



COLLEGE
ROSEMONT

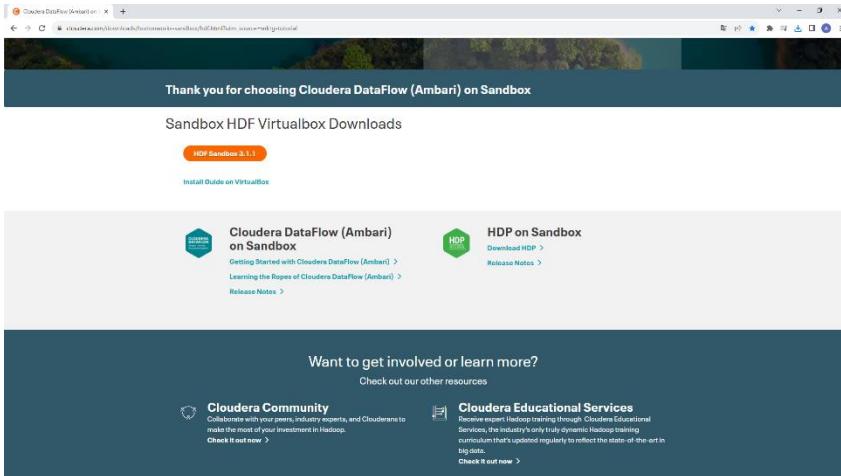
Valorisation de données

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

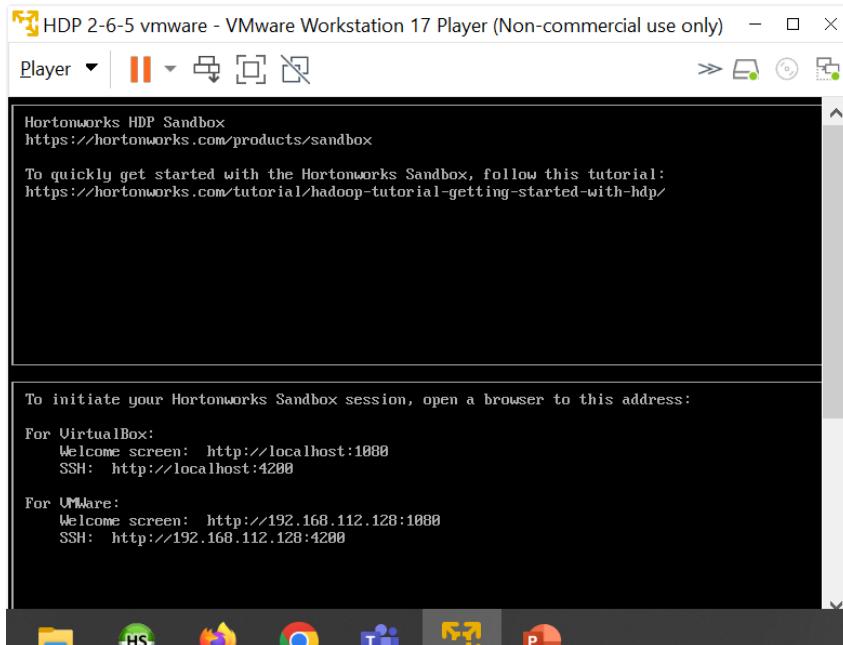
Machine virtuelle HortonWorks HDP 2.6.5



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

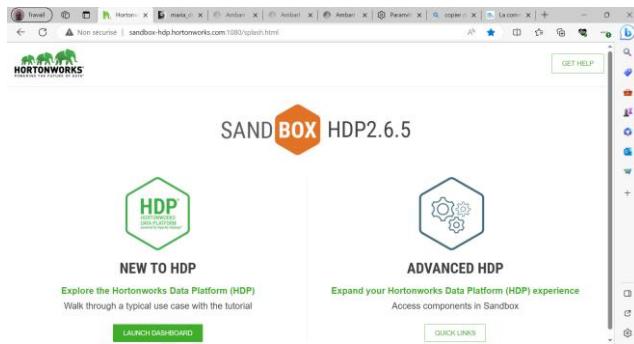
https://www.cloudera.com/downloads/hortonworks-sandbox/hdf.html?utm_source=mktg-tutorialm

Machine virtuelle avec HortonWorks HDP 2.6.5



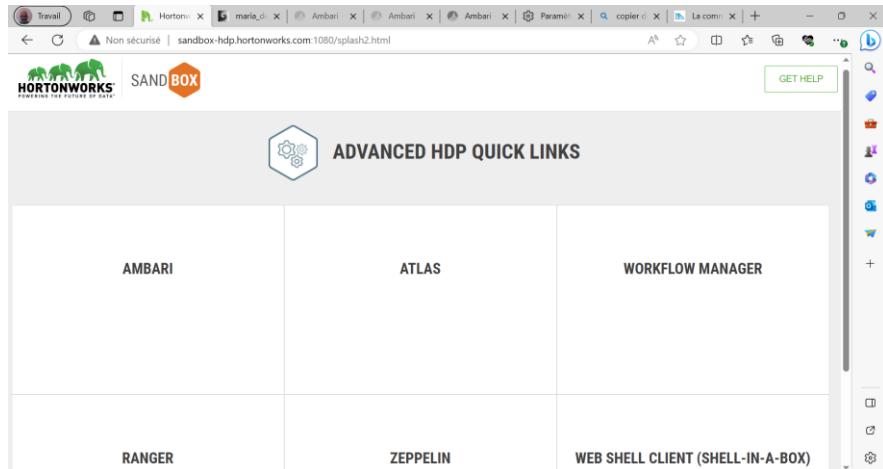
- Aller sur le serveur web dont l'ip a été générée par VMWare.

Connection aux pages web du serveur



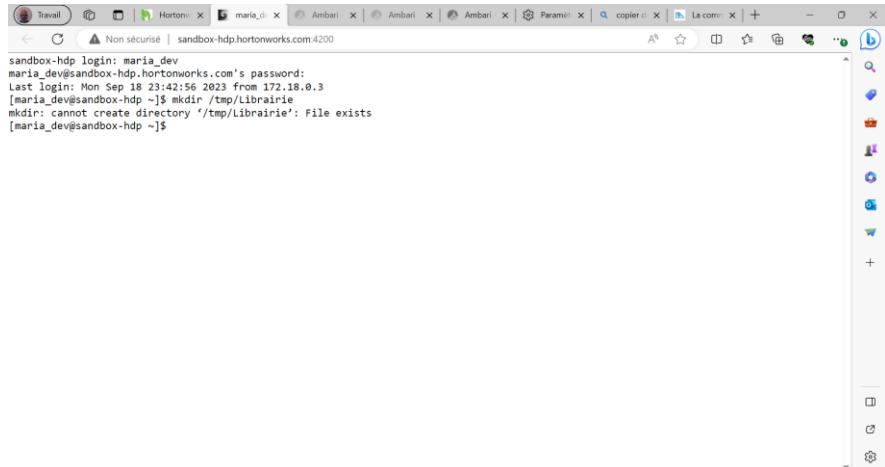
- Se connecter à <http://sandbox-hdp.hortonworks.com:1080/>
- Choisir l'option Advanced HDP

Connection à Ambari / shell



- Lancer le web shell

Connection au shell

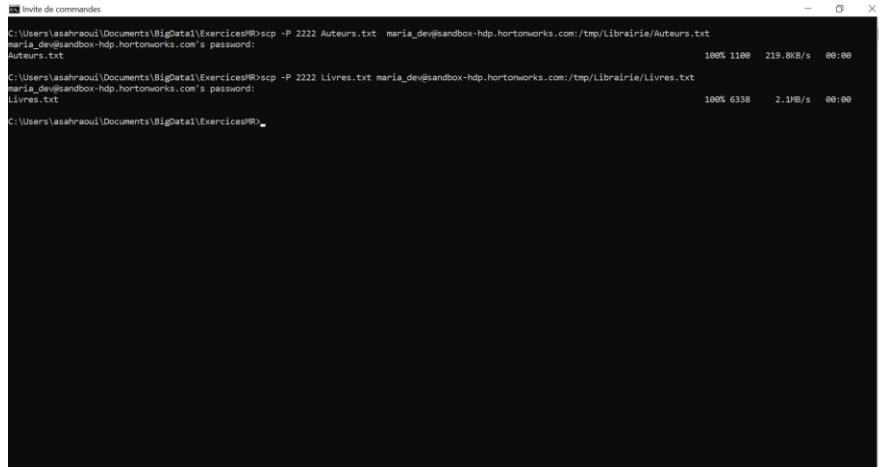


A screenshot of a web browser window titled "Non sécurisé | sandbox-hdp.hortonworks.com:4200". The address bar shows the URL. The main content area displays a terminal session. The session starts with a login prompt for "maria_dev" on "sandbox-hdp.hortonworks.com". It shows the user's last login information (Mon Sep 18 23:42:56 2023 from 172.18.0.3) and then attempts to run the command "mkdir /tmp/Librairie". The output indicates that the directory already exists.

```
sandbox-hdp login: maria_dev
maria_dev@ sandbox-hdp.hortonworks.com's password:
Last login: Mon Sep 18 23:42:56 2023 from 172.18.0.3
[maria_dev@sandbox-hdp ~]$ mkdir /tmp/Librairie
mkdir: cannot create directory '/tmp/Librairie': File exists
[maria_dev@sandbox-hdp ~]$
```

- Créer un dossier Librairie dans le dossier /tmp

Copie de fichiers vers le serveur



```
Invite de commandes

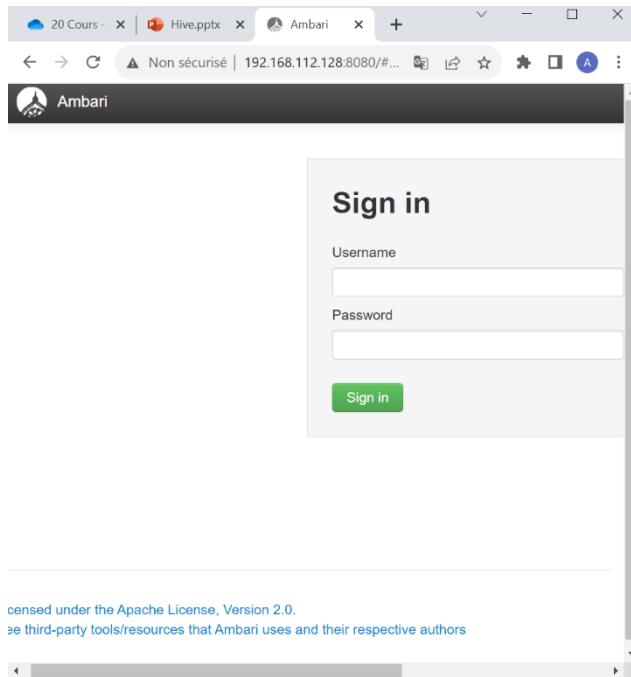
C:\Users\asahraoui\Documents\BigData1\ExercicesMR>scp -P 2222 Auteurs.txt maria_dev@sandbox-hdp.hortonworks.com:/tmp/Librairie/Auteurs.txt
maria_dev@ sandbox-hdp.hortonworks.com's password:
Auteurs.txt

C:\Users\asahraoui\Documents\BigData1\ExercicesMR>scp -P 2222 Livres.txt maria_dev@sandbox-hdp.hortonworks.com:/tmp/Librairie/Livres.txt
maria_dev@ sandbox-hdp.hortonworks.com's password:
Livres.txt

C:\Users\asahraoui\Documents\BigData1\ExercicesMR>
```

- 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

The screenshot shows the Ambari Metrics dashboard with the following details:

- Left sidebar (Services):** HDFS, YARN, MapReduce2, Tez, Hive, HBase, Pig, Sqoop, Oozie, ZooKeeper, Falcon, Storm, Flume, Ambari Infra, Alias, Kafka, Knox, Parrot.
- Middle section:**
 - HDFS Disk Usage:** 42% (Circular gauge)
 - DataNodes Live:** 1/1
 - HDFS Links:** NameNode, Secondary NameNode, 1 DataNodes
 - Memory Usage:** No Data Available
 - CPU Usage:** No Data Available
 - Cluster Load:** No Data Available
 - NameNode Heap:** 21% (Circular gauge)
 - NameNode RPC:** 0.19 ms
 - NameNode CPU WIO:** n/a
 - NameNode Uptime:** 4.2 d
 - HBase Master Heap:** n/a
 - HBase Links:** No Active Master, 1 RegionServers, n/a
 - HBase Ave Load:** n/a
 - HBase Master Uptime:** n/a
- Top navigation bar:** Ambari, Sandbox, 0 ops, 0 alerts, Dashboard, Services, Hosts, Alerts, Admin, maria_dev.
- Right sidebar (dropdown menu):** Files View, Hive View (selected), Pig View, Tez View, Workflow Manager.

Lancer
Hive View
2.0

Requête (DDL) pour créer une base

The screenshot shows the Ambari Hive interface. On the left, a query editor window titled 'CreerLibrairie2' contains the following Hive DDL script:

```
1 -- création de la bd Librairie
2 create database if not exists librairie;
3
4 -- sélectionner la bd de travail
5 use librairie;
6
7 -- création de la table auteur
8 create table if not exists auteurs (noauteur int, nom'auteur string)
9 row format delimited
10 fields terminated by ',';
11
12 -- charger les données des auteurs
13 load data local inpath '/tmp/Librairie/Auteurs.txt' overwrite into table auteurs;
14
15 -- création de la table livre
16 create table if not exists livres (nolivre int, titrelivre string, noauteur int)
17 row format delimited
18 fields terminated by ',';
19
20 -- charger les données des livres
21 load data local inpath '/tmp/Librairie/Livres.txt' overwrite into table livres;
```

Below the query editor are buttons for 'Execute', 'Save As', 'Insert UDF', and 'Visual Explain'. On the right, a sidebar shows a database named 'librairie' with two tables: 'auteurs' and 'livres'. The sidebar also includes icons for various Ambari services like HDFS, YARN, and Tez.

- 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

The screenshot shows the Hortonworks Data Studio interface. The top navigation bar has several tabs: Travail, Hortonworks, maria_db, Ambari, Ambari, Paramètres, supprimé, La commande, and a plus sign. Below the navigation is a header with 'Non sécurisé' and the URL 'sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance'. The main area is titled 'HIVE' and has tabs for QUERY, JOBS, TABLES, SAVED QUERIES, UDFs, and SETTINGS. A notifications icon is also present. On the left, there's a 'DATABASE' section with a search bar containing 'librairie' and a 'Browse' dropdown. The central part shows a table named 'auteurs' with two columns: 'noauteur' (int) and 'nomauteur' (string). The right sidebar features various icons for data exploration and management.

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

Visualisation de tables

The screenshot shows the Ambari Hive interface. At the top, there are several tabs: Travail, Hortonworks, maria_db, Ambari, Ambari, Paramètres, supprimé, La commande, and a plus sign. Below the tabs, the URL is sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance. The main navigation bar includes Ambari, Sandbox, Rops, alerts, Dashboard, Services, Hosts, Alerts, Admin, and a user dropdown for maria_dev. On the left, there's a sidebar with icons for QUERY, JOBS, TABLES (selected), SAVED QUERIES, UDFs, and SETTINGS. The main content area is titled 'HIVE' and shows the 'librairie' database selected. It displays two tables: 'auteurs' and 'livres'. The 'auteurs' table is currently selected. The 'DETAILED INFORMATION' tab is active, showing the following data:

INFORMATION	VALUE
Database Name	librairie
Owner	maria_dev
Create Time	Tue Sep 19 18:48:22 UTC 2023
Last Access Time	UNKNOWN
Retention	0
Table Type	MANAGED_TABLE
Location	hdfs://sandbox-hdp.hortonworks.com:8020/apps/hive/warehouse/librairie.db/auteurs
Parameters	{ "transient_lastDdlTime": "1695149324" }

- 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

The screenshot shows the Ambari Hive interface. At the top, there are several tabs: Travail, Hortonworks, maria_db, Ambari, Ambari, Paramètres, supprimé, La commande, and a plus sign. Below the tabs, the URL is sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance. The main navigation bar includes Ambari, Sandbox, Rops, alerts, Dashboard, Services, Hosts, Alerts, Admin, and a user dropdown for maria_dev. On the left, there's a sidebar with icons for QUERY, JOBS, TABLES (selected), SAVED QUERIES, UDFs, and SETTINGS. The right sidebar has icons for +NEW JOB, +NEW TABLE, and NOTIFICATIONS. The central area is titled 'HIVE' and shows the 'librairie' database selected. It displays two tables: 'auteurs' and 'livres'. The 'livres' table is currently selected. The 'DETAILED INFORMATION' tab is active, showing the following data:

INFORMATION	VALUE
Database Name	librairie
Owner	maria_dev
Create Time	Tue Sep 19 18:48:23 UTC 2023
Last Access Time	UNKNOWN
Retention	0
Table Type	MANAGED_TABLE
Location	hdfs://sandbox-hdp.hortonworks.com:8020/apps/hive/warehouse/librairie.db/livres
Parameters	{ "transient_lastDdlTime": "1695149304" }

- 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.

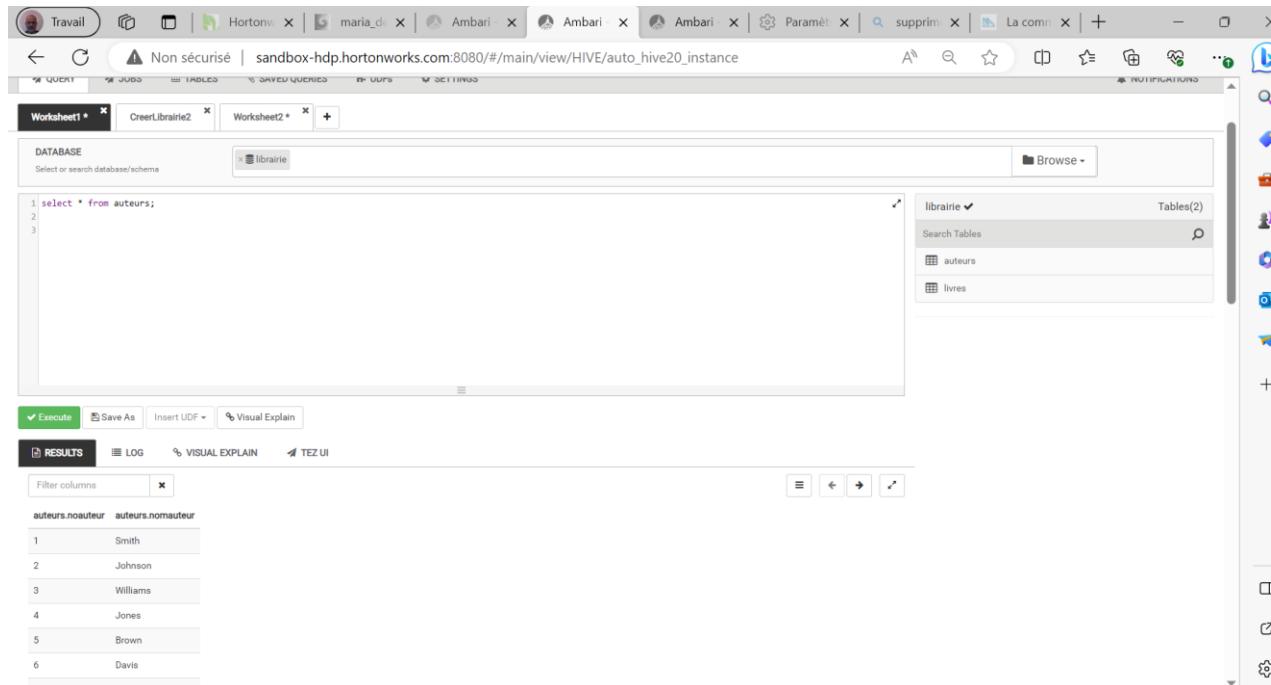
Visualisation de tables

The screenshot shows the Ambari Hive interface. At the top, there are several tabs: Travail, Hortonworks, maria_db, Ambari, Ambari, Paramètres, supprimé, La commande, and a plus sign. Below the tabs, the URL is sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance. The main navigation bar includes Ambari, Sandbox, Rops, alerts, Dashboard, Services, Hosts, Alerts, Admin, and a user dropdown for maria_dev. On the left, there's a sidebar with icons for QUERY, JOBS, TABLES (selected), SAVED QUERIES, UDFs, and SETTINGS. The right sidebar has icons for +NEW JOB, +NEW TABLE, and NOTIFICATIONS. The central area is titled 'HIVE' and shows the 'librairie' database selected. It displays two tables: 'auteurs' and 'livres'. The 'livres' table is highlighted. A detailed information panel for the 'livres' table is open, showing the following data:

INFORMATION	VALUE
Database Name	librairie
Owner	maria_dev
Create Time	Tue Sep 19 18:48:23 UTC 2023
Last Access Time	UNKNOWN
Retention	0
Table Type	MANAGED_TABLE
Location	hdfs://sandbox-hdp.hortonworks.com:8020/apps/hive/warehouse/librairie.db/livres
Parameters	{ "transient_lastDdlTime": "1695149304" }

- 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



The screenshot shows the Apache Tez UI interface. At the top, there's a navigation bar with various tabs like 'Travail', 'Hortonworks', 'maria_db', 'Ambari', 'Ambari', 'Paramètres', 'supprimé', 'La commande', and a '+' button. Below the navigation bar, there are two tabs: 'Worksheet1' and 'CréerLibrairie2'. A search bar is present above the main workspace. The main workspace contains a query editor with the following SQL code:

```
1 select * from auteurs;
```

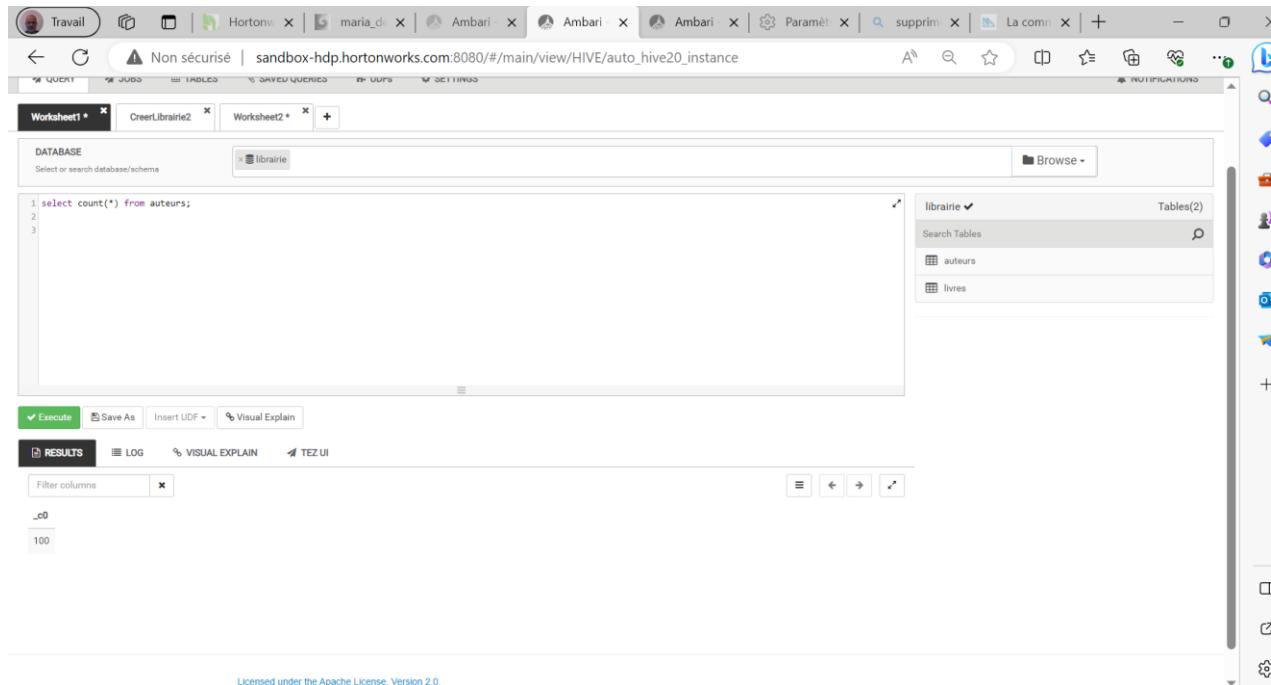
Below the query editor, there's a 'RESULTS' tab which is currently selected. The results show a table with two columns: 'auteurs.noauteur' and 'auteurs.nom'auteur'. The data is as follows:

auteurs.noauteur	auteurs.nom'auteur
1	Smith
2	Johnson
3	Williams
4	Jones
5	Brown
6	Davis
7	Miller

To the right of the main workspace, there's a sidebar with a 'librairie' section containing 'Tables(2)': 'auteurs' and 'livres'. The sidebar also includes icons for search, file, user, and other system functions.

- Afficher les données de la table auteurs.

Requête d'interrogation



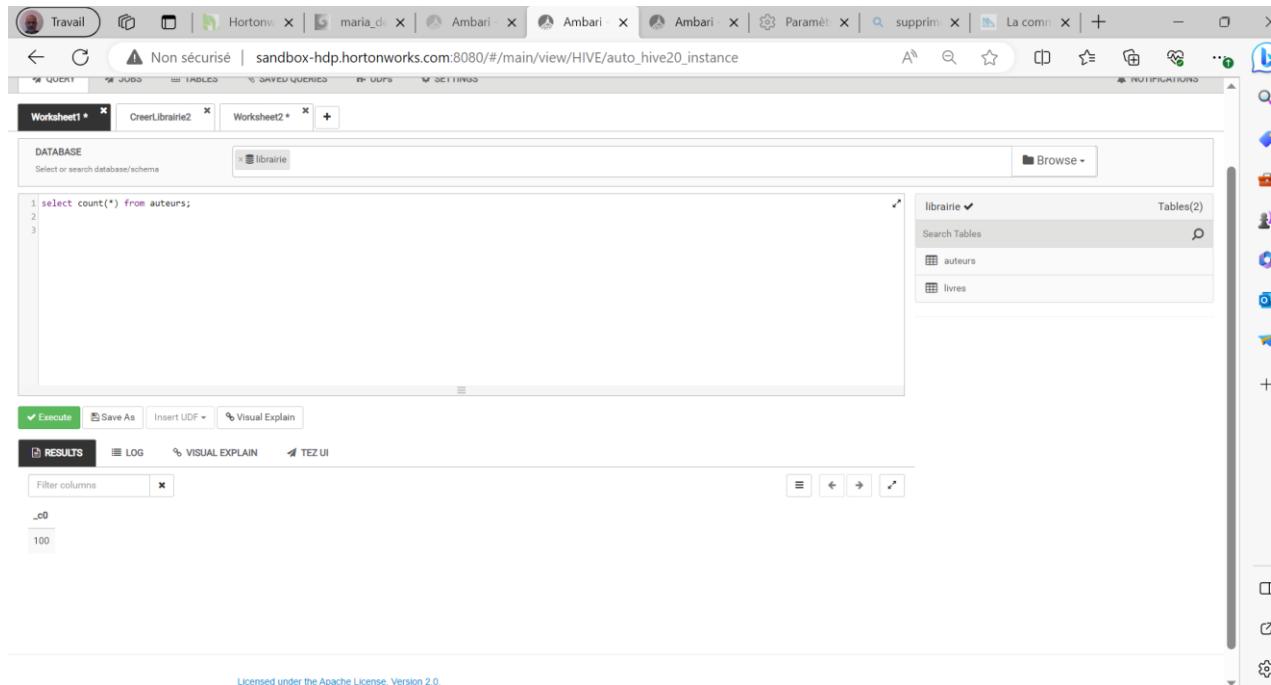
The screenshot shows the Apache Tez UI interface. At the top, there's a navigation bar with various tabs like 'Travail', 'Hortonworks', 'maria_db', 'Ambari', and 'Ambari'. Below the navigation bar, the main area has a title 'Non sécurisé | sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance'. There are two tabs open: 'Worksheet1' and 'Worksheet2'. In 'Worksheet1', a query is being typed into the editor:

```
1 select count(*) from auteurs;
```

The editor also includes buttons for 'Execute', 'Save As', 'Insert UDF', and 'Visual Explain'. To the right of the editor is a 'BROWSE' panel titled 'librairie' which lists tables: 'auteurs' and 'livres'. Below the editor is a 'RESULTS' section showing a single row with one column labeled 'c0' containing the value '100'. At the bottom of the interface, a small note reads 'Licensed under the Apache License, Version 2.0.'

- Compter les auteurs.

Requête d'interrogation



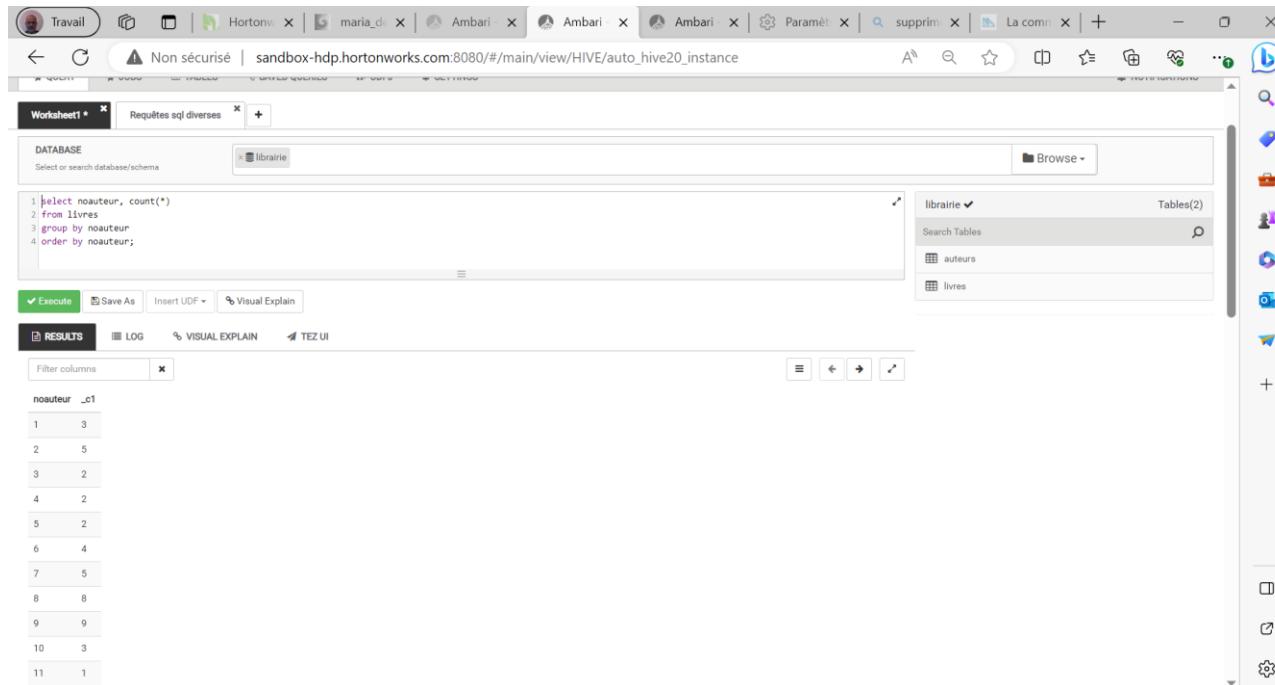
The screenshot shows the Apache Tez UI interface. At the top, there's a navigation bar with various tabs like 'Travail', 'Hortonworks', 'maria_db', 'Ambari', and 'Ambari'. Below the navigation bar, the main area has two tabs: 'Worksheet1' and 'Worksheet2'. A search bar labeled 'librairie' is present. The left pane contains a query editor with the following SQL code:

```
1 select count(*) from auteurs;
```

The right pane shows a results table titled 'librairie' with two rows: 'auteurs' and 'livres'. Below the table, the word 'Tables(2)' is displayed. At the bottom of the interface, there are buttons for 'Execute', 'Save As', 'Insert UDF', and 'Visual Explain'. The 'RESULTS' tab is selected, showing a single row with the value '100'.

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

Requête d'interrogation



The screenshot shows a browser window with multiple tabs open. The active tab is titled "Requêtes sql diverses" and contains the following SQL query:

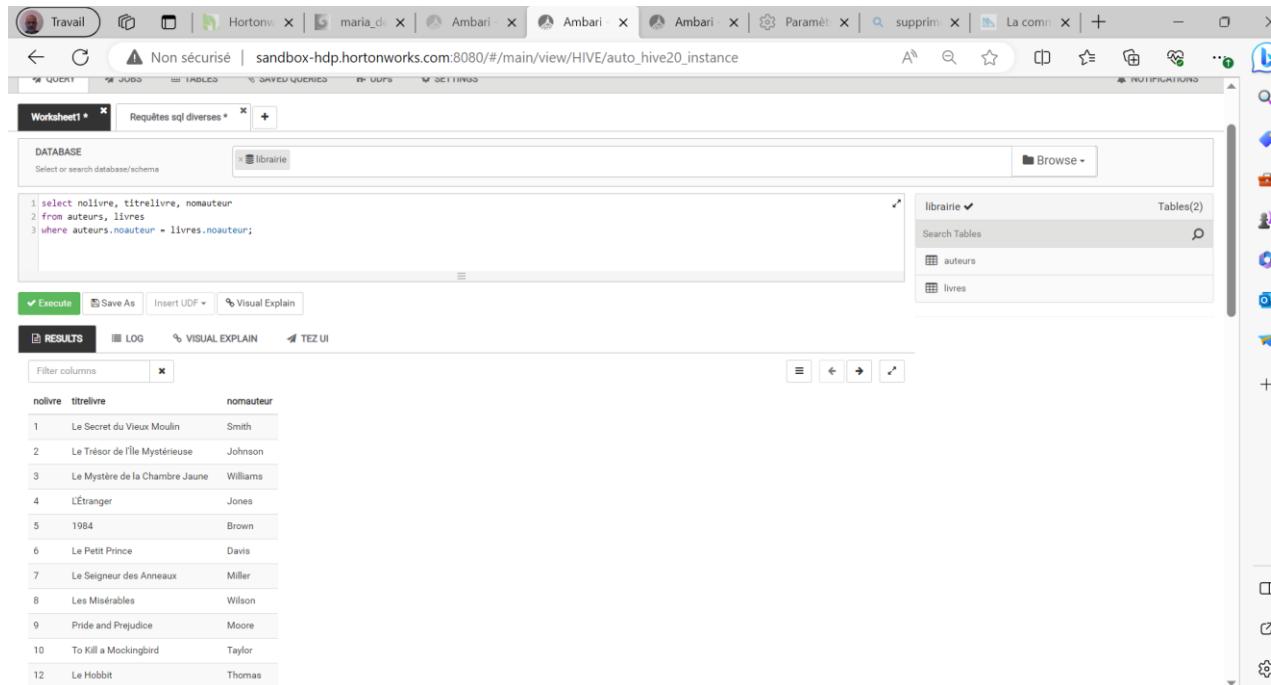
```
1 select noauteur, count(*)  
2 from livres  
3 group by noauteur;  
4 order by noauteur;
```

The results of this query are displayed in a table under the "RESULTS" tab:

noauteur	_c1
1	3
2	5
3	2
4	2
5	2
6	4
7	5
8	8
9	9
10	3
11	1

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

Requête de jointure



The screenshot shows the Apache Tez UI interface. In the top navigation bar, there are several tabs: Travail, Hortonworks, maria_db, Ambari, Ambari, Paramètres, supprimé, La commande, and a new tab. Below the tabs, the URL is sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance. The main area has a title "Worksheet1" and a sub-tab "Requêtes sql diverses". A search bar contains the text "librairie". On the left, a code editor displays the following SQL query:

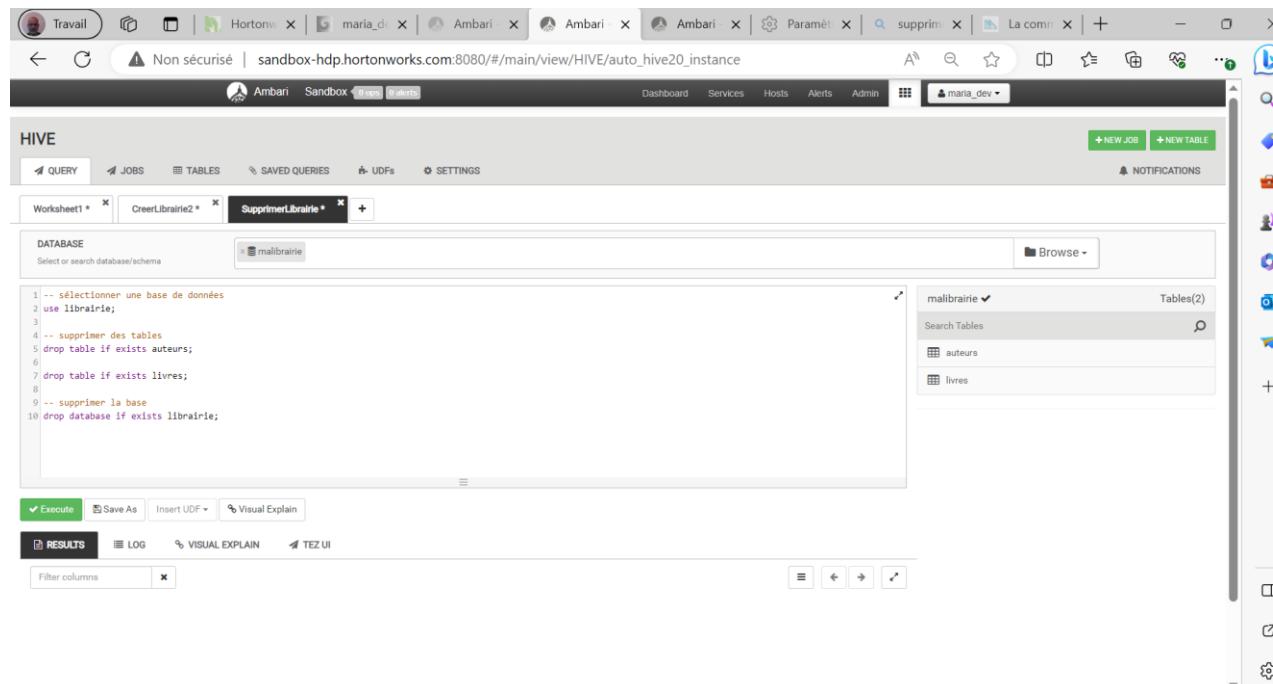
```
1 select nomlivre, titrelivre, nomauteur
2 from auteurs, livres
3 where auteurs.nomauteur = livres.nomauteur;
```

Below the code editor are buttons for "Execute", "Save As", "Insert UDF", and "Visual Explain". The "RESULTS" tab is selected. The results table has columns: nomlivre, titrelivre, and nom'auteur. The data is as follows:

nomlivre	titrelivre	nom'auteur
1	Le Secret du Vieux Moulin	Smith
2	Le Trésor de l'Île Mystérieuse	Johnson
3	Le Mystère de la Chambre Jaune	Williams
4	L'Étranger	Jones
5	1984	Brown
6	Le Petit Prince	Davis
7	Le Seigneur des Anneaux	Miller
8	Les Misérables	Wilson
9	Pride and Prejudice	Moore
10	To Kill a Mockingbird	Taylor
12	Le Hobbit	Thomas

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

Requêtes (DML) de suppression



The screenshot shows the Hortonworks Data Studio interface. In the top navigation bar, there are several tabs: Travail, Hortonworks, maria_db, Ambari, Ambari, Paramètres, supprim..., La comm..., and a search bar. Below the navigation bar, the main area is divided into two sections: a query editor on the left and a database browser on the right.

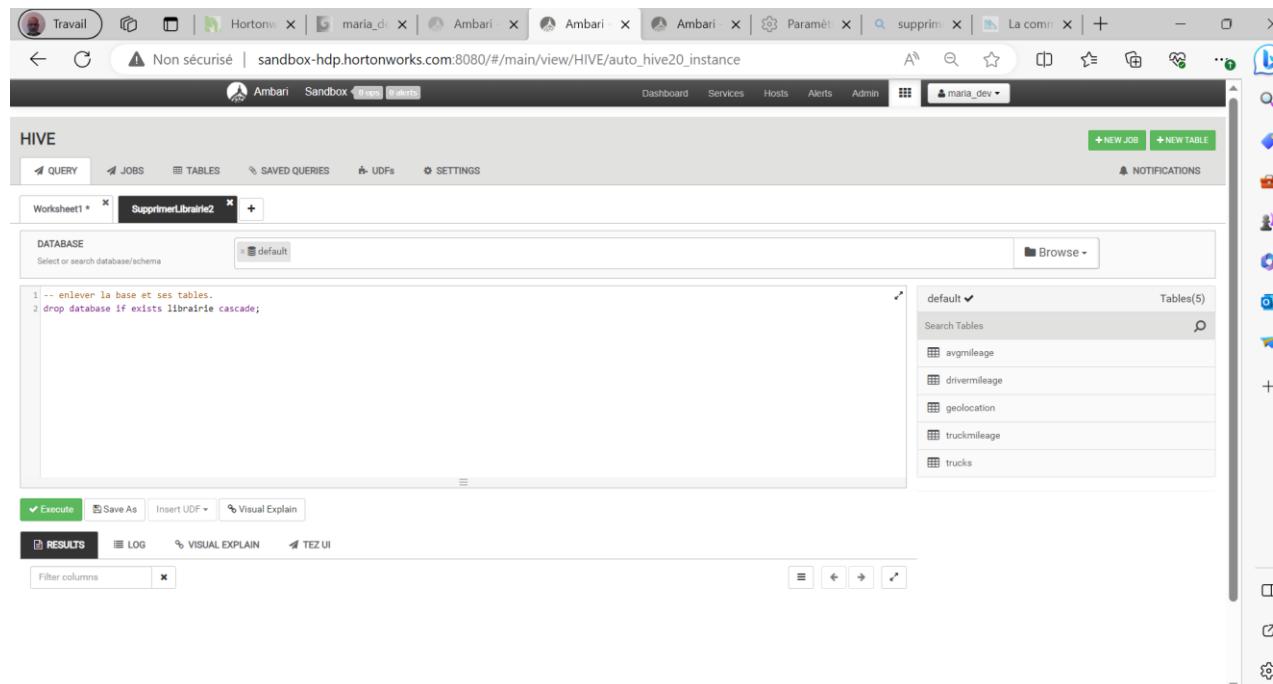
Query Editor: The title bar says "HIVE". It contains three tabs: Worksheet1*, CreerLibrairie2*, and SupprimerLibrairie*. The "SupprimerLibrairie*" tab is active. The code in the editor is:

```
1 -- sélectionner une base de données
2 use librairie;
3
4 -- supprimer des tables
5 drop table if exists auteurs;
6
7 drop table if exists livres;
8
9 -- supprimer la base
10 drop database if exists librairie;
```

Database Browser: The title bar says "maria_dev". It shows a tree view of the database structure under the schema "librairie". The "Tables(2)" node is expanded, showing "auteurs" and "livres".

- Supprimer la base librairie et ses tables.

Requêtes (DML) de suppression



The screenshot shows the Hortonworks Data Studio interface. In the top navigation bar, there are tabs for 'Travail', 'Ambari', 'Sandbox', 'Paramètres', 'supprim...', 'La comm...', and others. Below the navigation, the title bar says 'Non sécurisé | sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance'. The main area is divided into several sections: a 'HIVE' section with tabs for 'QUERY', 'JOBS', 'TABLES', 'SAVED QUERIES', 'UDFs', and 'SETTINGS'; a 'Worksheet1 *' tab containing a query editor with the following code:

```
1 -- enlever la base et ses tables,  
2 drop database if exists librairie cascade;
```

To the right of the query editor is a 'Tables(5)' browser window showing five tables: avgmileage, drivermileage, geolocation, truckmileage, and trucks. At the bottom of the interface, there are buttons for 'Execute', 'Save As', 'Insert UDF', 'Visual Explain', and tabs for 'RESULTS', 'LOG', 'VISUAL EXPLAIN', and 'TEZ UI'.

- 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 données 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

The screenshot shows the Ambari Hive interface. In the top navigation bar, there are several tabs: "Travail", "Hortonw", "maria_de", "Ambari", "Ambari", "Paramè", "supprim", "La comm", and a blank tab. Below the tabs, the URL is "Non sécurisé | sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance". The main area is titled "HIVE" and contains a "QUERY" tab. There are two tabs open: "Worksheet1" and "CreerLibrairie2". The "CreerLibrairie2" tab is active, showing the following DDL code:

```
1 -- création de la bd Librairie
2 create database if not exists librairie;
3
4 -- sélectionner la bd de travail
5 use librairie;
6
7 -- création de la table auteur
8 create table if not exists auteurs (noauteur int, nom'auteur string)
9 row format delimited
10 fields terminated by ',';
11
12 -- charger les données des auteurs
13 load data local inpath '/tmp/Librairie/Auteurs.txt' overwrite into table auteurs;
14
15 -- création de la table livre
16 create table if not exists livres (nolivre int, titrelivre string, noauteur int)
17 row format delimited
18 fields terminated by ',';
19
20 -- charger les données des livres
21 load data local inpath '/tmp/Librairie/Livres.txt' overwrite into table livres;
```

Below the code, there are buttons for "Execute", "Save As", "Insert UDF", and "Visual Explain". To the right of the code editor is a sidebar titled "librairie" which lists "Tables(2)": "auteurs" and "livres". The sidebar also includes a "Browse" button and a search bar.

- 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

- 1 74,La Trilogie de l'Empire,Raymond E. Feist#Janny Wurts
- 2 76,Le Cycle de l'Âge de la Mort,Margaret Weis#Tracy Hickman
- 3 78,La Trilogie de l'Éveil,Pauline Alphen

- Créer le fichier
livres2.txt

Copie de fichier

```
Invité de commandes
Microsoft Windows [version 10.0.19044.3324]
(c) Microsoft Corporation. Tous droits réservés.

C:\Users\asahraoui>cd Documents\BigData1\ExercicesMR

C:\Users\asahraoui\Documents\BigData1\ExercicesMR>scp -P 2222 Livres2.txt maria_dev@sandbox-hdp.hortonworks.com:/tmp/Librairie/Livres2.txt
maria_dev@sandbox-hdp.hortonworks.com's password:                                         100% 145   48.7KB/s  00:00
Livres2.txt

C:\Users\asahraoui\Documents\BigData1\ExercicesMR>scp -P 2222 Livres2.txt maria_dev@sandbox-hdp.hortonworks.com:/tmp/Librairie/Livres2.txt
maria_dev@sandbox-hdp.hortonworks.com's password:                                         100% 160   53.3KB/s  00:00
Livres2.txt

C:\Users\asahraoui\Documents\BigData1\ExercicesMR>
```

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

Requête DDL

The screenshot shows the Ambari Hive interface. On the left, a query editor window titled 'CreerLibrairie2' contains the following Hive DDL script:

```
1 -- création de la bd Librairie
2 create database if not exists librairie;
3
4 -- sélectionner la bd de travail
5 use librairie;
6
7 -- création de la table auteur
8 create table if not exists auteurs (noauteur int, nom'auteur string)
9 row format delimited
10 fields terminated by ',';
11
12 -- charger les données des auteurs
13 load data local inpath '/tmp/Librairie/Auteurs.txt' overwrite into table auteurs;
14
15 -- création de la table livre
16 create table if not exists livres (nolivre int, titrelivre string, noauteur int)
17 row format delimited
18 fields terminated by ',';
19
20 -- charger les données des livres
21 load data local inpath '/tmp/Librairie/Livres.txt' overwrite into table livres;
```

Below the query editor are buttons for 'Execute', 'Save As', 'Insert UDF', and 'Visual Explain'. To the right of the query editor is a sidebar titled 'librairie' showing tables: 'auteurs' and 'livres'. The top of the screen shows a browser bar with tabs for 'Travail', 'Hortonworks', 'maria_de...', 'Ambari', 'Ambari', 'Paramètres', 'supprim...', 'La comm...', and others.

- 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

The screenshot shows a web browser window with multiple tabs. The active tab is 'Ambari - Sandbox' at 'sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance'. The page displays the Ambari interface for managing a Hadoop cluster. On the left, there's a sidebar with icons for Ambari, Sandbox, Dashboard, Services, Hosts, Alerts, Admin, and a user dropdown for 'maria_dev'. The main area is titled 'HIVE' and contains tabs for QUERY, JOBS, TABLES, SAVED QUERIES, UDFs, and SETTINGS. A 'Worksheet1' tab is open, showing a query editor with the following SQL:

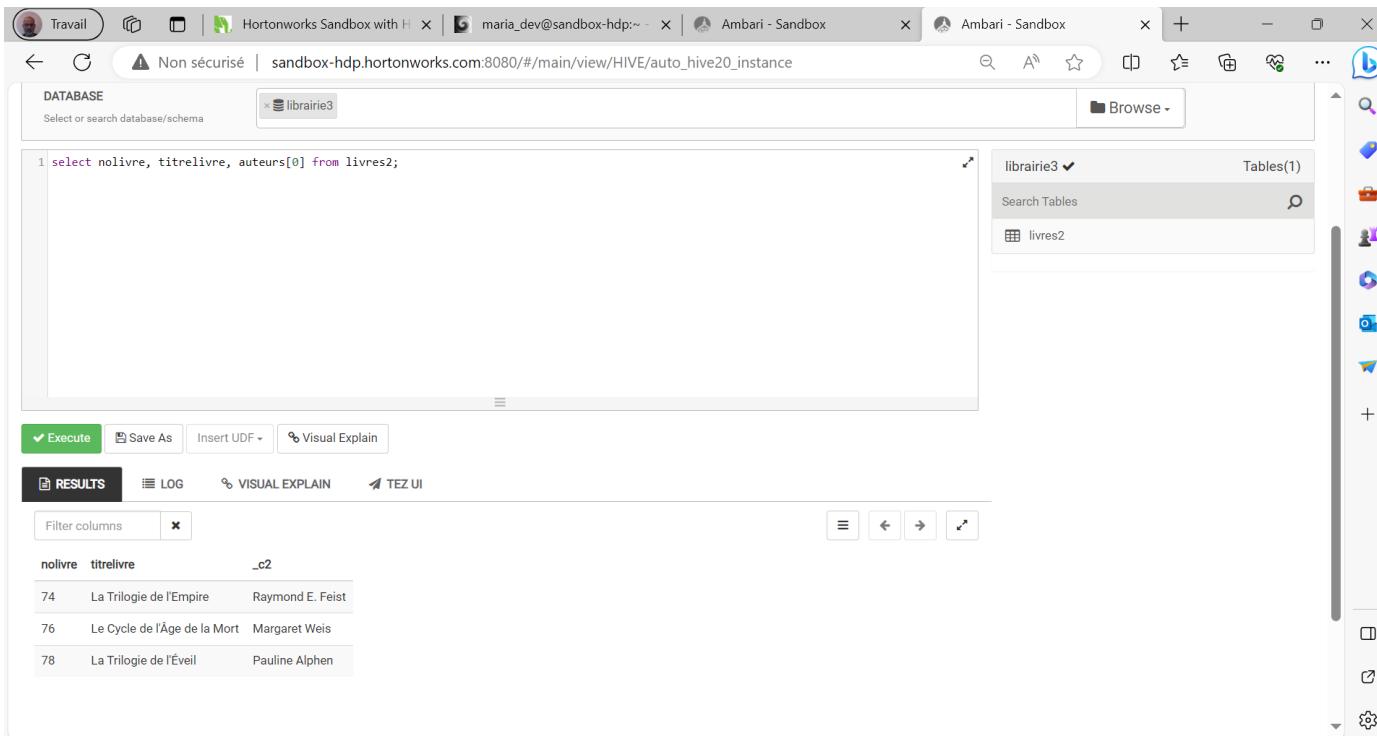
```
1 select * from livres2;
```

Below the query editor, there are buttons for Execute, Save As, Insert UDF, and Visual Explain. The 'RESULTS' tab is selected, displaying the output of the query:

livres2.nolivre	livres2.titrelivre	livres2.auteurs
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"]

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

Requête Interrpgation



The screenshot shows a web-based interface for running SQL queries against a Hive database. The URL is http://sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance. The database selected is 'librairie3'. The query entered is:

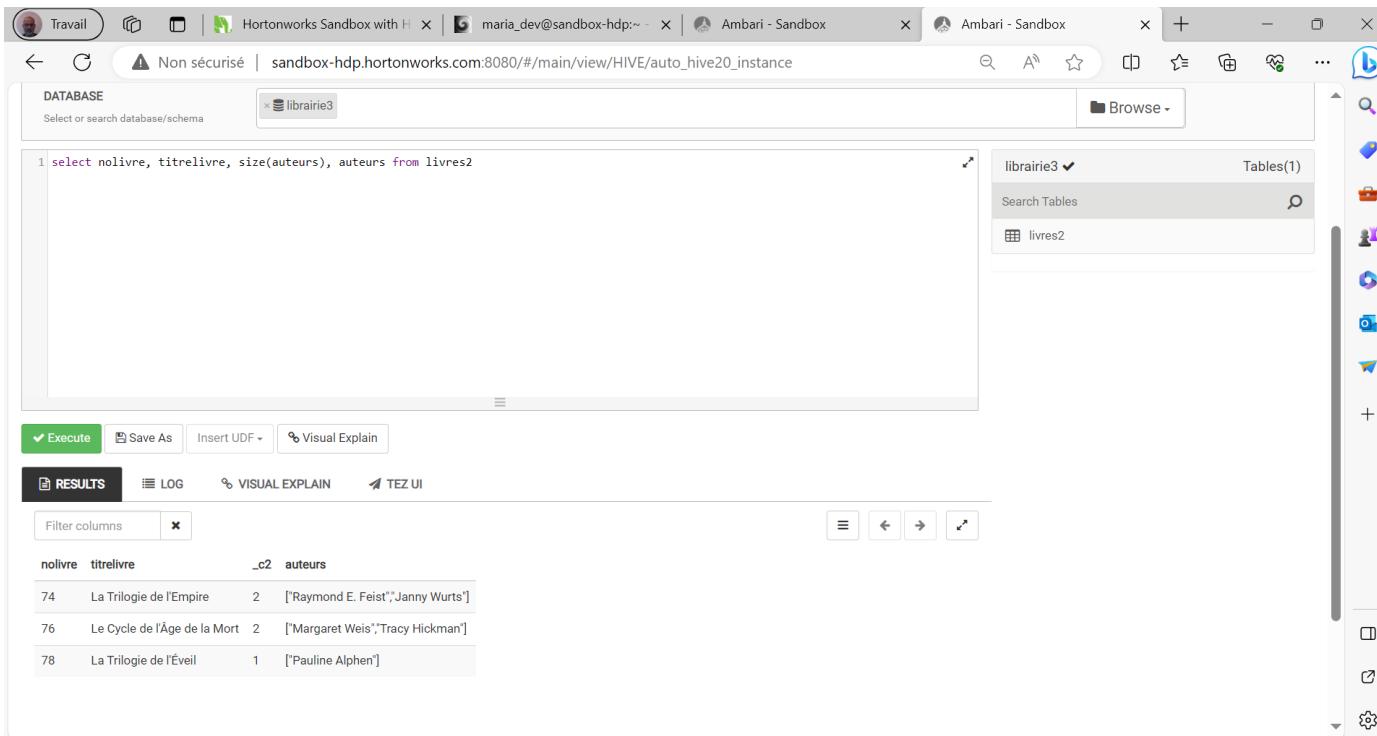
```
1 select nolivre, titrelivre, auteurs[0] from livres2;
```

The results table displays the following data:

nolivre	titrelivre	auteurs[0]
74	La Trilogie de l'Empire	Raymond E. Feist
76	Le Cycle de l'Âge de la Mort	Margaret Weis
78	La Trilogie de l'Éveil	Pauline Alphen

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

Requête Interrpgation



The screenshot shows a web browser window with multiple tabs. The active tab is 'Non sécurisé | sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance'. The page displays a Hive query results table. The query is:

```
1 select nolivre, titrelivre, size(auteurs), auteurs from livres2
```

The results table has columns: nolivre, titrelivre, _c2, auteurs. The data is:

nolivre	titrelivre	_c2	auteurs
74	La Trilogie de l'Empire	2	["Raymond E. Feist","Janny Wurts"]
76	Le Cycle de l'Âge de la Mort	2	["Margaret Weis","Tracy Hickman"]
78	La Trilogie de l'Éveil	1	["Pauline Alphen"]

- Exécuter la requête ci-dessus 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)

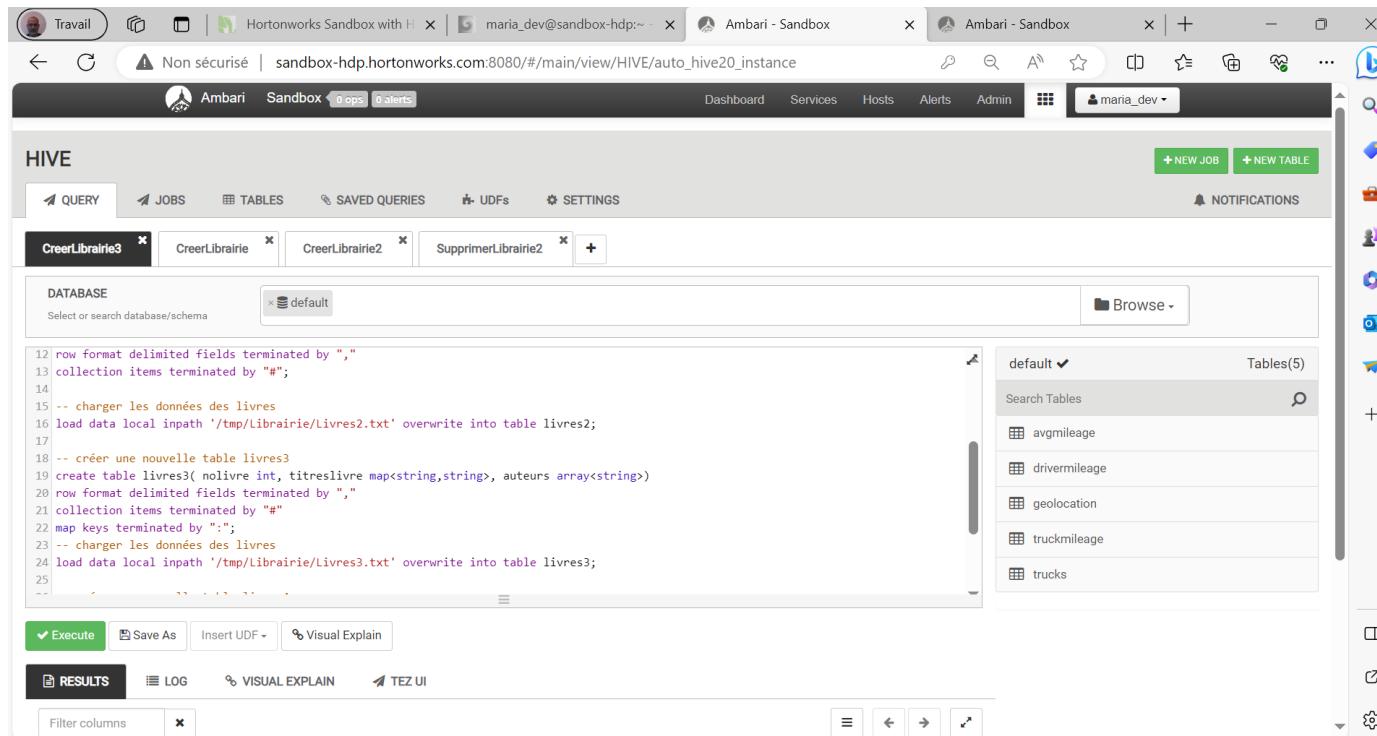
{"Français" : La trilogie de l'Empire, "Anglais" : "The Empire Trilogy" }

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



The screenshot shows a browser window with three tabs open:

- Hortonworks Sandbox with HDFS
- maria_dev@sandbox-hdp:~ ->
- Ambari - Sandbox

The middle tab, "maria_dev@sandbox-hdp:~ ->", displays the Ambari-Sandbox interface. The main area is titled "HIVE" and contains a query editor. The query editor has tabs for "QUERY", "JOBS", "TABLES", "SAVED QUERIES", "UDFs", and "SETTINGS". The "TABLES" tab is selected, showing a list of tables in the "default" database: avgmileage, drivermileage, geolocation, truckmileage, and trucks. Below the table list is a sidebar with various icons for file operations.

The query editor contains the following Hive DDL code:

```
12 row format delimited fields terminated by ","
13 collection items terminated by "#";
14
15 -- charger les données des livres
16 load data local inpath '/tmp/Librairie/Livres2.txt' overwrite into table livres2;
17
18 -- créer une nouvelle table livres
19 create table livres3(nolivre int, titreslivre map<string,string>, auteurs array<string>)
20 row format delimited fields terminated by ","
21 collection items terminated by "#"
22 map keys terminated by ";";
23 -- charger les données des livres
24 load data local inpath '/tmp/Librairie/Livres3.txt' overwrite into table livres3;
25
```

At the bottom of the query editor, there are buttons for "Execute", "Save As", "Insert UDF", and "Visual Explain".

- Taper la requête ci-dessous 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

```
1 74,Français:La Trilogie de l'Empire#Anglais:The Empire Trilogy,Raymond E. Feist#Janny Wurts
2 76,Français:Le Cycle de l'Âge de la Mort#Anglais:The Death Gate Cycle,Margaret Weis#Tracy Hickman
3 78,Français:La Trilogie de l'Éveil,Pauline Alphen
```

Copie de fichiers

```
invite de commandes
C:\Users\asahraoui\Documents\BigData1\ExercicesMR>scp -P 2222 Livres3.txt maria_dev@sandbox-hdp.hortonworks.com:/tmp/Librairie/Livres3.txt
maria_dev@sandbox-hdp.hortonworks.com's password:
Livres3.txt
100% 217 54.3KB/s 00:00
C:\Users\asahraoui\Documents\BigData1\ExercicesMR>
```

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

Requête DDL

The screenshot shows the Ambari Sandbox interface with the Hive query editor open. The query editor contains the following DDL code:

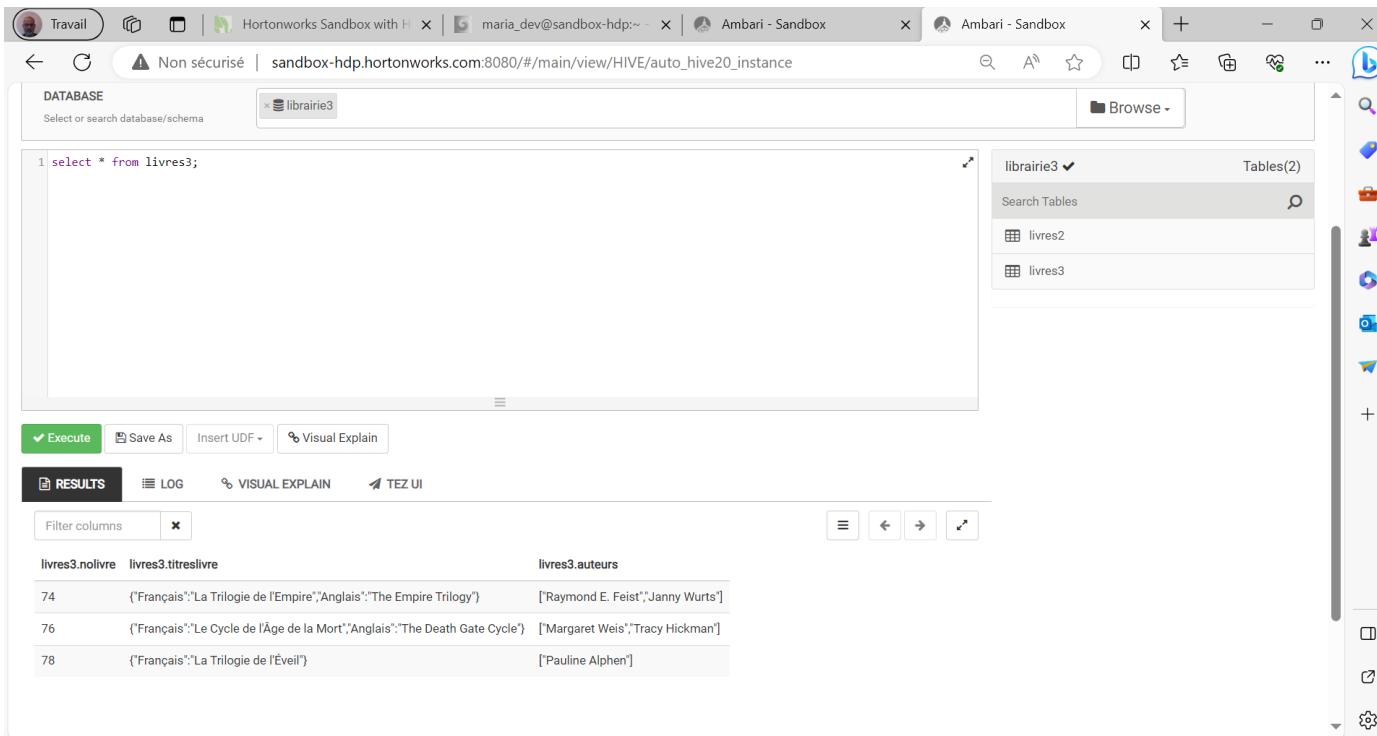
```
12 row format delimited fields terminated by ","
13 collection items terminated by "#";
14
15 -- charger les données des livres
16 load data local inpath '/tmp/Librairie/Livres2.txt' overwrite into table livres2;
17
18 -- créer une nouvelle table livres3
19 create table livres3( moliere int, titrelivre string, auteurs array<string>, titreOriginal map<string,string>)
20 row format delimited fields terminated by ","
21 collection items terminated by "#"
22 map keys terminated by ":";
23
24 -- charger les données des livres
25 load data local inpath '/tmp/Librairie/Livres3.txt' overwrite into table livres3;
```

The sidebar on the right shows the following database structure:

- default (Tables 5)
 - avgmileage
 - drivermileage
 - geolocation
 - truckmileage
 - trucks

- 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



The screenshot shows the Apache Tez UI interface for executing Hive queries. The top navigation bar includes tabs for 'Travail', 'Hortonworks Sandbox with H...', 'maria_dev@sandbox-hdp:~ -', 'Ambari - Sandbox', and 'Ambari - Sandbox'. The main area has tabs for 'RESULTS', 'LOG', 'VISUAL EXPLAIN', and 'TEZ UI'. A sidebar on the right lists databases ('librairie3', 'livres2', 'livres3') and provides icons for search, refresh, and other functions.

In the 'RESULTS' tab, a query is being run:

```
1 select * from livres3;
```

The results table displays the following data:

livres3.nolivre	livres3.titreslivre	livres3.auteurs
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"]

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

Requête Interrogation

The screenshot shows a web-based interface for running Hive queries. The URL in the address bar is `sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance`. The query entered in the editor is:

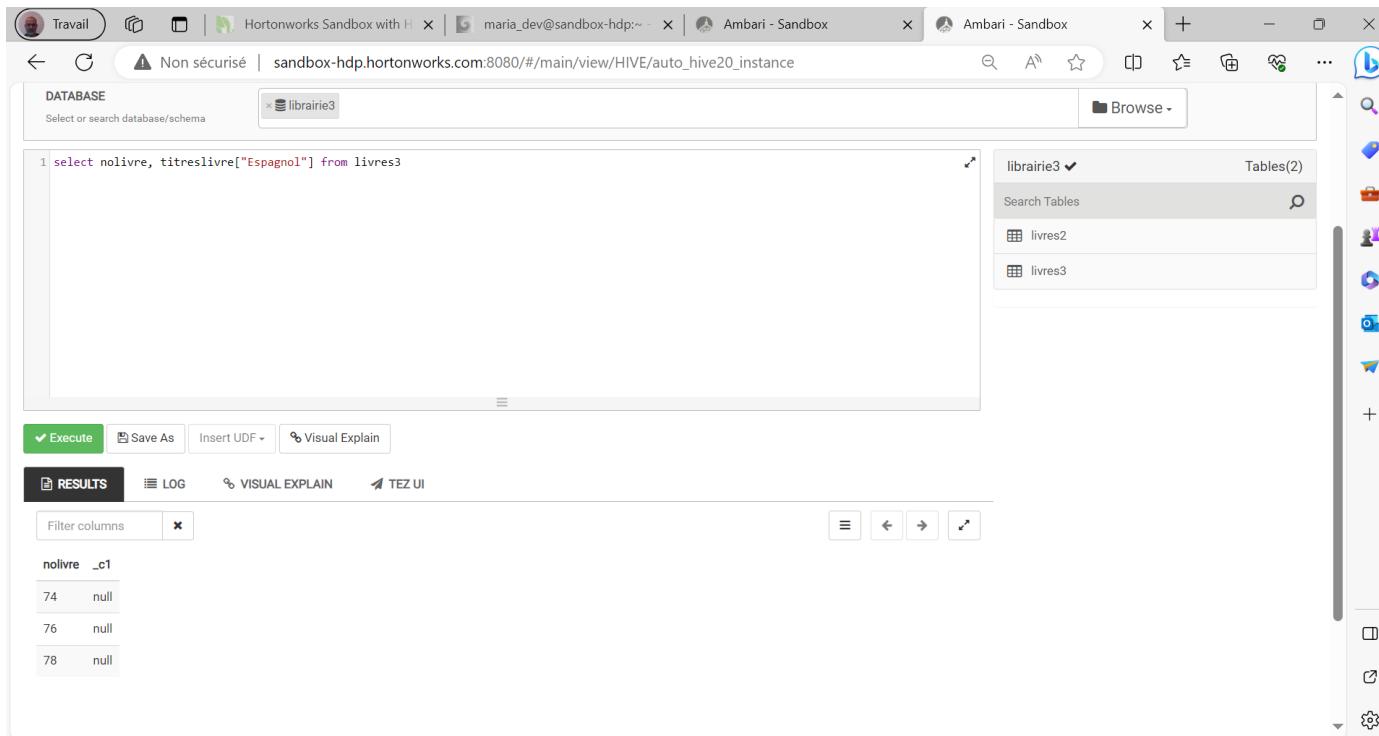
```
1 select nolivre, titreslivre["Anglais"] from livres3
```

The results table displays the following data:

nolivre	_c1
74	The Empire Trilogy
76	The Death Gate Cycle
78	null

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

Requête Interrogation



The screenshot shows the Hortonworks Sandbox Hive interface. The top navigation bar includes tabs for 'Travail', 'Hortonworks Sandbox with H...', 'maria_dev@sandbox-hdp:~ -', and 'Ambari - Sandbox'. The main window has a title 'Non sécurisé | sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance'. The left panel displays a query editor with the following SQL:

```
1 select nolivre, titreslivre["Espagnol"] from livres3
```

The right panel shows the results of the query, which currently displays no data (null values). The sidebar on the right contains various icons for database management.

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

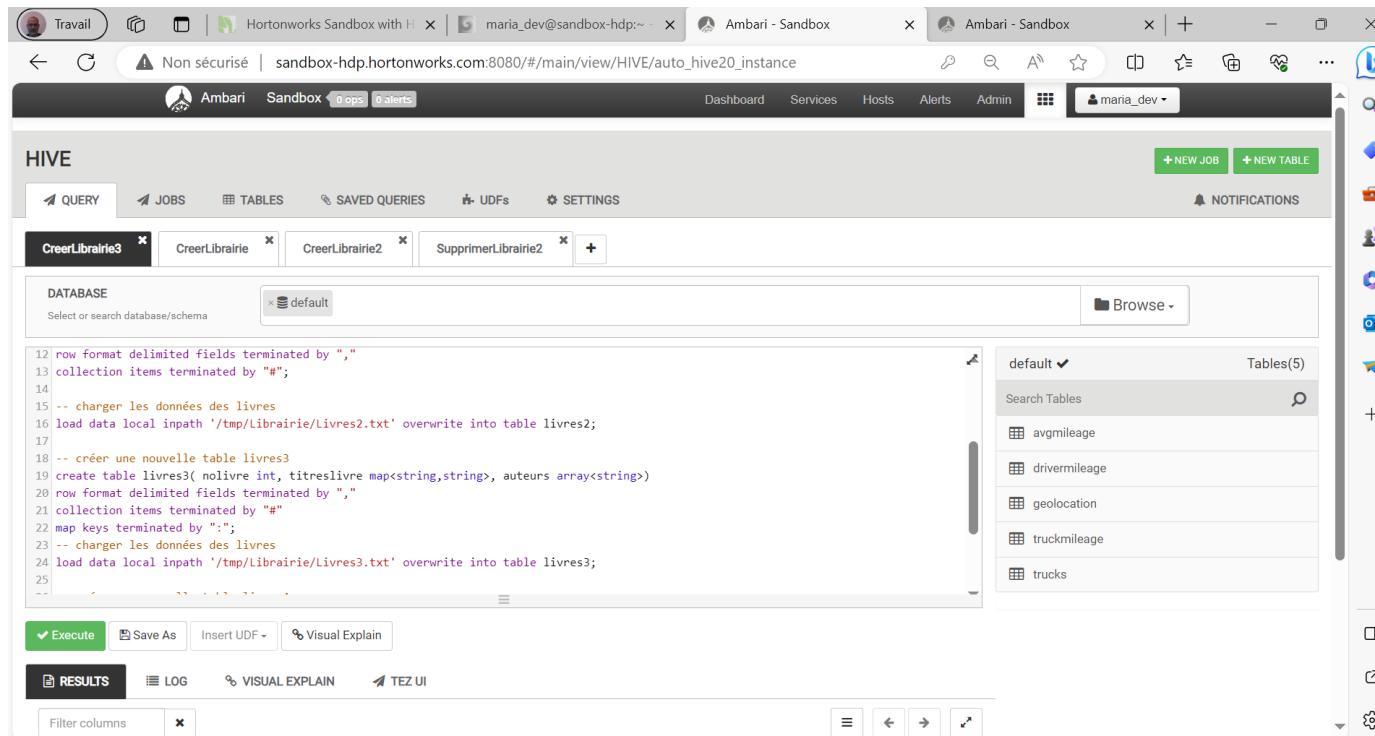
Types complexes

- 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 donnée 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



The screenshot shows a browser window with three tabs open:

- Hortonworks Sandbox with HDFS
- maria_dev@sandbox-hdp:~ ->
- Ambari - Sandbox

The main content area is a Hive query editor titled "HIVE". It has tabs for QUERY, JOBS, TABLES, SAVED QUERIES, UDFs, and SETTINGS. The QUERY tab is active, showing the following DDL script:

```
12 row format delimited fields terminated by ","
13 collection items terminated by "#";
14
15 -- charger les données des livres
16 load data local inpath '/tmp/Librairie/Livres2.txt' overwrite into table livres2;
17
18 -- créer une nouvelle table livres
19 create table livres3(nolivre int, titreslivre map<string,string>, auteurs array<string>)
20 row format delimited fields terminated by ","
21 collection items terminated by "#"
22 map keys terminated by ";";
23 -- charger les données des livres
24 load data local inpath '/tmp/Librairie/Livres3.txt' overwrite into table livres3;
25
```

Below the script are buttons for Execute, Save As, Insert UDF, and Visual Explain. The RESULTS tab is selected. To the right, there is a sidebar with a "Tables(5)" section containing a list of tables: avgmileage, drivermileage, geolocation, truckmileage, and trucks. A "Search Tables" input field is also present.

- Taper la requête ci-dessous 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

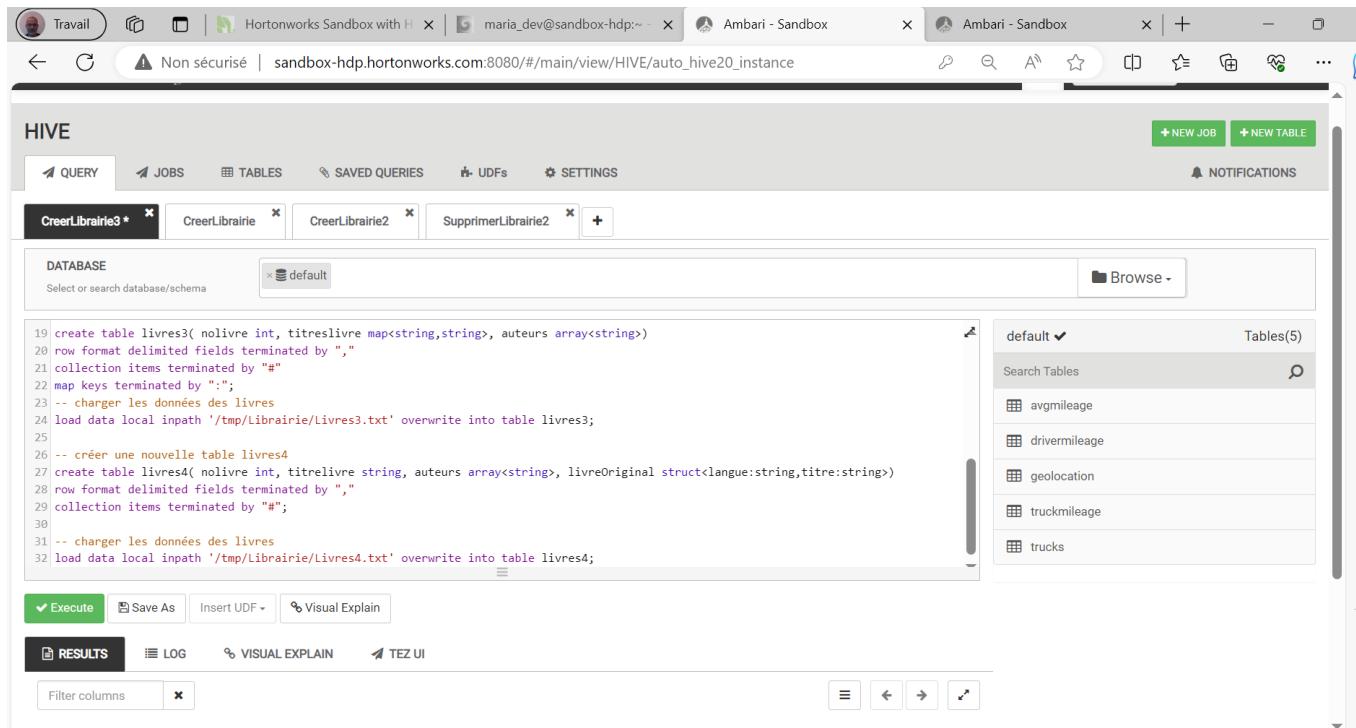
```
≡ Livres4.txt
1 74,La Trilogie de l'Empire,Raymond E. Feist#Janny Wurts,Anglais#The Empire Trilogy
2 76,Le Cycle de l'Âge de la Mort,Margaret Weis#Tracy Hickman,Anglais#The Death Gate Cycle
3 78,La Trilogie de l'Éveil,Pauline Alphen
```

Copie de fichiers

```
invite de commandes
C:\Users\asahraoui\Documents\BigData1\ExercicesMR>scp -P 2222 Livres4.txt maria_dev@sandbox-hdp.hortonworks.com:/tmp/Librairie/Livres4.txt
maria_dev@sandbox-hdp.hortonworks.com's password:
Livres4.txt
100% 217 108.6KB/s 00:00
C:\Users\asahraoui\Documents\BigData1\ExercicesMR>
```

- Copier le fichier `Livres4.txt` dans `/tmp/Librairie`

Requête DDL



The screenshot shows a browser window with two tabs open. The left tab is titled "Hortonworks Sandbox with HDFS" and the right tab is titled "Ambari - Sandbox". The main content area is a Hive query editor titled "HIVE". It has tabs for "QUERY", "JOBS", "TABLES", "SAVED QUERIES", "UDFs", and "SETTINGS". A search bar at the top says "Non sécurisé | sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance". Below the tabs, there are several tabs for queries: "CreerLibrairie3", "CreerLibrairie", "CreerLibrairie2", "SupprimerLibrairie2", and a "+" button. A "DATABASE" dropdown is set to "default". The main pane contains the following Hive DDL code:

```
19 create table livres3( nolivre int, titreslivre map<string,string>, auteurs array<string>)
20 row format delimited fields terminated by ","
21 collection items terminated by "#"
22 map keys terminated by ":";
23 -- charger les données des livres
24 load data local inpath '/tmp/Librairie/Livres3.txt' overwrite into table livres3;
25
26 -- créer une nouvelle table livres4
27 create table livres4( nolivre int, titrelivre string, auteurs array<string>, livreOriginal struct<langue:string,titre:string>)
28 row format delimited fields terminated by ","
29 collection items terminated by "#";
30
31 -- charger les données des livres
32 load data local inpath '/tmp/Librairie/Livres4.txt' overwrite into table livres4;
```

Below the code, there are buttons for "Execute", "Save As", "Insert UDF", and "Visual Explain". The "RESULTS" tab is selected. At the bottom, there are buttons for "Filter columns" and navigation icons. To the right of the editor, there is a sidebar titled "NOTIFICATIONS" with a green bell icon. The sidebar also contains a "Tables(5)" section with a search bar and a list of tables: avgmileage, drivermileage, geolocation, truckmileage, and trucks.

- 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

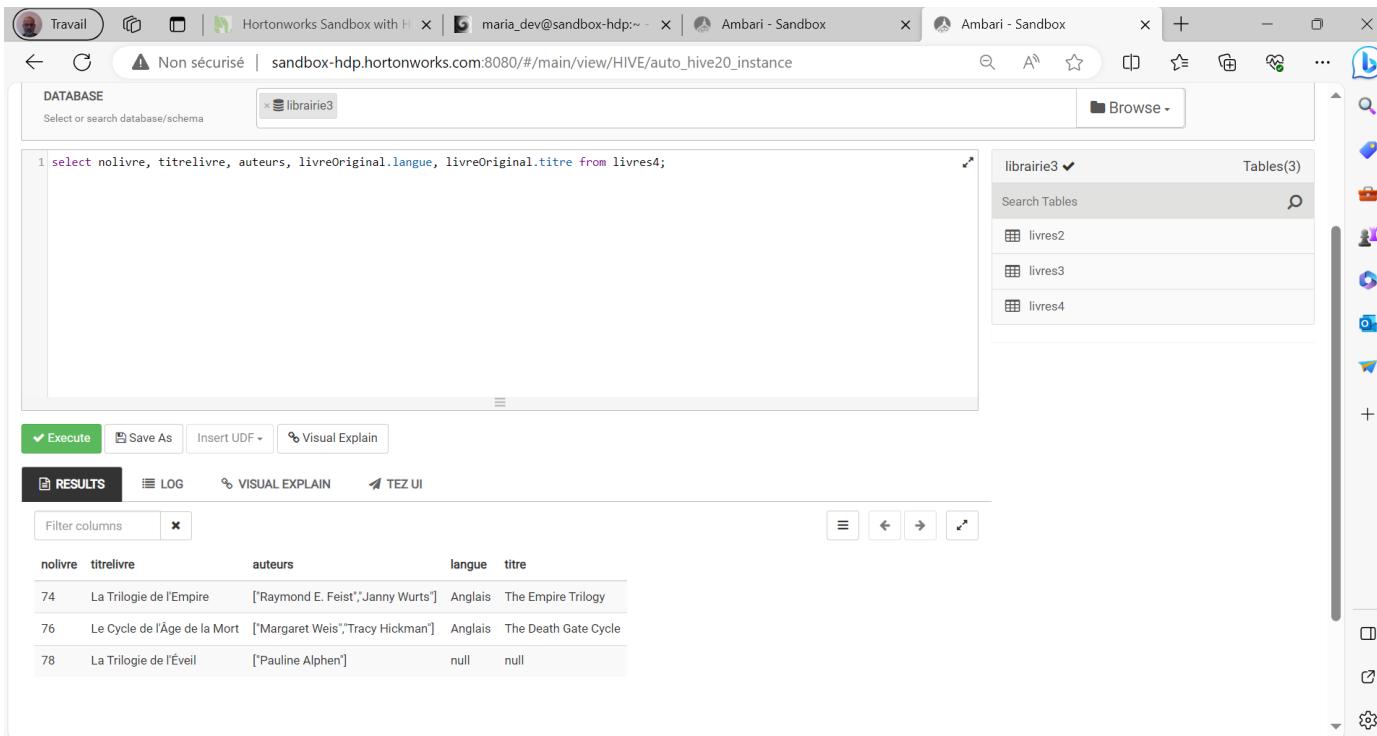
Requête Interrogation

```
1 select * from livres4
```

livres4.nolivre	livres4.titrelivre	livres4.auteurs	livres4.livreoriginal
74	La Trilogie de l'Empire	["Raymond E. Feist","Janny Wurts"]	{"langue":"Anglais","titre":"The Empire Trilogy"}
76	Le Cycle de l'Âge de la Mort	["Margaret Weis","Tracy Hickman"]	{"langue":"Anglais","titre":"The Death Gate Cycle"}
78	La Trilogie de l'Éveil	["Pauline Alphen"]	null

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

Requête Interrogation



Non sécurisé | sandbox-hdp.hortonworks.com:8080/#/main/view/HIVE/auto_hive20_instance

DATABASE librairie3

```
1 select nolivre, titrelivre, auteurs, livreOriginal.langue, livreOriginal.titre from livres4;
```

Execute Save As Insert UDF Visual Explain

RESULTS LOG VISUAL EXPLAIN TEZ UI

nolivre	titrelivre	auteurs	langue	titre
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"]	null	null

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