



COLLÈGE
ROSEMONT

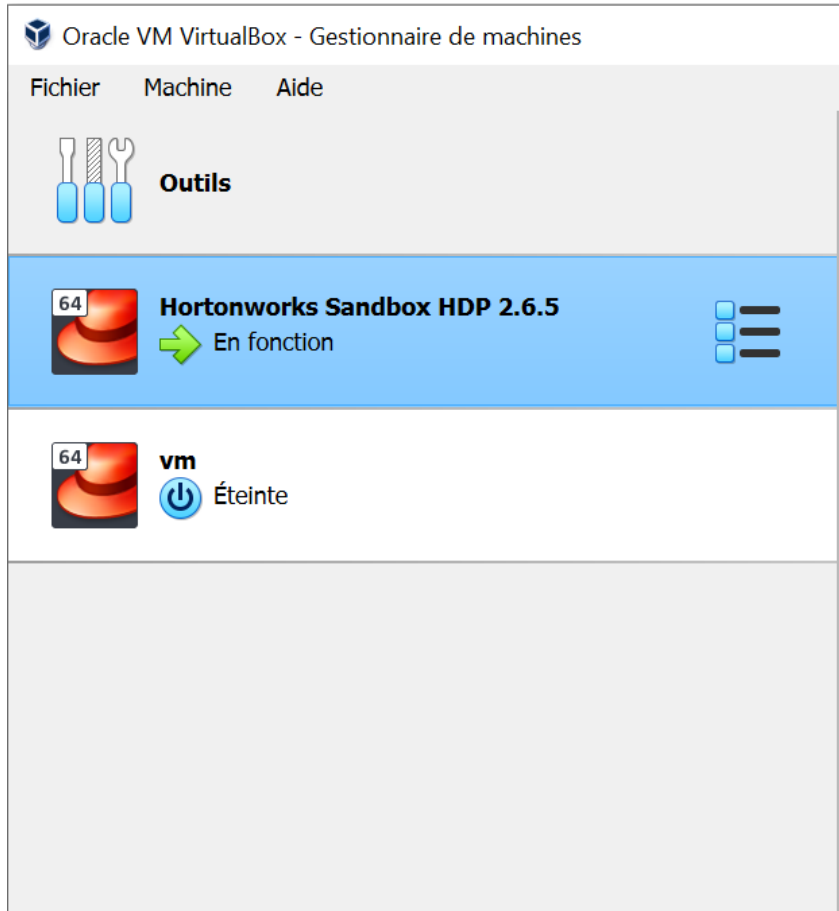
Valorisation de données

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**

Connexion à Ambari par le serveur local



Sign in

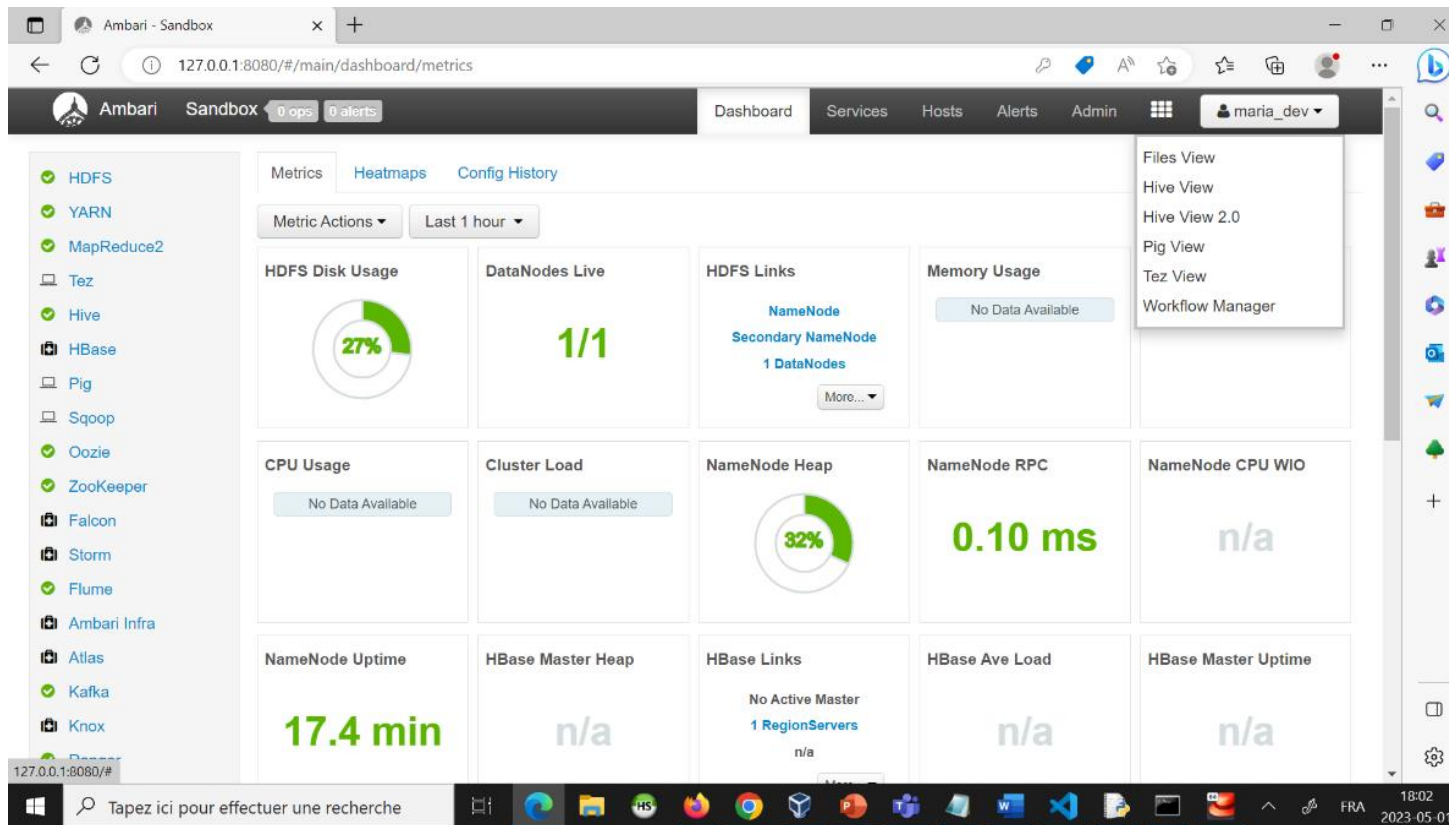
Username

Password

Sign in

Se connecter à
Ambari avec l'utilisateur
maria_dev (password
maria_dev)

Ambari



Lancer
Files View

Gestion de fichiers sur HDFS

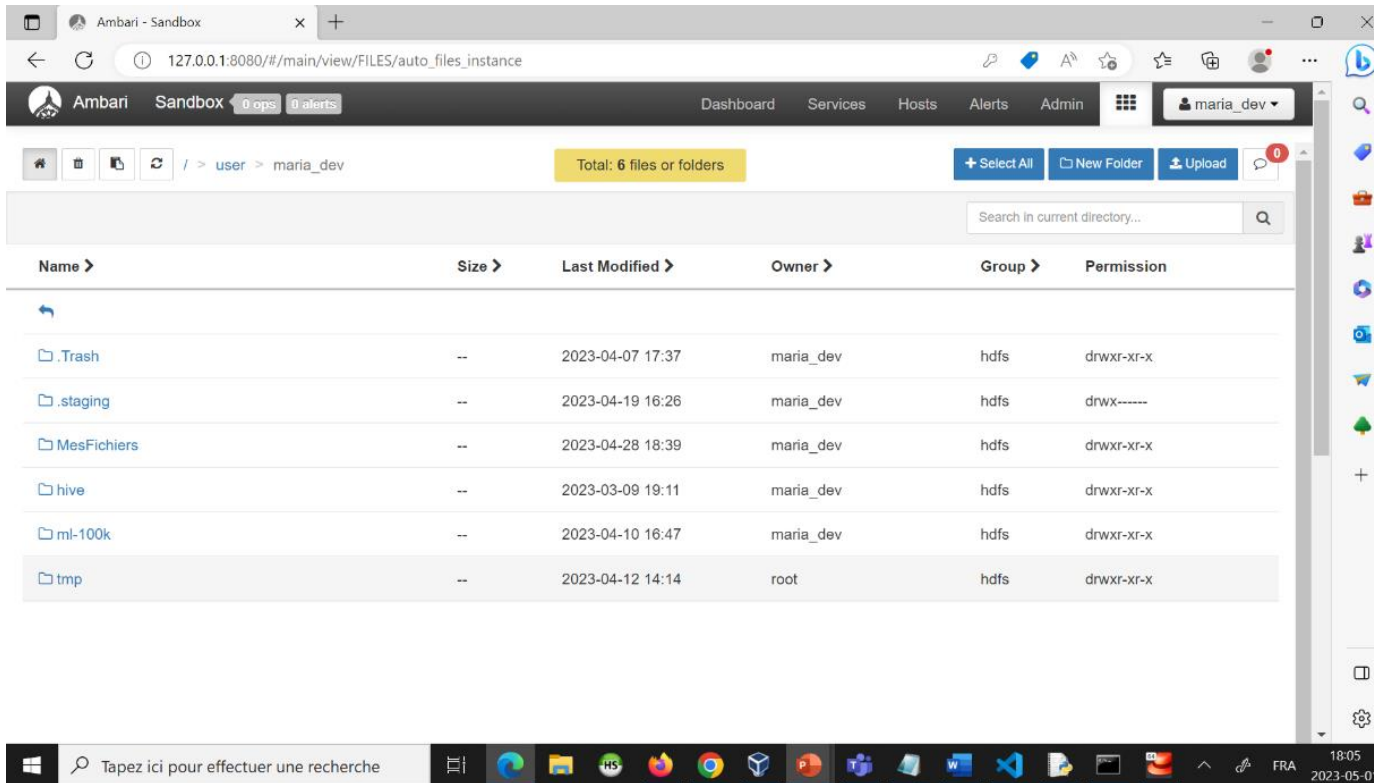
The screenshot shows the Ambari Sandbox web interface. The breadcrumb navigation indicates the current path is / > user > maria_dev. A yellow status bar shows 'Total: 6 files or folders'. Below this is a table listing the contents of the directory:

Name >	Size >	Last Modified >	Owner >	Group >	Permission
↶					
📁 .Trash	--	2023-04-07 17:37	maria_dev	hdfs	drwxr-xr-x
📁 .staging	--	2023-04-19 16:26	maria_dev	hdfs	drwx-----
📁 MesFichiers	--	2023-04-28 18:39	maria_dev	hdfs	drwxr-xr-x
📁 hive	--	2023-03-09 19:11	maria_dev	hdfs	drwxr-xr-x
📁 ml-100k	--	2023-04-10 16:47	maria_dev	hdfs	drwxr-xr-x
📁 tmp	--	2023-04-12 14:14	root	hdfs	drwxr-xr-x

The interface includes standard file management actions like '+ Select All', 'New Folder', and 'Upload'. The bottom of the image shows a Windows taskbar with various application icons and a search bar.

Ouvrir **user**
puis **maria_dev**

Gestion de fichiers sur HDFS



The screenshot shows the Ambari Sandbox web interface for file management. The browser address bar shows the URL `127.0.0.1:8080/#/main/view/FILES/auto_files_instance`. The interface includes a navigation bar with tabs for Dashboard, Services, Hosts, Alerts, and Admin. The user is logged in as `maria_dev`. The main content area displays a file browser for the `user > maria_dev` directory, showing a total of 6 files or folders. A search bar is available for searching the current directory. Below the search bar is a table listing the files and folders.

Name >	Size >	Last Modified >	Owner >	Group >	Permission
↶					
📁 .Trash	--	2023-04-07 17:37	maria_dev	hdfs	drwxr-xr-x
📁 .staging	--	2023-04-19 16:26	maria_dev	hdfs	drwx-----
📁 MesFichiers	--	2023-04-28 18:39	maria_dev	hdfs	drwxr-xr-x
📁 hive	--	2023-03-09 19:11	maria_dev	hdfs	drwxr-xr-x
📁 ml-100k	--	2023-04-10 16:47	maria_dev	hdfs	drwxr-xr-x
📁 tmp	--	2023-04-12 14:14	root	hdfs	drwxr-xr-x

Créer le dossier
MesFichiers et
charger
VictorHugo.txt

Hive View

The screenshot shows the Ambari web interface in a browser window. The address bar displays the URL `127.0.0.1:8080/#/main/view/FILES/auto_files_instance`. The top navigation bar includes links for Dashboard, Services, Hosts, Alerts, and Admin. A user profile dropdown for 'maria_dev' is open, showing a menu with the following options: Files View, Hive View, Hive View 2.0, Pig View, Tez View, and Workflow Manager. The 'Hive View' option is highlighted. Below the navigation bar, the breadcrumb path is `/ > user > maria_dev > MesFichiers`, and a yellow box indicates 'Total: 1 files or folders'. A table lists the files in the directory:

Name >	Size >	Last Modified >	Owner >	Group >	Permissions
VictorHugo.txt	5.3 kB	2023-05-01 18:40	maria_dev	hdfs	-rw-r--r--

The Windows taskbar at the bottom shows the time as 18:41 on 2023-05-01.

Lancer
Hive View

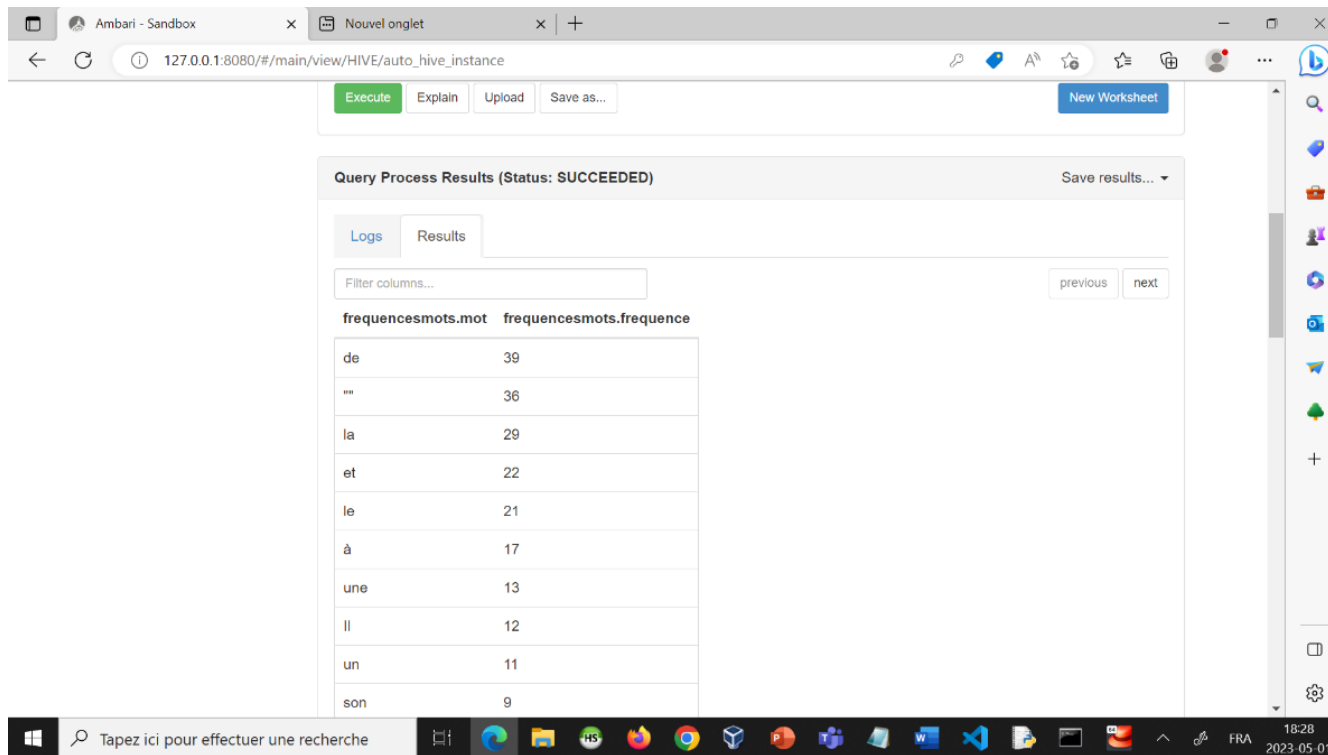
Requête Hive

```
1 --Requête pour créer une tables contenant les mots d'un fichier texte
2 --et les fréquences de ces mots dans le texte
3 DROP TABLE IF EXISTS MonTexte;
4 DROP TABLE IF EXISTS FrequencesMots;
5 CREATE TABLE MonTexte (line STRING);
6 LOAD DATA INPATH '/user/maria_dev/MesFichiers/VictorHugo.txt' OVERWRITE INTO TABLE MonTexte;
7 CREATE TABLE FrequencesMots AS
8 SELECT mot, count(1) AS frequence FROM
9 (SELECT explode(split(line, ' ')) AS mot FROM MonTexte) temp
10 GROUP BY mot
11 ORDER BY mot;
12 SELECT * from FrequencesMots ORDER BY frequence DESC;
```

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

Exécuter la requête et observer le résultat.

Résultats Requête



Query Process Results (Status: SUCCEEDED)

Save results...

Logs Results

Filter columns...

previous next

frequencesmots.mot	frequencesmots.frequency
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 ordre
décroissant des
fréquences**

Table frequencemots

Ambari - Sandbox | Nouvel onglet | 127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance

Ambari | Sandbox | 0 ops | 0 alerts | Dashboard | Services | Hosts | Alerts | Admin | maria_dev

Hive | Query | Saved Queries | History | UDFs | Upload Table

Database Explorer

default

Search tables...

Databases

- default
- docs
- frequencemots
- montexte
- movies
- ratings
- word counts
- foodmart

Query Editor

HiveQryCompteMots

```
1 --Requête pour créer une tables contenant les mots d'un fichier texte
2 --et les fréquences de ces mots dans le texte
3 DROP TABLE IF EXISTS FrequencesMots;
4 DROP TABLE IF EXISTS MonTexte;
5 CREATE TABLE MonTexte (line STRING);
6 LOAD DATA INPATH '/user/maria_dev/MesFichiers/VictorHugo.txt' OVERWRITE INTO TABLE MonTexte;
7 CREATE TABLE FrequencesMots AS
8 SELECT mot, count(1) AS frequency FROM
9   (SELECT explode(split(line, ' ')) AS mot FROM MonTexte) temp
10 GROUP BY mot
11 ORDER BY mot;
12 SELECT * from FrequencesMots ORDER BY frequency DESC;
```

Execute | Explain | Upload | Save as... | New Worksheet

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

Table frequencemots

The screenshot displays the Ambari web interface in a browser window. The address bar shows the URL `127.0.0.1:8080/#/main/view/HIVE/auto_hive_instance`. The interface is divided into several sections:

- Database Explorer:** Located on the left, it shows a tree view of databases. The 'default' database is selected, and a list of tables is visible, including 'docs', 'frequencemots', 'montexte', 'movies', 'ratings', 'word counts', and 'foodmart'.
- Query Editor:** The central area where a SQL query is entered. The query is `SELECT * FROM frequencemots LIMIT 100;`. Below the editor are buttons for 'Execute', 'Explain', 'Upload', 'Save as...', and 'New Worksheet'.
- Query Process Results:** At the bottom, it shows the status 'SUCCEEDED' and a 'Save results...' button. There are tabs for 'Logs' and 'Results'.
- Right Sidebar:** Contains various tool icons, including 'SQL', 'TEZ', and a mail icon with a red notification badge.

The Windows taskbar at the bottom shows the search bar with the text 'Tapez ici pour effectuer une recherche' and several application icons. The system clock indicates the time is 18:35 on 2023-05-01.

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>