Урок 3. Работа с данными через Hive и Hue. Обзор PIG и Impala

Выполнил Колеганов Н.Д.

Задание

[исследовательское задание] Разобраться с тем, что такое bucketing.

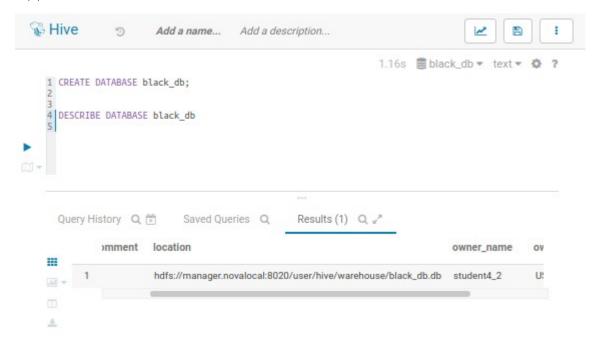
- 1. Скачать любой датасет из списка ниже https://www.kaggle.com/shuyangli94/food-com-recipes-and-user-interactions https://www.kaggle.com/datasnaek/youtube-new https://www.kaggle.com/akhilv11/border-crossing-entry-data https://www.kaggle.com/tristan581/17k-apple-app-store-strategy-games https://www.kaggle.com/gustavomodelli/forest-fires-in-brazil
- 2.Загрузить этот датасет в HDFS в свою домашнюю папку
- 3. Создать собственную базу данных в HIVE
- 4.Создать EXTERNAL таблицы внутри базы данных с использованием всех загруженных файлов. Один файл одна таблица.
- 5.Сделать любой запрос по загруженным данным используя груповые и агрегатные функции.
- 6.Сделать любой запрос по загруженным данным используя JOIN.
- 7.[Продвинутый вариант] Сделать все вышеперечисленное с использованием JSON SerDe. Подсказка: см в сторону команды «ADD JAR» Если последнее задание выполнено -- то пункты 2-5 не обязательно делать.

Решение

Bucketing - это метод оптимизации, который использует **сегменты** (и **столбцы**), чтобы определить разбиение данных и избежать перемешивания данных.

Мотивация состоит в том, чтобы оптимизировать производительность запроса на соединение, избегая перемешивания (или *обмена*) таблиц, участвующих в соединении. Вucketing приводит к меньшему количеству обменов (этапов).

Создал БД black_db



Создал таблицу forest

STORED AS INPUTFORMAT

```
create table black_db.forest

(

year__ int,

state string,

month_ string,

number int,

date_ string
)

ROW FORMAT SERDE

'org.apache.hadoop.hive.serde2.OpenCSVSerde'
```

'org.apache.hadoop.mapred.TextInputFormat'

OUTPUTFORMAT

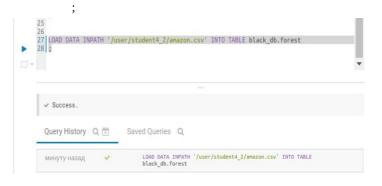
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'

TBLPROPERTIES (

```
'serialization.null.format' = ",
'skip.header.line.count' = '1')
```

```
create table black_db.forest
8 9
    year_
                     string,
string,
10
     state
11
     month
12
    number
                             int,
13
                  string
     date_
14
15
     ROW FORMAT SERDE
     'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS INPUTFORMAT
16
17
18
           'org.apache.hadoop.mapred.TextInputFormat'
19
     OUTPUTFORMAT
20
            'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat'
21
22
23
24
     TBLPROPERTIES (
           'serialization.null.format' = '',
           'skip.header.line.count' = '1')
25
      Success.
                                        Saved Queries Q
   Query History Q 🛱
                                               create table black_db.forest ( year_ int, state string, month_ string, number int, date_ string ) ROW FORMAT SERDE 
'org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS INPUTFORMAT
   2 минуты назад
                                               'org.apache.hadoop.mapred.TextInputFornat' OUTPUTFORNAT
'org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextDutputFornat'
TBLPROPERTIES ('serialization.null.format' = '',
'skip.header.line.count' = '1')
```

LOAD DATA INPATH '/user/student4_2/amazon.csv' INTO TABLE black_db.forest



Co	olumns Details	Sample Ana	ılysis		
	forest.year_	forest.state	forest.month_	forest.number	forest.date_
	1998	Acre	Janeiro	0	1998-01-01
2	1999	Acre	Janeiro	0	1999-01-01
3	2000	Acre	Janeiro	0	2000-01-01
4	2001	Acre	Janeiro	0	2001-01-01
5	2002	Acre	Janeiro	0	2002-01-01
6	2003	Acre	Janeiro	10	2003-01-01
7	2004	Acre	Janeiro	0	2004-01-01
8	2005	Acre	Janeiro	12	2005-01-01
9	2006	Acre	Janeiro	4	2006-01-01
10	2007	Acre	Janeiro	0	2007-01-01
11	2008	Acre	Janeiro	0	2008-01-01
12	2009	Acre	Janeiro	0	2009-01-01
13	2010	Acre	Janeiro	1	2010-01-01
14	2011	Acre	Janeiro	0	2011-01-01
15	2012	Acre	Janeiro	0	2012-01-01

Создал таблицу appstore

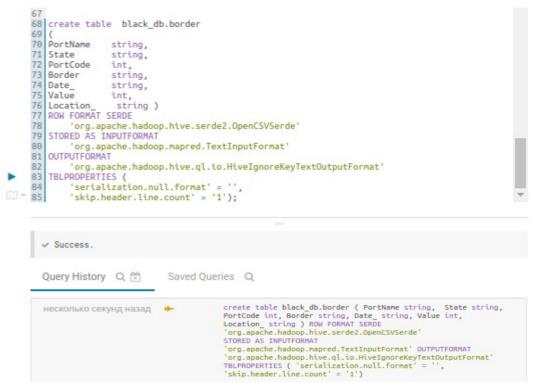
```
create table black_db.appstore
            string,
URL
ID
         int,
Name
                 string,
Subtitle
                    int,
IconURL string,
AvrUsrRate float,
UsrRateCount float,
Price float,
InAppPurchase string,
Description string,
Developer string,
AgeRate string,
Lang
              string,
Size
              int,
PrimGenre string,
Genres string,
OriginalRelease string,
CurRelease string
)
ROW FORMAT SERDE
   'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS INPUTFORMAT
   'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT
   'org. apache. hadoop. hive. ql. io. Hive Ignore Key Text Output Format'\\
TBLPROPERTIES (
   'serialization.null.format' = ",
   'skip.header.line.count' = '1')
LOAD DATA INPATH '/user/student4_2/appstore_games.csv' INTO TABLE black_db.appstore
                            Query History Q 🛱
                                                              Saved Queries Q
                                                                    create table black db.appstore ( URL string, ID int, Name string, Subtitle int, IconURL string, AvrUsrRate float, UsrRateCount float, Price float, InAppPurchase string, Description string, Developer string, AgeRate string, Lang string, Size_int, PrimGenre string, Genres string, OriginalRelease string, CurRelease string) ROM FORMAT SEMB() org.apache.hadoop.hive.serde2.OpenCSVSerde' STORED AS INPUTFORMAT 'org.apache.hadoop.hive.gl.io.HiveIgnoreKeyTextOutputFormat' TBLPROPERTIES ('serialization.null.format' = '', 'skip.header.line.count' = '1')
                            минуту назад
                            LOAD DATA INPATH '/user/student4_2/appstore_games.csv' INTO TABLE black_db.appstore
                                                             Saved Queries Q
                           Query History Q P
                                                                            LOAD DATA INPATH '/user/student4_2/appstore_games.csv' INTO TABLE black_db.appstore
                           несколько секунд назад
```

Columns Details Sample Analysis

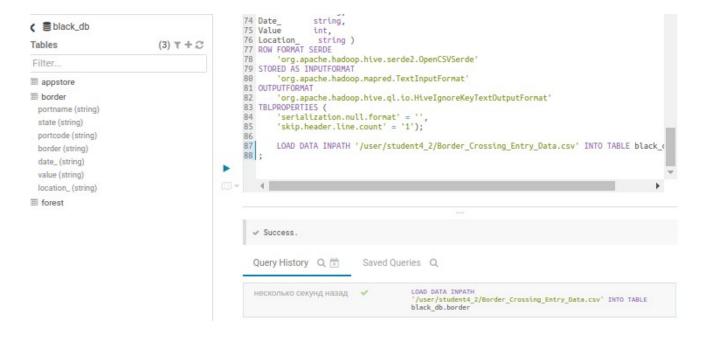
	appstore.url	appstore.id	appstore.name
1	https://apps.apple.com/us/app/sudoku/id284921427	284921427	Sudoku
2	https://apps.apple.com/us/app/reversi/id284926400	284926400	Reversi
3	https://apps.apple.com/us/app/morocco/id284946595	284946595	Morocco
4	https://apps.apple.com/us/app/sudoku-free/id285755462	285755462	Sudoku (Free)
5	https://apps.apple.com/us/app/senet-deluxe/id285831220	285831220	Senet Deluxe
6	https://apps.apple.com/us/app/sudoku-classic-number-puzzle/id286210009	286210009	Sudoku - Classic number puzzl
7	https://apps.apple.com/us/app/gravitation/id286313771	286313771	Gravitation
8	https://apps.apple.com/us/app/colony/id286363959	286363959	Colony
9	https://apps.apple.com/us/app/carte/id286566987	286566987	Carte
10	https://apps.apple.com/us/app/barrels-o-fun/id286682679	286682679	"Barrels O' Fun"
11	https://apps.apple.com/us/app/quaddraxx/id287563734	287563734	Quaddraxx
12	https://apps.apple.com/us/app/lumen-lite/id288096268	288096268	Lumen Lite
13	https://apps.apple.com/us/app/bubblepop/id288669794	288669794	BubblePop
14	https://apps.apple.com/us/app/marple/id288689440	288689440	Marple
15	https://apps.apple.com/us/app/tetravex-lite/id288918962	288918962	Tetravex Lite

Создал таблицу border

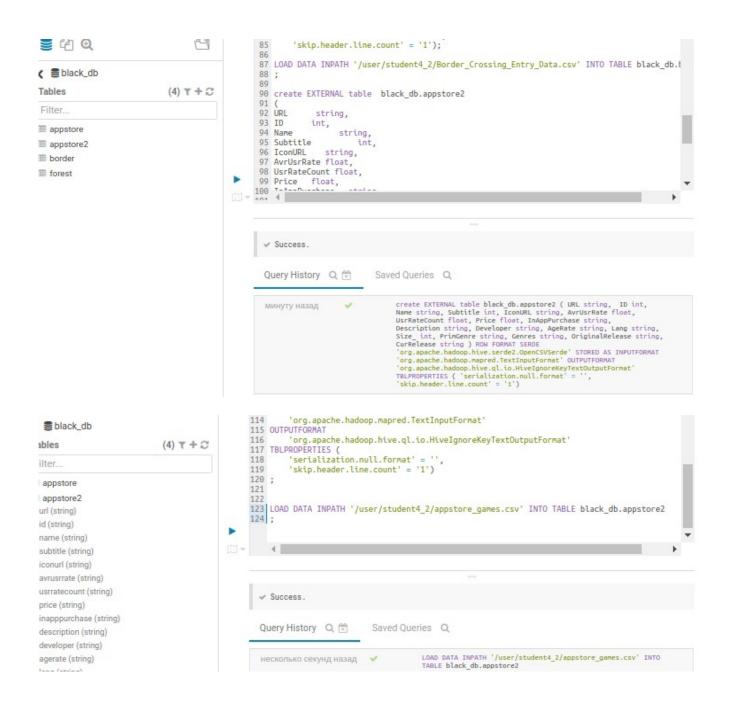
```
create table black_db.border
PortName string,
State string,
PortCode int,
Border string,
Date_
         string,
Value
         int,
Location_ string)
ROW FORMAT SERDE
  'org.apache.hadoop.hive.serde2.OpenCSVSerde'
STORED AS INPUTFORMAT
  'org.apache.hadoop.mapred.TextInputFormat'
OUTPUTFORMAT
  'org. a pache. hadoop. hive. ql. io. Hive Ignore Key Text Output Format'\\
TBLPROPERTIES (
  'serialization.null.format' = ",
  'skip.header.line.count' = '1');
```



LOAD DATA INPATH '/user/student4_2/Border_Crossing_Entry_Data.csv' INTO TABLE black_db.border



Потом заметил что нужно делать external таблицы и сделал тоже самое с ключевым словом EXTERNAL.



Примеры запросов:

select id,name,originalrelease from black_db.appstore;

