# ASSIGNMENT 9.2

## Problem Statement:

**Problem 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.

**Problem 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 for problem 1:

Let's create a table in HBase named 'clicks' with column family 'hits'. We can make use of 'VERSIONS' option for it to support retrieval of last 5 versions of data related 'hits' column family.

Here is the command to do the same on HBase shell:

hbase> create 'clicks', {NAME=>'hits', VERSIONS=>5}

```
hbase(main):003:0> create 'clicks', {NAME=>'hits', VERSIONS=>5}
0 row(s) in 1.0630 seconds

=> Hbase::Table - clicks
hbase(main):004:0> list
TABLE
clicks
1 row(s) in 0.0280 seconds

=> ["clicks"]
hbase(main):005:0> scan 'clicks'
ROW                              COLUMN+CELL
0 row(s) in 0.1380 seconds

hbase(main):006:0>
```

## Solution for problem 2:

**Step 1:** Let's insert few rows into this table using 'put' command as shown below:

hbase> put 'clicks','192.168.124.1','hits:hits_count','20'

hbase> put 'clicks','192.168.124.1','hits:website','www.acadgild.com'

hbase> put 'clicks','192.168.124.2','hits:hits_count','30'

hbase> put 'clicks','192.168.124.2','hits:website','www.google.com'

hbase> put 'clicks','192.168.124.3','hits:hits_count','25'

hbase> put 'clicks','192.168.124.3','hits:website','www.facebook.com'

hbase> put 'clicks','192.168.124.4','hits:hits_count','40'

hbase> put 'clicks','192.168.124.4','hits:website','www.acadgild.com'

In the above commands, I have added two columns such as ‘hits_count’ and ‘website’ for column family ‘hits’. An IP address is used as row key as given in the problem statement. Now we can see 4 rows getting inserted into ‘clicks’ table by executing scan command:

hbase> scan ‘clicks’

```
hbase(main):026:0> put 'clicks','192.168.124.1','hits:hits_count','20'
0 row(s) in 0.0340 seconds

hbase(main):027:0> put 'clicks','192.168.124.1','hits:website','www.acadgild.com'
0 row(s) in 0.0160 seconds

hbase(main):028:0>
hbase(main):029:0* put 'clicks','192.168.124.2','hits:hits_count','30'
0 row(s) in 0.0140 seconds

hbase(main):030:0> put 'clicks','192.168.124.2','hits:website','www.google.com'
0 row(s) in 0.0160 seconds

hbase(main):031:0>
hbase(main):032:0* put 'clicks','192.168.124.3','hits:hits_count','25'
0 row(s) in 0.0100 seconds

hbase(main):033:0> put 'clicks','192.168.124.3','hits:website','www.facebook.com'
0 row(s) in 0.0210 seconds

hbase(main):034:0>
hbase(main):035:0* put 'clicks','192.168.124.4','hits:hits_count','40'
0 row(s) in 0.0130 seconds

hbase(main):036:0> put 'clicks','192.168.124.4','hits:website','www.acadgild.com'
0 row(s) in 0.0180 seconds

hbase(main):037:0> scan 'clicks'
ROW                           COLUMN+CELL
 192.168.124.1                column=hits:hits_count, timestamp=1514265531592, value=20
 192.168.124.1                column=hits:website, timestamp=1514265531659, value=www.acadgild.com
 192.168.124.2                column=hits:hits_count, timestamp=1514265531750, value=30
 192.168.124.2                column=hits:website, timestamp=1514265531808, value=www.google.com
 192.168.124.3                column=hits:hits_count, timestamp=1514265531882, value=25
 192.168.124.3                column=hits:website, timestamp=1514265531958, value=www.facebook.com
 192.168.124.4                column=hits:hits_count, timestamp=1514265532091, value=40
 192.168.124.4                column=hits:website, timestamp=1514265534298, value=www.acadgild.com
4 row(s) in 0.1210 seconds
```

**Step 2:** Update few records and see the changes by scanning the whole table.

Let’s try to update the record with row key: ‘192.168.124.3’.

hbase> put 'clicks','192.168.124.3','hits:hits_count','20'

hbase> put 'clicks','192.168.124.3','hits:website','www.linkedin.com'

hbase> put 'clicks','192.168.124.3','hits:hits_count','25'

hbase> put 'clicks','192.168.124.3','hits:website','www.acadgild.com'

hbase> put 'clicks','192.168.124.3','hits:hits_count','40'

hbase> put 'clicks','192.168.124.3','hits:website','www.google.com'

hbase> put 'clicks','192.168.124.3','hits:hits_count','30'

hbase> put 'clicks','192.168.124.3','hits:website','www.cloudera.com'

We can notice from these commands that the record with row key ‘192.168.124.3’ has got updated four times and now it has five versions in total including the one which was been inserted in the beginning. Let’s scan the table to see which value this record holds:

hbase>  scan ‘clicks’

```
hbase(main):038:0> put 'clicks','192.168.124.3','hits:hits_count','20'
0 row(s) in 0.0290 seconds

hbase(main):039:0> put 'clicks','192.168.124.3','hits:website','www.linkedin.com'
0 row(s) in 0.0150 seconds

hbase(main):040:0> put 'clicks','192.168.124.3','hits:hits_count','25'
0 row(s) in 0.0070 seconds

hbase(main):041:0> put 'clicks','192.168.124.3','hits:website','www.acadgild.com'
0 row(s) in 0.0080 seconds

hbase(main):042:0> put 'clicks','192.168.124.3','hits:hits_count','40'
0 row(s) in 0.0100 seconds

hbase(main):043:0> put 'clicks','192.168.124.3','hits:website','www.google.com'
0 row(s) in 0.0070 seconds

hbase(main):044:0> put 'clicks','192.168.124.3','hits:hits_count','30'
0 row(s) in 0.0040 seconds

hbase(main):045:0> put 'clicks','192.168.124.3','hits:website','www.cloudera.com'
0 row(s) in 0.0160 seconds

hbase(main):046:0> scan 'clicks'
ROW                           COLUMN+CELL
 192.168.124.1                column=hits:hits_count, timestamp=1514265531592, value=20
 192.168.124.1                column=hits:website, timestamp=1514265531659, value=www.acadgild.com
 192.168.124.2                column=hits:hits_count, timestamp=1514265531750, value=30
 192.168.124.2                column=hits:website, timestamp=1514265531808, value=www.google.com
 192.168.124.3                column=hits:hits_count, timestamp=1514266344956, value=30
 192.168.124.3                column=hits:website, timestamp=1514266350313, value=www.cloudera.com
 192.168.124.4                column=hits:hits_count, timestamp=1514265532091, value=40
 192.168.124.4                column=hits:website, timestamp=1514265534298, value=www.acadgild.com
4 row(s) in 0.0980 seconds

hbase(main):047:0>
```

As we can see from the screenshot above, the record holds only the latest updated value. Now if we want to see all five versions data for this column family, we can get it by following command:
hbase> scan 'clicks',{COLUMN=>'hits',VERSIONS=>5}

```
hbase(main):047:0> scan 'clicks',{COLUMN=>'hits',VERSIONS=>5}
ROW                           COLUMN+CELL
 192.168.124.1                column=hits:hits_count, timestamp=1514265531592, value=20
 192.168.124.1                column=hits:website, timestamp=1514265531659, value=www.acadgild.com
 192.168.124.2                column=hits:hits_count, timestamp=1514265531750, value=30
 192.168.124.2                column=hits:website, timestamp=1514265531808, value=www.google.com
 192.168.124.3                column=hits:hits_count, timestamp=1514266344956, value=30
 192.168.124.3                column=hits:hits_count, timestamp=1514266344843, value=40
 192.168.124.3                column=hits:hits_count, timestamp=1514266344735, value=25
 192.168.124.3                column=hits:hits_count, timestamp=1514266344641, value=20
 192.168.124.3                column=hits:hits_count, timestamp=1514265531882, value=25
 192.168.124.3                column=hits:website, timestamp=1514266350313, value=www.cloudera.com
 192.168.124.3                column=hits:website, timestamp=1514266344910, value=www.google.com
 192.168.124.3                column=hits:website, timestamp=1514266344779, value=www.acadgild.com
 192.168.124.3                column=hits:website, timestamp=1514266344702, value=www.linkedin.com
 192.168.124.3                column=hits:website, timestamp=1514265531958, value=www.facebook.com
 192.168.124.4                column=hits:hits_count, timestamp=1514265532091, value=40
 192.168.124.4                column=hits:website, timestamp=1514265534298, value=www.acadgild.com
4 row(s) in 0.1450 seconds
```

Now let’s try to update the same record again and then fetch all versions of data from the same column family:

hbase> put 'clicks','192.168.124.3','hits:hits_count','40'

hbase> put 'clicks','192.168.124.3','hits:website','www.acadgild.com'

hbase> scan 'clicks'

```
hbase(main):048:0> put 'clicks','192.168.124.3','hits:hits_count','40'
0 row(s) in 0.0200 seconds

hbase(main):049:0> put 'clicks','192.168.124.3','hits:website','www.acadgild.com'
0 row(s) in 0.0320 seconds

hbase(main):050:0> scan 'clicks'
ROW                           COLUMN+CELL
 192.168.124.1                column=hits:hits_count, timestamp=1514265531592, value=20
 192.168.124.1                column=hits:website, timestamp=1514265531659, value=www.acadgild.com
 192.168.124.2                column=hits:hits_count, timestamp=1514265531750, value=30
 192.168.124.2                column=hits:website, timestamp=1514265531808, value=www.google.com
 192.168.124.3                column=hits:hits_count, timestamp=1514266647764, value=40
 192.168.124.3                column=hits:website, timestamp=1514266649192, value=www.acadgild.com
 192.168.124.4                column=hits:hits_count, timestamp=1514265532091, value=40
 192.168.124.4                column=hits:website, timestamp=1514265534298, value=www.acadgild.com
4 row(s) in 0.1090 seconds
```

The record with row key ‘192.168.124.3’ has been updated with values for columns from latest put commands. Let’s try to get all six versions for this column family:

hbase> scan 'clicks',{COLUMN=>'hits',VERSIONS=>6}

```
hbase(main):051:0> scan 'clicks',{COLUMN=>'hits',VERSIONS=>6}
ROW                           COLUMN+CELL
 192.168.124.1                column=hits:hits_count, timestamp=1514265531592, value=20
 192.168.124.1                column=hits:website, timestamp=1514265531659, value=www.acadgild.com
 192.168.124.2                column=hits:hits_count, timestamp=1514265531750, value=30
 192.168.124.2                column=hits:website, timestamp=1514265531808, value=www.google.com
 192.168.124.3                column=hits:hits_count, timestamp=1514266647764, value=40
 192.168.124.3                column=hits:hits_count, timestamp=1514266344956, value=30
 192.168.124.3                column=hits:hits_count, timestamp=1514266344843, value=40
 192.168.124.3                column=hits:hits_count, timestamp=1514266344735, value=25
 192.168.124.3                column=hits:hits_count, timestamp=1514266344641, value=20
 192.168.124.3                column=hits:website, timestamp=1514266649192, value=www.acadgild.com
 192.168.124.3                column=hits:website, timestamp=1514266350313, value=www.cloudera.com
 192.168.124.3                column=hits:website, timestamp=1514266344910, value=www.google.com
 192.168.124.3                column=hits:website, timestamp=1514266344779, value=www.acadgild.com
 192.168.124.3                column=hits:website, timestamp=1514266344702, value=www.linkedin.com
 192.168.124.4                column=hits:hits_count, timestamp=1514265532091, value=40
 192.168.124.4                column=hits:website, timestamp=1514265534298, value=www.acadgild.com
4 row(s) in 0.0860 seconds
```

As we can see in the above screenshot, it displays only the last five data versions for columns of given column family and I have lost the initial version of it. Because I had mentioned the maximum number of versions this column family can hold as five while creating the table.