

MOURAH Amel

M1 Informatique

Pojet Big Data : Creation d'une architecture Big Data et analyse de données

Introduction

Des volumes considérables de données sont créés tous les jours à partir des données utilisateurs générés automatiquement sur Internet. Réseaux sociaux, appareils mobiles, messagerie électronique, blogs, vidéos, transactions bancaires et autres interactions utilisateur pilotent désormais les campagnes Marketing, les études sociodémographiques, les enquêtes de polices, les intentions électorales et autres, en établissant une nouvelle dimension appelée **BigData**.

Les opérateurs de modélisation traditionnels basés sur le **standard SQL** ont de bonnes performances lors du traitement de petites quantités de données relationnelles mais ces outils sont très limités face à l'expansion des données en volume et en complexité, une chose qui actuellement ne peut être gérée que par des techniques de modélisation non-relationnelles.

Hadoop MapReduce est considéré comme la technique de traitement la plus efficace, comparée aux bases de données SQL, il dispose d'une performance proportionnelle à la complexité des données volumineuses. C'est un outil efficace pour résoudre les problèmes de données massives.

Dans ce contexte, mon projet consiste la réalisation d'un **cluster Hadoop** qui est une grappe de serveurs permettant d'effectuer des analyses de données Big Data rapidement et efficacement, en répartissant la tâche entre les différents ordinateurs qui composent la grappe, cela sur **AWS** (Amazon Web Services) grâce au service **Amazon EMR** Qui facilite la création et la gestion de clusters élastiques et entièrement configurés d'instances **Amazon EC2** exécutant Hadoop. Mes données pour ce projet sont un jeu de donnée issues des **Datasets Kaggle** <https://www.kaggle.com/lava18/google-play-store-apps>.

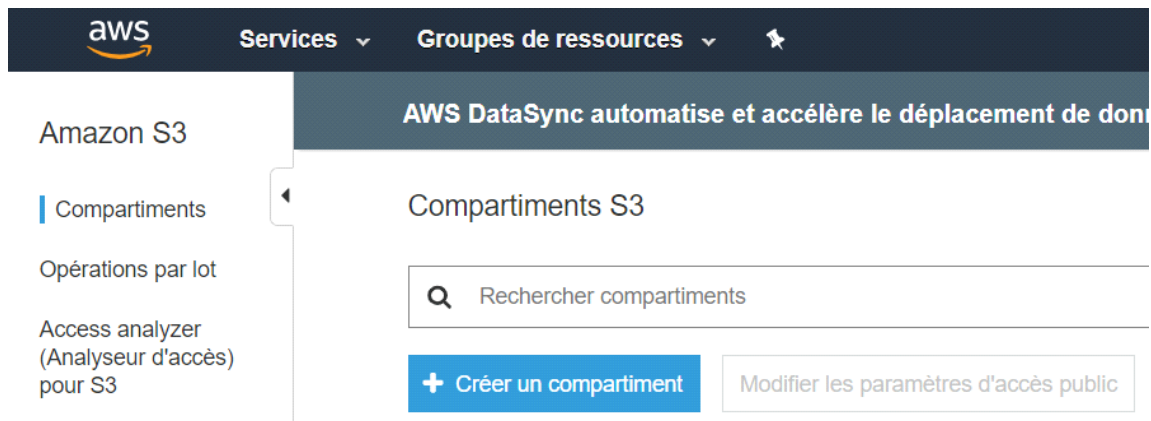
Dans ce qui suit je vais exposer les étapes de la réalisation de ce projet en détails.

Etapes de la réalisation

- **Chargement des données à traiter dans AWS**

Dans un premier temps je vais stocker mes données sur amazon dans le service S3 de AWS

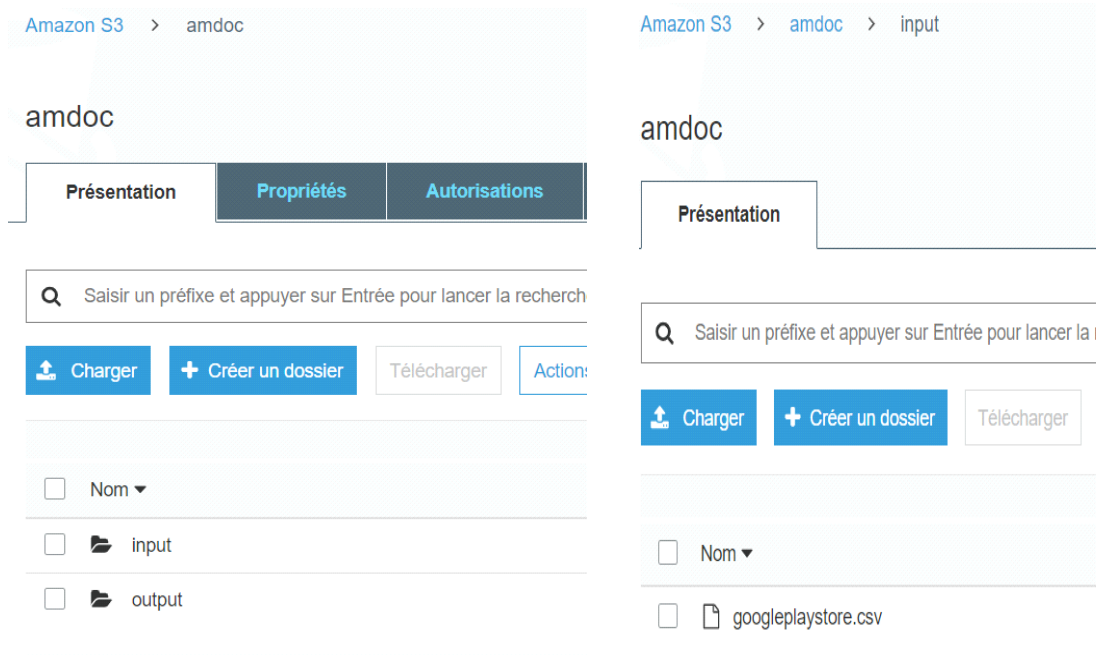
Je crée d'abord un compartiment :



- Figure 1 : création du compartiment -

Après avoir créé le compartiment nous allons l'utiliser pour créer dedans deux sous dossier :

- Un dossier **input** qui va contenir nos données à traiter : le jeu de données **googleplaystore.csv** télécharger depuis Kaggle Datasets .
- Un dossier **output** vide dans lequel on va mettre le résultat des traitements qu'on va effectuer sur nos données.



- Figure 2 : contenu du compartiment amdocr -

- **Création du cluster**

Dans un deuxième temps nous allons créer notre cluster dans le service AWS EMR section “cluster” comme suit :

Pour créer un cluster nous avons besoin d’une paire de clés EC2, nous allons donc créer une clé dans la section **EC2--> RÉSEAU ET SÉCURITÉ --> paires de clés**, en choisissant le format ppk et la télécharger.

The image shows the 'Create key pair' page in the AWS Management Console. The breadcrumb navigation at the top reads 'EC2 > Key pairs > Create key pair'. The main heading is 'Create key pair'. Below it, a description states: 'A key pair, consisting of a private key and a public key, is a set of security credentials that you use to prove your identity when connecting to an instance.' The 'Name' field is a text input containing 'macled'. Below the field, a note says: 'The name can be up to 255 characters long. Valid characters include _ , -, a-z, A-Z, and 0-9.' The 'File format' section has two radio buttons: 'pem' (selected) and 'ppk' (selected). Below 'pem' is the text 'For use with OpenSSH'. Below 'ppk' is the text 'For use with PuTTY'. At the bottom right, there are two buttons: 'Cancel' and 'Create key pair' (orange).


- Figure 3 : création de la clé EC2 -

Maintenant que nous avons notre clé nous allons créer notre cluster avec cette clé

Configuration générale

Nom du cluster

☒ Journalisation ⓘ

Dossier S3 

Mode de lancement ☒ Cluster ⓘ ☐ Exécution d'étape ⓘ

Configuration des logiciels

Configuration du matériel

Type d'instance

Nombre d'instances (1 nœud maître et 2 nœuds principaux)

Sécurité et accès

Paire de clés EC2

Autorisations ☒ Par défaut ☐ Personnalisé

Utilisez les rôles IAM par défaut. Si des rôles sont absents, nous les créons automatiquement pour vous avec des stratégies gérées par défaut.

Rôle EMR [EMR_DefaultRole](#) ⓘ

Profil d'instance EC2 [EMR_EC2_DefaultRole](#) ⓘ

- Figure 4 : création du cluster -

Notre cluster contient par défaut un nœud maître et 2 nœuds esclaves.

Après que notre cluster est créé, dans la partie détails nous accédons à **Sécurité et accès** --> **Groupes de sécurité pour le principal**

Un groupe de sécurité du cluster se compose d'un ensemble de règles qui contrôlent l'accès au cluster. Des règles individuelles identifient un groupe de sécurité Amazon EC2 qui est autorisé à accéder au cluster. Lorsque on associe un groupe de sécurité à un cluster, les règles définies dans celui-ci contrôlent l'accès au cluster.

Pour qu'on puisse accéder au cluster depuis notre poste il faut modifier les 2 groupes de sécurité maître et esclave.

On sélectionne le **groupe** --> **Entrant** --> **modifier** --> **ajouter une règle** --> **type = SSH**, **protocole = TCP**, **plage de port = 22**, **source = mon IP** --> **Enregistrer**

<input type="checkbox"/>	Name	ID du groupe	Nom du groupe
<input checked="" type="checkbox"/>		sg-0e13104b8cffb1101	ElasticMapReduce-master
<input type="checkbox"/>		sg-0fa86d75161e5b7b6	ElasticMapReduce-slave

Groupe de sécurité: sg-0e13104b8cffb1101

Type	Protocole	Plage de ports	Source
Tous les TCP	TCP	0 - 65535	Personnali sg-0e13104b8cffb1101
Tous les TCP	TCP	0 - 65535	Personnali sg-0fa86d75161e5b7b6
SSH	TCP	22	Mon IP 82.224.244.89/32

- Figure 5 : création du compartiment -

Enfin, pour se connecter au nœud maitre à l'aide de SSH et exécuter des requêtes sur nos données il suffit de sélectionner notre **cluster --> afficher les détails --> Récapitulatif --> DNS public principal (SSH)** et suivre les instructions.

SSH

Se connecter au nœud maître à l'aide de SSH

Vous pouvez vous connecter au nœud maître Amazon EMR à l'aide de SSH pour exécuter les requêtes interactives, examiner les fichiers journaux, soumettre les commandes Linux, etc. [En savoir plus](#)

Windows

Mac/Linux

1. Téléchargez PuTTY.exe sur votre ordinateur à partir de : <http://www.chiark.greenend.org.uk/~sgtatham/putty/download.html>
2. Lancez PuTTY.
3. Dans la liste Category, cliquez sur Session.
4. Dans le champ Host Name, tapez **hadoop@ec2-34-254-151-107.eu-west-1.compute.amazonaws.com**
5. Dans la liste Category, développez Connection > SSH, puis cliquez sur Auth.
6. Pour le fichier de clés privées utilisé pour l'authentification, cliquez sur Browse et sélectionnez le fichier de clés privées (**macie.ppk**) utilisé pour lancer le cluster.
7. Cliquez sur Open.
8. Cliquez sur Yes pour ignorer l'alerte de sécurité.

- Figure 1 : connexion au nœud maitre à l'aide de SSH-

En suivant les étapes une ligne de commande nous apparait

```
EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMM RRRRRRRRRRRR
E::::::::::::::::::::E M::::::::M      M::::::::M R:::::::::
EE::::EEEEEEEE::::E M::::::::M      M::::::::M R::::RRRRRR
  E::::E      EEEEE M::::::::M      M::::::::M RR::::R
E::::E      M::::M:M:M      M::M::::M      R:::R
E::::EEEEEEEEEE M::::M M::M M::M M::::M      R::RRRRRR
E::::::::::E M::::M M::M:M:M      M::::M      R:::::::::
E::::EEEEEEEEEE M::::M      M::::M      M::::M      R::RRRRRR
E::::E      M::::M      M::M      M::::M      R:::R
E::::E      EEEEE M::::M      MMM      M::::M      R:::R
EE::::EEEEEEEE::::E M::::M      M::::M      R:::R
E::::::::::E M::::M      M::::M      RRR::::R
EEEEEEEEEEEEEEEEEEEE MMMMMMM      MMMMMMM RRRRRRR

[hadoop@ip-172-31-10-186 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf
properties Async: false
hive> create database playstore
> create database playstore;
FAILED: ParseException line 2:0 missing EOF at 'create' near
hive> create database playstore;
OK
Time taken: 0.27 seconds
```

- Figure 6 : ligne de commande Hive -

- **Traitement et analyse de données**

Dans la ligne de commande :

Crée une base de données

hive> create database if not exists playstore ;

Crée une table pour récupérer notre jeu de données depuis son emplacement dans AWS

hive> use playstore;

create external table googlepls(

App String ,

Category String ,

Rating Decimal ,

Reviews Integer ,

Size String ,

Installs Decimal ,

Type String ,

Price Decimal ,

Content String ,

Genres String ,

LastUp String ,

Curre Decimal ,

Version String) row format delimited fields terminated by ',' location 's3://amdoc/input/';

Effectuer des requêtes sur nos données

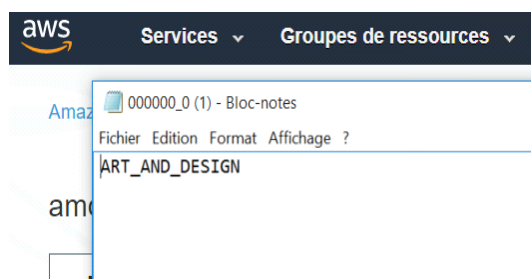
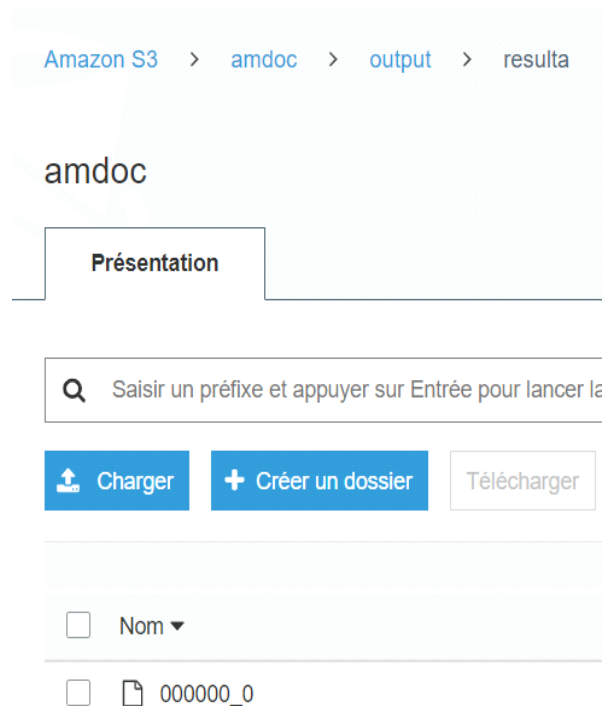
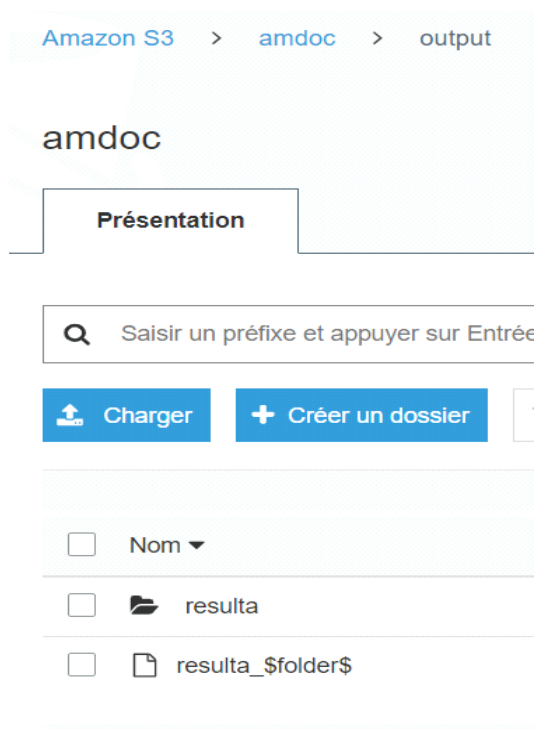
INSERT OVERWRITE DIRECTORY 's3://amdoc/output/resulta/' select Category from googleps where App="Infinite Painter" ;

```
hive> INSERT OVERWRITE DIRECTORY 's3://amdoc/output/resulta/' select Category from googleps where App="Infinite Painter" ;
Query ID = hadoop_20200106134511_b0c3cca0-fd6e-4baf-a4b4-f6825a998c44
Total jobs = 1
Launching Job 1 out of 1
Tez session was closed. Reopening...
Session re-established.
Status: Running (Executing on YARN cluster with App id application_1578311552119_0005)

-----
      VERTICES      MODE      STATUS  TOTAL  COMPLETED  RUNNING  PENDING  FAILED  KILLED
-----
Map 1 ..... container      SUCCEEDED      1          1          0          0          0          0
-----
VERTICES: 01/01  [=====>>>] 100%  ELAPSED TIME: 9.41 s
-----
Moving data to directory s3://amdoc/output/resulta
OK
Time taken: 15.385 seconds
hive> █
```

- Figure 7 : requête Hive -

Le résultat dans le dossier output dans S3



7	Smoke Effect Photo Maker - Smoke Editor	ART_AND_DESIGN	3.8
8	Infinite Painter	ART_AND_DESIGN	4.1
9	Garden Coloring Book	ART_AND_DESIGN	4.4
10	Kids Paint Free - Drawing Fun	ART_AND_DESIGN	4.7

- Figure 7 : résultat de la requête dans S3 output -

