

## Démo du notre projet

**Pour activer le serveur de hadoop + et le service de hive**

```
startCDH.sh
```

**Pour consulter le fichier journal.txt**

```
cat journal.txt
```

**Pour stocker le fichier journal.txt dans le hdfs**

```
hdfs dfs -put out.txt hadoop
```

**Pour consulter le contenu de fichier journal.txt dans le hdfs**

```
hdfs dfs -cat hadoop/journal.txt
```

**Pour connecte au hive**

```
hive
```

**Création la base pds**

```
create database pds;
```

**Pour afficher les bases de données existes**

```
show databases;
```

**Pour utiliser la base pds**

```
use pds ;
```

**Pour créer la table globale**

```
create table journal(id int , name string , country string , date_j string , title string ,  
author string , content string )  
ROW FORMAT DELIMITED  
FIELDS TERMINATED BY ':'  
STORED AS TEXTFILE;
```

**Pour extraire les données de fichier journal.txt (stocker dans hdfs) vers la table hive**

```
load data inpath 'hadoop/journal.txt' into table journal ;
```

**Creation le table de fact et les tables de dimension**

**Creation table de fact**

```
create table journal_fact(id int , name string , title string , content string );
```

**Creation les table de dimension**

```
create table author_dim(id int , author string);  
create table country_dim (id int , country string) ;  
create table date_dim (id int , date_j string );
```

## **Pour remplir la table de faits et les tables dimensions**

### **Remplir la table de faits**

```
insert into table journal_fact select id, name , title , content from journal ;
```

### **Remplir les tables dimension**

```
insert into table date_dim select id, date_j from journal ;
```

```
insert into table country_dim select id, country from journal ;
```

```
insert into table author_dim select id, author from journal ;
```

## **Pour analyser les données et stocker les résultats dans un fichier**

```
hive -e "select a.author , count(*) from pds.author_dim a INNER JOIN pds.journal_fact j  
on j.id = a.id group by a.author" > out3.txt
```

```
hive -e "select d.date_j , count(*) from pds.date_dim d INNER JOIN pds.journal_fact j on  
j.id = d.id group by d.date_j" > out2.txt
```

```
hive -e "select c.country , count(*) from pds.country_dim c INNER JOIN pds.journal_fact j  
on j.id = c.id group by c.country" > out1.txt
```

## **Pour consulter la resultat dans les fichiers**

```
cat out1.txt
```

```
cat out2.txt
```

```
cat out3.txt
```