Google Cloud platform

Hive

Project phase 4

The reason I used Hive on GCP is, Cloudera VM (old version) did not support Serde library which is able to sort and store data into the table. Our dataset has quite long string value in each column which is needed to be partitioned properly and Google Cloud Platform is quite convenient enough to deal with our dataset due to the property of various new library support.

Step 1. **Use Hive on google cloud platform** : beeline -u jdbc:hive2://localhost:10000/default -n kinznzl@bigdata-lab-m -d org.apache.hive.jdbc.HiveDriver --incremental=true

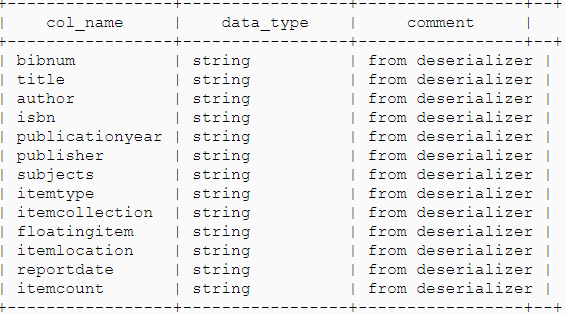
( \* incremental allows to use a lot of buffers in case I use select several columns )

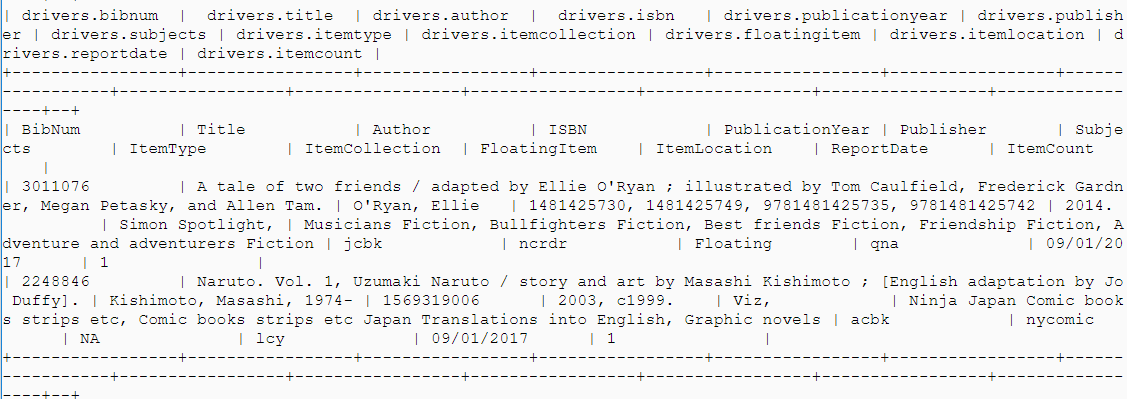
Step 2. **Create two tables including information of two csv files**.

create external table drivers(BibNum string, Title string, Author string, ISBN string, PublicationDate string, Publisher string, Subjects string, ItemType string, ItemCollection string, FloatingItem string, ItemLocation string, ReportDate string, ItemCount string) row format serde ‘org.apache.hadoop.hive.serde2.openCSVSerde’ stored as textfile load data location ‘gs://cpsc651-1024474/inventory/’;

desc drivers;

select \* from drivers limit 3;

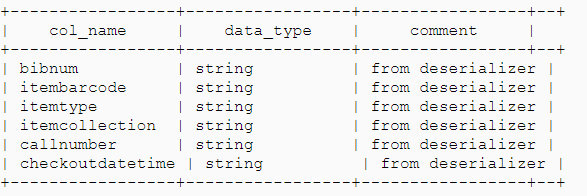


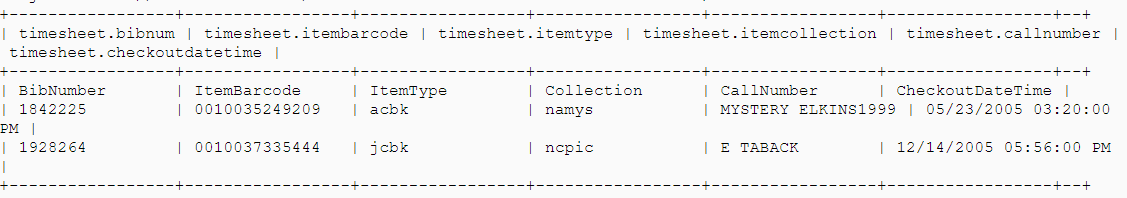


create external table timesheet(BibNum string, ItemBarcode string, ItemType string, ItemCollection , CallNumber string, CheckoutDateTime string) row format serde ‘org.apache.hadoop.hive.serde2.openCSVSerde’ stored as textfile load data location ‘gs://cpsc651-1024474/checkouts/’;

desc timesheet;

select \* from timesheet limit 3;





Step 3. **Make query to join each table ( drivers and timesheet ) .**

**As a result, I can extract the most popular author and how many books are borrowed upon author.**

select t2.Author,

count(t1.BibNum) as cnt

from timesheet t1

Inner join drivers t2 on trim(t1.BibNum) = trim(t2.BibNum)

where length(trim(t2.Author)) > 0

group by t2.author

order by cnt desc

limit 5;

WHERE LENGTH(trim(t2.Author)) > 0 # it checks if there is at least one Letter on author name

