HBASE

Task1

1.What is NoSQL data base?

NOT sql database, We can store huge amount of data and it is alternative to the rdbms.

2.How does data get stored in NoSQl database?

Column-based Store- Each storage block contains data from only one column, {Example- HBase, Cassandra}

Document-based Store- It stores documents made up of tagged elements. {Example- MongoDB}Json kind of data

3.What is a column family in HBase?

Column families are the base storage mechanism in HBase. An HBase table is comprised of one or more column families, each of which is stored in a separate set of regionfiles sharing a common key. The Rowkey is an identifier , c.f. could be category cq : subcategory and value could be anything To retrieve the saved value you should know identifier and category and sub category .

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

There is no hard limit to number of columns in HBase , we can have more than 1 million columns but usually three column families are recommended (not more than three).

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

Column qualifiers are mutable and they may vary between rows. They do not have data types and they are always treated as arrays of bytes.

6. How does data get managed in HBase?

It will store hfile in each region. Region is full again it split into number of regions. Like tree structureBased on the space the regions will be create on region server (on the data node)

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

Internally hbase is store in the form column family, first it will go to the WAl in region server ,from wal file to the region one of the memstore and it will store into the hfile.The process of wal file storage is up to persistence storage only. Where memstore is the cache based working. Hfile are the main data storage parts in each region of the region server. Rowkey, table based we can the fetch the data from each region.

Task2

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.

```
hbase(main):002:0> create 'clicks',{NAME=>'hits',VERSIONS=>5}
```

clicks is the table name

hits is the column family name

versions is to store recent data and it’s uses the numer.

columns=column qualifiers

```
hbase(main):007:0> 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_ENCODING => 'NONE
', TTL => 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true',
BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.0790 seconds
```

We can insert data into the table like below

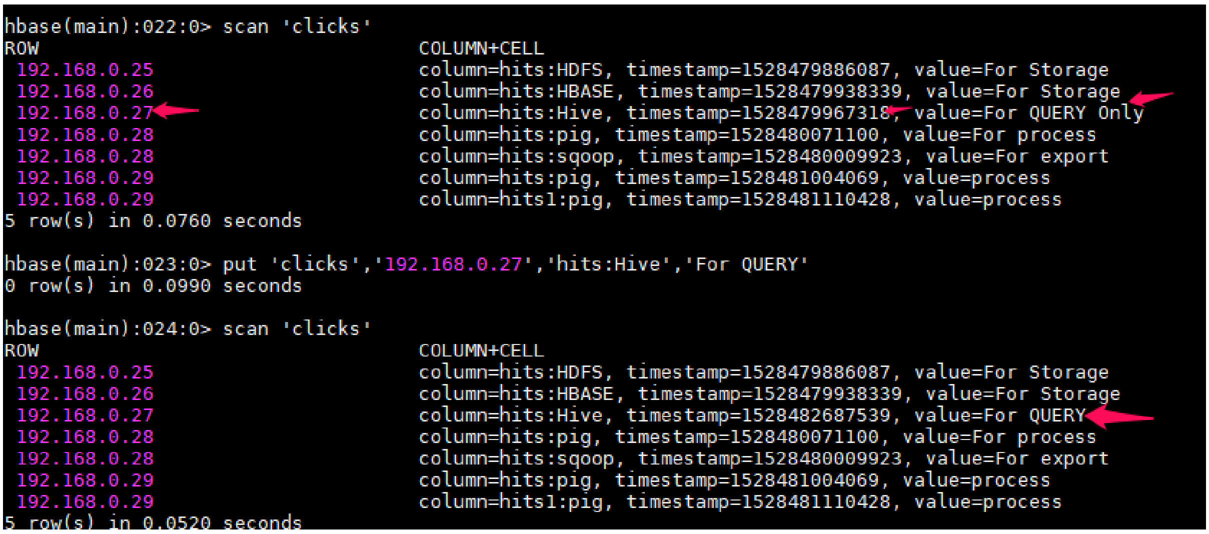
```
hbase(main):053:0> put 'clicks','192.168.0.29','hits:pig','process'
```

I am creating new column family on exist table

```
hbase(main):049:0> alter 'clicks',{NAME=>'hits1',VERSIONS=>6}
```

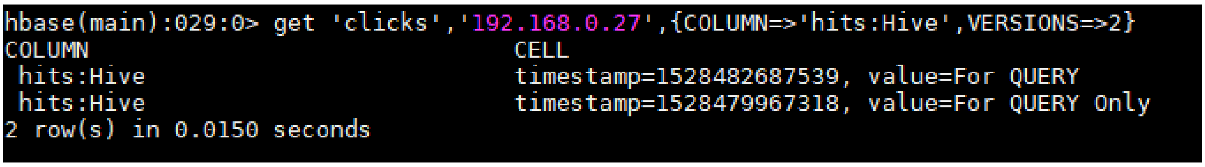
2. Add few records in the table and update some of them. Use IP Address as row-key. Scan the table to view if all the previous versions are getting displayed.

SCAN the table for updated values in the below image have updated the rowkey **192.168.0.27**



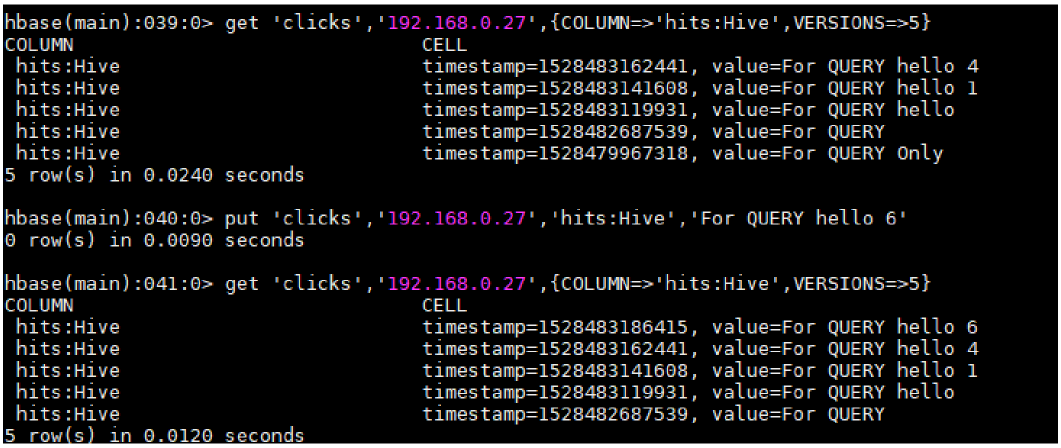
```
hbase(main):022:0> scan 'clicks'
ROW                                  COLUMN+CELL
 192.168.0.25                        column=hits:HDFS, timestamp=1528479886087, value=For Storage
 192.168.0.26                        column=hits:HBASE, timestamp=1528479938339, value=For Storage
 192.168.0.27                        column=hits:Hive, timestamp=1528479967318, value=For QUERY Only
 192.168.0.28                        column=hits:pig, timestamp=1528480071100, value=For process
 192.168.0.28                        column=hits:sqoop, timestamp=1528480009923, value=For export
 192.168.0.29                        column=hits:pig, timestamp=1528481004069, value=process
 192.168.0.29                        column=hits1:pig, timestamp=1528481110428, value=process
5 row(s) in 0.0760 seconds

hbase(main):023:0> put 'clicks','192.168.0.27','hits:Hive','For QUERY'
0 row(s) in 0.0990 seconds

hbase(main):024:0> scan 'clicks'
ROW                                  COLUMN+CELL
 192.168.0.25                        column=hits:HDFS, timestamp=1528479886087, value=For Storage
 192.168.0.26                        column=hits:HBASE, timestamp=1528479938339, value=For Storage
 192.168.0.27                        column=hits:Hive, timestamp=1528482687539, value=For QUERY
 192.168.0.28                        column=hits:pig, timestamp=1528480071100, value=For process
 192.168.0.28                        column=hits:sqoop, timestamp=1528480009923, value=For export
 192.168.0.29                        column=hits:pig, timestamp=1528481004069, value=process
 192.168.0.29                        column=hits1:pig, timestamp=1528481110428, value=process
5 row(s) in 0.0520 seconds
```

Updated and previous values we can see using versioning concept.



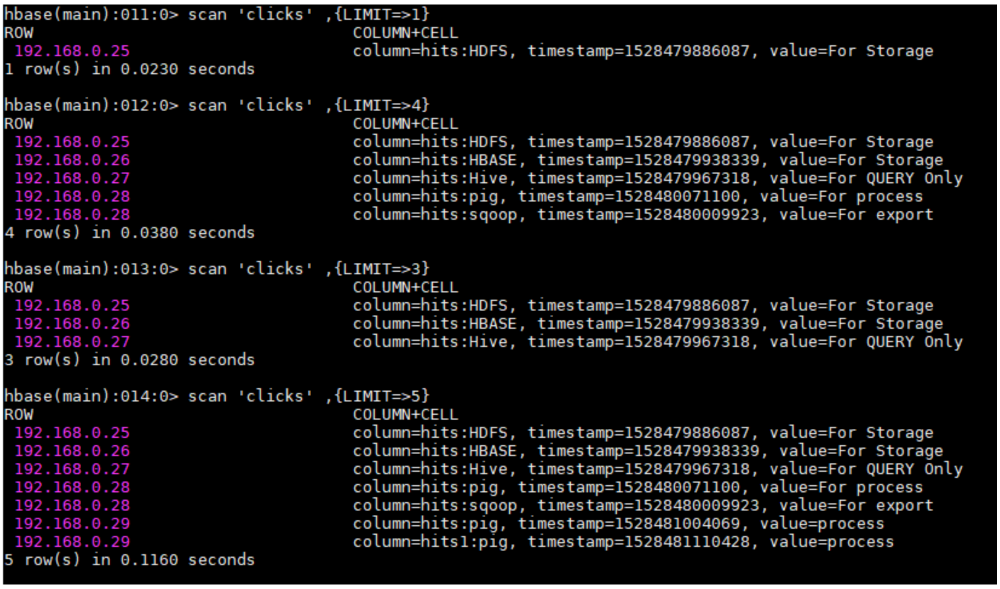
```
hbase(main):029:0> get 'clicks','192.168.0.27',{COLUMN=>'hits:Hive',VERSIONS=>2}
COLUMN                               CELL
 hits:Hive                           timestamp=1528482687539, value=For QUERY
 hits:Hive                           timestamp=1528479967318, value=For QUERY Only
2 row(s) in 0.0150 seconds
```

Version is set to 5 only, IF I updated again on the same row and qualifier the first old will data removed.



```
hbase(main):039:0> get 'clicks','192.168.0.27',{COLUMN=>'hits:Hive',VERSIONS=>5}
COLUMN                               CELL
 hits:Hive                           timestamp=1528483162441, value=For QUERY hello 4
 hits:Hive                           timestamp=1528483141608, value=For QUERY hello 1
 hits:Hive                           timestamp=1528483119931, value=For QUERY hello
 hits:Hive                           timestamp=1528482687539, value=For QUERY
 hits:Hive                           timestamp=1528479967318, value=For QUERY Only
5 row(s) in 0.0240 seconds

hbase(main):040:0> put 'clicks','192.168.0.27','hits:Hive','For QUERY hello 6'
0 row(s) in 0.0090 seconds

hbase(main):041:0> get 'clicks','192.168.0.27',{COLUMN=>'hits:Hive',VERSIONS=>5}
COLUMN                               CELL
 hits:Hive                           timestamp=1528483186415, value=For QUERY hello 6
 hits:Hive                           timestamp=1528483162441, value=For QUERY hello 4
 hits:Hive                           timestamp=1528483141608, value=For QUERY hello 1
 hits:Hive                           timestamp=1528483119931, value=For QUERY hello
 hits:Hive                           timestamp=1528482687539, value=For QUERY
5 row(s) in 0.0120 seconds
```

Limit



```
hbase(main):011:0> scan 'clicks' ,{LIMIT=>1}
ROW                                 COLUMN+CELL
 192.168.0.25                       column=hits:HDFS, timestamp=1528479886087, value=For Storage
1 row(s) in 0.0230 seconds

hbase(main):012:0> scan 'clicks' ,{LIMIT=>4}
ROW                                 COLUMN+CELL
 192.168.0.25                       column=hits:HDFS, timestamp=1528479886087, value=For Storage
 192.168.0.26                       column=hits:HBASE, timestamp=1528479938339, value=For Storage
 192.168.0.27                       column=hits:Hive, timestamp=1528479967318, value=For QUERY Only
 192.168.0.28                       column=hits:pig, timestamp=1528480071100, value=For process
 192.168.0.28                       column=hits:sqoop, timestamp=1528480009923, value=For export
4 row(s) in 0.0380 seconds

hbase(main):013:0> scan 'clicks' ,{LIMIT=>3}
ROW                                 COLUMN+CELL
 192.168.0.25                       column=hits:HDFS, timestamp=1528479886087, value=For Storage
 192.168.0.26                       column=hits:HBASE, timestamp=1528479938339, value=For Storage
 192.168.0.27                       column=hits:Hive, timestamp=1528479967318, value=For QUERY Only
3 row(s) in 0.0280 seconds

hbase(main):014:0> scan 'clicks' ,{LIMIT=>5}
ROW                                 COLUMN+CELL
 192.168.0.25                       column=hits:HDFS, timestamp=1528479886087, value=For Storage
 192.168.0.26                       column=hits:HBASE, timestamp=1528479938339, value=For Storage
 192.168.0.27                       column=hits:Hive, timestamp=1528479967318, value=For QUERY Only
 192.168.0.28                       column=hits:pig, timestamp=1528480071100, value=For process
 192.168.0.28                       column=hits:sqoop, timestamp=1528480009923, value=For export
 192.168.0.29                       column=hits:pig, timestamp=1528481004069, value=process
 192.168.0.29                       column=hits1:pig, timestamp=1528481110428, value=process
5 row(s) in 0.1160 seconds
```