Big Data And Hadoop

Session 17 – Assignment 2

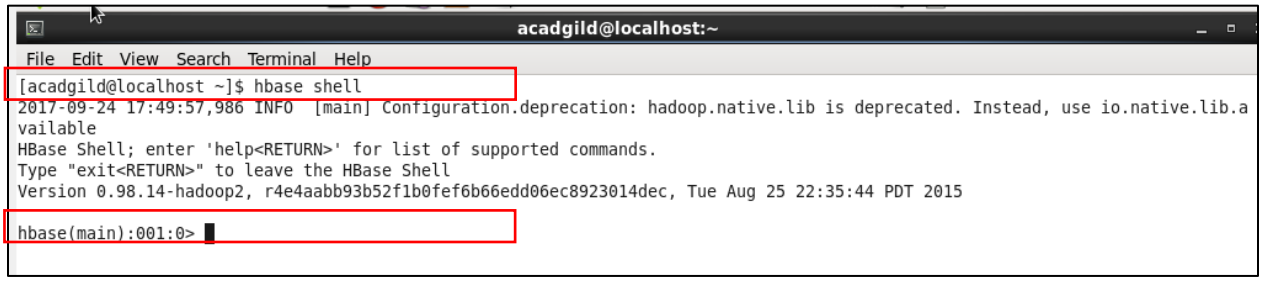
Problem Statement 1:

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.

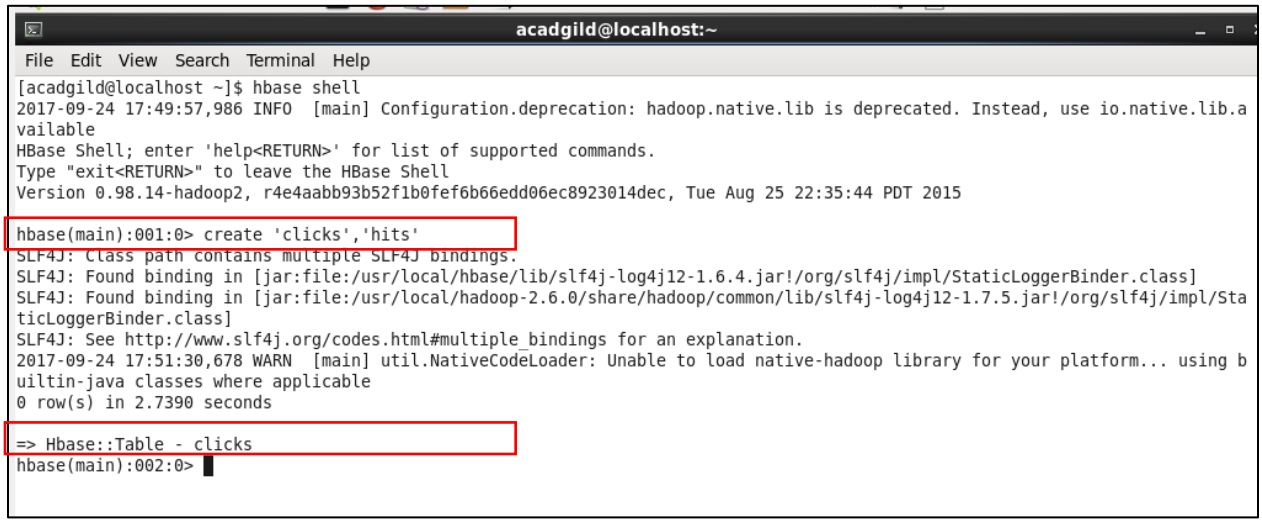
Solution:

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:

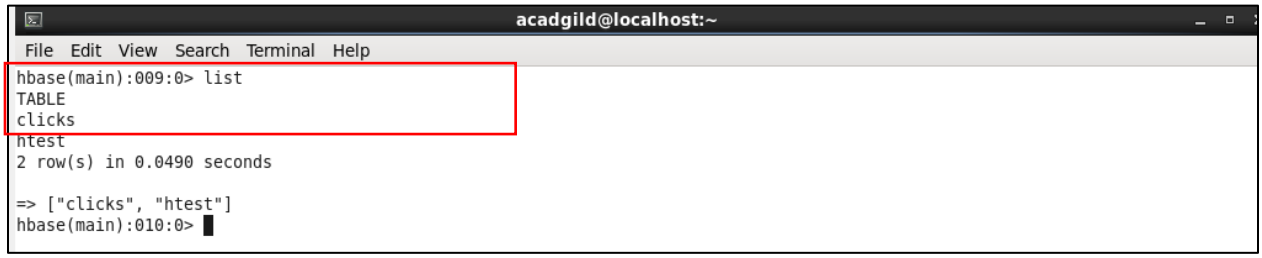


Step 3: Now we create the table named **clicks** using the command **create 'clicks', 'hits'** as follows:



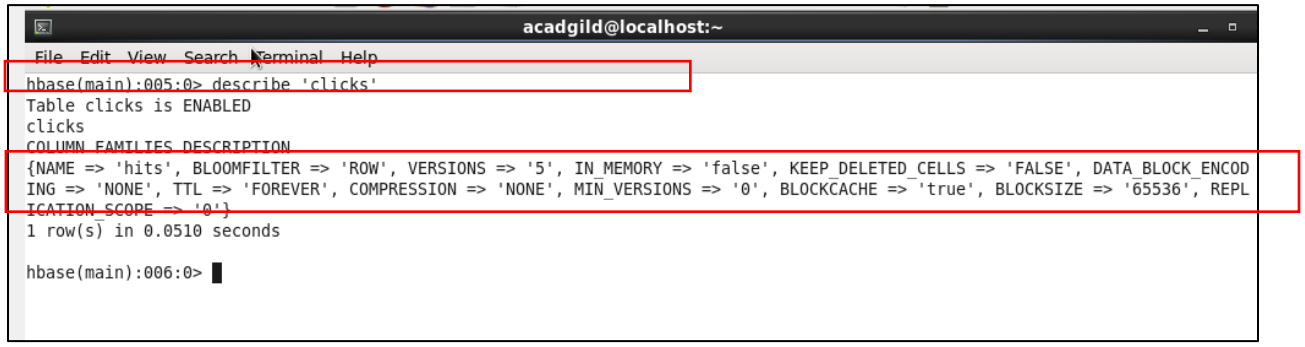
In this command, we also mention the name of the column family to be **hits.** Hence, a table named **clicks** with the column family **hits** got created.

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.



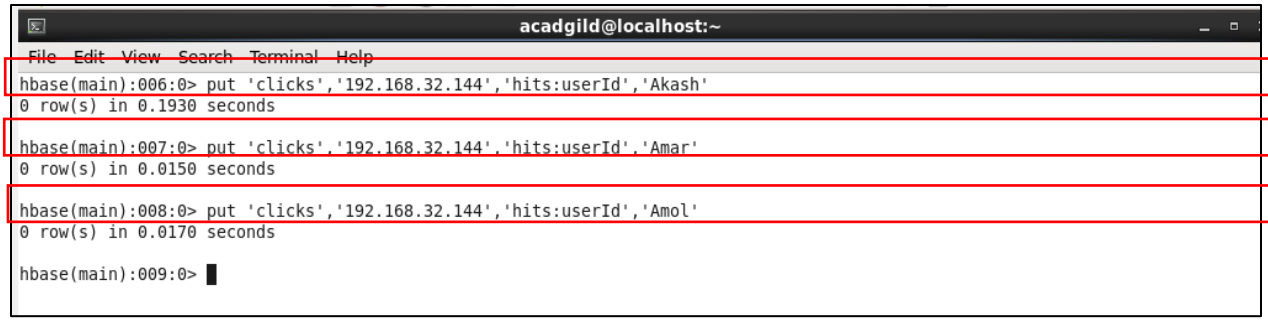
Problem Statement 2:

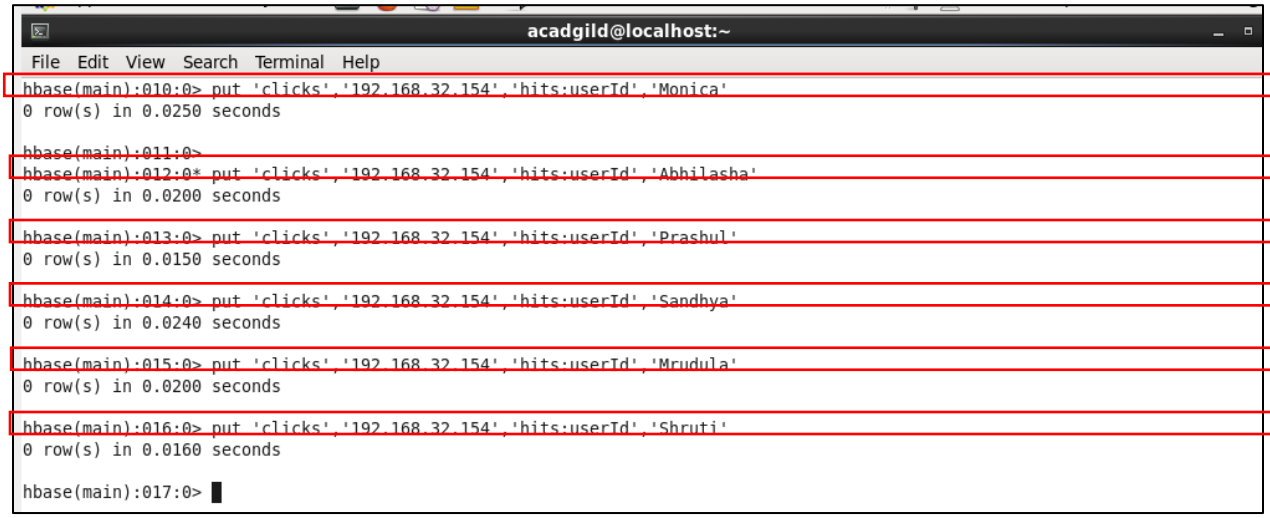
Add few records in the table and update some of them. Use IP Address as row-key. Scan the table to view if all the previous versions are getting displayed.

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:

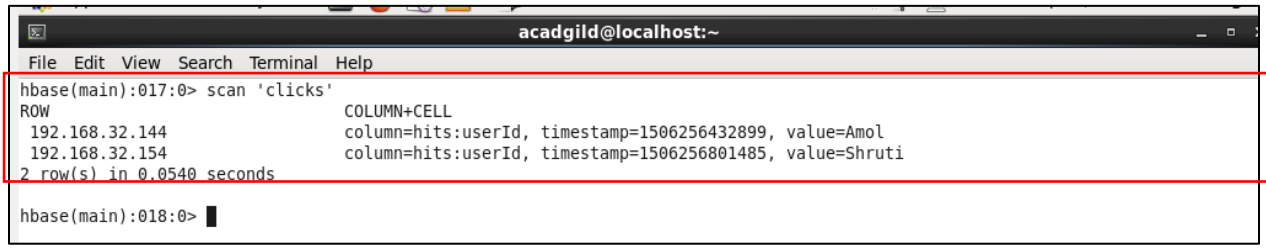




Here, we have used **put** command to insert/update data in **clicks** table. The column family **hits** has a column named **userId.**

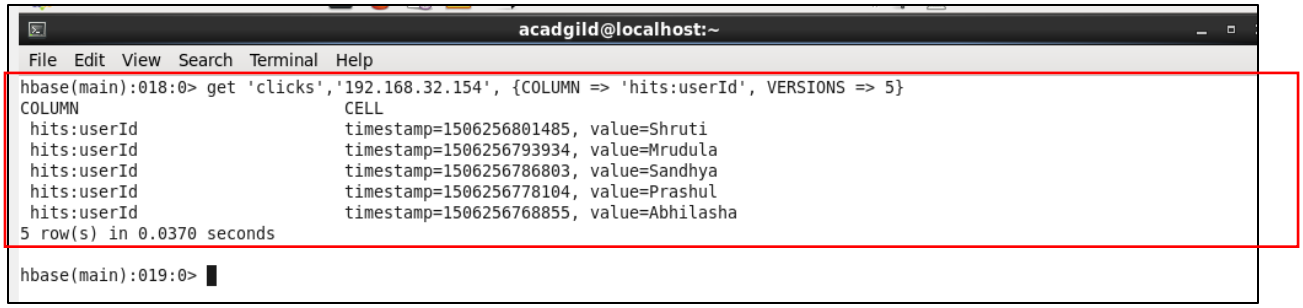
Same command is used to insert as well as update data into the table. If an entry for a row key and column in the column family is already inserted, that record gets updated.

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.

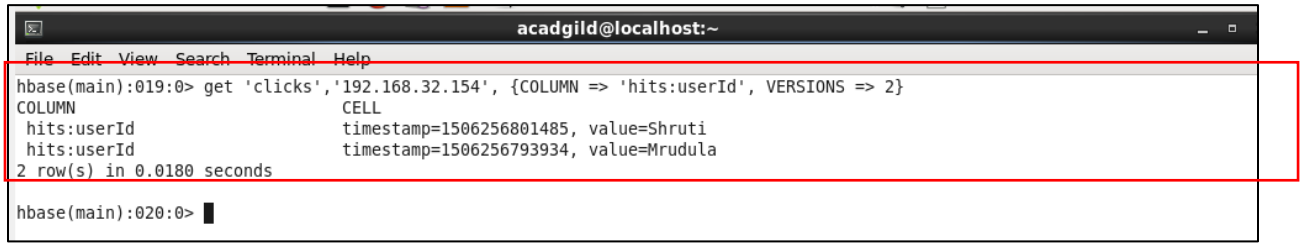


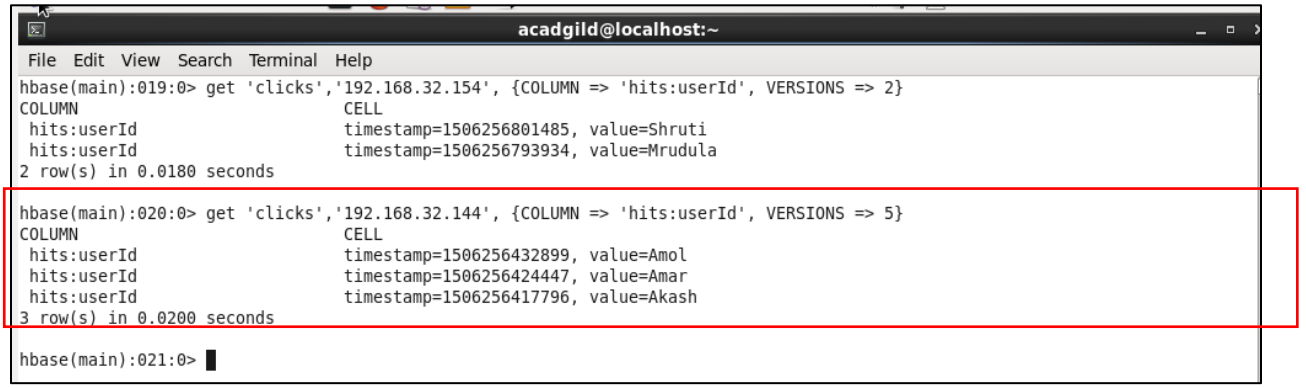
In order to see the effect of versions specified in the schema, we see it using the following command:

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.



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.