### Big Data And Hadoop

## Session 17 – Assignment 2

### **Problem Statement 1:**

Create an HBase table named 'clicks' with a column family 'hits' such that it should be able to store last 5 values of qualifiers inside 'hits' column family.

#### Solution:

Step 1: We first start HBase using the command **start-hbase.sh** as follows:

Step 2: Next we start the hbase shell using the command hbase shell as follows:

```
| acadgild@localhost:~ | acadgild@localhost:~ | acadgild@localhost:~ | File Edit View Search Terminal Help |
| [acadgild@localhost ~]$ hbase shell |
| 2017-09-24 17:49:57,986 INFO [main] Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.a vailable |
| HBase Shell; enter 'help<RETURN>' for list of supported commands. |
| Type "exit<RETURN>" to leave the HBase Shell |
| Version 0.98.14-hadoop2, r4e4aabb93b52flb0fef6b66edd06ec8923014dec, Tue Aug 25 22:35:44 PDT 2015 |
| hbase(main):001:0> |
```

Step 3: Now we create the table named clicks using the command create 'clicks', 'hits' as follows:

```
acadgild@localhost:~
 File Edit View Search Terminal Help
[acadgild@localhost ~]$ hbase shell
2017-09-24 17:49:57,986 INFO [main] Configuration.deprecation: hadoop.native.lib is deprecated. Instead, use io.native.lib.a
vailable
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 0.98.14-hadoop2, r4e4aabb93b52f1b0fef6b66edd06ec8923014dec, Tue Aug 25 22:35:44 PDT 2015
hbase(main):001:0> create 'clicks', 'hits'
SLF4J: Class path contains multiple SLF4J bindings
SLF4J: Found binding in [jar:file:/usr/local/hbase/lib/slf4j-log4j12-1.6.4.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.6.0/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Sta
ticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
2017-09-24 17:51:30,678 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
0 row(s) in 2.7390 seconds
=> Hbase::Table - clicks
hbase(main):002:0>
```

In this command, we also mention the name of the column family to be **hits.** Hence, a table named **clicks** with the column family **hits** got created.

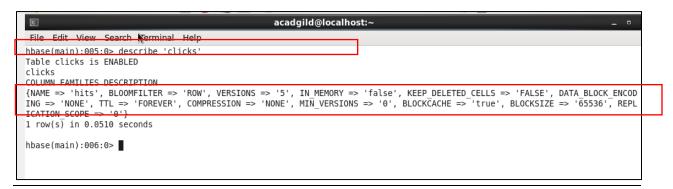
Step 4: To verify if the table has been created, we use the command **list** to get the list of all the tables present as follows:



Step 5: Now, we use alter command to add version specification to the table schema. In the command executed, the number mentioned against **VERSIONS** specifies how many updates for the particular column family are to be maintained. Here, we have set it to 5. So when the values in this column family is updated, the last 5 values for the same are also maintained. This helps in maintaining the history of change.



Step 6: **describe** command gives the schema of the table. Its output will help us verify that version has been set.



## **Problem Statement 2:**

Add few records in the table and update some of them. Use IP Address as row-key. Scan the table to view if all the previous versions are getting displayed.

# **Solution:**

In the above problem statement, we made use of the hbase shell and created the table clicks.

Now we add/ update data into it as follows:





Here, we have used **put** command to insert/update data in **clicks** table. The column family **hits** has a column named **userId**.

Same command is used to insert as well as update data into the table. If an entry for a row key and column in the column family is already inserted, that record gets updated.

If we use the command **scan 'clicks'**, it will give the latest records. We have used **put** command around 9 times. However, only two unique row keys were used, hence rest of the times, the existing records got updated. **Scan** command shows the latest values only. Hence, only 2 rows in the output.



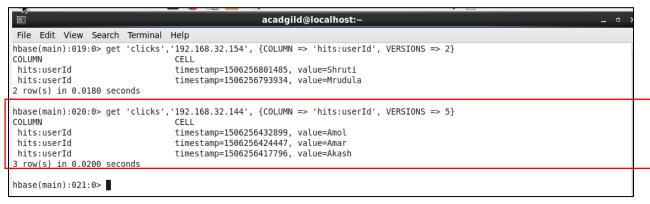
In order to see the effect of versions specified in the schema, we see it using the following command:

In the command, because we specified versions equal to 5, it showed us last 5 updates made for the rwo key with ip address 192.168.32.154.

```
acadgild@localhost:~
File Edit View Search Terminal Help
hbase(main):018:0> get 'clicks','192.168.32.154', {COLUMN => 'hits:userId', VERSIONS => 5}
COLUMN
                                CELL
hits:userId
                                timestamp=1506256801485, value=Shruti
hits:userId
                                timestamp=1506256793934, value=Mrudula
hits:userId
                                timestamp=1506256786803, value=Sandhya
hits:userId
                                timestamp=1506256778104, value=Prashul
hits:userId
                                timestamp=1506256768855, value=Abhilasha
5 row(s) in 0.0370 seconds
hbase(main):019:0>
```

If we specify the versions equal to 2, it will give us data of last two updates as follows:





In the above screenshot, although the versions specified in the command is equal to 5, it showed only 3 records for ip address 192.168.32.144 because it had history of only 3 updates.