Evaluation hive

1. Ajouter le fichier dans HDFS

Un ensemble d'instruction pour télécharger les données et les placer dans Hive

a) bash command

mkdir data

```
wget https://datasets.imdbws.com/name.basics.tsv.gz
```

sudo cp name.basic.tsv data/

chown formation data.tsv

docker exec -it hive-server bash

hdfs dfs -mkdir /user/dataset

hdfs dfs -put /data/data.tsv /user/dataset

hive

b) Commande hive

Une proposition d'architecture des différentes tables (schémas et choix d'optimisation du stockage justifiés)

J'ai opté pour la création des buckets car il m'a été difficile de trouver une colonne avec les éléments pouvant être regrouper car les données sont distinctes. J'ai opté pour la création d'une table dont les données seront partitionnées en fonction de la valeur de birthYear et du nombre de bucket(pour éviter d'avoir trop de partition).

```
CREATE DATABASE evaluation;
USE evaluation;

CREATE TABLE person
(
   nconst string,
   primaryName string,
```

```
   birthYear INT,
   deathYear INT,
   primaryProfession array<string>  ,
   knownForTitles array<string>

)
ROW FORMAT DELIMITED
STORED AS TEXTFILE;

LOAD DATA INPATH '/user/dataset/data.tsv'
INTO TABLE person;

Describe person ;
SET hive.exec.dynamic.partition=true;
SET hive.exec.dynamic.partition.mode=nonstrict;

CREATE TABLE person_partitions_buckets
(
   nconst string,
   primaryName string,
   deathYear INT,
   primaryProfession array<string>,
   knownForTitles array<string>

)
PARTITIONED BY (birthYear INT)
CLUSTERED BY (deathYear) INTO 32 BUCKETS
ROW FORMAT DELIMITED
LINES TERMINATED BY "\n"
STORED AS TEXTFILE;
```

```
INSERT INTO TABLE person_partitions_buckets partition(birthYear)

SELECT nconst, primaryName,  deathYear, primaryProfession, knownForTitles, birthYear

FROM person;


Commande bash permettant de visualiser les différentes partitions

hdfs dfs -ls /user/hive/warehouse/evaluation.db/person_partitions_buckets ;

# on explore le premier bucket

hdfs dfs -ls
/user/hive/warehouse/evaluation.db/person_partitions_buckets/birthyear=__HIVE_DEFAULT_PARTI
TION__

hdfs dfs  -cat
/user/hive/warehouse/evaluation.db/person_partitions_buckets/birthyear=__HIVE_DEFAULT_PARTI
TION__/000000_0
```

Quelques requêtes pour démontrer l'utilisation de la base

```
describe person;


SELECT * FROM person LIMIT 5;

SELECT birthYear, COUNT(*) FROM person_partitions_buckets GROUP BY birthYear;

SELECT nconst, birthYear FROM person WHERE birthYear >1800 LIMIT 5;

SELECT nconst, primaryName, birthYear,deathYear FROM person ORDER BY deathYear LIMIT 5;

SELECT  count(*)  from person GROUP BY birthYear;
```