THURSDAY, OCTOBER 26, 2023

HADOOP

MASTODON

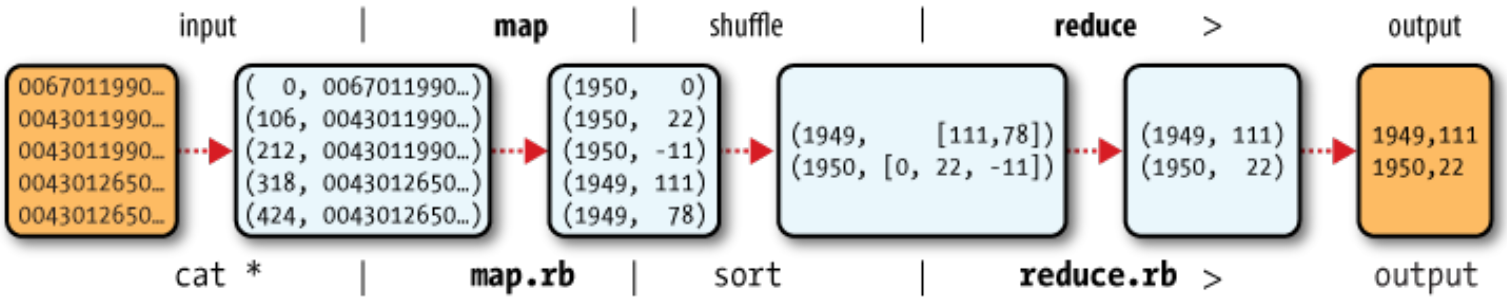
RAPPORT

AYMANE SABRI

DEVELOPER DE DONNÉES

Objectif du projet :

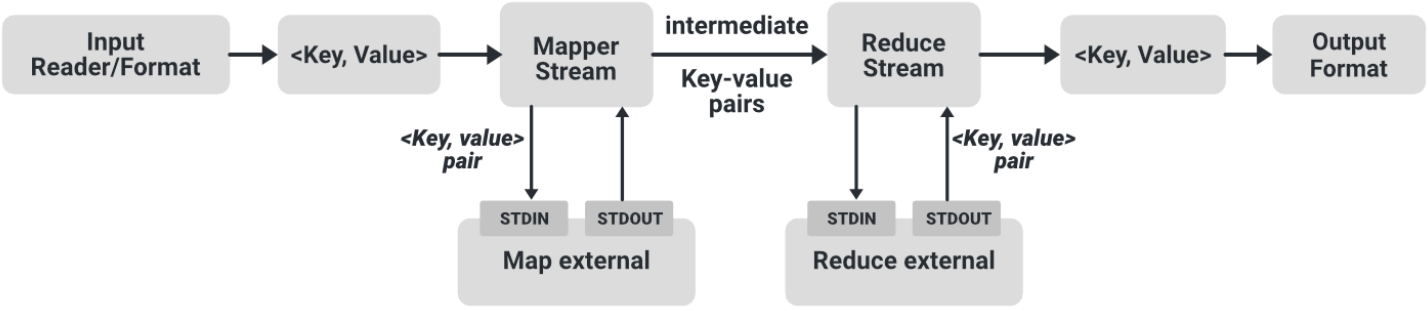
L'objectif principal de ce projet est de développer une compréhension approfondie des concepts fondamentaux de l'Hadoop en utilisant les différents outils Data . Nous nous sommes engagés à mettre en œuvre un scénario d'extraction, de transformation et de chargement de données.



I. Introduction :

1. Contexte du Projet :

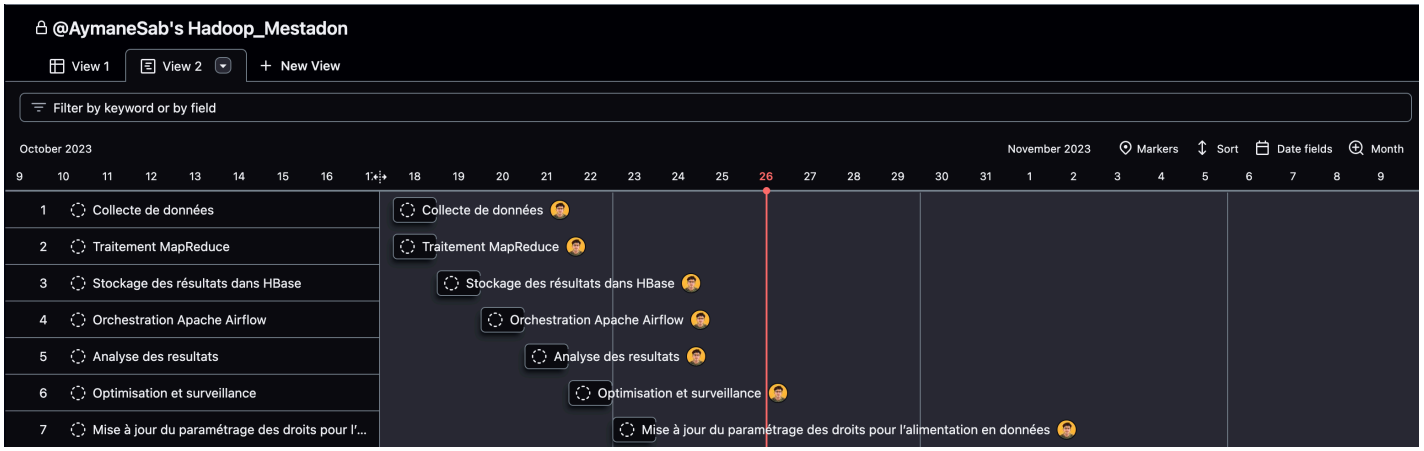
Le projet consiste à mettre en place un pipeline automatisé pour collecter, traiter et analyser des données brutes à partir de la plateforme Mastodon. Les données seront stockées dans un système de fichiers distribué HDFS, traitées à l'aide de MapReduce, et les résultats seront stockés dans HBase. Un DAG Apache Airflow orchestrera le workflow, permettant d'effectuer des analyses sur les utilisateurs, le contenu, la langue, la région, l'engagement des médias, les balises et les mentions. Le projet inclut également la gestion des droits d'accès, la programmation des mises à jour et la mise en conformité avec le RGPD pour le traitement des données personnelles.



2. Planification du plan du réalisation de projet :

Dans l'objectif de bien mener cd projet, j'ai commencé par établir le planning à suivre durant la période de brief .

2.1.Etapes Suivie :



II. Installation :

1. Hadoop :

```
(base) hadoop@sabri-Parallels-Virtual-Platform:~$ jps
4678 NameNode
5110 SecondaryNameNode
275107 Jps
6504 ThriftServer
5401 ResourceManager
4829 DataNode
5534 NodeManager
(base) hadoop@sabri-Parallels-Virtual-Platform:~$
```

2. Hbase :

```
(base) hadoop@sabri-Parallels-Virtual-Platform:~$ jps
276310 HMaster
6504 ThriftServer
276472 Jps
(base) hadoop@sabri-Parallels-Virtual-Platform:~$
```

3. AirFlow :

```
(airflow-environment) hadoop@sabri-Parallels-Virtual-Platform:~/AirFlow/airflow-environment$ airflow webserver -p 8080
[2023-10-26T07:13:47.043+0100] {configuration.py:2067} INFO - Creating new FAB webserver config file in: /home/hadoop/airflow/webserver_config.py
  ____________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /    _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
Running the Gunicorn Server with:
Workers: 4 sync
Host: 0.0.0.0:8080
Timeout: 120
Logfiles: - -
Access Logformat:
==================================================================
```

```
(airflow-environment) hadoop@sabri-Parallels-Virtual-Platform:~/AirFlow/airflow-environment$ airflow scheduler
  ______________       _____________
 ____    |__( )_________  __/__  /________      __
____  /| |_  /__  ___/_  /_ __  /_  __ \_ | /| / /
___  ___ |  / _  /    _  __/ _  / / /_/ /_ |/ |/ /
 _/_/  |_/_/  /_/    /_/    /_/  \____/____/|__/
[2023-10-26T07:15:09.266+0100] {executor_loader.py:117} INFO - Loaded executor: SequentialExecutor
```

II. Mastodon Analysis :

1. Retrieving Mastodon Data :

Browse Directory

/Mostodon/Raw Go!

Show 25 entries Search:

Permission	Owner	Group	Size	Last Modified	Replication	Block Size	Name
-rw-r--r--	hadoop	supergroup	263.05 KB	Oct 19 23:37	1	128 MB	mastodon_data_2023-10-19.json
-rw-r--r--	hadoop	supergroup	5.97 MB	Oct 22 18:56	1	128 MB	mastodon_data_2023-10-22.json
-rw-r--r--	hadoop	supergroup	55.93 MB	Oct 23 23:59	1	128 MB	mastodon_data_2023-10-23.json
-rw-r--r--	hadoop	supergroup	257.99 KB	Oct 24 12:47	1	128 MB	mastodon_data_2023-10-24.json
-rw-r--r--	hadoop	supergroup	0 B	Oct 25 20:45	1	128 MB	mastodon_data_2023-10-25.json

Showing 1 to 5 of 5 entries Previous 1 Next

Hadoop, 2023.

ents

{"id":"111286705686071926","created_at":"2023-10-23T22:32:00.000Z","in_reply_to_id":null,"in_reply_to_account_id":null,"sensitive":false,"spoiler_text":"","visibility":"public","language":"en","url":"https://twitter.com/NCAABaseball/status/1716583278243688662","replies_count":0,"reblogs_count":0,"favourites_count":0,"edited_at":null,"content":"Off season drip \ud83d\udca7#NCAABaseball x \ud83d\udcf8 @Ohio_Baseball@twitter.com","reblog":null,"account":{"id":"110730189051479019","username":"NCAABaseball","acct":"NCAABaseball@sportsbots.xyz","display_name":"NCAA Baseball \ud83e

2. Map Reduce :

- Mapper Results

File contents


```
{
 "HasMediaAttachments": {
  "111286112546404847": 1,
  "111286112589624420": 1,
  "111286112713706157": 1,
  "111286112736823290": 1,
  "111286112744423098": 1,
  "111286112771203848": 1,
```


- Reducer Results


```
"Tag": {
  "03Dic": 3,
  "0_archivebox": 1,
  "100daysofart": 1,
  "100daysofsketching": 1,
  "100girlfriends": 1,
  "100\u4e07\u4eba": 1,
  "100\u4e07\u4eba\u9054\u6210": 1,
  "10films": 1,
  "111mm": 1,
  "12ancestors": 1,
  "12daysofhalloween": 1,
  "147sf": 3,
  "14thamendment": 1,
  "15minutesfromhome": 2,
  "1980s": 1,
  "19thcentury": 1,
  "1password": 5,
  "1st_revue": 1,
```

3. Insertion dans Apache Hbase :

- Creation des tables

```
6 row(s)
Took 1.8085 seconds
=> ["Language", "MastodonGrowth", "Mentions", "Tags", "URLShare", "User"]
hbase:004:0>
```

User Tables

6 table(s) in set.

Table	Description
Language	'Language', {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}, {NAME => 'Language', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}
MastodonGrowth	'MastodonGrowth', {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}, {NAME => 'DateMetrics', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}
Mentions	'Mentions', {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}, {NAME => 'MentionsInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}
Tags	'Tags', {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}, {NAME => 'TagsInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}
URLShare	'URLShare', {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}, {NAME => 'ShareCount', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}, {NAME => 'URLInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}
User	'User', {TABLE_ATTRIBUTES => {METADATA => {'hbase.store.file-tracker.impl' => 'DEFAULT'}}}, {NAME => 'UserInfo', INDEX_BLOCK_ENCODING => 'NONE', VERSIONS => '1', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL => 'FOREVER', MIN_VERSIONS => '0', REPLICATION_SCOPE => '0', BLOOMFILTER => 'ROW', IN_MEMORY => 'false', COMPRESSION => 'NONE', BLOCKCACHE => 'true', BLOCKSIZE => '65536 B (64KB)'}

- Insertion des tables

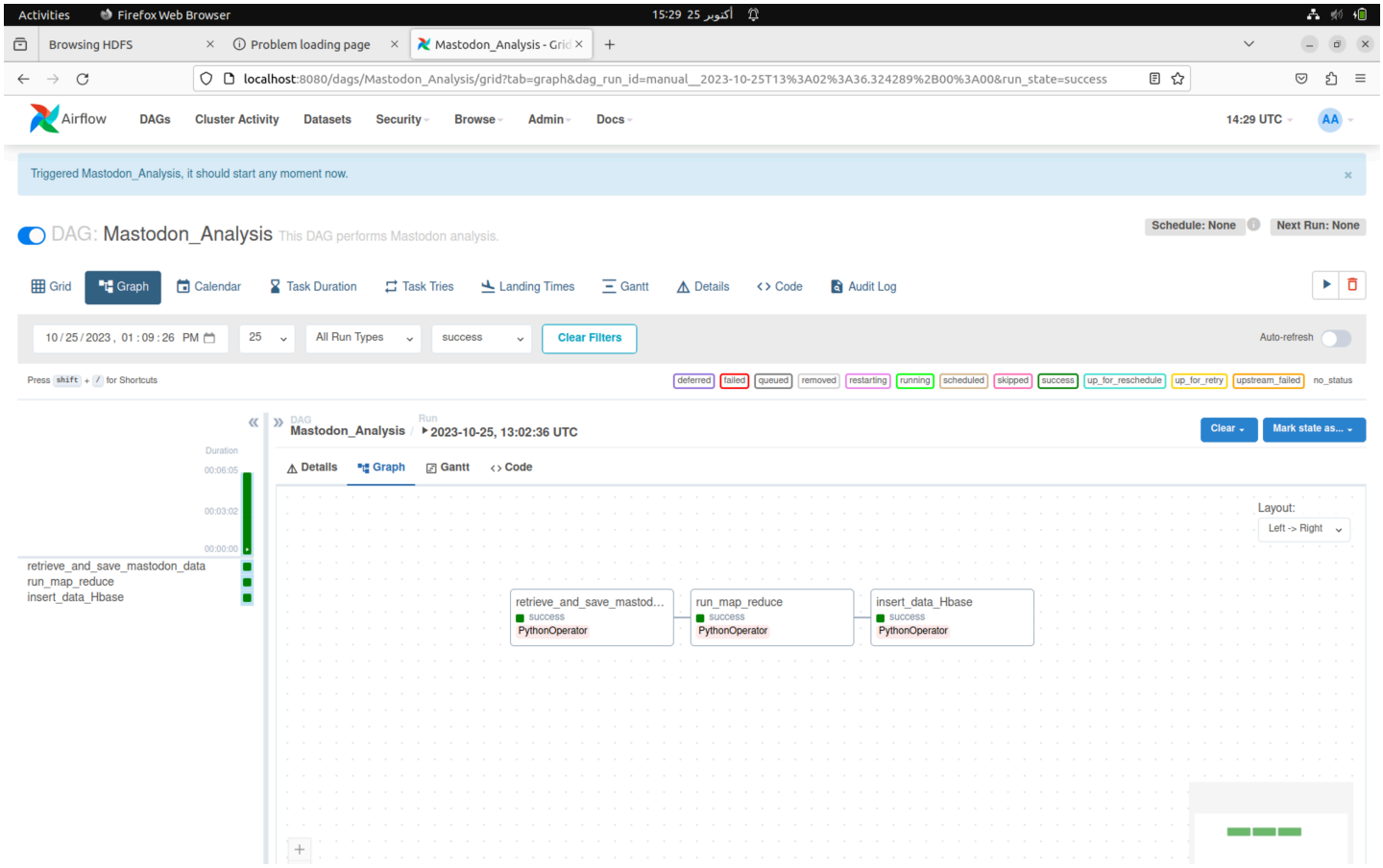
```
(base) hadoop@sabri-Parallels-Virtual-Platform:~/Mastodon/Hbase$ python3 HbaseClient.py
Connected to HBase Succefully
Connected to table: User
Connected to table: Language
Connected to table: URLShare
Connected to table: Mentions
Connected to table: Tags
Connected to table: MastodonGrowth
Connected to table: Year
Connection closed
(base) hadoop@sabri-Parallels-Virtual-Platform:~/Mastodon/Hbase$
```

- Verification des tables

```
ar                                  column=Language:Count, timestamp=2023-10-25T20:46:05.761, value=11
az                                  column=Language:Count, timestamp=2023-10-25T20:46:05.763, value=2
be                                  column=Language:Count, timestamp=2023-10-25T20:46:05.764, value=6
bg                                  column=Language:Count, timestamp=2023-10-25T20:46:05.766, value=2
br                                  column=Language:Count, timestamp=2023-10-25T20:46:05.768, value=1
ca                                  column=Language:Count, timestamp=2023-10-25T20:46:05.769, value=48
cs                                  column=Language:Count, timestamp=2023-10-25T20:46:05.772, value=41
cy                                  column=Language:Count, timestamp=2023-10-25T20:46:05.776, value=3
da                                  column=Language:Count, timestamp=2023-10-25T20:46:05.780, value=9
de                                  column=Language:Count, timestamp=2023-10-25T20:46:05.781, value=1031
el                                  column=Language:Count, timestamp=2023-10-25T20:46:05.783, value=9
en                                  column=Language:Count, timestamp=2023-10-25T20:46:05.785, value=4671
```

```
watch@ovo.st                        column=MentionsInfo:Count, timestamp=2023-10-25T20:46:16.356, value=1
wikipedia@wikis.world               column=MentionsInfo:Count, timestamp=2023-10-25T20:46:16.358, value=2
wordcampgermany@wp-social.net       column=MentionsInfo:Count, timestamp=2023-10-25T20:46:16.360, value=1
zarame_senbee@mstdn.jp              column=MentionsInfo:Count, timestamp=2023-10-25T20:46:16.362, value=1
zenn_dev_trend_bot@silicon.moe      column=MentionsInfo:Count, timestamp=2023-10-25T20:46:16.365, value=2
zetthochdrei                        column=MentionsInfo:Count, timestamp=2023-10-25T20:46:16.368, value=1
zhenyi                              column=MentionsInfo:Count, timestamp=2023-10-25T20:46:16.369, value=1
zughy_boi                           column=MentionsInfo:Count, timestamp=2023-10-25T20:46:16.371, value=1
```

4. AirFlow :


Activities
Firefox Web Browser
15:29 25 أكتوبر
Browsing HDFS
Problem loading page
Mastodon_Analysis - Gric
localhost:8080/dags/Mastodon_Analysis/grid?tab=graph&dag_run_id=manual__2023-10-25T13%3A02%3A36.324289%2B00%3A00&run_state=success
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
14:29 UTC
AA
Triggered Mastodon_Analysis, it should start any moment now.
DAG: Mastodon_Analysis
This DAG performs Mastodon analysis.
Schedule: None
Next Run: None
Grid
Graph
Calendar
Task Duration
Task Tries
Landing Times
Gantt
Details
Code
Audit Log
10 / 25 / 2023 , 01 : 09 : 26 PM
25
All Run Types
success
Clear Filters
Auto-refresh
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
DAG
Mastodon_Analysis
Run
2023-10-25, 13:02:36 UTC
Clear
Mark state as...
Duration
00:06:05
00:03:02
00:00:00
Details
Graph
Gantt
Code
Layout:
Left -> Right
retrieve_and_save_mastodon_data
run_map_reduce
insert_data_Hbase
retrieve_and_save_mastod...
success
PythonOperator
run_map_reduce
success
PythonOperator
insert_data_Hbase
success
PythonOperator