Une image contenant Graphique, logo, graphisme, cercle

Description générée automatiquementMESIIN480323 : No SQL 19/01

Project Report

On

CASSANDRA

**reuters.json**

Gabriel Chaix & Malo Leroy & Charles Terrier

École Supérieure d’Ingénieurs Léonard de Vinci

Année 2023 / 2024

Index

[Introduction and Problem Statement 2](#_Toc157616834)

[I. Data Importation 3](#_Toc157616835)

[II. Data cleaning 4](#_Toc157616836)

[III. Querys 5](#_Toc157616837)

[A. Simple querys (6) 5](#_Toc157616838)

[1. List of reuters 5](#_Toc157616839)

[2. List of texts titles 5](#_Toc157616840)

[3. Titles, id and dateline of texts where id “120” 5](#_Toc157616841)

[4. Count of reuters whose place is France. 5](#_Toc157616842)

[5. Count of “texts” publications, 6](#_Toc157616843)

[6. Count of texts WHERE title contains “News”, 6](#_Toc157616844)

[B. Complex querys (2) 7](#_Toc157616845)

[1. Count the number of articles having the topic ‘earn’. 7](#_Toc157616846)

[2. Give the number articles written by each author ? 7](#_Toc157616847)

[C. Hard query (1) 8](#_Toc157616848)

[1. How much article there is by place 8](#_Toc157616849)

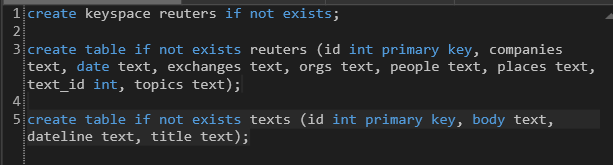
# Introduction and Problem Statement

Reuters is a news agency founded in 1851 in London. It is one of the global and generalist news agencies, an activity that represents a part of its revenue, mainly devoted to financial information. It is one of the largest news agencies in the world. We want to create a Cassandra database with reuters articles and query it.

# Data Importation

In a first manner, we must create our database using TablePlus. To do this, we run the following queries to build the database and then create the two tables that compose our database: “reuters” and “texts”.

Table creation - Figure 1.1



Once this is done, we must fill the tables with our dataset content. After downloading our [json file](https://learning.devinci.fr/pluginfile.php/217984/mod_folder/intro/reuters.json.zip) we added brackets [] at the start and the end and used replace all to add a comma at the end of each line, in order to make it a valid json.

Une image contenant texte, capture d’écran

Description générée automatiquementThen here is the importation script we’ve wrote

←

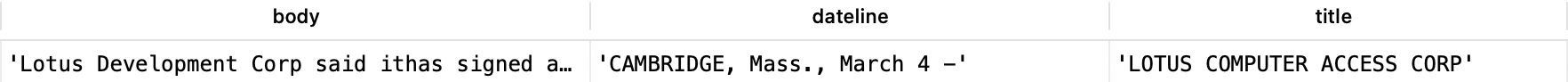
After importing all the rows, we make sure that each rows is correctly imported.

# Data cleaning

Une image contenant texte, capture d’écran

Description générée automatiquementThe first step we’ve take into cleaning had been to update all the empty entry to NULL. To address this problem, we wrote a python script.

We also observe that ever text attribute had quotes, due to the importation script, since we did not success in modifying the importation script, we added a similar python function to take them off.



# Querys

## Simple querys (6)

### List of reuters

Une image contenant texte, capture d’écran, Police, logiciel

Description générée automatiquement

### List of texts titles

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

### Titles, id and dateline of texts where id “120”

Une image contenant texte, logiciel, Page web, Site web

Description générée automatiquement

### Count of reuters whose place is France.

Une image contenant texte, logiciel, multimédia, Logiciel multimédia

Description générée automatiquement

### Count of “texts” publications,

Une image contenant texte, capture d’écran, logiciel, Système d’exploitation

Description générée automatiquement

### Count of texts WHERE title contains “News”,

In order to use the LIKE statement, we need to create a SASI index, let’s enable it in cassandra.yaml file.

Une image contenant texte, capture d’écran, Police, algèbre

Description générée automatiquement

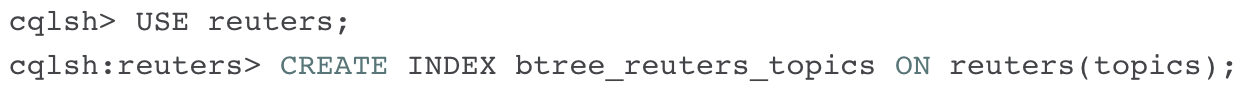
Then we can create our custom index and use LIKE statement on the desire column.Une image contenant texte, logiciel, ordinateur, Système d’exploitation

Description générée automatiquement

## Complex querys (2)

### Count the number of articles having the topic ‘earn’.

In order to use the WHERE =’’ statement, we had to create a btree indexation on the topic column of reuters table.



Une image contenant texte, capture d’écran, affichage, logiciel

Description générée automatiquement

Une image contenant texte, logiciel, multimédia, Logiciel multimédia

Description générée automatiquement

### Give the number articles written by each author ?

To achieve this query, we need to make a join between reuiters and text tables. To do this in CQL, we unormalize the reuiters table by adding to it an attribute for author name. We create a new table and fill and clean it using our python script:

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Now that author is part of the primary key, we can group by author.

Une image contenant texte, capture d’écran, logiciel, Logiciel multimédia

Description générée automatiquement

## Hard query (1)

### How much article there is by place

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, capture d’écran, Police

Description générée automatiquement

Une image contenant texte, capture d’écran, logiciel

Description générée automatiquement