Big Data Hadoop 'Session 10: HBASE BASICS

Assignment 1'

Task 1

Answer in your own words with example.

1.What is NoSQL data base?

- NoSQL database is design that can accommodate a wide variety of data models, including key-value, document, columnar and graph formats.
- NoSQL, which stand for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built.
- NoSQL databases are especially useful for working with large sets of distributed data.
- Eg Hbase,Mongodb

2.How does data get stored in NoSQl database?

- In NoSQL, data can be stored in a schema-less or free-form fashion. Any data can be stored in any record.

Eg

- ***Document databases*** *(e.g. CouchDB, MongoDB)*
- ***Key-value stores*** *(e.g. Redis, Riak)*
- ***Wide column stores*** *(e.g. HBase, Cassandra)*
- ***Graph databases*** *(e.g. Neo4j)*

3.What is a column family in HBase?

- Columns in Apache HBase are grouped into *column families*. All column members of a column family have the same prefix.
- Column families must be declared up front at schema definition time. It is like category
- Eg.here CustomerName is column family

Row Key	Column Family: {Column Qualifier:Version:Value}
00001	CustomerName: {'FN': 1383859182496:'John',

4.How many maximum number of columns can be added to HBase table?

- No Limit in no. of column in hbase

5.Why columns are not defined at the time of table creation in HBase?

- Unlike columns in a relational database, which reserve empty space for columns with no values, HBase columns simply don't exist for rows where they have no values.
- This not only saves space, but means that different rows need not have the same columns; we can use whatever columns you need for our data on a per-row basis.

6.How does data get managed in HBase?

- Table: HBase organizes data into tables. Table names are Strings and composed of characters
- Row: Within a table, data is stored according to its row. Rows are identified uniquely by their row key. Row keys do not have a data type and are always treated as a byte[ ] (byte array).
- Column Family: Data within a row is grouped by column family. Column families also impact the physical arrangement of data stored in HBase. For this reason, they must be defined up front and are not easily modified.
- Column Qualifier: Data within a column family is addressed via its column qualifier.Column qualifiers need not be specified in advance.
- Timestamp: Values within a cell are versioned. Versions are identified by their version number, which by default is the timestamp of when the cell was written. The default number of cell versions is three.

7.What happens internally when new data gets inserted into HBase table?

- Timestamp value is generated against each put operation
- *eg*

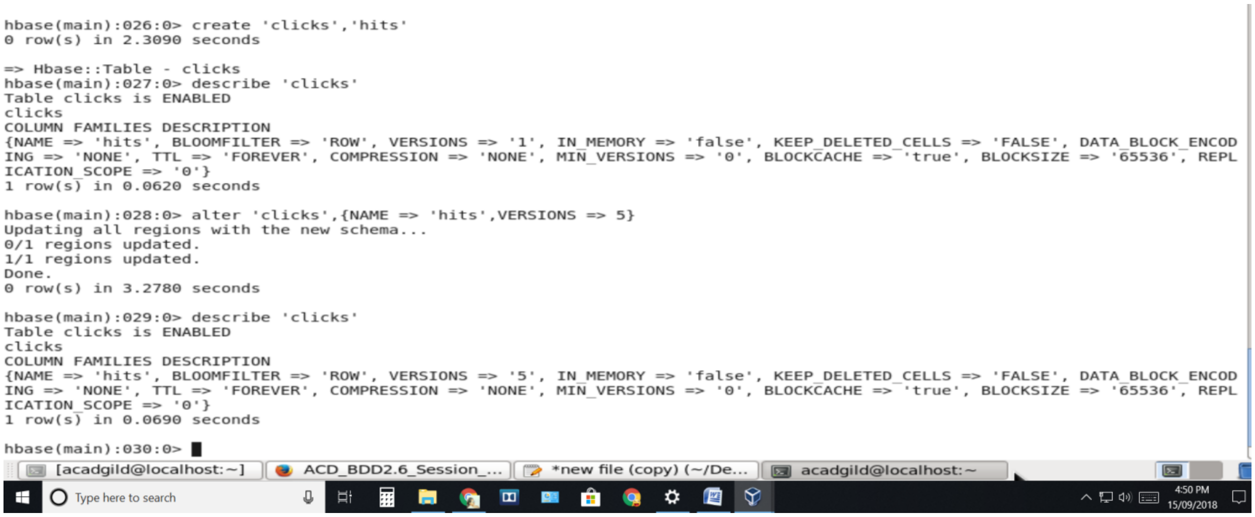
```
hbase> put 'test', 'row1', 'cf:a', 'value1'
scan 'test'
ROW                     COLUMN+CELL
 row1                   column=cf:a, timestamp=1403759475114, value=value1
```

Task 2

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.

- created table clicks with Column family hits and versions 5
- Described table to check version is 5 now.

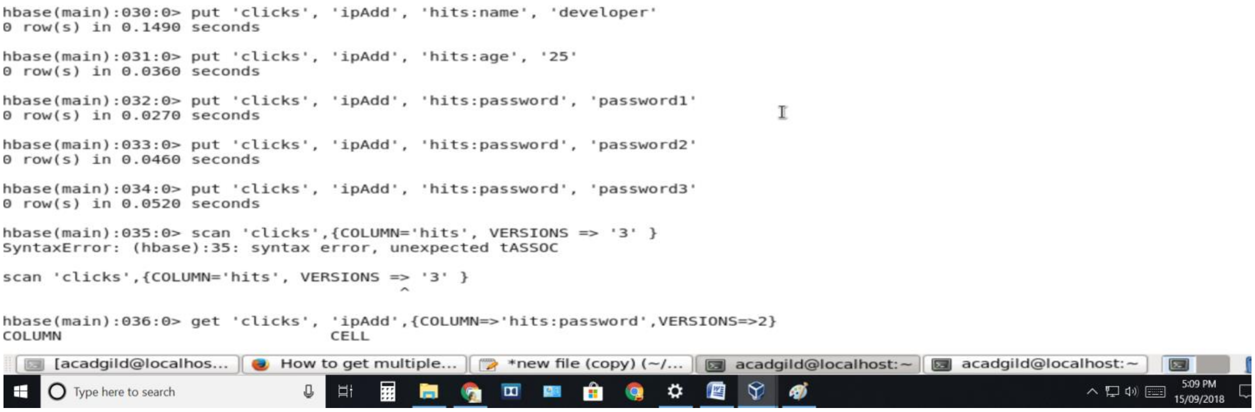


```
hbase(main):026:0> create 'clicks','hits'
0 row(s) in 2.3090 seconds

=> Hbase::Table - clicks
hbase(main):027:0> describe 'clicks'
Table clicks is ENABLED
clicks
COLUMN FAMILIES DESCRIPTION
{NAME => 'hits', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCOD
ING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPL
ICATION_SCOPE => '0'}
1 row(s) in 0.0620 seconds

hbase(main):028:0> alter 'clicks',{NAME => 'hits',VERSIONS => 5}
Updating all regions with the new schema...
0/1 regions updated.
1/1 regions updated.
Done.
0 row(s) in 3.2780 seconds

hbase(main):029:0> describe 'clicks'
Table clicks is ENABLED
clicks
COLUMN FAMILIES DESCRIPTION
{NAME => 'hits', BLOOMFILTER => 'ROW', VERSIONS => '5', IN_MEMORY => 'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCOD
ING => 'NONE', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPL
ICATION_SCOPE => '0'}
1 row(s) in 0.0690 seconds

hbase(main):030:0> █
```

2. Add few records in the table and update some of them. Use IP Address as row-key.

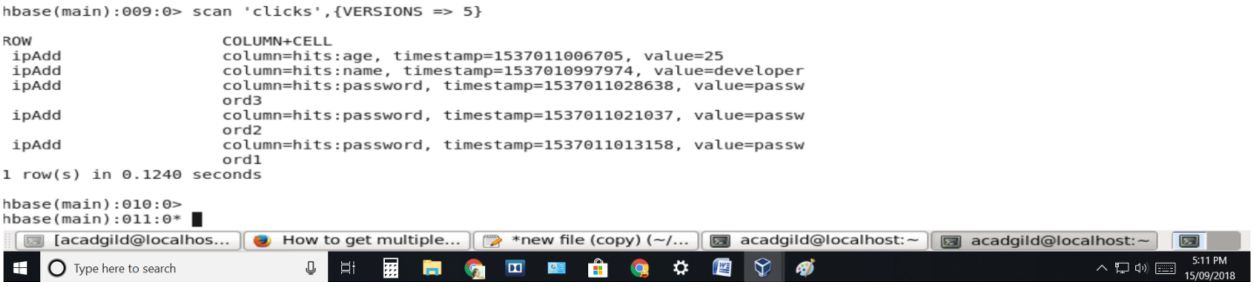
Scan the table to view if all the previous versions are getting displayed.

- Added new entry in table 'clicks'
- Update hits:password



```
hbase(main):030:0> put 'clicks', 'ipAdd', 'hits:name', 'developer'
0 row(s) in 0.1490 seconds

hbase(main):031:0> put 'clicks', 'ipAdd', 'hits:age', '25'
0 row(s) in 0.0360 seconds

hbase(main):032:0> put 'clicks', 'ipAdd', 'hits:password', 'password1'
0 row(s) in 0.0270 seconds

hbase(main):033:0> put 'clicks', 'ipAdd', 'hits:password', 'password2'
0 row(s) in 0.0460 seconds

hbase(main):034:0> put 'clicks', 'ipAdd', 'hits:password', 'password3'
0 row(s) in 0.0520 seconds

hbase(main):035:0> scan 'clicks',{COLUMN='hits', VERSIONS => '3' }
SyntaxError: (hbase):35: syntax error, unexpected tASSOC

scan 'clicks',{COLUMN='hits', VERSIONS => '3' }
                                        ^

hbase(main):036:0> get 'clicks', 'ipAdd',{COLUMN=>'hits:password',VERSIONS=>2}
COLUMN                            CELL
```

- *Scanned the table 'clicks' to view if all the previous versions are getting displayed.*



```
hbase(main):009:0> scan 'clicks',{VERSIONS => 5}

ROW                          COLUMN+CELL
 ipAdd                       column=hits:age, timestamp=1537011006705, value=25
 ipAdd                       column=hits:name, timestamp=1537010997974, value=developer
 ipAdd                       column=hits:password, timestamp=1537011028638, value=passw
                             ord3
 ipAdd                       column=hits:password, timestamp=1537011021037, value=passw
                             ord2
 ipAdd                       column=hits:password, timestamp=1537011013158, value=passw
                             ord1
1 row(s) in 0.1240 seconds

hbase(main):010:0>
hbase(main):011:0* █
```