



Projet BigData NoSQL



SUCUPIRA Lia, ITTEL Etienne,
LI Cyril, PEYROT Guillaume



Sommaire

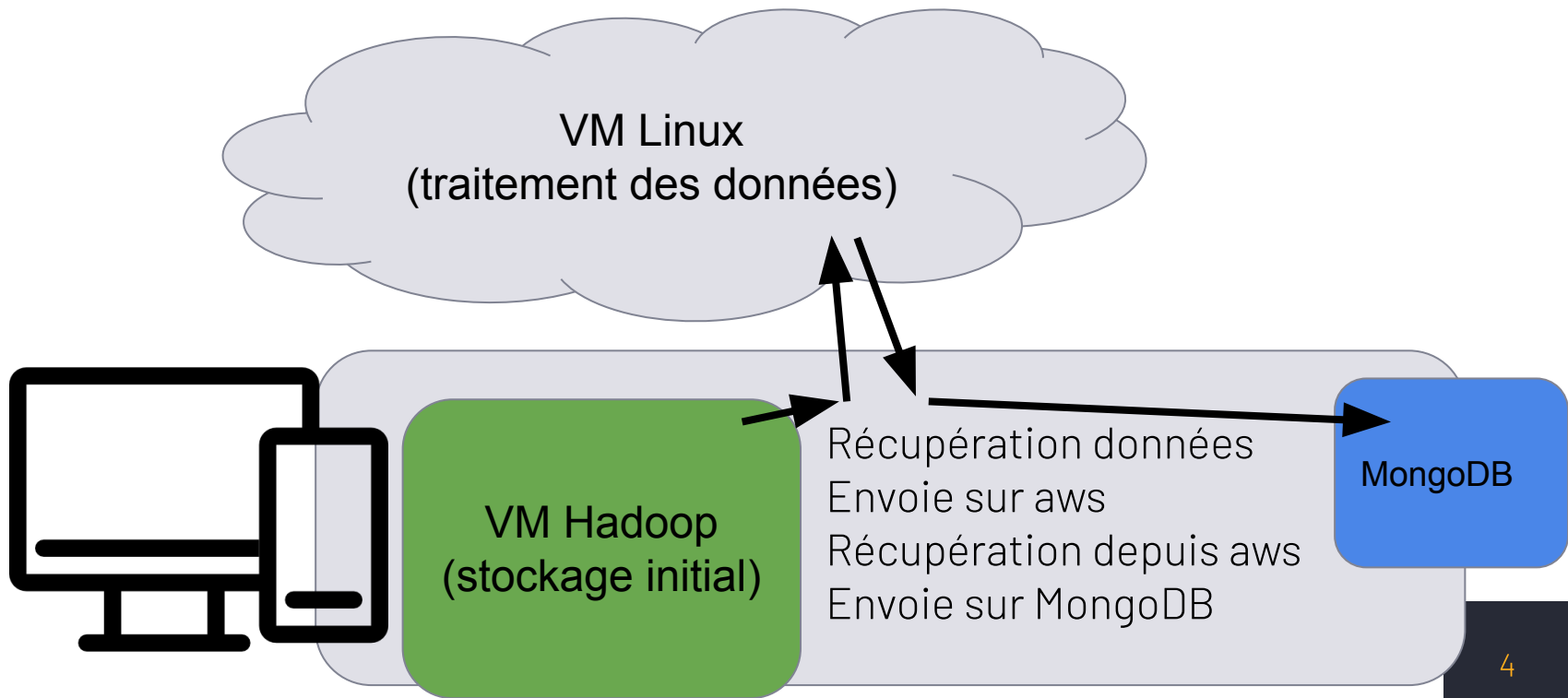
- Attente des clients
- VM Hadoop et HDFS
- VM Linux sur le Cloud AWS
- Model d'Apprentissage
- Base MongoDB





Attente des Clients

Attente des clients





VM Hadoop et HDFS



VM Hadoop et HDFS

The screenshot displays the Ambari web interface for the HDFS service. The left sidebar shows the navigation menu with 'Dashboard', 'Services', and 'HDFS' selected. The main content area is titled 'Summary' and includes tabs for 'SUMMARY', 'HEATMAPS', 'CONFIGS', and 'METRICS'. A 'Views' dropdown menu is open, showing 'Files View' and 'Workflow Manager'. Below the tabs, the 'Files View' shows a directory listing for the 'Sandbox' directory. The listing includes columns for Name, Size, Last Modified, Owner, Group, Permission, Erasure Coding, and Encrypted. The table lists various directories such as 'app-logs', 'apps', 'ats', 'atsv2', 'hdp', 'livy2-recovery', 'mapred', 'mr-history', 'ranger', 'spark2-history', 'tmp', 'user', and 'warehouse'.

Ambari

Dashboard

Services

HDFS

YARN

MapReduce2

Services / HDFS / Summary

SUMMARY HEATMAPS CONFIGS METRICS

Summary

Files View

Sandbox

Views

Files View

Workflow Manager

ACTIONS

Quick Links

Total: 13 files or folders

Select All New Folder Upload

Search in current directory...

Name	Size	Last Modified	Owner	Group	Permission	Erasure Coding	Encrypted
app-logs	--	2018-11-29 18:56	yarn	hadoop	drwxrwxrwx		No
apps	--	2018-11-29 20:01	hdfs	hdfs	drwxr-xr-x		No
ats	--	2018-11-29 18:25	yarn	hadoop	drwxr-xr-x		No
atsv2	--	2018-11-29 18:26	hdfs	hdfs	drwxr-xr-x		No
hdp	--	2018-11-29 18:26	hdfs	hdfs	drwxr-xr-x		No
livy2-recovery	--	2018-11-29 18:55	livy	hdfs	drwx-----		No
mapred	--	2018-11-29 18:26	mapred	hdfs	drwxr-xr-x		No
mr-history	--	2018-11-29 18:26	mapred	hadoop	drwxrwxrwx		No
ranger	--	2018-11-29 19:54	hdfs	hdfs	drwxr-xr-x		No
spark2-history	--	2020-01-30 10:18	spark	hadoop	drwxrwxrwx		No
tmp	--	2018-11-29 20:01	hdfs	hdfs	drwxrwxrwx		No
user	--	2018-11-29 20:21	hdfs	hdfs	drwxr-xr-x		No
warehouse	--	2018-11-29 18:51	hdfs	hdfs	drwxr-xr-x		No

Ambari

Dashboard

Services

- HDFS
- YARN
- MapReduce2
- Tez

Ambari

Dashboard

Services

- HDFS
- YARN
- MapReduce2
- Tez
- Hive
- HBase
- Pig
- Sqoop

/ Files View

Settings 0 Notifications 0 Grid User: maria_dev

Sandbox

/ > user > maria_dev

Total: 1 files or folders

+ Select All

New Folder

Upload

Search in current directory...



Name >	Size >	Last Modified >	Owner >	Group >	Permission	Erasure Coding	Encrypted
--------	--------	-----------------	---------	---------	------------	----------------	-----------



/ Files View

Settings 0 Notifications 0 Grid User: maria_dev

Sandbox

/ > user > maria_dev

Total: 2 files or folders

+ Select All

New Folder

Upload

Search in current directory...



Name >	Size >	Last Modified >	Owner >	Group >	Permission	Erasure Coding	Encrypted
--------	--------	-----------------	---------	---------	------------	----------------	-----------



.Trash	--	2020-01-30 10:19	maria_dev	hdfs	drwxr-xr-x		No
train.csv	20.1 MB	2020-01-30 10:21	maria_dev	hdfs	-rw-r--r--		No



Machine AWS

Utilisation de la VM

- Bibliotheque python : paramiko
- Connection par SSH
- Connection sftp:
 - Envoi du fichier
 - Appel du traitement
 - Recupération du fichier
 - Suppresion du premier fichier

Configuration de la VM

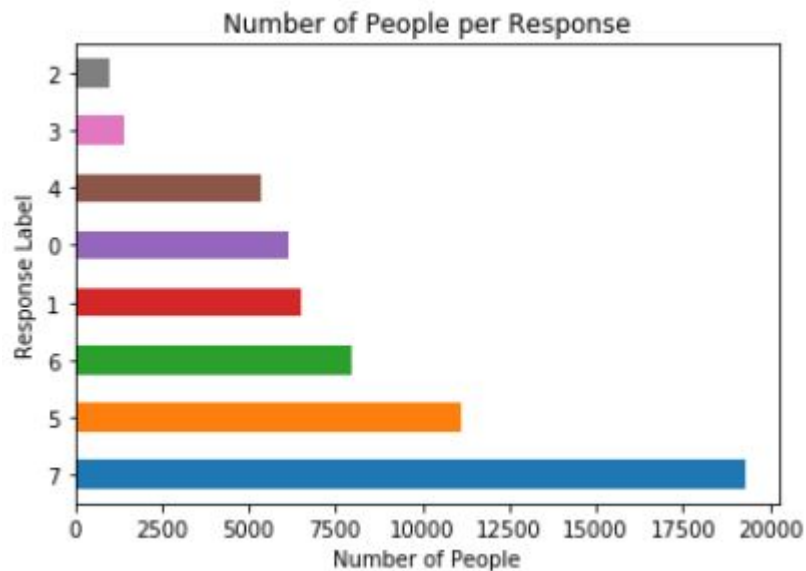
- Script Python création Instance
- Type = t2.micro
- Paire de clef
- Connexion par SSH



Dataset



Dataset



- 128 features
- 18 colonnes quantitatives
- 58881 samples
- Réponse (de 1 à 8)

Données d'apprentissage et de tests

Train: 39450 samples

Test: 19431 samples

Distribution des données

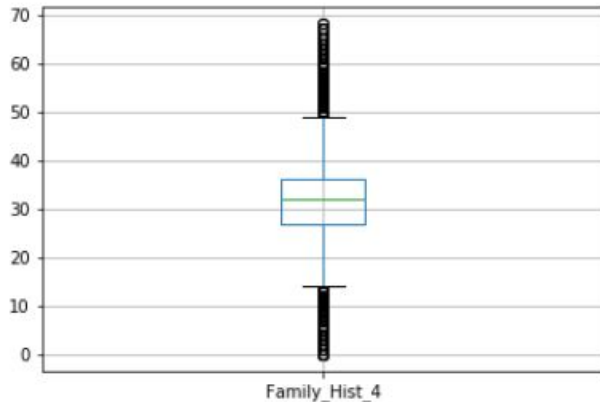
Response	Samples
0	4134
1	4318
2	697
3	952
4	3599
5	7454
6	5352
7	12944

Pre-traitement

- Remplacement de valeurs manquantes
- Labélisation
- Corrélation (On retire les features avec une corrélation supérieur à 0.95)
- Scaling (moyenne = 0 et variance =1)

Quantitative Features

- Outliers (Zone Interquartile)



Categorical Features

- Features avec des valeurs unique
- Features avec des proportions insignifiante (1% des données concernées)

Après le traitement

- 81 colonnes
- Train: 7578 lignes

Réponse	Samples
0	670
1	874
2	154
3	185
4	771
5	1388
6	1126
7	2386



Modèle d'apprentissage

KNN

- Nombre de composante du PCA : 13
- Nombres de voisins : 60
- Temps d'exécution : 0.10 secondes

Results on the test set:

	precision	recall	f1-score	support
0	0.37	0.08	0.13	2002
1	0.33	0.18	0.23	2166
2	0.00	0.00	0.00	333
3	0.00	0.00	0.00	488
4	0.28	0.14	0.19	1721
5	0.29	0.29	0.29	3646
6	0.30	0.22	0.25	2664
7	0.51	0.91	0.66	6411
avg / total	0.36	0.42	0.36	19431

Extended accuracy : 0.607

SVM

- Nombre de composante du PCA : 16
- C=1.0
- Temps d'exécution : 15.2 secondes

Results on the test set:

	precision	recall	f1-score	support
0	0.31	0.03	0.06	2002
1	0.33	0.20	0.25	2166
2	0.00	0.00	0.00	333
3	0.00	0.00	0.00	488
4	0.25	0.24	0.25	1721
5	0.29	0.33	0.31	3646
6	0.35	0.20	0.26	2664
7	0.55	0.90	0.68	6411
avg / total	0.37	0.43	0.37	19431

Extended accuracy : 0.639

Logistic Regression

- Nombre de composante du PCA : 73
- $C = 1.0$
- solver = lbfgs
- Temps d'exécution : 0.31 secondes

Results on the test set:

	precision	recall	f1-score	support
0	0.34	0.17	0.23	2002
1	0.33	0.17	0.22	2166
2	0.18	0.06	0.09	333
3	0.19	0.09	0.12	488
4	0.34	0.32	0.33	1721
5	0.33	0.32	0.33	3646
6	0.38	0.26	0.31	2664
7	0.58	0.90	0.70	6411
avg / total	0.41	0.46	0.42	19431

Extended accuracy : 0.674

Neural Network

- Fonction d'activation: tanh
- Couche cachée : 1 avec 100 neurones
- Temps d'exécution : 24.9 secondes

Results on the test set:

	precision	recall	f1-score	support
0	0.33	0.18	0.23	2072
1	0.28	0.19	0.22	2185
2	0.18	0.03	0.05	321
3	0.22	0.10	0.13	445
4	0.33	0.36	0.34	1792
5	0.33	0.37	0.35	3623
6	0.39	0.25	0.31	2645
7	0.61	0.85	0.71	6348
avg / total	0.42	0.46	0.42	19431

Extended accuracy : 0.644

Comparison Models

	Extended Accuracy	Precision	Recall
KNN	0.607	0.36	0.42
SVM	0.639	0.37	0.43
Logistic Regression	0.674	0.41	0.46
Neural Network	0.644	0.40	0.43



Tests d'indépendance

Test d'indépendance du khi-deux

- But : Mesurer la corrélation entre 2 variables X et Y
- X : Genre, Religion, Groupe ethnique
- Y : Attribution du niveau de sûreté

$$E_{ij} = \frac{O_{i+} \times O_{+j}}{N}$$

où

$$O_{i+} = \sum_{j=1}^J O_{ij} \text{ (nombre de données pour lesquelles } X = i \text{)}$$

et

$$O_{+j} = \sum_{i=1}^I O_{ij} \text{ (nombre de données pour lesquelles } Y = j \text{)}$$

$$T = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Test d'indépendance du khi-deux

- Taux d'erreur : 5%
- **Genre** : Loi du Khi 2 à 7 degrés de libertés
 - Indépendance si **$T < 14.067$**
- **Religion** : Loi du Khi 2 à 35 degrés de libertés
 - Indépendance si **$T < 45.902$**
- **Groupe ethnique** : Loi du Khi 2 à 49 degrés de libertés
 - Indépendance si **$T < 66.339$**

Genre

	T seuil	T obtenu	Indépendance
Sans prétraitement	14.067	820.758	Non
Régression Linéaire	14.067	1346.821	Non
Réseau de neurones	14.067	342.973	Non

Religion

	T seuil	T obtenu	Indépendance
Sans prétraitement	45.902	286.904	Non
Régression Linéaire	45.902	78.203	Non
Réseau de neurones	45.902	31.155	Oui

Groupe ethnique

	T seuil	T obtenu	Indépendance
Sans prétraitement	66.339	114.953	Non
Régression Linéaire	66.339	83.755	Non
Réseau de neurones	66.339	44.257	Oui

Test d'indépendance

- Algorithme sans biais pour la religion et le groupe ethnique
- Algorithme avec un biais réduit pour le genre
- Réseaux de neurones > Régression linéaire
- Importance du prétraitement



Choix du modèle

Précision

	Extended Accuracy	Precision	Recall
KNN	0.607	0.36	0.42
SVM	0.639	0.37	0.43
Logistic Regression	0.674	0.41	0.46
Neural Network	0.644	0.40	0.43

Groupe ethnique

	Genre	Religion	Groupe ethnique
Sans prétraitement	Bias	Bias	Bias
Régression Linéaire	Bias	Bias	Bias
Réseau de neurones	Indépendance	Indépendance	Indépendance

Notre modèle

- En prenant la **précision** et le le **biais** des données sensibles
 - **Réseau de neurones**
- Indépendance à la religion et au groupe ethnique
- 64% de précision



Base MongoDB



Base MongoDB

MongoDB Compass - localhost:27017/Big_Data

Connect View Help

Local

4 DBS 2 COLLECTIONS

☆ FAVORITE

HOST
localhost:27017

CLUSTER
Standalone

EDITION
MongoDB 4.2.3 Community

Filter your data

Big_Data

- Insurance0
- Insurance1
- Insurance2
- Insurance3
- Insurance4
- Insurance5
- Insurance6
- Insurance7

> admin

> config

+

Collections

CREATE COLLECTION

Collection Name	Documents	Avg. Document Size	Total Document Size	Num. Indexes	Total Index Size	Properties
Insurance0	20	1.9 KB	37.4 KB	1	4.1 KB	
Insurance1	44	1.9 KB	82.2 KB	1	4.1 KB	
Insurance2	2	1.9 KB	3.7 KB	1	4.1 KB	
Insurance3	3	1.9 KB	5.6 KB	1	4.1 KB	
Insurance4	33	1.9 KB	61.6 KB	1	4.1 KB	
Insurance5	120	1.9 KB	224.2 KB	1	4.1 KB	
Insurance6	41	1.9 KB	76.6 KB	1	4.1 KB	
Insurance7	237	1.9 KB	442.7 KB	1	4.1 KB	



Merci!

questions?