درس : Big Data
استاد : خانم دکتر صفری
دانشجو : رقیه یزدانی
تمرین : HBase

HBase

ایجاد بستر لازم برای امکان کار با HBase

قبلاً فایل cloudera-quickstart-vm-5.13.0-0-vmware که درواقع ماشین مجازی cloudera می‌باشد، دانلود شده و با استفاده از نرم افزار VMWare Player 15.5 امکان کار با آن فراهم شده است و در تمرین Hive، از Hue Interface که از طریق url زیر در دسترس بود، استفاده شد.

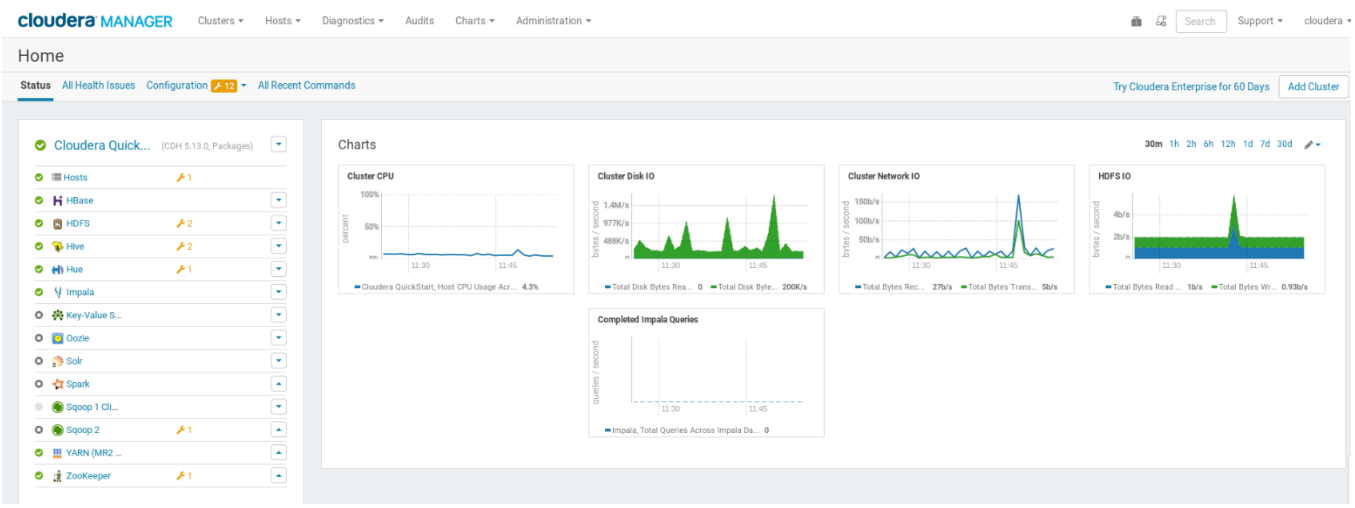
http://localhost:8888/

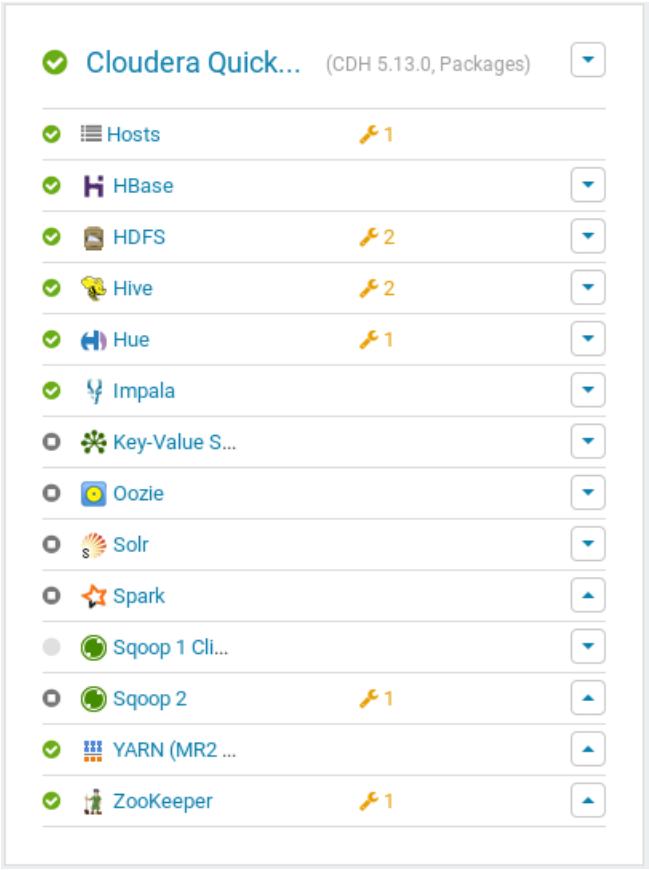
برای این تمرین، لازم است که cloudera manager(CM) را launch کرده باشیم. برای اینکار لازم است که کانفیگ ماشین مجازی cloudera را تغییر داده و اندازه RAM آنرا، به حداقل 8GB و تعداد cpu را به حداقل، 2 عدد افزایش دهیم. (بصورت پیش فرض، 4GB RAM و 1 cpu دارد. بعد از انجام این تغییرات، ماشین مجازی را start می‌کنیم.

روی desktop یک آیکن به نام Launch Cloudera Express وجود دارد که با کلیک بر روی آن، عملیات آماده‌سازی cloudera manager انجام می‌شود، اینکار تنها یکبار نیاز است و در دفعات بعدی که ماشین مجازی را start می‌کنیم، نیازی به تکرار آن نیست. پس از تکمیل عملیات که چند دقیقه ای طول می‌کشد، در browser با کلیک بر روی آیکن cloudera manager یا وارد کردن url زیر می‌توان به آن متصل شد.

http://localhost:7180/

کلمه کاربری و رمز عبور، cloudera می‌باشد. صفحه اول آن، مشابه تصویری است که در ادامه می‌آید. در سمت چپ تصویر، سرویسهای مختلفی که امکان کار با آنها از طریق CM فراهم شده است، دیده می‌شود از جمله HBase، HDFS، Hive، Impala و ... . همچنین آیکن تیک سبزرنگ کنار آنها، نشان می‌دهد که درحال حاضر، آن سرویس در وضعیت start بوده و قابل استفاده است در غیراینصورت آن سرویس down بوده و نمی‌تواند به درخواستها پاسخ دهد.





پس از launch کردن CM و یا در دفعات بعدی، پس از روشن کردن VM و اتصال به CM، این سرویسها، پس از گذشت چند دقیقه بصورت اتوماتیک بالا می‌آیند اما اگر چنین نشد برای امکان کار با HBase لازم است که سرویسهای زیر به ترتیب start شوند.

- Start HDFS
- Start YARN
- Start Hive

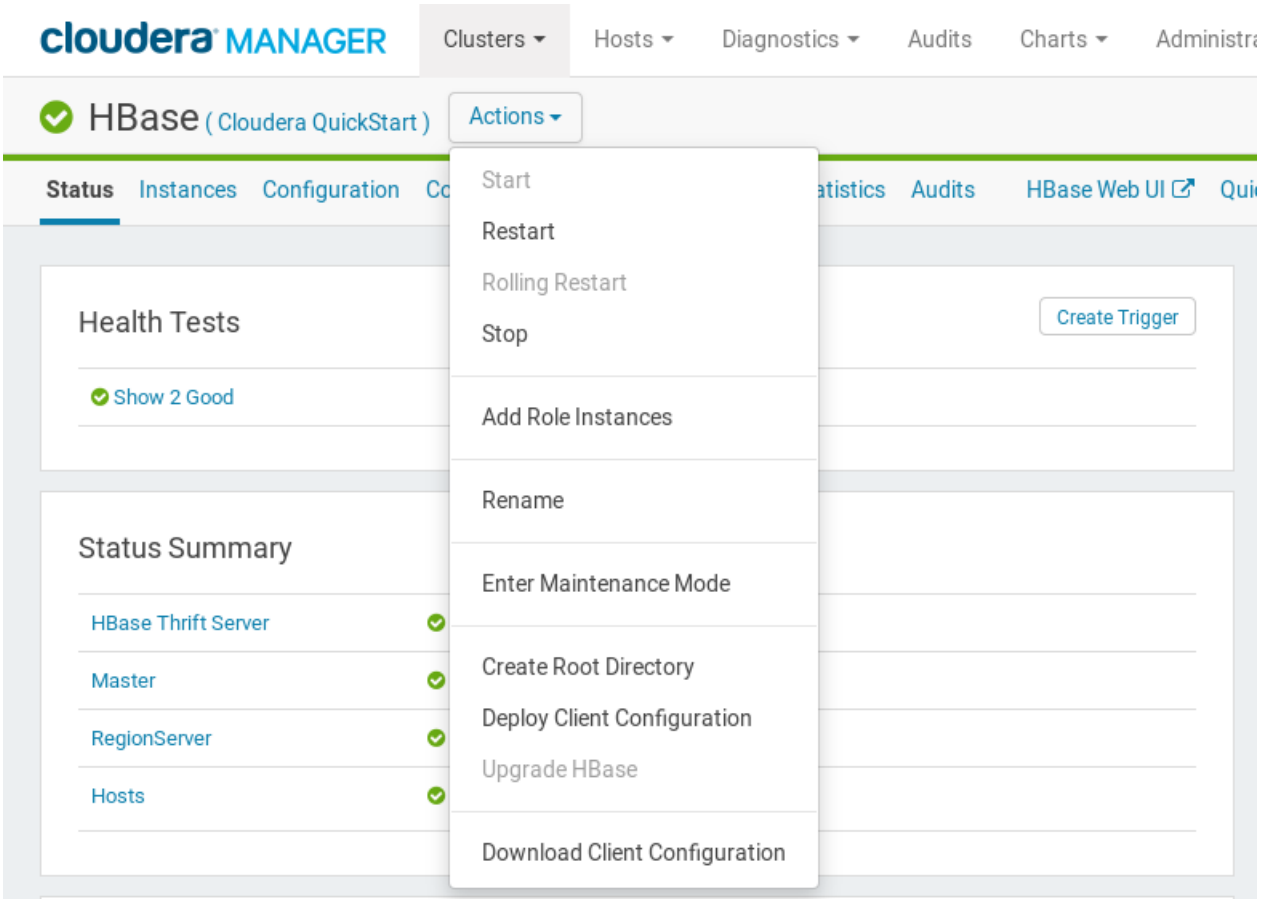
- Start ZooKeeper
- Start HBase

برای start کردن آنها، می‌توان از طریق خط فرمان و با نوشتن command اینکار را انجام داد و یا با کلیک بر روی هر سرویس، وارد صفحه مربوطه شده و با کلیک بر روی دکمه Action، آن سرویس را start، stop یا restart کرد. اگر هنگام بالا آمدن هر یک از سرویسها، خطایی تولید شود، می‌توان با خواندن StdErr یا Stdout و یا Full Log در صفحه‌ای که ظاهر می‌شود، علت خطا را متوجه شده و در جهت رفع آن اقدام نمود.

سرویسهای مربوط به HBase شامل سه سرویس زیر می باشد :

- Master
- Region Server
- Thrift Server

که هر سه باید در وضعیت Good Health باشند، تا بتوان انجام تمرین را شروع کرد.



ایجاد dataset مربوط به تمرین

برای این تمرین، قرار است با دیتاست analytics_demo کار شود. با توجه به اینکه دیتاست پیش فرضی در HBase وجود ندارد، پس از جستجو در اینترنت، لینک زیر که امکان ایجاد دیتای تستی جدول analytics را فراهم می‌کند، پیدا شد:

https://gethue.com/hadoop-tutorial-how-to-create-example-tables-in-hbase/

در این لینک؛ دو فایل create_schema.py و load-data.sh را در اختیار قرار می‌دهد، که اولی یک برنامه پایتون است که داده‌های تستی را تولید کرده و در فایل hbase-analytics.tsv ذخیره می‌کند. فایل دوم هم، داده های موجود در این فایل تولید شده را در جدول analytics از HBase بارگذاری می‌کند.

مراحل کار بصورت زیر است :

- انتقال فایل create_schema.py به مسیر /tmp از ماشین مجازی cloudera
- دادن دسترسی اجرا به آن و سپس اجرای آن

```
$ chmod +x create_schema.py

./ create_schema.py
```

- اتصال به hbase و ایجاد جدول analytics با سه Column Family به نامهای hour و day و total. این کار هم از طریق اینترفیس قابل انجام است و هم از طریق خط فرمان. در زیر دستورات لازم برای اجرا از طریق خط فرمان، آمده است :

```
$ hbase shell

hbase> create 'analytics', 'hour','day','total'

hbase> scan ‘analytics’
```

- بارگزاری داده‌های فایل تولید شده (/tmp/ hbase-analytics.tsv) در جدول ایجاد شده.

برای اینکار ابتدا فایل load_data.sh به مسیر /tmp از ماشین مجازی cloudera منتقل شده و دسترسی اجرا به آن داده می‌شود. سپس، باید، فایل حاوی دیتا، را به hdfs منتقل نموده و سپس فایل load_data.sh را اجرا کرد. لازم به ذکر است که بارگذاری دیتا در جدولی در HBase در واقع از طریق MapReduce انجام می‌شود. دستورات استفاده شده، بصورت زیر می‌باشد :

```
$ chmod +x download load_data.sh

$ hdfs dfs -copyFromLocal /tmp/hbase-analytics.tsv /tmp/
```

```
$ ./load_data.sh
```

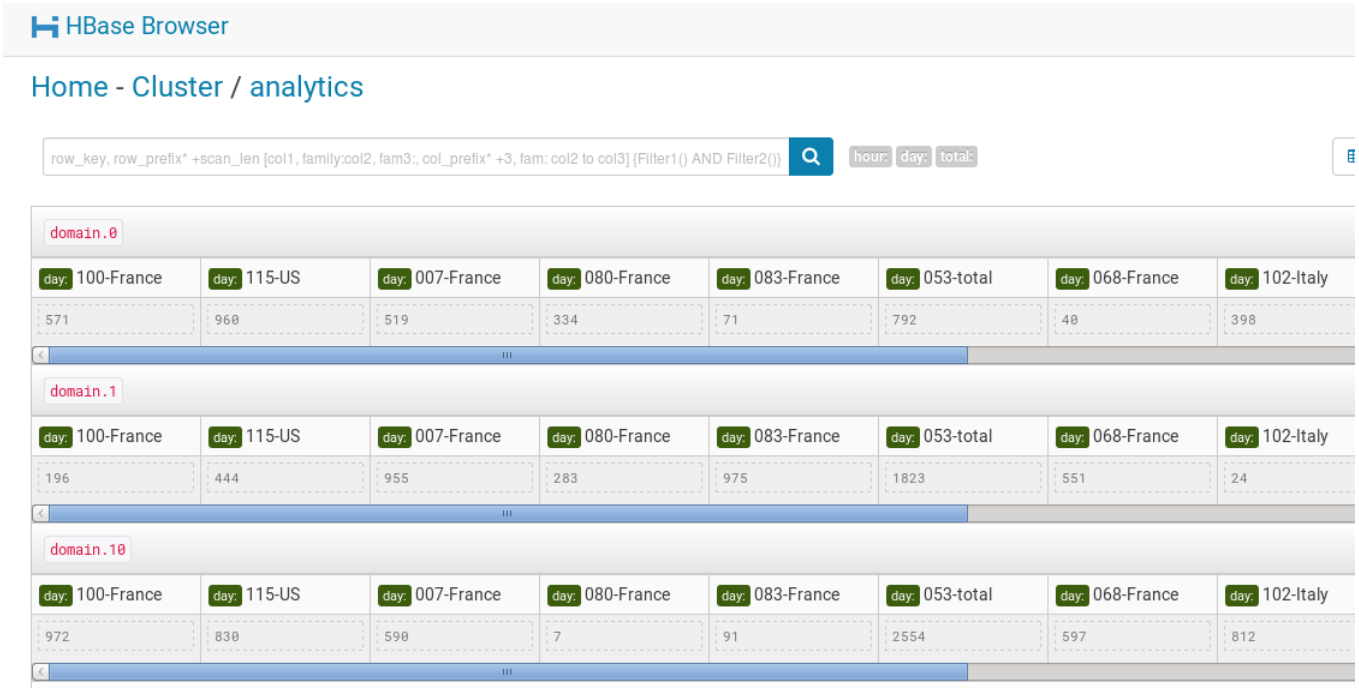
حال می‌توان از طریق Hue Interface داده های این جدول را در HBase Browser مشاهده کرد. اگر سرویس Hue پایین باشد، میتوان آن را از طریق اینترفیس، مشابه سرویسهای قبلی، بالا آورد و یا در خط فرمان، دستور زیر را اجرا کرد :

```
$ sudo service hue start
```

پس از بالا آمدن سرویس Hue، از طریق url زیر می‌توان به Hbase Browser متصل شد :

http://localhost:8888/hue/hbase

در این صفحه، جدول analytics دیده می‌شود که پس از کلیک بر روی آن، داده های آن بصورت زیر نمایش داده می‌شود :



از طریق این اینترفیس، می‌توان بین row های مختلف جدول جستجو کرد، یک row جدید ایجاد کرد، داده های یک row را ویرایش کرد مثلاً یک ستون جدید به یک CF از یک row اضافه یا کم کرد، و ...

توضیح ساختار دیتای جدول analytics

در این جدول سه CF(Cloumn Family) وجود دارد به نامهای hour، day و total. هر سطر، یک row key دارد مثلاً domain.1 یا domain.15 یا ... .

هر سطر اطلاعات site visit مربوط به هر ساعت (1 تا 24) از روزهای مختلف سال (1 تا 365) را در ناحیه مربوطه (domain) به ازای سه کشور مختلف نشان می‌دهد و اطلاعات سرجمع(total) را هم در سطح سال، روز و ساعت و کشور فراهم می‌کند.

در شکل زیر بصورت شماتیک، این ساختار نشان داده شده است :

hour	1-24									day	1-365						total			
1				2				...	24	1				2	...	365	Fran	Italy	US	total
Fran	Italy	US	total	Fran	Italy	US	total	...	...	Fran	Italy	US	total							
10	5	100	115	...						100	50	1000	1150	...		...	...	...	...	...

CF hour از هر row، به ازای هر ساعت (از 1 تا 24) و هر کشور (France ، Italy، US) که تشکیل یک Column را می‌دهد، تعداد site visit را نشان می‌دهد و به ازای هر ساعت، یک ستون با نام total هم دارد که جمع site visit کل کشورها را در آن ساعت نشان می‌دهد.

CF day از هر row، به ازای هر روز از سال (از 1 تا 365) و هر کشور (France ، Italy، US) که تشکیل یک Column را می‌دهد، تعداد site visit را نشان می‌دهد و به ازای هر روز، یک ستون با نام total هم دارد که جمع site visit کل کشورها را در آن روز نشان می‌دهد.

CF total از هر row، به ازای هر کشور (France ، Italy، US)، تعداد site visit را در آن ناحیه، نشان می‌دهد و ستون total در آن، جمع site visit کل کشورها را در آن ناحیه نشان می‌دهد.

به عنوان مثال در سطر domain.0، ستون با نام 'hour:16-Italy' دارای value برابر 31 است که بدین معناست که در ساعت 16، سایت کشور Italy، 31 بار visit شده است. بطور مشابه ستون با نام 'day:200-France' دارای مقدار 58 است، یعنی در روز 200 از سال، سایت کشور فرانسه 58 بار visit شده است.

تمرین 1

1. Utilizing the analytics_demo data set via HBase data browsing, using the prefix search capability - within domain 15 **which country France, Italy or US had the most visits?** (Hint: use the row key search for each country; write the result for each country)

a) US
b) Italy
c) France

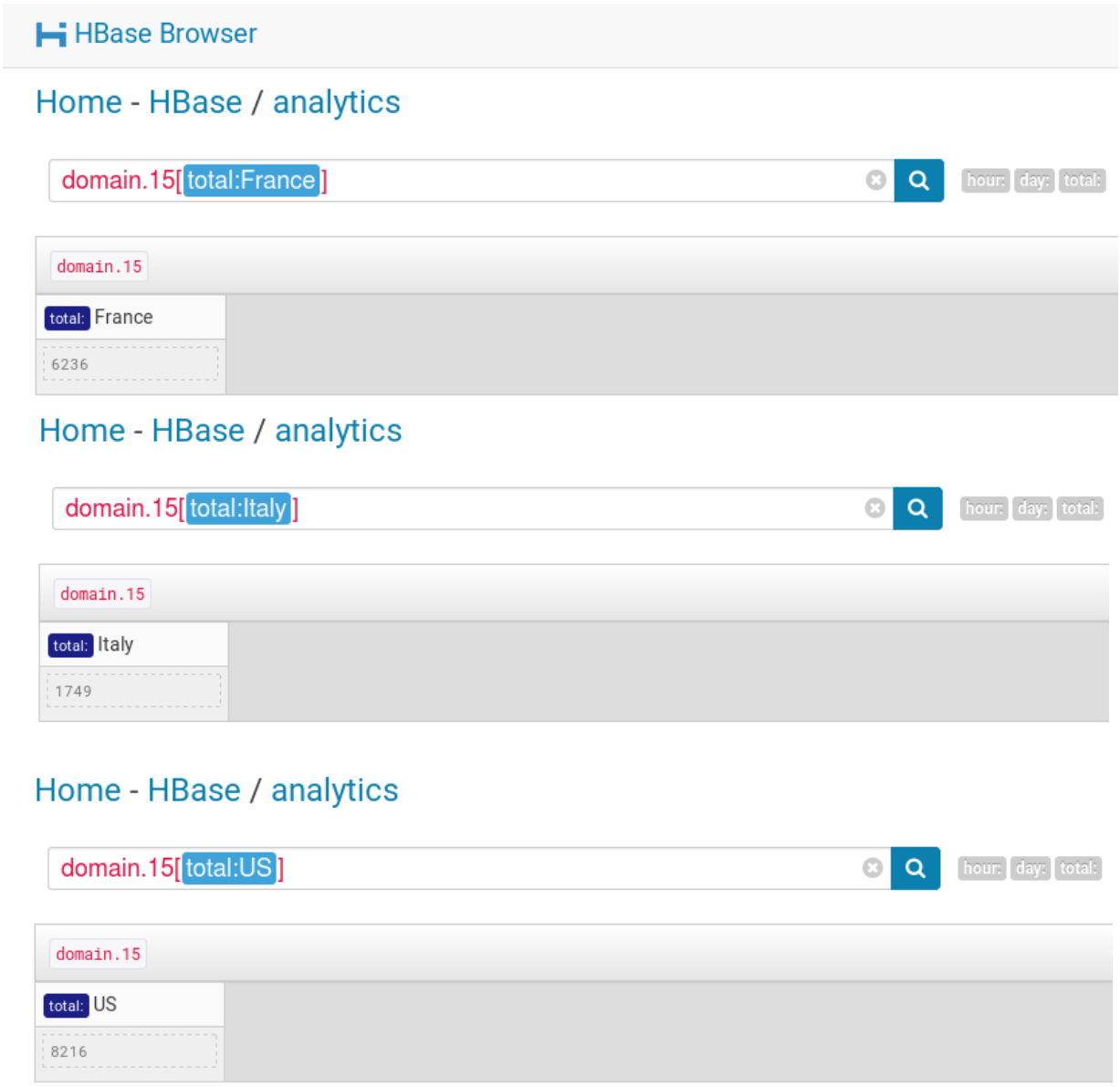
برای حل این تمرین در HBase Browser می‌توان در قسمت جستجو، سه query زیر را جستجو کرد:

domain.15 [total:France]

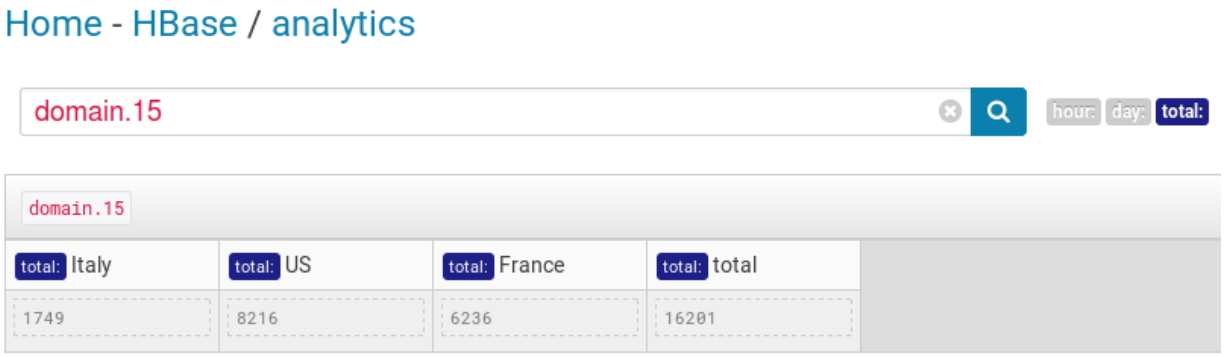
domain.15 [total:Italy]

domain.15 [total:US]

خروجی هر یک از query های بالا در ادامه آمده است :



البته می‌توان به جای نوشتن query های جدا، با انتخاب آیکن total CF بصورت زیر عمل کرد :



با توجه به نتیجه نمایش داده شده، گزینه اول (US) صحیح است.

تمرین 2

2. How many different domains are there in the analytics_demo data set (**Using the Hue web interface**)? ***Hint***: domains are rows, not individual cells in the rows. Only count the default shown, don't select all.

a) 10
b) 100
c) 15
d) 55

اگر در قسمت جستجو، query خاصی نوشته نشود و بصورت پیش فرض، 10 سطر نشان داده می‌شود، بنابراین گزینه اول صحیح است.

تمرین 3

3. Utilizing the analytics_demo data set via HBase data browsing - **which domain has the most (total)** site visits for the US? (hint: sort columns by descending order)

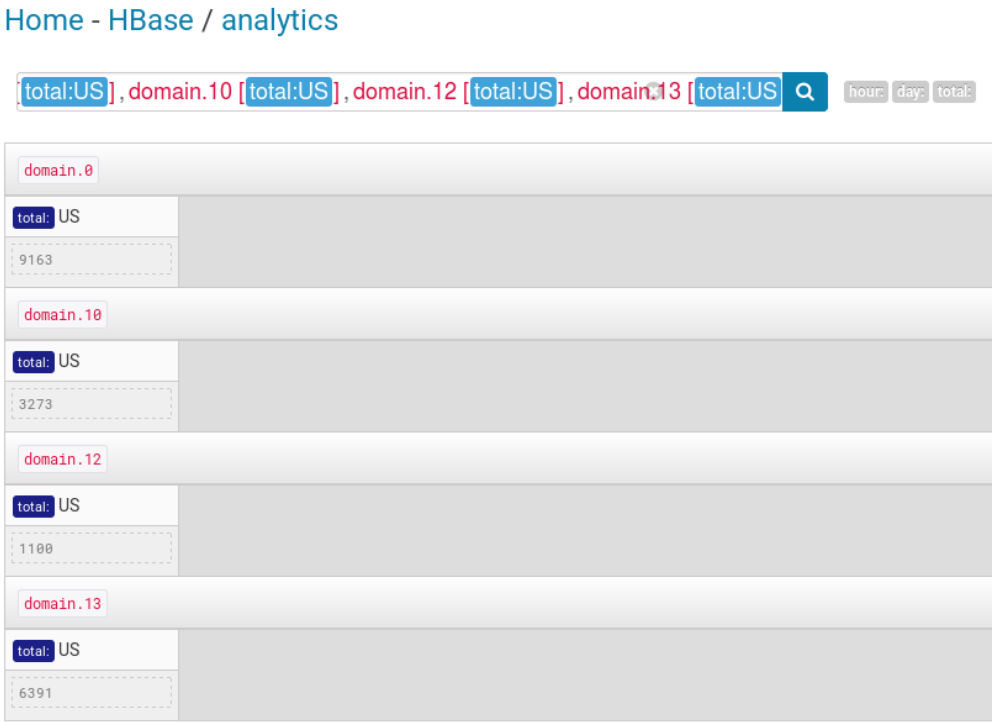
a) domain.10
b) domain.0
c) domain.12
d) domain.13

برای حل این تمرین در HBase Browser می‌توان در قسمت جستجو، query زیر را جستجو کرد:

domain.0 [total:US],domain.10 [total:US],domain.12 [total:US],domain.13 [total:US]

در این query، به ازای سطر domain.0، از total CF، ستون US (که با qualify کردن بصورت total:US نشان داده می‌شود) انتخاب شده است و به همین ترتیب برای سه سطر دیگر؛ query ها نوشته شده و با ‘,’ از هم جدا شده اند.

نتیجه query نشان می‌دهد که domain.0 بیشترین site visit را برای US داشته است، پس گزینه دوم صحیح است.



تمرین 4

4. Which 2 domains are the least visited (smallest total number of site visits) for all three countries together?

a) domain.16 and domain.15
b) domain.17 and domain.12
c) domain.16 and domain.12
d) domain.14 and domain.0

بطور مشابه، query های زیر اجرا می‌شود :

domain.16[total],domain.15[total]

domain.17[total],domain.12[total]

domain.16[total],domain.12[total]

domain.14[total],domain.0[total]

سپس بر اساس نتیجه هر query، ستونهای total آن دو سطر را جمع می‌زنیم که برای اولین query داریم 16201 + 10715 و برای query دوم داریم 10834 + 20619 و برای سوم داریم 10834 + 10715 و در نهایت برای query چهارم داریم 17375 + 21097. بنابراین، گزینه سوم، کمترین site visit را داشته و پاسخ سوال می‌باشد.

مثال نحوه کار با HBase از طریق Hive

در HBase Browser یا hbase shell تنها می‌توان عملیات محدودی بعنوان دستورات DDL یا DML با جداول انجام داد و نمی‌توان query با فرمت معمول SQL روی آنها اجرا کرد. به همین خاطر برای اینکار از Hive یا Impala استفاده می‌شود.

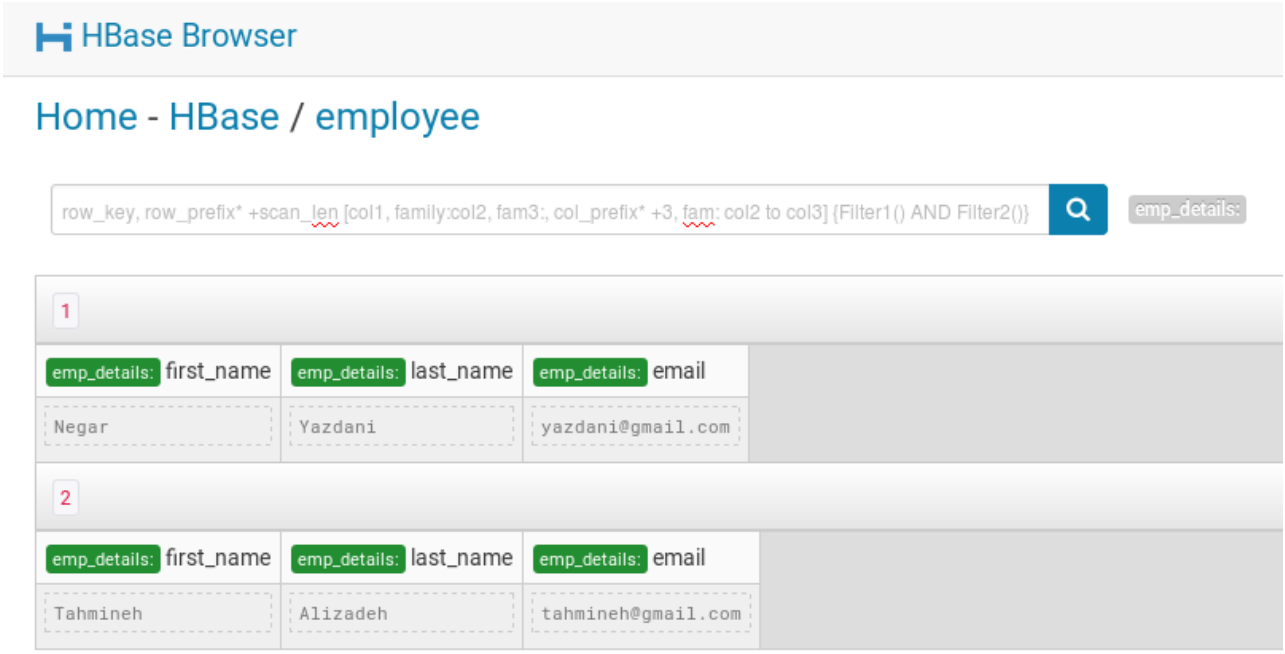
روال کار بدین صورت است که یک external table در Hive ایجاد شده و به جدول متناظرش در Hbase متصل می‌شود. حال می‌توان با این external table از طریق Hive یا Impala کار کرد و تغییرات در جدول اصلی Hbase اعمال می‌شود. با insert کردن دیتا در external table درواقع دیتا در جدول Hbase نوشته می‌شود و همینطور در مورد حذف یا تغییر دیتا. بنابراین اگر این external table را drop کنیم، جدول اصلی در Hbase کماکان وجود دارد.

مثال :

به hbase shell متصل شده و جدول employee را با یک CF به نام emp_details در hbase ایجاد می‌کنیم. سپس یا دستور put برای سطر با شماره 1، مقادیر سه ستون در emp_details را می نویسیم. و سپس با دستور scan اطلاعات نوشته شده را می‌بینیم:

```
[cloudera@quickstart Desktop]$ hbase shell
20/06/04 20:00:20 INFO Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.0-cdh5.13.0, rUnknown, Wed Oct  4 11:16:18 PDT 2017

hbase(main):001:0> Create 'employee','emp_details'
NoMethodError: undefined method `Create' for #<Object:0x947d127>

hbase(main):002:0> create 'employee','emp_details'
0 row(s) in 2.7760 seconds

=> Hbase::Table - employee
hbase(main):003:0> put 'employee',1,'emp_details:first_name','Negar'
0 row(s) in 0.2800 seconds

hbase(main):004:0> put 'employee',1,'emp_details:last_name','Yazdani'
0 row(s) in 0.0200 seconds

hbase(main):005:0> put 'employee',1,'emp_details:email','yazdani@gmail.com'
0 row(s) in 0.0110 seconds

hbase(main):006:0> scan 'employee'
ROW                                                   COLUMN+CELL
 1                                                    column=emp_details:email, timestamp=1591326175780, value=yazdani@gmail.com
 1                                                    column=emp_details:first_name, timestamp=1591326127901, value=Negar
 1                                                    column=emp_details:last_name, timestamp=1591326148571, value=Yazdani
1 row(s) in 0.1320 seconds

hbase(main):007:0>
```

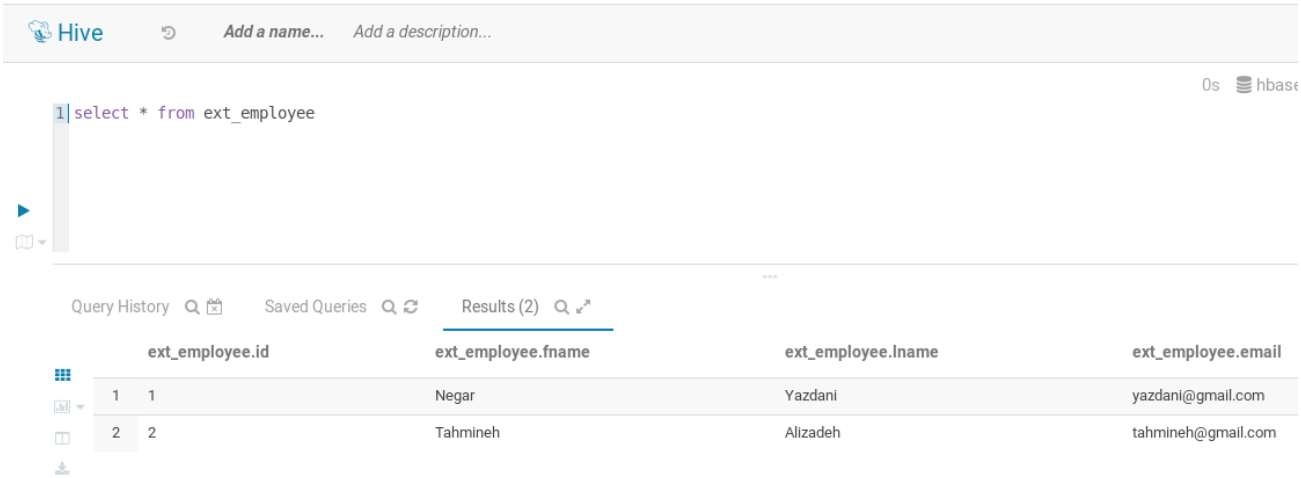
سطر دوم را هم به همین ترتیب اضافه می‌کنیم. این اطلاعات را در Hbase Browser هم میتوان دید :



حال به hive متصل شده و external table معادل جدول بالا را می‌سازیم :

```
[cloudera@quickstart Desktop]$ hive

Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-common-1.1.0-cdh5.13.0.jar!/hive-log4j.properties
WARNING: Hive CLI is deprecated and migration to Beeline is recommended.
hive> use hbasedb;
OK
Time taken: 2.235 seconds
hive> CREATE EXTERNAL TABLE ext_employee (
    > id int,
    > fname string,
    > lname string,
    > email string)
    > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler'
    > WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,emp_details:first_name, emp_details:last_name, emp_details:email ")
    > TBLPROPERTIES("hbase.table.name" = "employee")
    > ;
OK
Time taken: 2.445 seconds
hive>
```

پس از اجرای دستور بالا می‌توان از این جدول در Hive Query Editor از Hue Interface کوئری گرفت.



حال مثلاً می توان در این جدول از طریق hive یک رکورد جدید ایجاد کرد. قرار گرفتن دیتا در hbase در واقع از طریق اجرای یک MapReduce job انجام می‌شود :

```
hive> insert into table ext_employee values(3,'Maryam','Ahsani','maryam@gmail.com');
Query ID = cloudera_20200604205959_846285d9-67cf-46fd-8a17-bf8b64f853de
Total jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_1591320049324_0001, Tracking URL = http://quickstart.cloudera:8088/proxy/application_1591320049324_0001/
Kill Command = /usr/lib/hadoop/bin/hadoop job  -kill job_1591320049324_0001
Hadoop job information for Stage-0: number of mappers: 1; number of reducers: 0
2020-06-04 20:59:24,227 Stage-0 map = 0%,  reduce = 0%
2020-06-04 20:59:28,408 Stage-0 map = 100%,  reduce = 0%, Cumulative CPU 1.83 sec
MapReduce Total cumulative CPU time: 1 seconds 830 msec
Ended Job = job_1591320049324_0001
MapReduce Jobs Launched:
Stage-Stage-0: Map: 1   Cumulative CPU: 1.83 sec   HDFS Read: 12257 HDFS Write: 0 SUCCESS
Total MapReduce CPU Time Spent: 1 seconds 830 msec
OK
Time taken: 18.368 seconds
hive>
```

و نتیجه را از طریق hbase shell مشاهده کرد :

```
hbase(main):007:0> scan 'employee'
ROW                                                   COLUMN+CELL
 1                                                    column=emp_details:email, timestamp=1591326175780, value=yazdani@gmail.com
 1                                                    column=emp_details:first_name, timestamp=1591326127901, value=Negar
 1                                                    column=emp_details:last_name, timestamp=1591326148571, value=Yazdani
 2                                                    column=emp_details:email, timestamp=1591328620466, value=tahmineh@gmail.com
 2                                                    column=emp_details:first_name, timestamp=1591328583737, value=Tahmineh
 2                                                    column=emp_details:last_name, timestamp=1591328599538, value=Alizadeh
 3                                                    column=emp_details:email, timestamp=1591329567984, value=maryam@gmail.com
 3                                                    column=emp_details:first_name, timestamp=1591329567984, value=Maryam
 3                                                    column=emp_details:last_name, timestamp=1591329567984, value=Ahsani
3 row(s) in 0.0240 seconds

hbase(main):008:0>
```

نمونه query در HBase Browser

domain.15 [total:Italy][total:France][total:US]

domain.0 [total:US],domain.10 [total:US],domain.12 [total:US],domain.13 [total:US]

domain.1+10[total:US]

domain.100, domain.200+5

domain.100, domain.200+5[hour:]

domain.100, domain.200+5[hour:16-total]

domain.100*+3

domain.100+5{ColumnPrefixFilter(‘083’)}

domain.100+5{ColumnPrefixFilter(‘10’) AND }