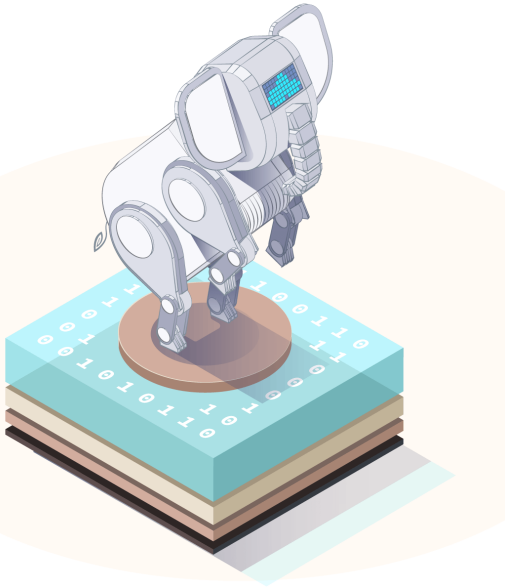
OVH Academy
Lab Guide

Discovery scenario

OVH Analytics Data Platform

We will size and deploy a big data Hadoop cluster, then perform a use case allowing us to discover basic features : user and rights, dashboard, ingest data, queries on data.



ALL the credentials, useful links, and code,

We be available on

https://github.com/baaastijn/workshops

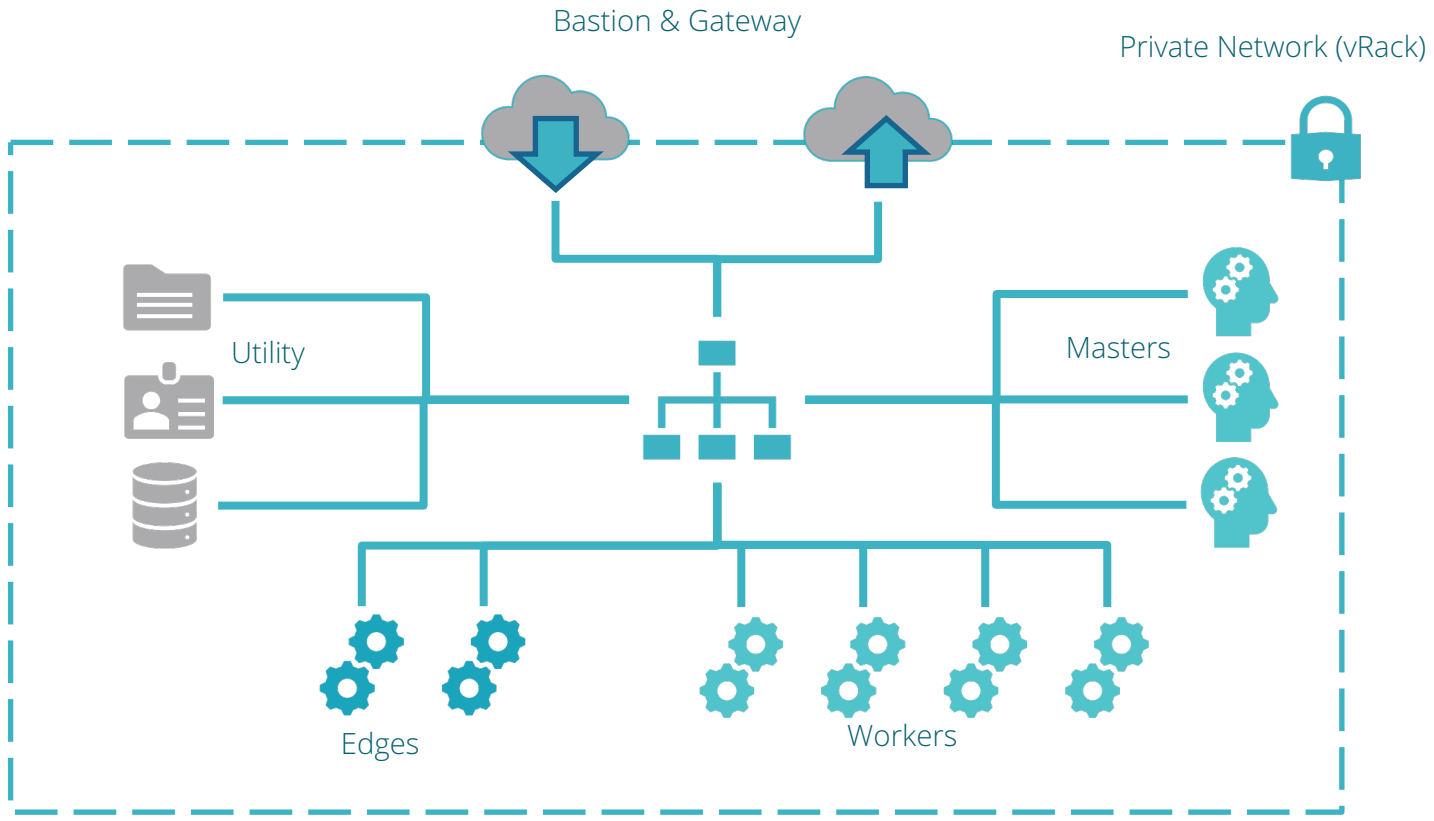
Discover the ADP product !

1. Infrastructure
2. Launch a cluster
3. Discover the Ambari dashboard
4. First steps with user management
5. First steps with data security

Step 1: Estimate your needs, size you cluster

We will explain it live :

- ✓ Size the Workers
- ✓ Size the Edges
- ✓ Site the Storage



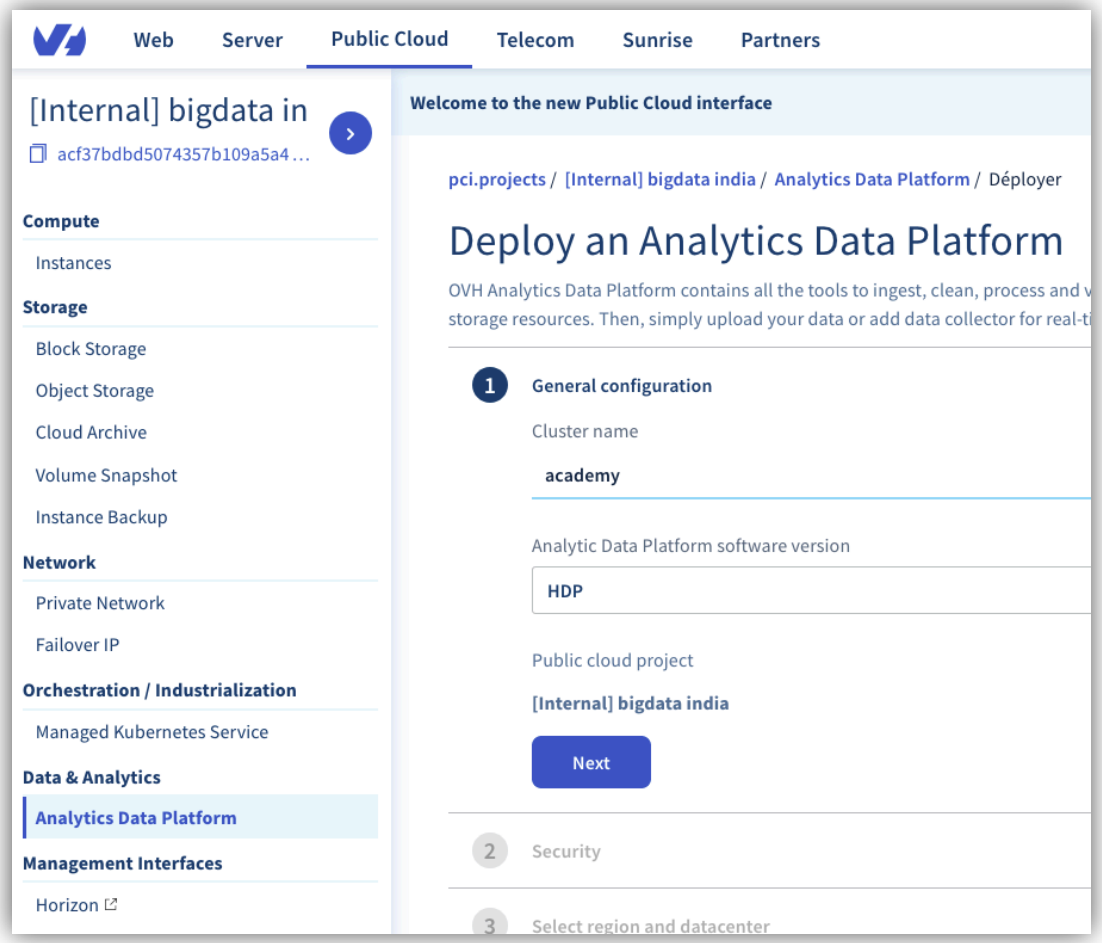
Step 2 : launch your OVH cluster (live demo)

A cluster deployment takes approximately 40 minutes.

https://www.ovh.com/fr/public-cloud/big-data-hadoop/

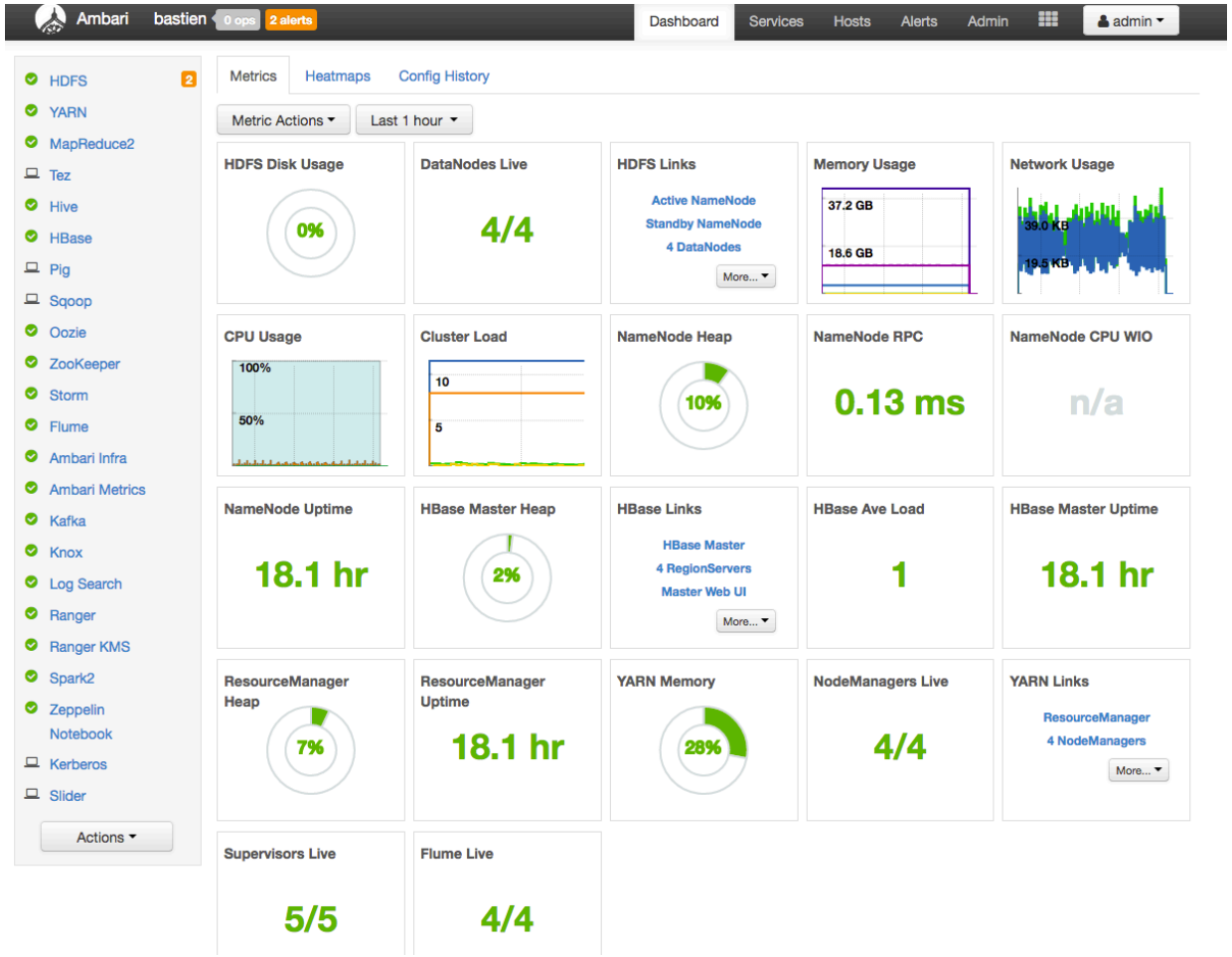
OVH will deploy ADP cluster for you ☺

You will find the credentials on the OVH control panel



Step 3 : Connect to you cluster, discover Ambari

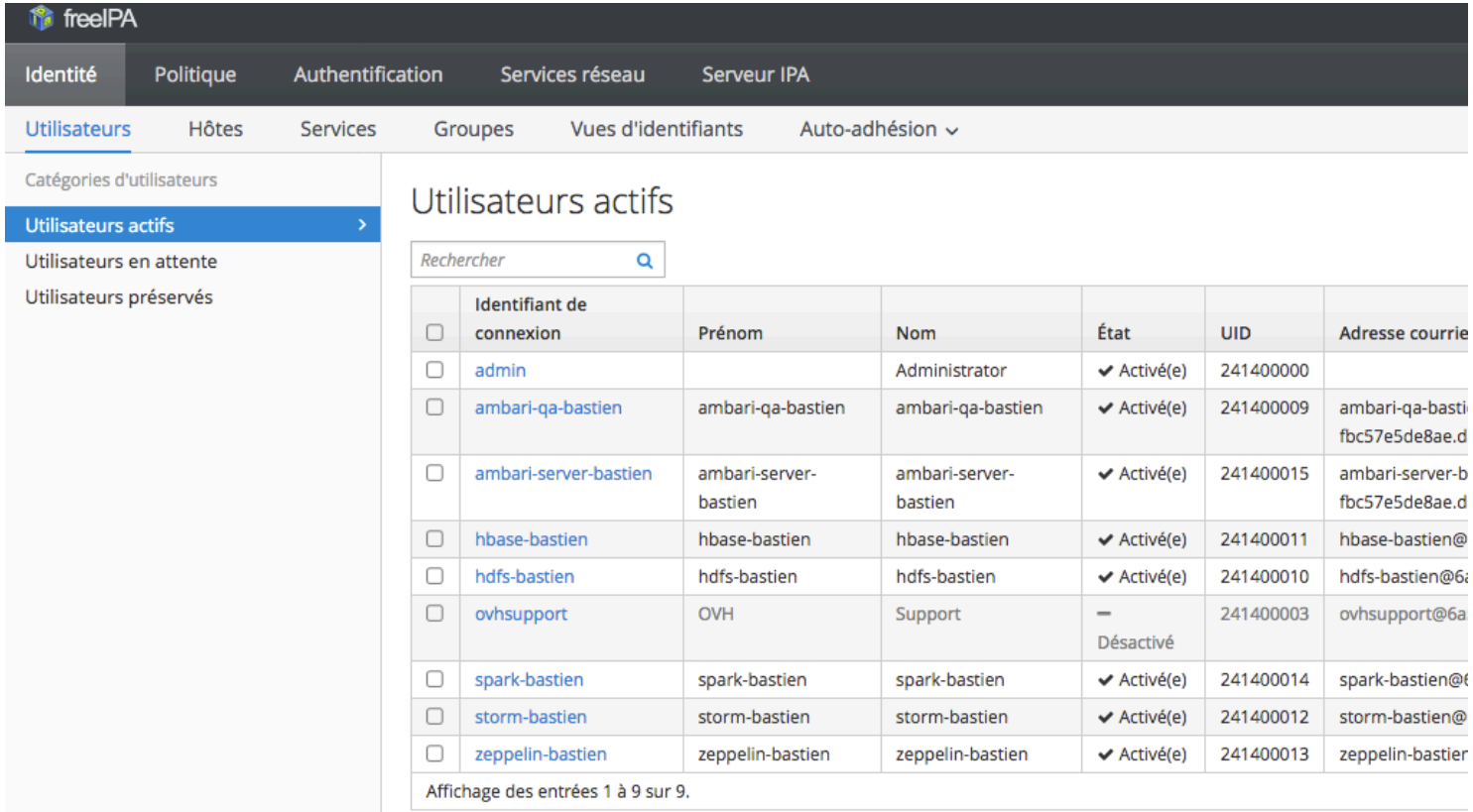
1. Browse the Control Panel and find the Ambari cluster URL
2. Connect to **Ambari** dashboard with your given credentials
3. Discover Ambari dashboard :
    1. Left menu, top menu
    2. HDFS infos and config
    3. Alerts



Step 4 : Discover FreeIPA, the users manager

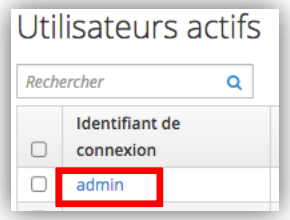
Ambari is a dashboard to manage your cluster

- To manage the users, please connect to FreeIPA (find the URL in control panel)
- Login/password : same as Ambari

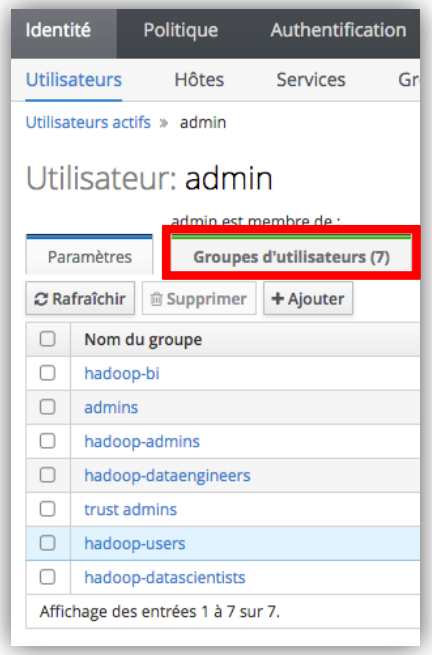


Modify your account

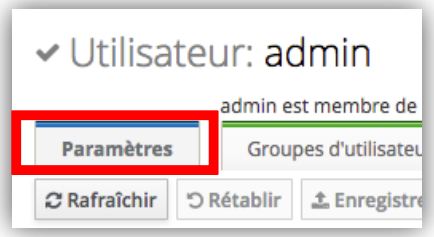
1. Click on your account



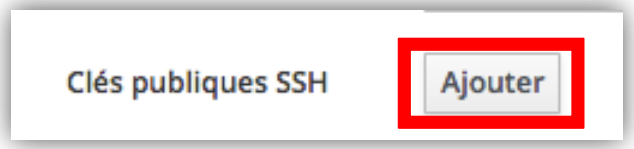
2. Check on which groups your are



3. in the “paramètres” tab



4. Add your SSH public Key if you have one

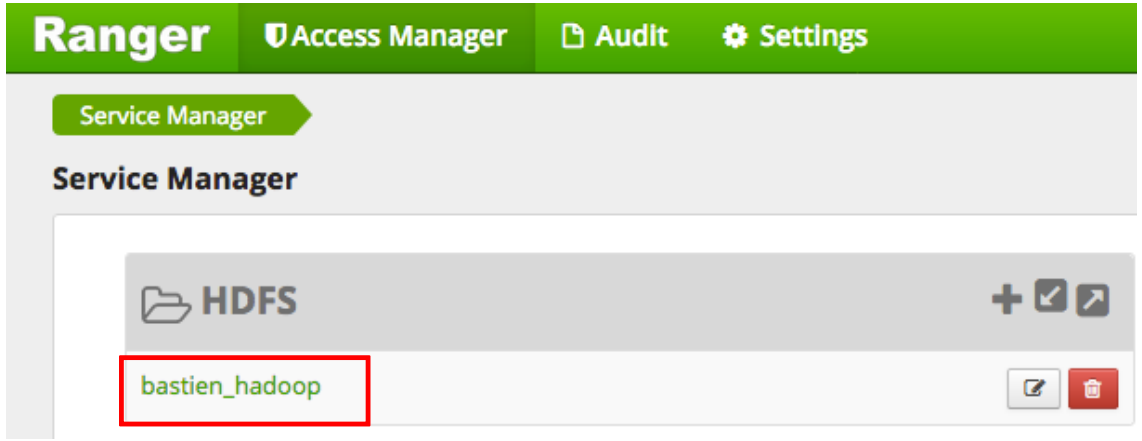


Step 5 : discover data management

FreeIPA is for managing users, Ranger is for data security

1. To manage the data, please connect to Apache Ranger (find the URL in control panel)
   - Login/password : same as Ambari

2. Then, discover for example HDFS rights ➔ live demo



Use case : let's play with real estate market in France !

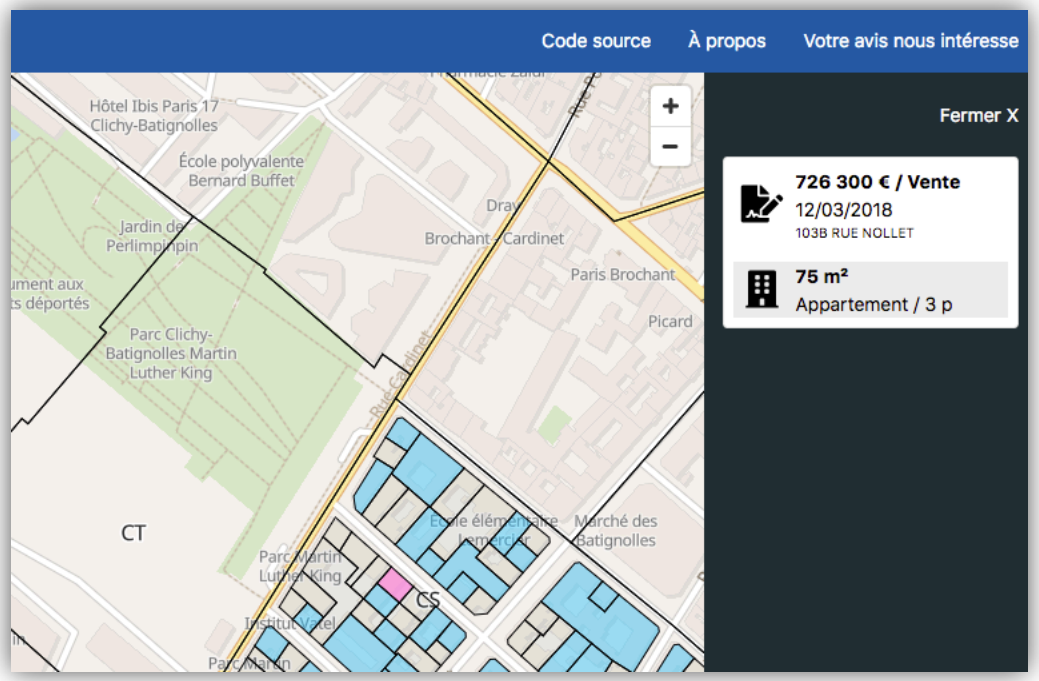
France government releases a lot of data ("open data")

In 2019, they proposed the DVF datas, the "Demandes de Valeurs Foncières".

= Amount of money for each real estate transaction in France since 5 years.

Example : in blue the sales in 2018 in Paris 75017 , with the prices

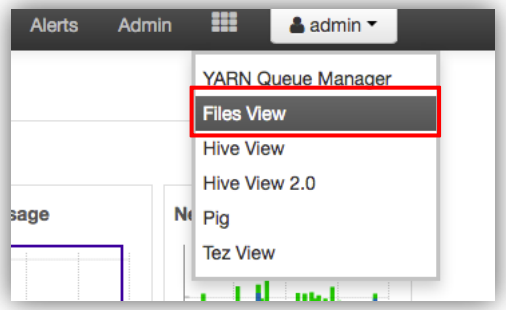
https://app.dvf.etalab.gouv.fr/



Step 1 (1/2) : ingest static data

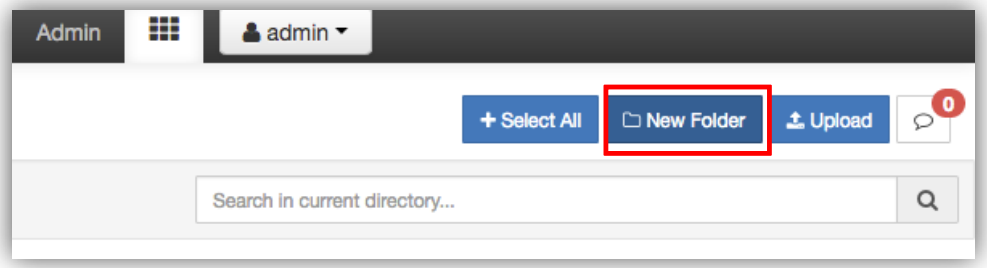
1. Go in github link, find the DVF website, download DVF data 2018

2. Go back in Ambari dashboard
3. Click on Files View :



4. go in /user/<your_login>
5. Create a new directory called “ovh”

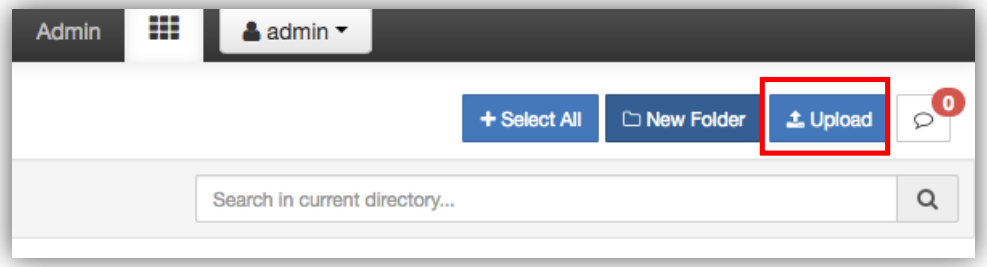
You should have : /user/<your_login>/ovh



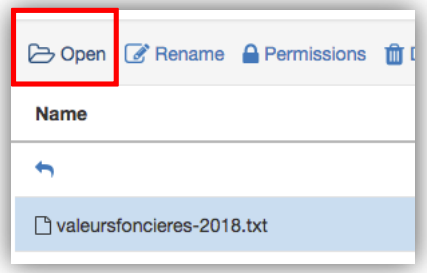
Step 1 (2/2) : ingest static data

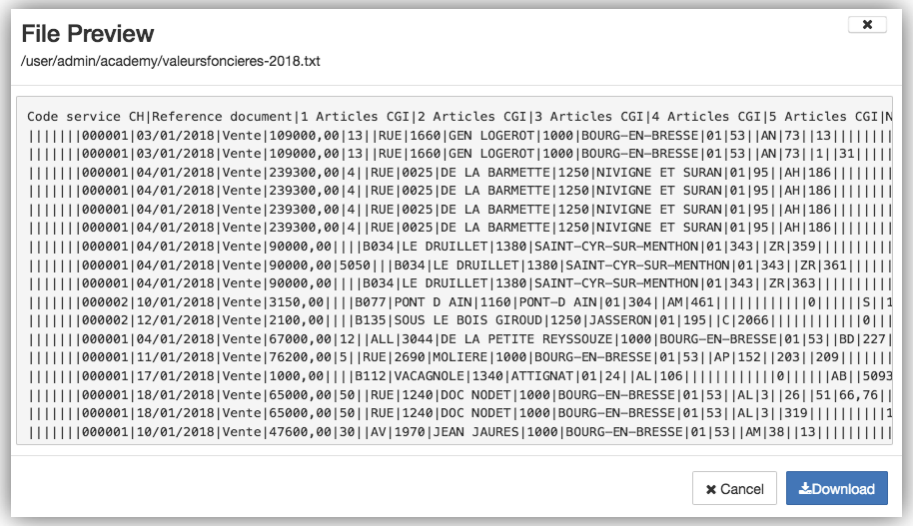
Now we will ingest static data, the TXT file previously downloaded.

1. Click on upload and add the TXT file « valeursfoncieres-2018.txt »



2. Once uploaded, you can select your file then click on « Open » to have a file preview



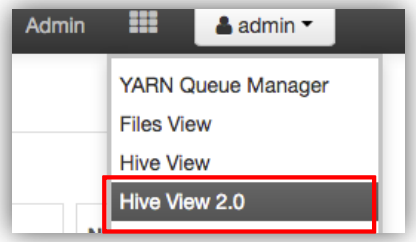


Step 2 : Prepare the data with Hive

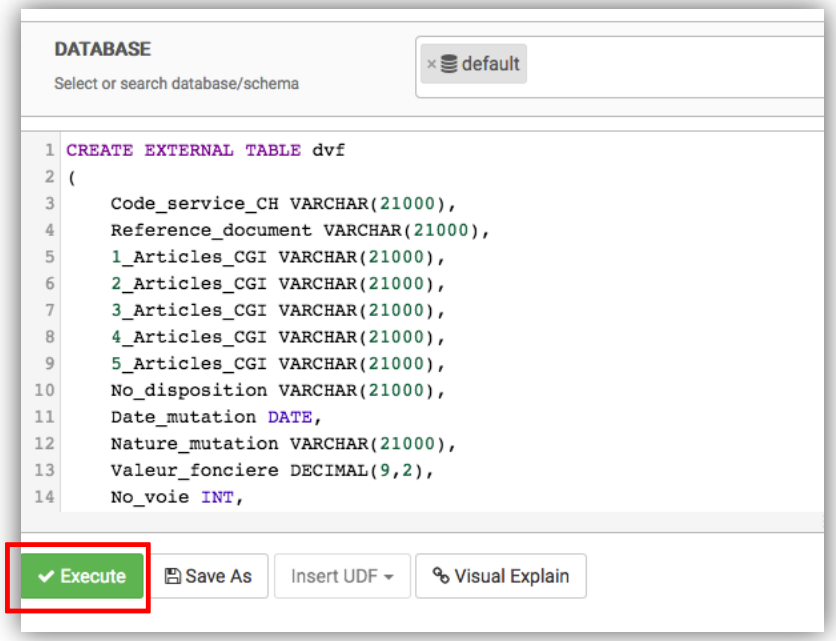
Hive allow you to query the data with a language very similar to SQL

We will create a new table in Hive to “load” the TXT file data.

1. Go in Hive view 2.0



2. Copy paste the SQL code to create a table
    1. WARNING : change the path with your login



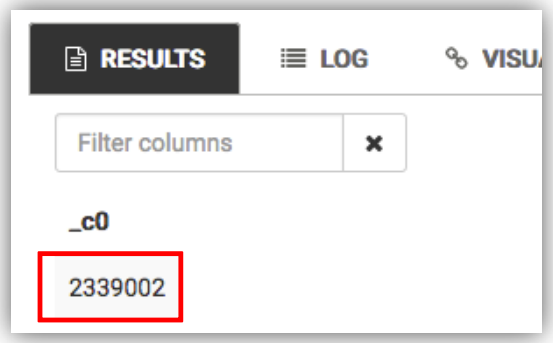
3. Click on “Execute”

Step 3 : simple queries to verify if everything is ok

For each step, enter the query in HIVE window then execute it (see previous step)

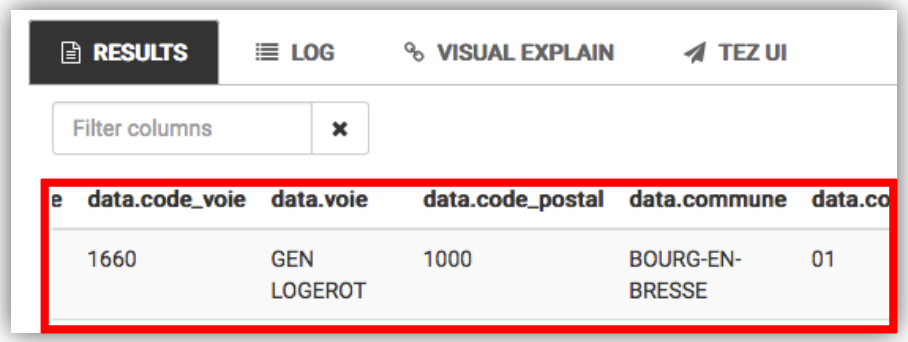
1. Let's check the amount of lines in the file

```
SELECT count(*) FROM dvf_<your_login>
```



2. Let's verify the data inside hive

```
SELECT * FROM dvf_<your_login> LIMIT 20
```

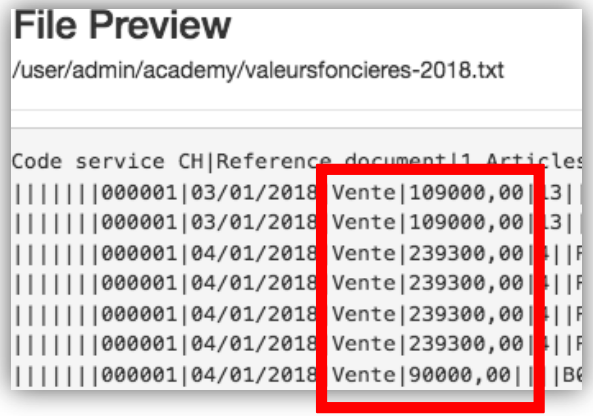


3. And more !

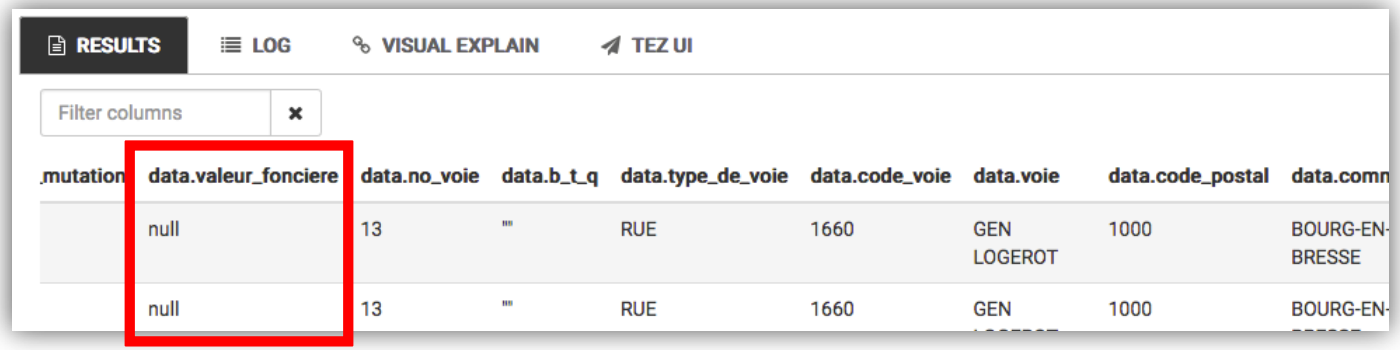
```
SELECT * from dvf_<your_login> WHERE code_postal = "75017" AND  type_local = "Appartement" LIMIT 100
```

?! Where are the prices for transactions ?

- In the files view, we had the amount of money for each transaction



- In Hive, we don't have it, NULL everywhere

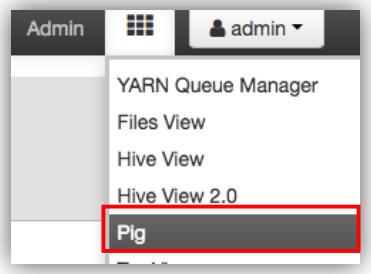


Issue : in the file; prices are with commas (virgules), instead of periods (points)

➔ let's fix it with Apache Pig !

Step 4 : create script in Apache Pig

1. First, create a new folder in "Files view" in /user/<your_login>/cleaned/"
2. Then go in Pig :



3. Create a new script called "commas" :

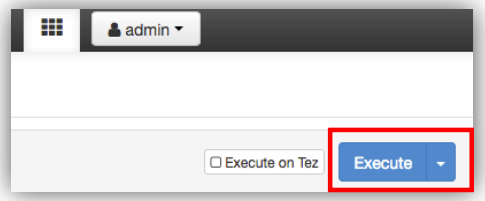
4. Add 3 lines of code (see github) :

5. Save

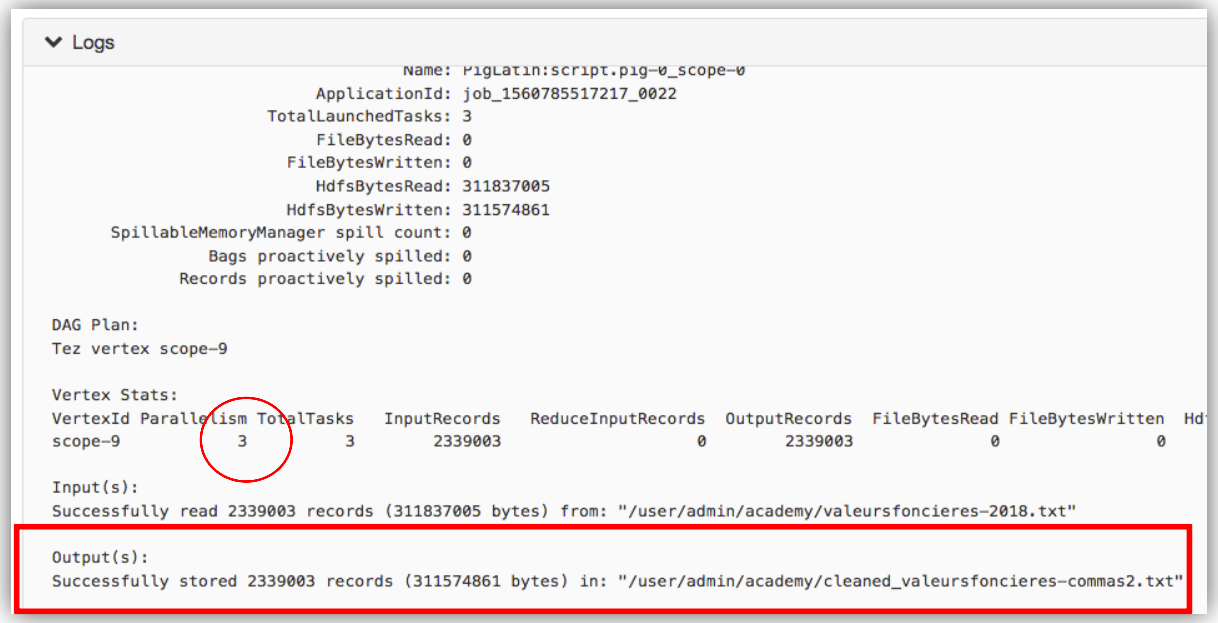
Step 5 : launch your script

The script will load your data, replace “,” by “.” then store it in another folder

1. Execute the script :



2. Watch the logs, wait for completion



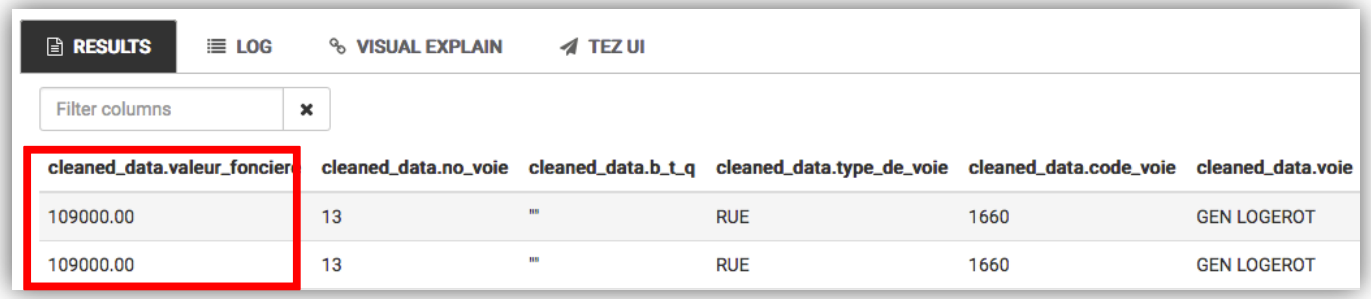
3. Verify your files in “Files view”
4. Your TXT file is now a folder with 3 sub files containing cleaned data

STEP 6 : Calculate average apartment price

1. Go back in “Hive View 2.0”
2. Create a new table called “cleaned_data” by copy/paste code (in Github)

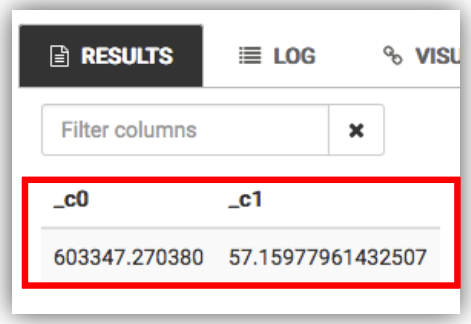
WARNING : change the path with your login

3. Verify that ”Valeur_Fonciere” is now here, with periods (points)



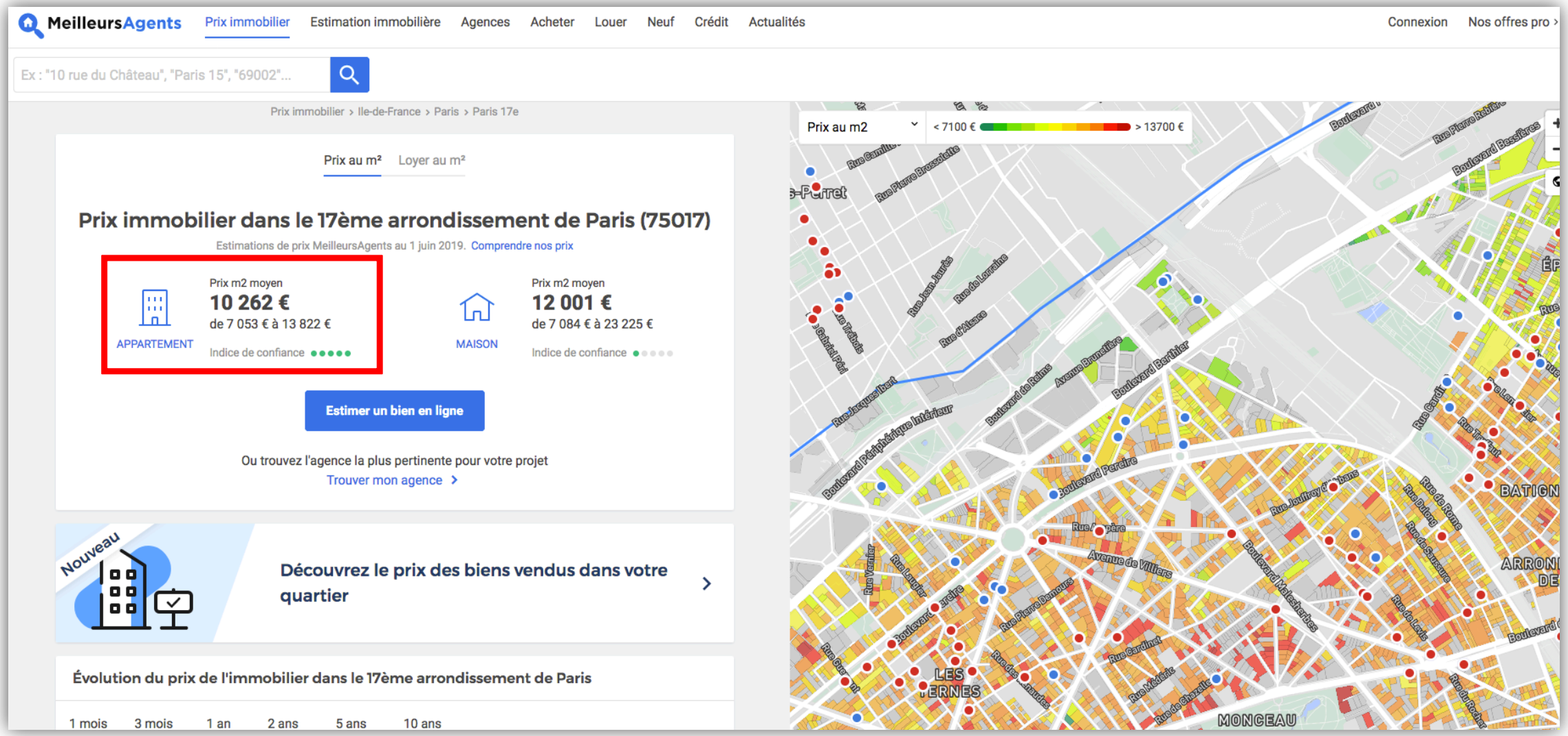
4. Query the data to find the average price of appartment sold in 2018 in Paris 17

```
SELECT AVG(valeur_fonciere), AVG(Surface_reelle_bati) FROM cleaned_data_<your_login> WHERE code_postal = "75017" AND type_local =
"Appartement"
```



Average price is 603’347€
Average square meter is 57 m2
= 10’526€ per square meter in 2018

This use case is real for businesses !



https://www.meilleursagents.com/prix-immobilier/paris-17eme-arrondissement-75017/

(bonus) STEP 7 : connection in SSH to your Edge node

Warning : make sure you added your SSH Public key in your profile in FreeIPA.

Propagation time +- 20mins

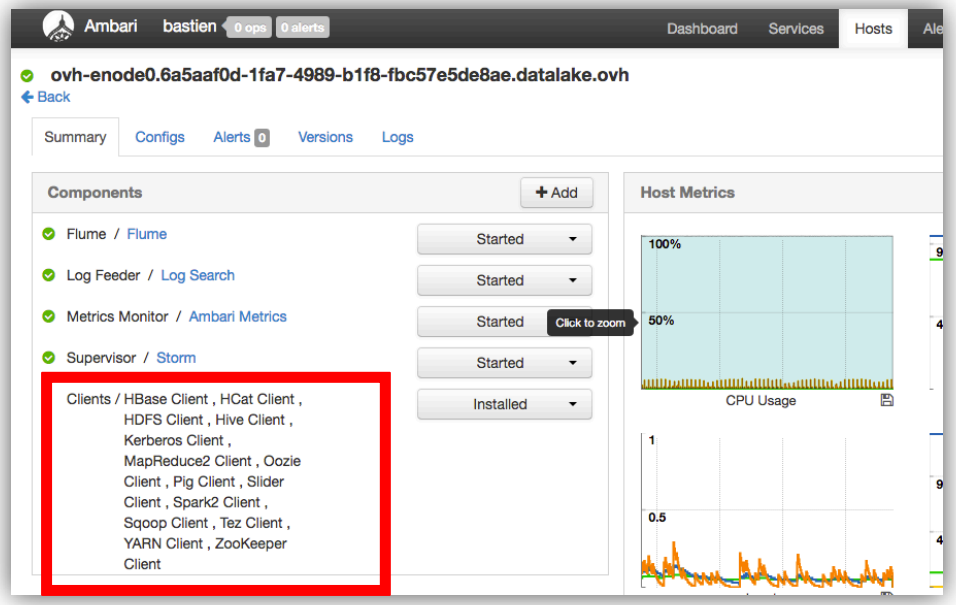
1. Connect to the cluster Bastion

```
ssh -A <user_login>@{bastion_IP}
```

2. Then connect to your Edge node, from your bastion

```
ssh <user_login>@ovh-enode0
```

3. You can now use all the clients available on your Edge node



You can find the list in Ambari / Hosts

(bonus) Step 8 : submit a Hive job via CLI

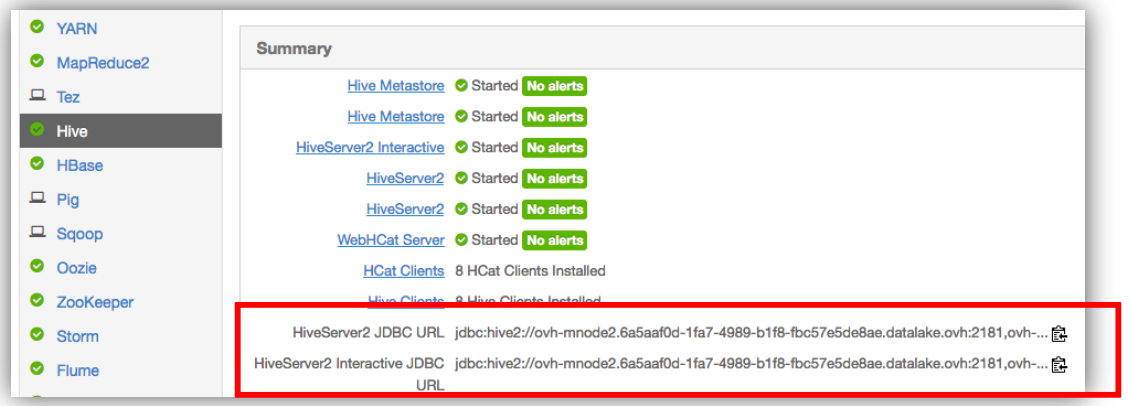
1. Once you are logged in SSH on you Edge node, launch Apache Beeline

```
beeline
```

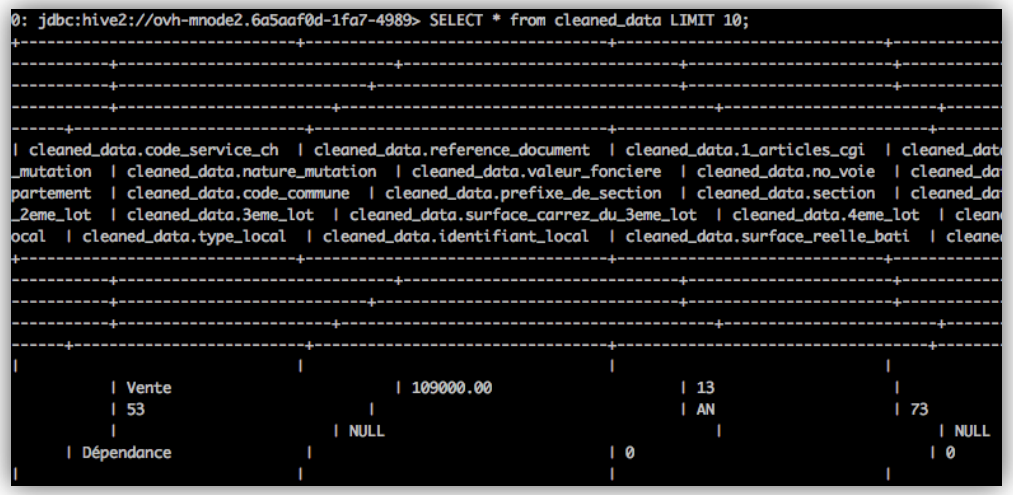
2. Then connect to Hive

```
!connect <Hive_server_URL>
```

(You will find your Hive server URL in Ambari)



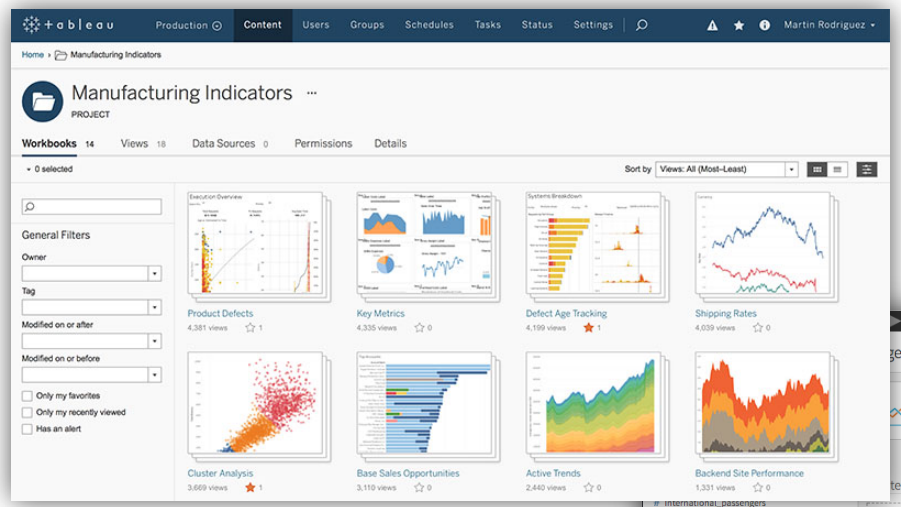
3. Now you can submit queries, exactly like inside the Hive UI

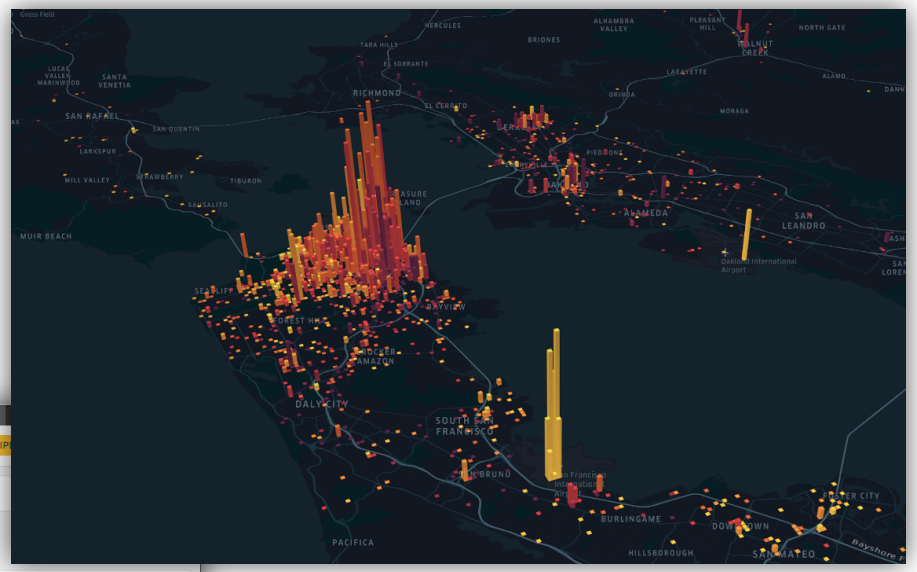


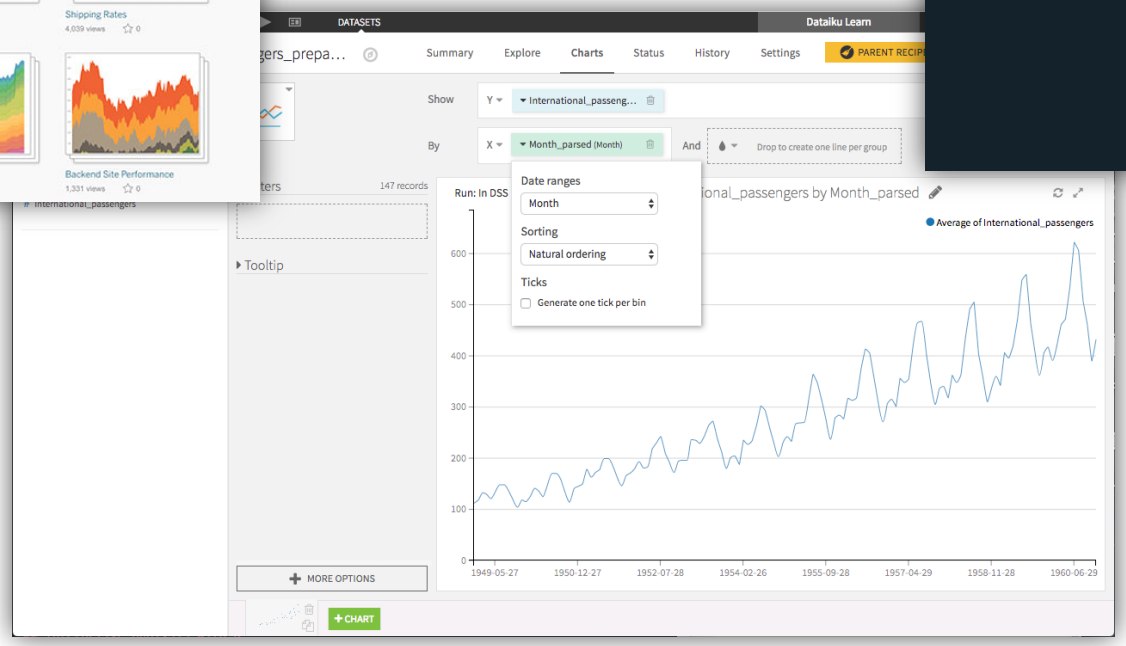
Now Imagine the same queries from various softwares...

Big data queries are usually made from third-party softwares

You “connect” them to a big data cluster







THANK YOU ! LAB DONE ☺

Offer page : https://www.ovh.com/fr/public-cloud/big-data-hadoop/

Credits : Bastien Verdebout / https://twitter.com/BastienOvh