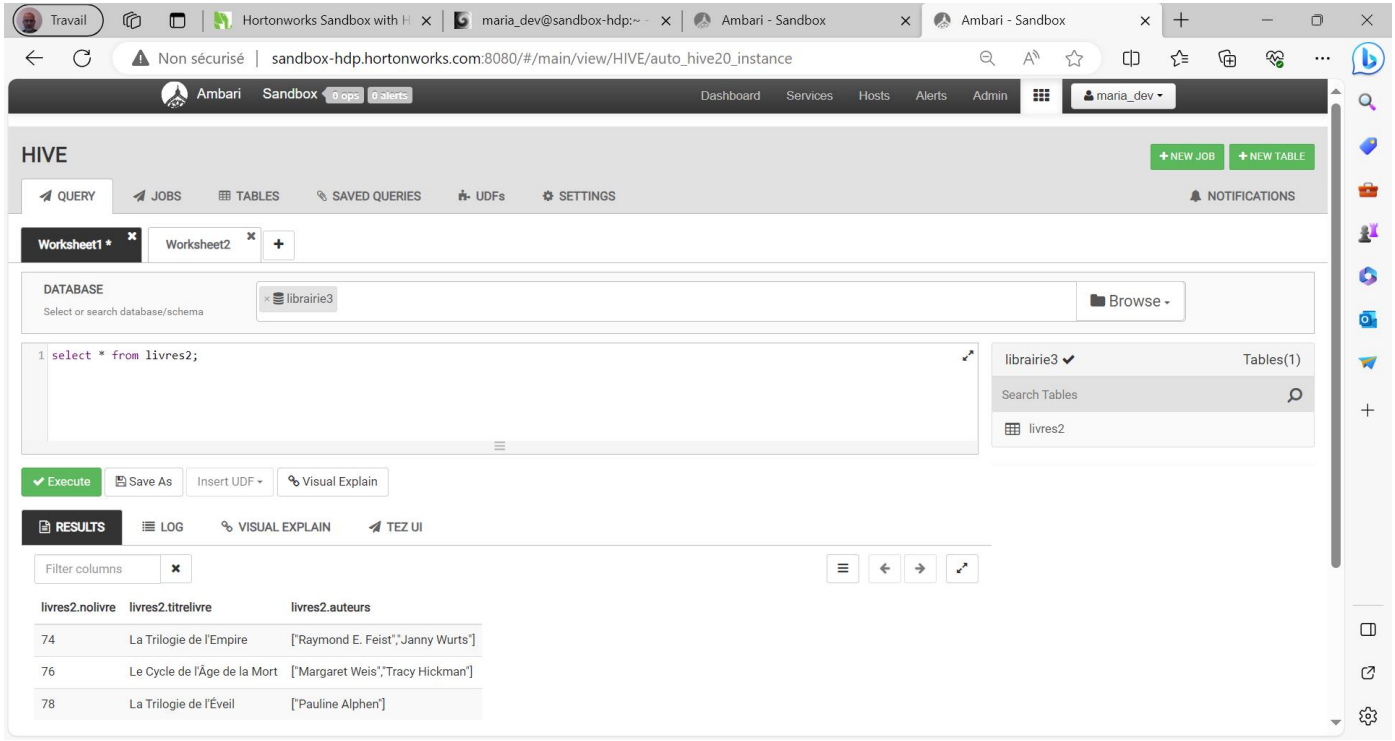
- Copier le fichier livres2.txt dans /tmp/Librairie

Requête DDL



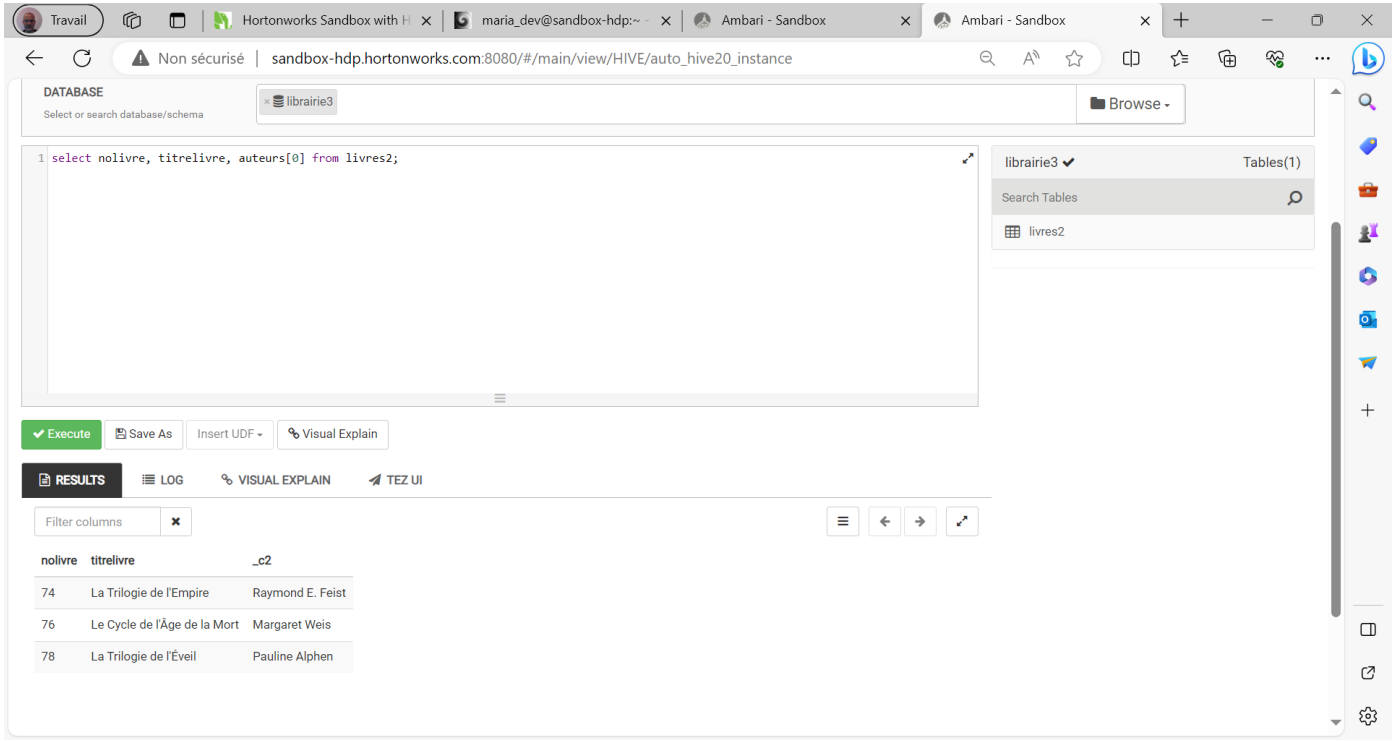
- Exécuter la requête pour créer la BD Librairie3 et sa table Livres2 à partir du fichier stockée dans le dossier local /tmp/Librairie

Requête Interrpgation



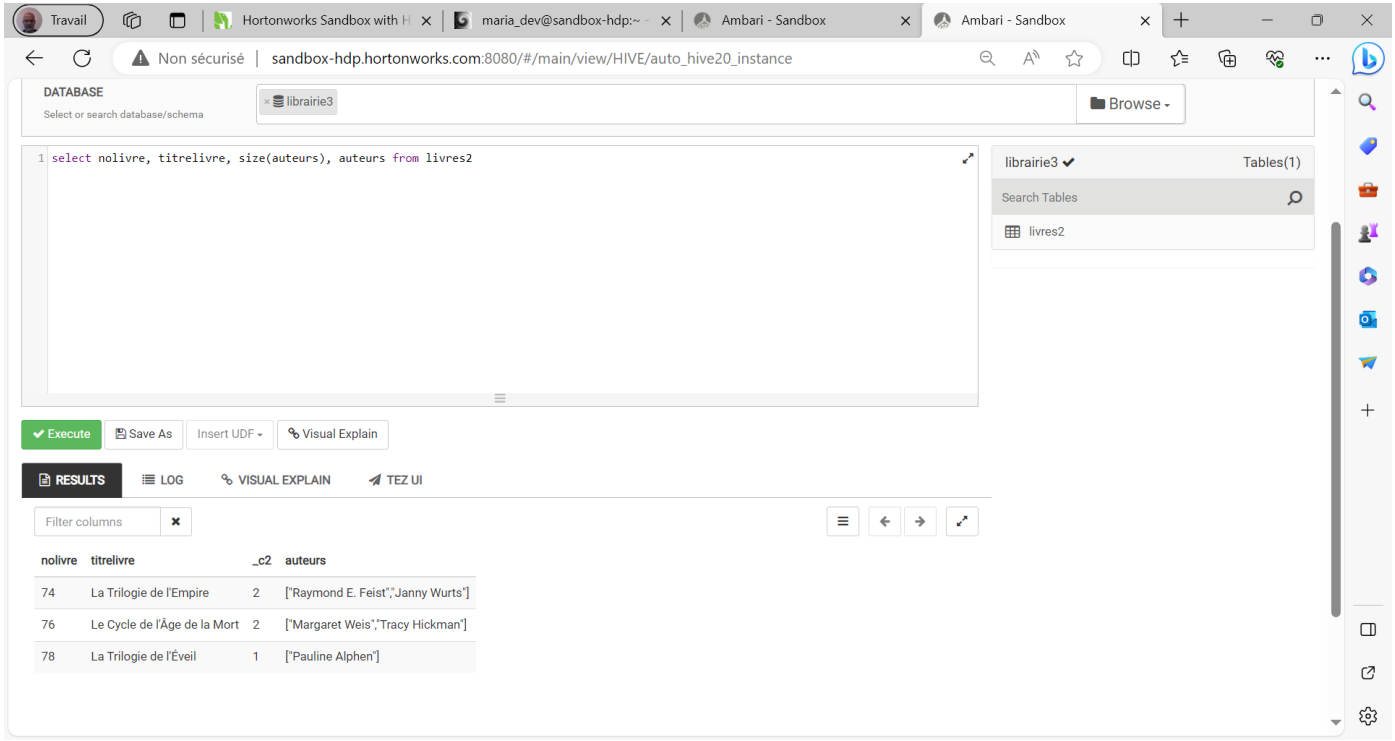
- Exécuter la requête ci-contre pour afficher le contenu de la table livres2

Requête Interrpgation



- Exécuter la requête ci-contre pour afficher le premier auteur de chaque livre

Requête Interrpgation



- Exécuter la requête ci-contre pour afficher le nombre d'auteurs de chaque livre.

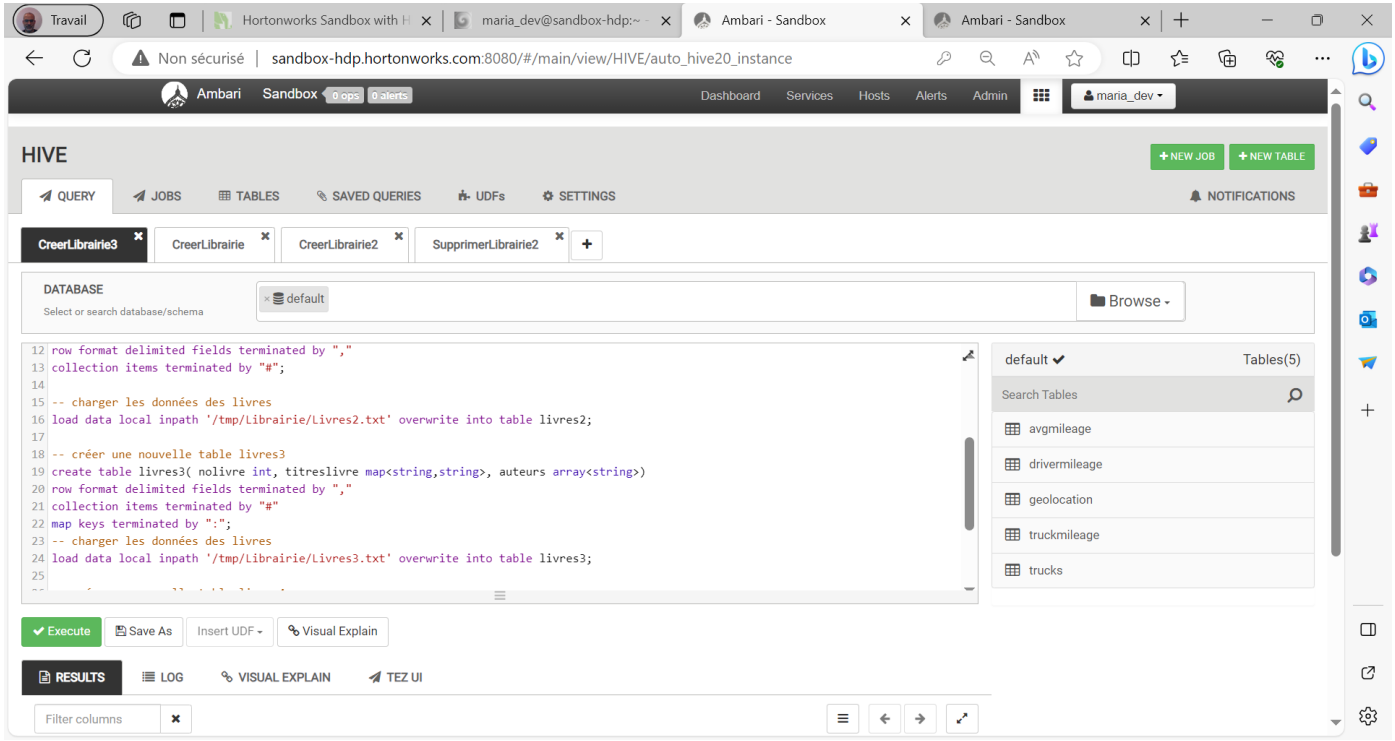
Types complexes

- map : est une collection non ordonnées de une ou plusieurs paires de clé : valeur.
  - La clé et la valeur ont leur propre type.
  - L'accès à une valeur se fait via la clé.

titres map(string, string)
{"Français" : La trilogie de l'Empire, "Anglais " : " The Empire Trilogy " }
titres["Français"] permet d'obtenir le titre en français
titres ["Anglais"] permet d'obtenir le titre en anglais

https://www.guru99.com/introduction-hive.html

Requête DDL



- Taper la requête ci-contre pour créer la BD Librairie3 et ses tables Livres2 et Livres3 à partir des fichiers stockés dans le dossier local /tmp/Librairie

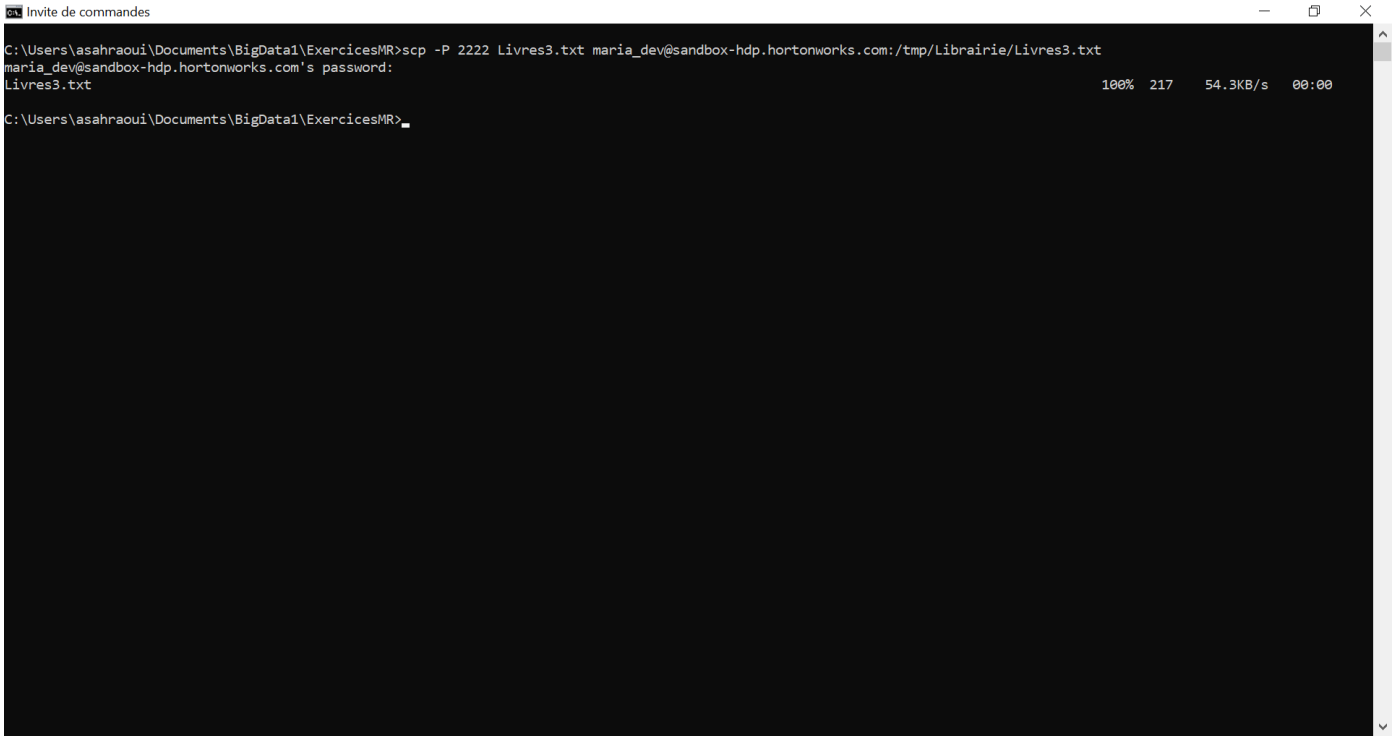
Fichier à charger

- Créer le fichier livres3.txt

Livres3.txt

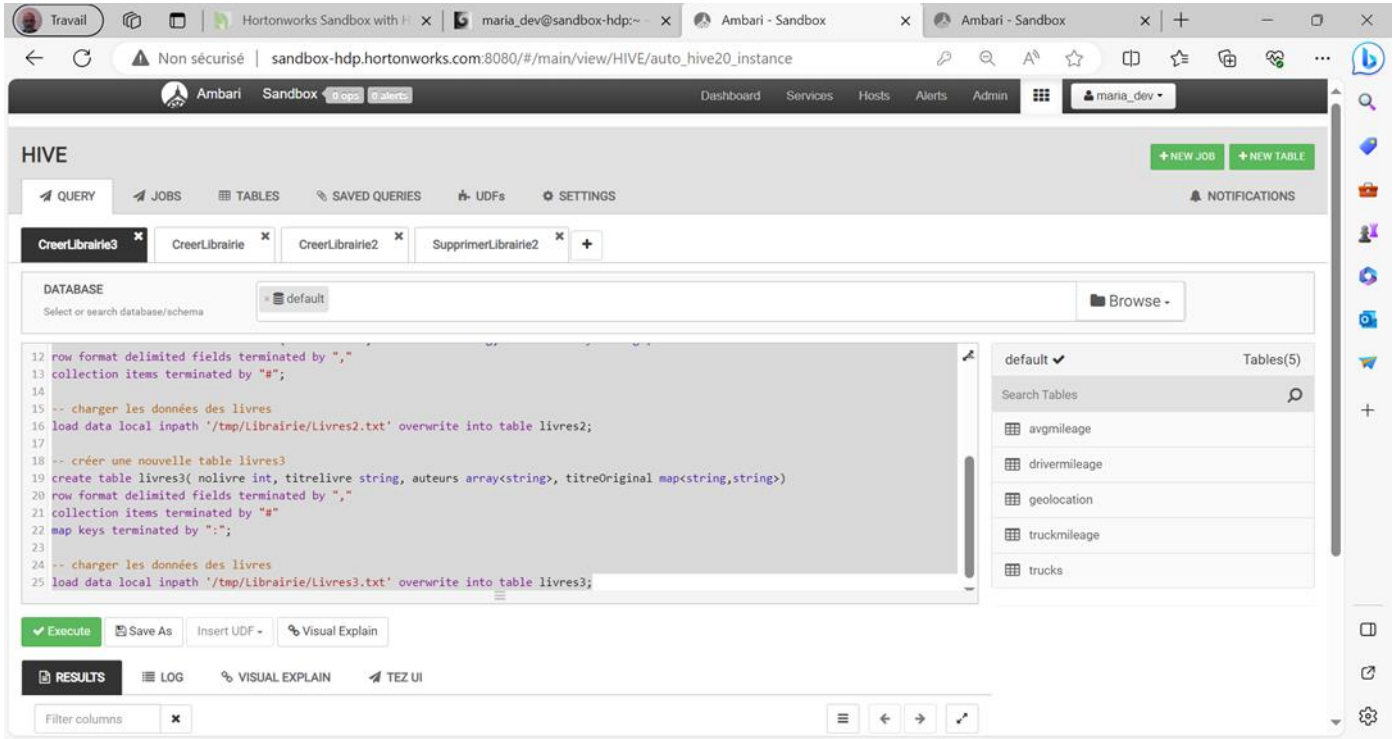
```
74,Français:La Trilogie de l'Empire#Anglais:The Empire Trilogy,Raymond E. Feist#Janny Wurts
76,Français:Le Cycle de l'Âge de la Mort#Anglais:The Death Gate Cycle,Margaret Weis#Tracy Hickman
78,Français:La Trilogie de l'Éveil,Pauline Alphen
```

Copie de fichiers



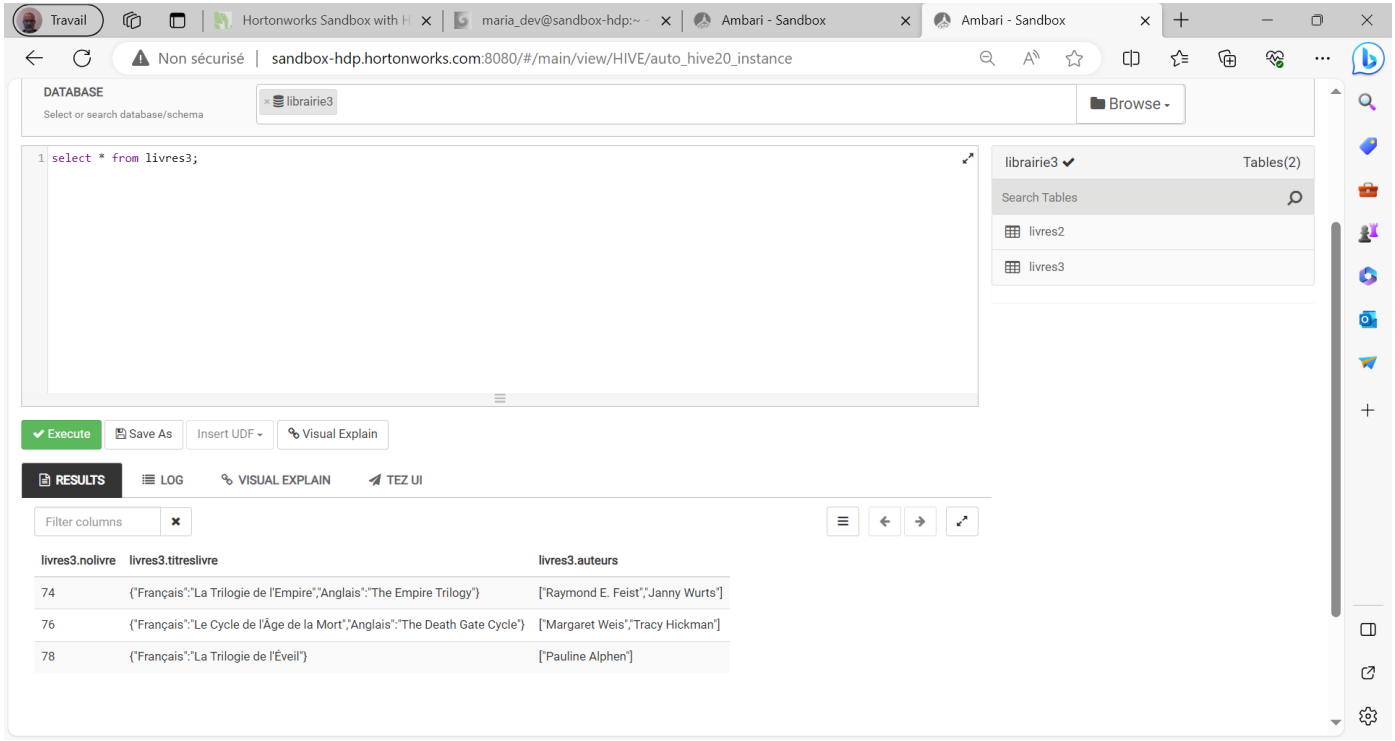
- Copier le fichier livres3.txt dans /tmp/Librairie

Requête DDL



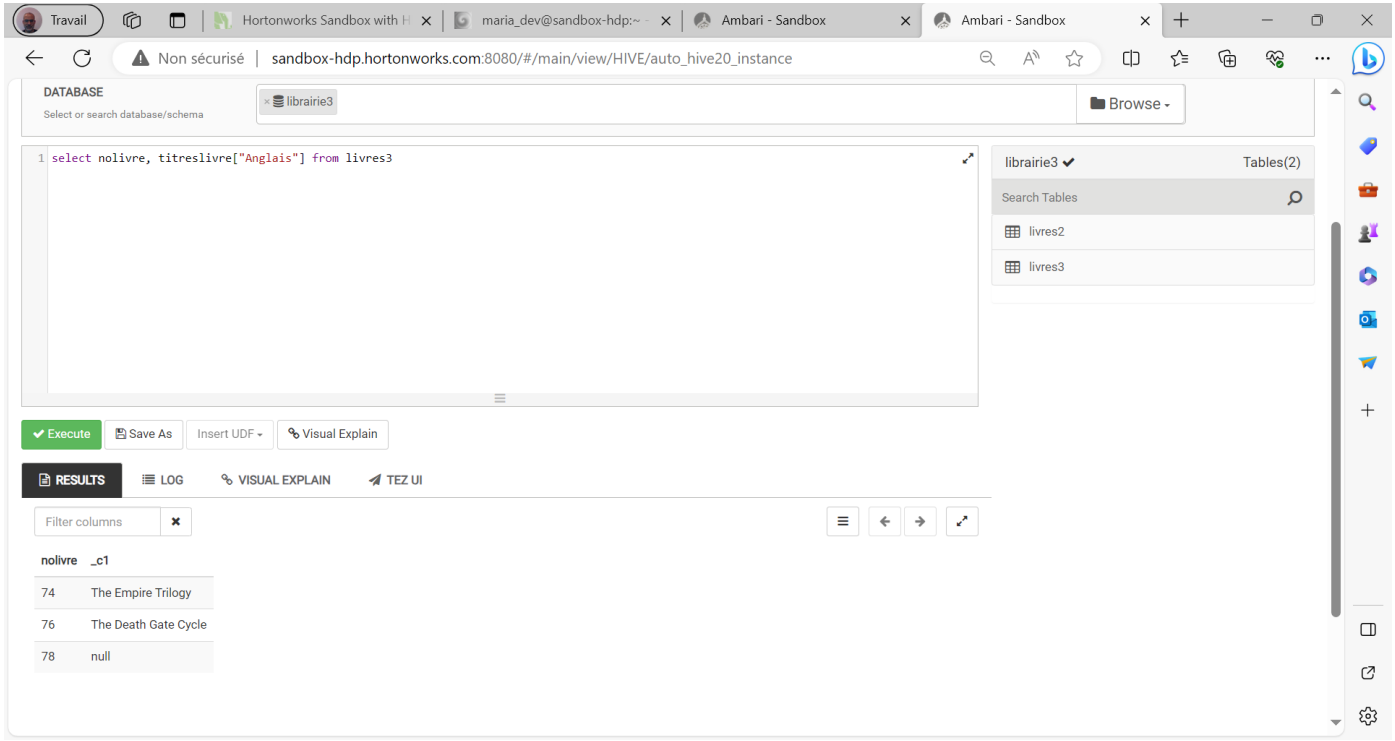
- Exécuter la requête pour créer la BD Librairie3 et ses tables Livres2 et Livres3 à partir des fichiers stockés dans le dossier local /tmp/Librairie

Requête Interrogation



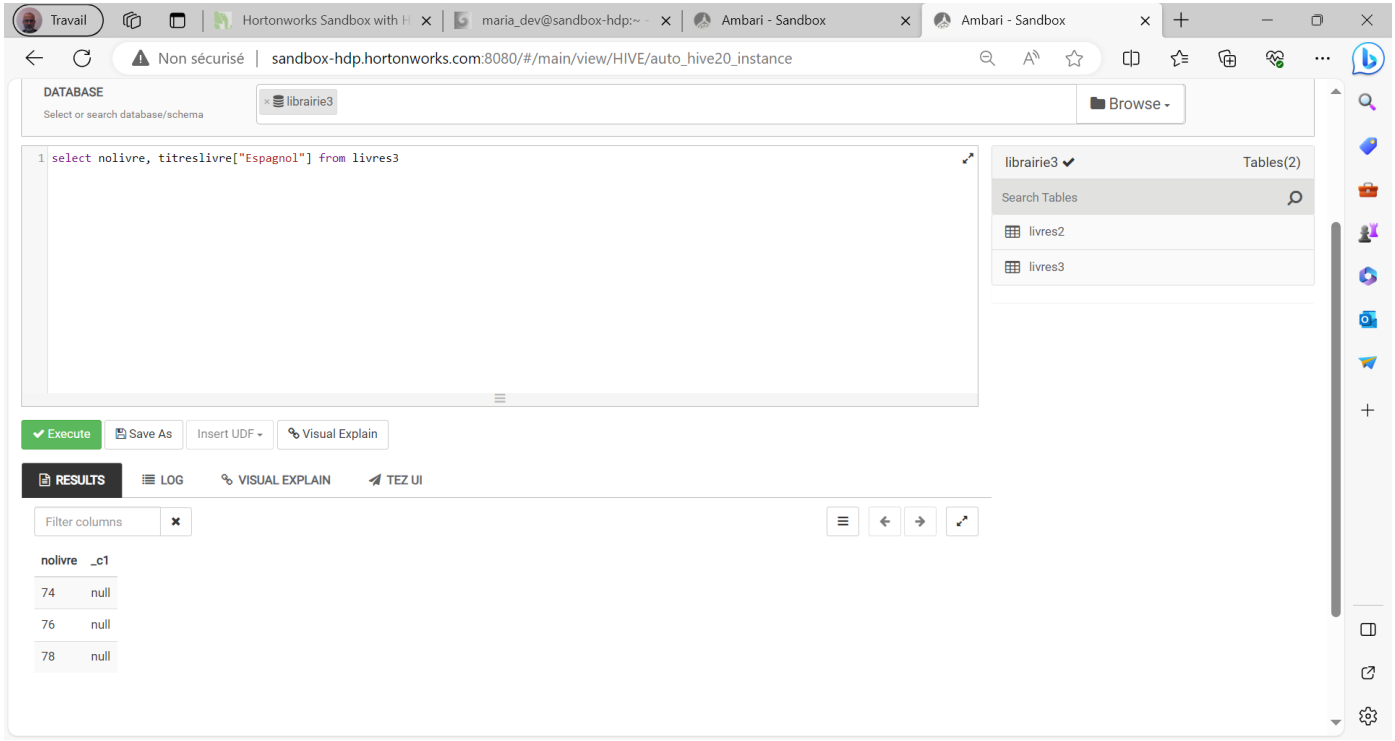
- Exécuter la requête ci-contre pour afficher le contenu de la table livres3

Requête Interrogation



- Exécuter la requête ci-contre pour afficher le titre original en Anglais

Requête Interrogation



- Exécuter la requête ci-contre pour afficher le titre original en Espagnol

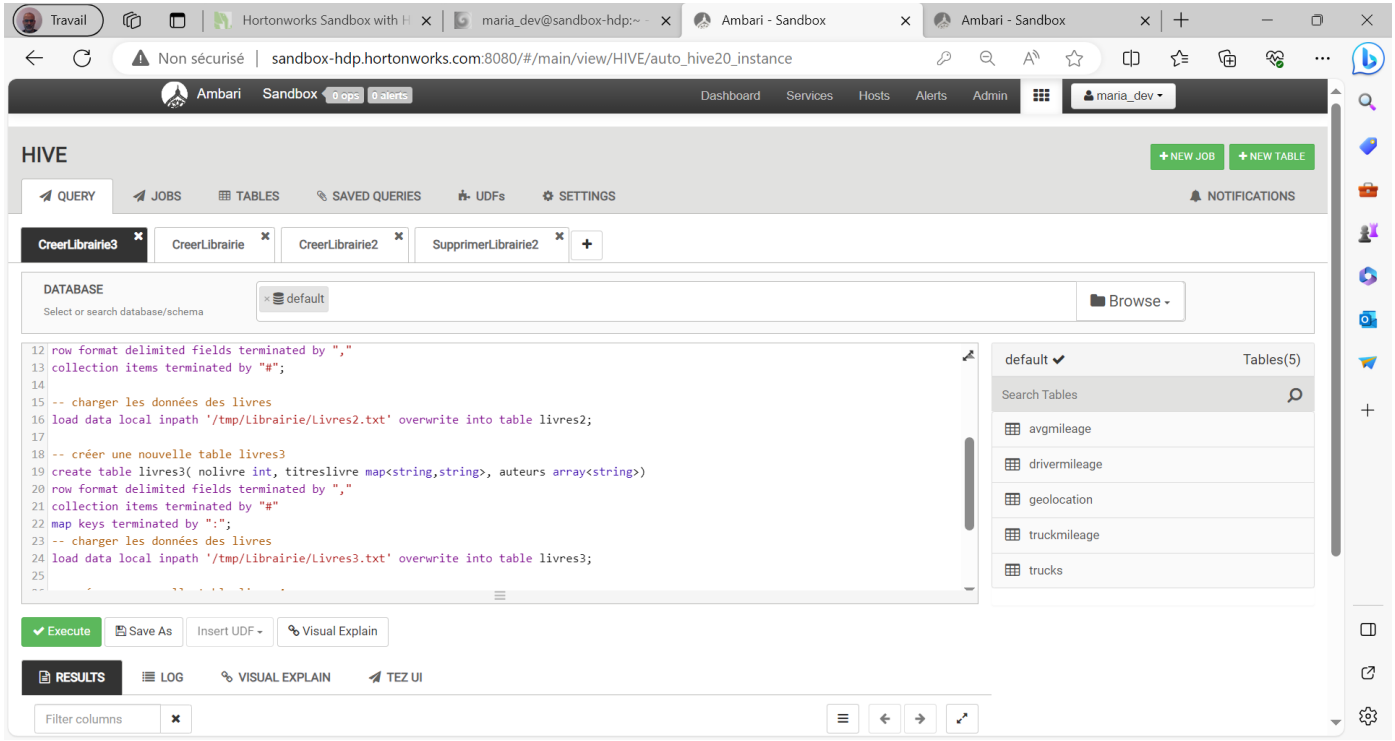
Types complexes

- struct : est un ensemble de deux ou plusieurs. Chaque donnée est référencée par un nom et peut avoir un type différent ou non des types des autres données.
  - L'accès à une donné se fait via son nom.

livreOriginal struct<langue string, titre string>
("Anglais " : " The Empire Trilogy ")
titres["Français"] permet d'obtenir le titre en français
titres ["Anglais"] permet d'obtenir le titre en anglais

https://www.guru99.com/introduction-hive.html

Requête DDL



- Taper la requête ci-contre pour créer la BD Librairie3 et ses tables Livres2 , Livres3 et livres4 à partir des fichiers stockés dans le dossier local /tmp/Librairie

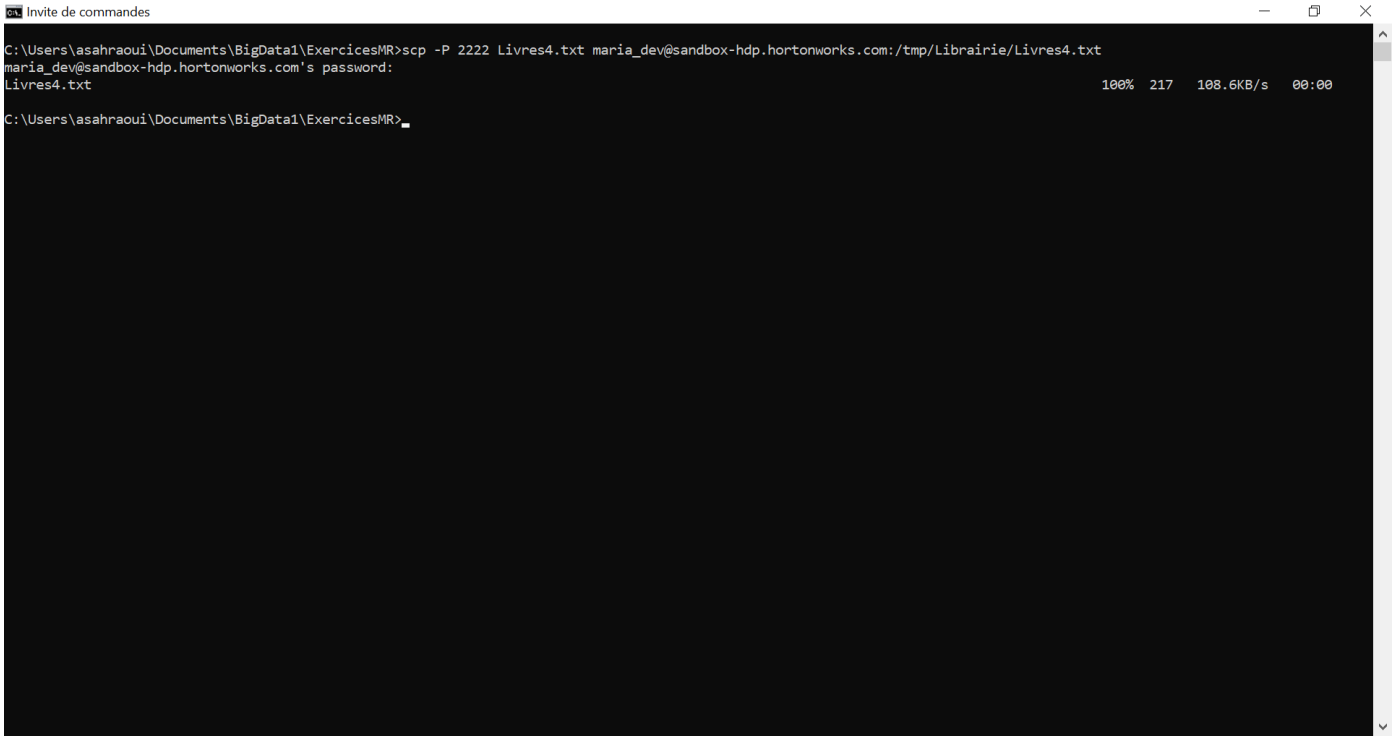
Fichier à charger

- Créer le fichier livres4.txt

Livres4.txt

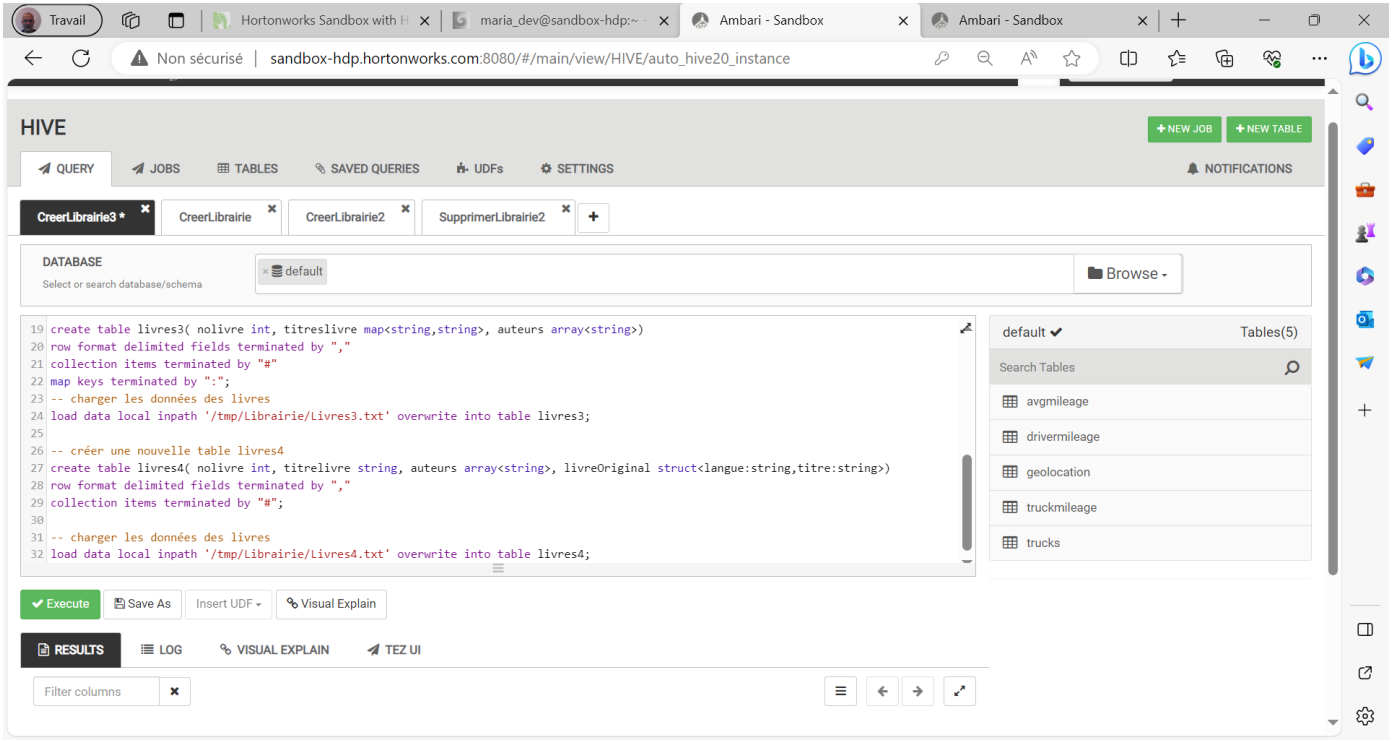
```
74,La Trilogie de l'Empire,Raymond E. Feist#Janny Wurts,Anglais#The Empire Trilogy
76,Le Cycle de l'Âge de la Mort,Margaret Weis#Tracy Hickman,Anglais#The Death Gate Cycle
78,La Trilogie de l'Éveil,Pauline Alphen
```

Copie de fichiers



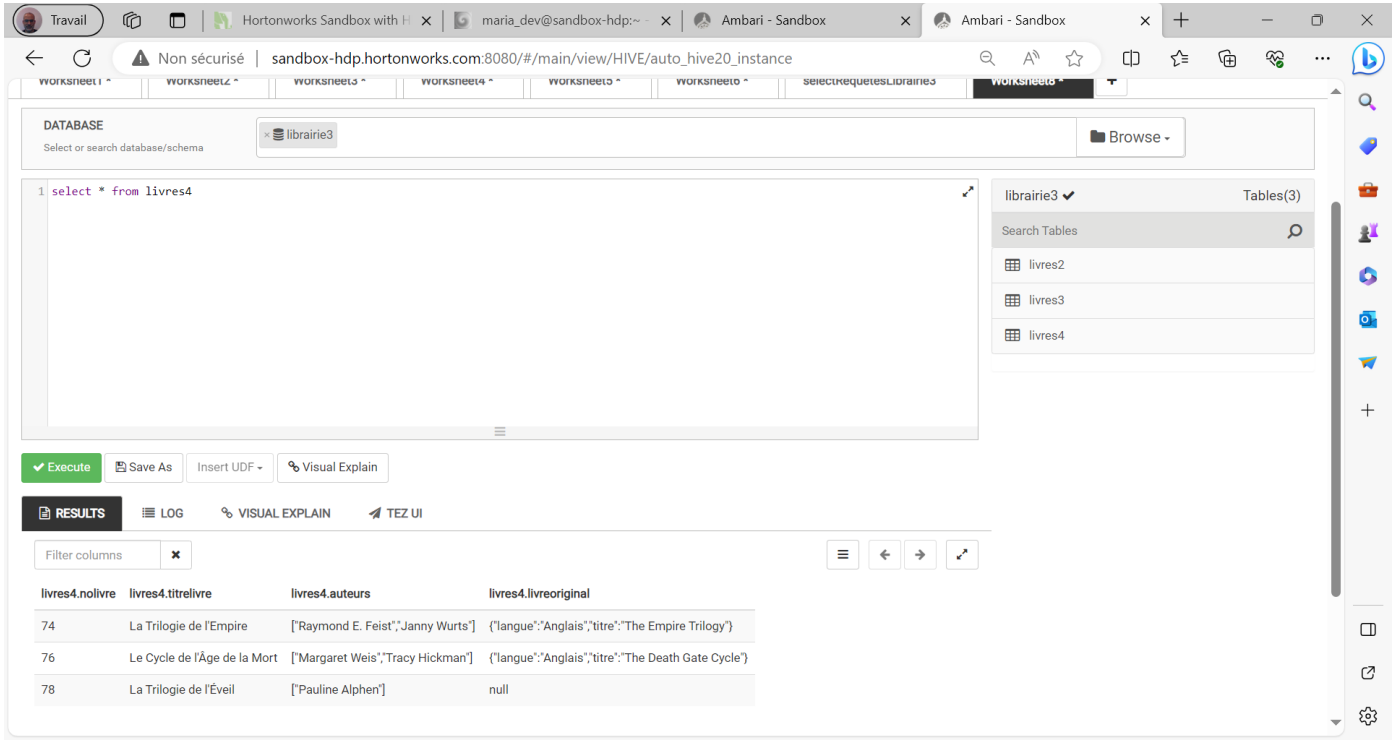
- Copier le fichier livres4.txt dans /tmp/Librairie

Requête DDL



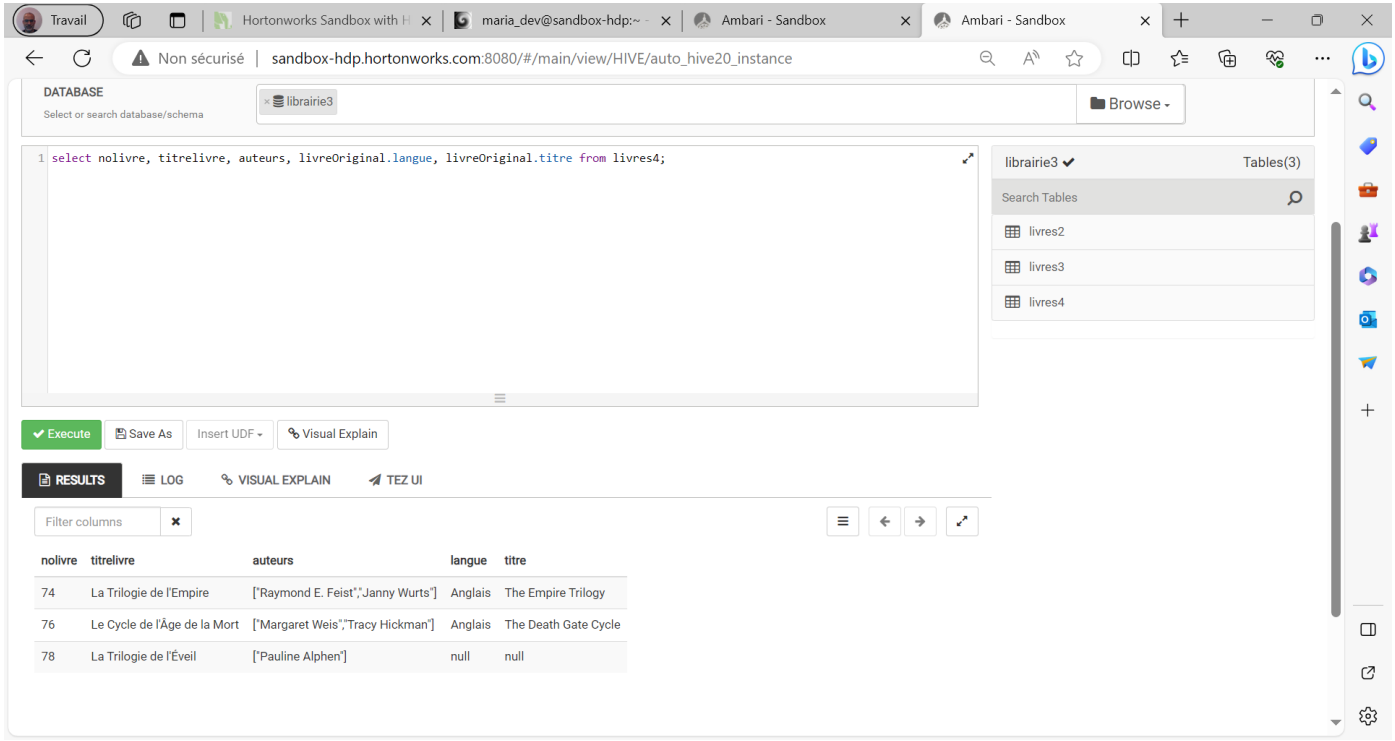
- Exécuter la requête pour créer la BD Librairie3 et ses tables Livres2, Livres3 et Livres4 à partir des fichiers stockés dans le dossier local /tmp/Librairie

Requête Interrogation



- Exécuter la requête ci-contre pour afficher le contenu de la table livres4

Requête Interrogation



- Exécuter la requête ci-contre pour afficher la langue et le titre de la version originale du livre.