Session 10:HBASE BASICS

Task 1

1.What is NoSQL data base?

Ans : It is not only SQL, which is used for storing unstructured data unlike RDBMS. Main advantage is schema less . Can perform curd operationsCassandra , Mongo, Hbase are examples of nosql data bases.

2.How does data get stored in NoSQl database?

Data is stored as row key in nosql data base mainly for storing unstructured data

3.What is a column family in HBase?

Column is mainly used to avoid joins and all related data we can keep in one column family for one row key.

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

We can add as number of columns in table as we want. No limit.

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

Only column family is declared at schema level and because columns are added dynamically into hbase tables

6.How does data get managed in HBase?

Data store in tables and these tables are stored in regions, when tables become large , regions are splint toto multiple region and these are managed by region servers

The Hadoop DataNode stores the data that the Region Server is managing

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

Data is stored as row key In the tables, when a data is added into a particular table like htest and column family cf, we are adding the column qualifier and the value using put in the hbase shell at run time.

EX put 'htest', 'r1', 'cf:c2', 'v2'

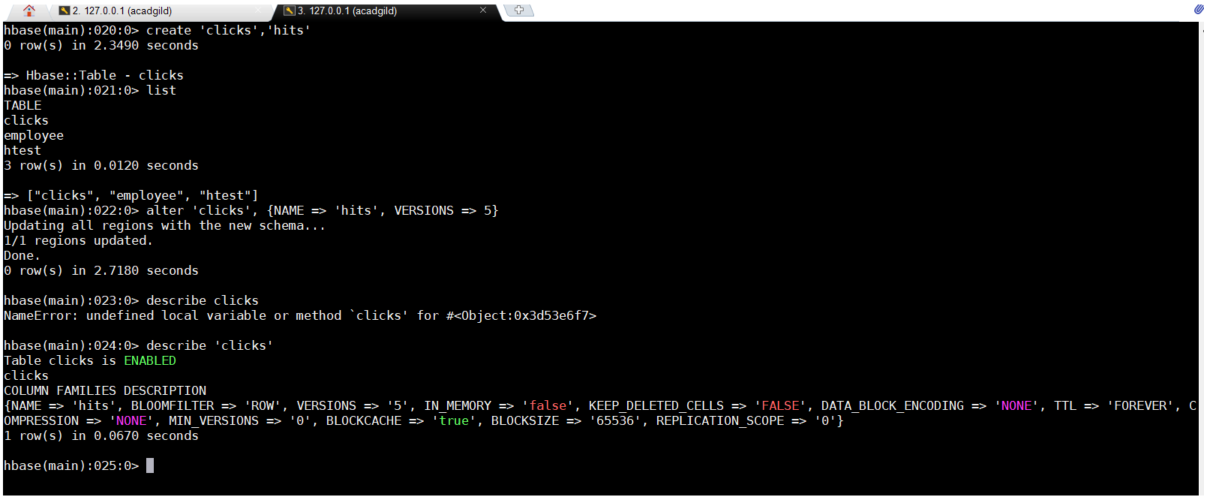
Data store in tables and these tables are stored in regions, when tables become large , regions are splint toto multiple region and these are managed by region servers

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.

create 'clicks','hits'

alter 'clicks', {NAME => 'hits', VERSIONS => 5}
describe 'clicks'



```
hbase(main):020:0> create 'clicks','hits'
0 row(s) in 2.3490 seconds

=> Hbase::Table - clicks
hbase(main):021:0> list
TABLE
clicks
employee
htest
3 row(s) in 0.0120 seconds

=> ["clicks", "employee", "htest"]
hbase(main):022:0> alter 'clicks', {NAME => 'hits', VERSIONS => 5}
Updating all regions with the new schema...
1/1 regions updated.
Done.
0 row(s) in 2.7180 seconds

hbase(main):023:0> describe clicks
NameError: undefined local variable or method `clicks' for #<Object:0x3d53e6f7>

hbase(main):024:0> 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', C
OMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.0670 seconds

hbase(main):025:0>
```

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.

put 'clicks', 'IP_address', 'hits:c2', 'x'

put 'clicks', 'IP_address', 'hits:c3', 'y'

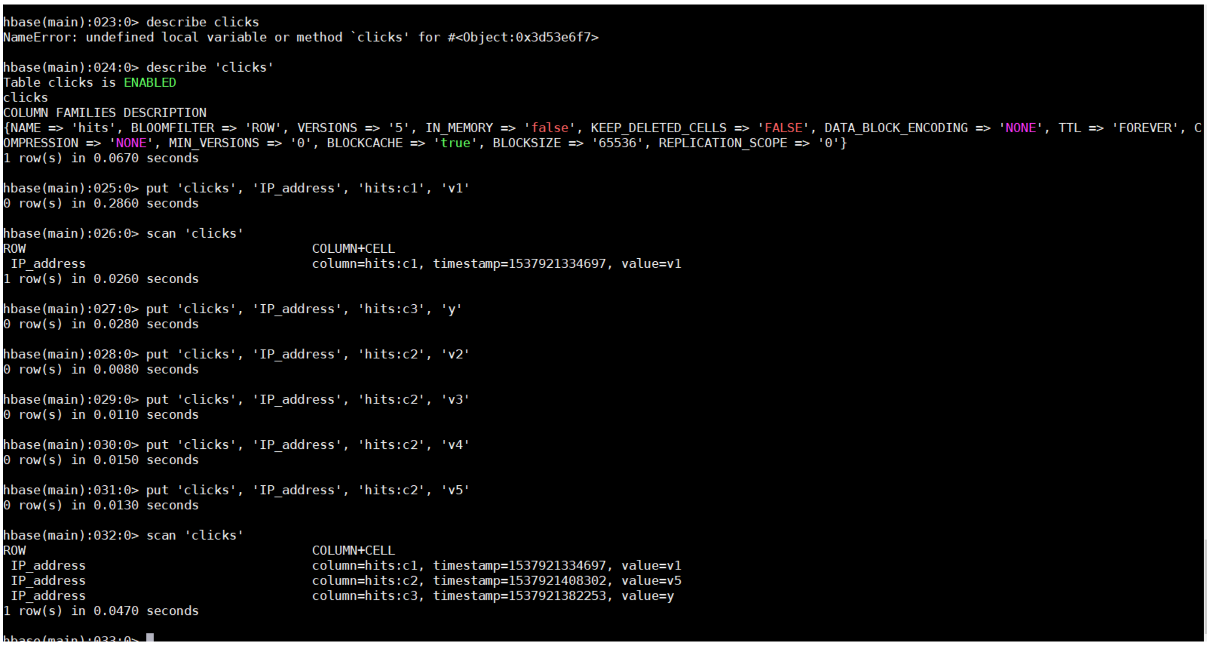
put 'clicks', 'IP_address', 'hits:c2', 'v2'

put 'clicks', 'IP_address', 'hits:c2', 'v3'

put 'clicks', 'IP_address', 'hits:c2', 'v4'

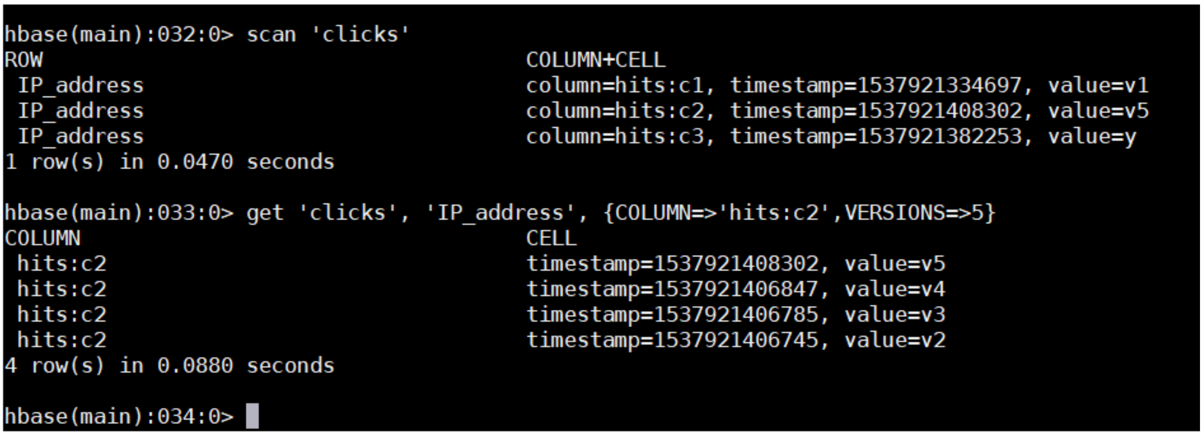
put 'clicks', 'IP_address', 'hits:c2', 'v5'

scan 'clicks'



```
hbase(main):023:0> describe clicks
NameError: undefined local variable or method `clicks' for #<Object:0x3d53e6f7>

hbase(main):024:0> 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', C
OMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKCACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
1 row(s) in 0.0670 seconds

hbase(main):025:0> put 'clicks', 'IP_address', 'hits:c1', 'v1'
0 row(s) in 0.2860 seconds

hbase(main):026:0> scan 'clicks'
ROW                                      COLUMN+CELL
 IP_address                              column=hits:c1, timestamp=1537921334697, value=v1
1 row(s) in 0.0260 seconds

hbase(main):027:0> put 'clicks', 'IP_address', 'hits:c3', 'y'
0 row(s) in 0.0280 seconds

hbase(main):028:0> put 'clicks', 'IP_address', 'hits:c2', 'v2'
0 row(s) in 0.0080 seconds

hbase(main):029:0> put 'clicks', 'IP_address', 'hits:c2', 'v3'
0 row(s) in 0.0110 seconds

hbase(main):030:0> put 'clicks', 'IP_address', 'hits:c2', 'v4'
0 row(s) in 0.0150 seconds

hbase(main):031:0> put 'clicks', 'IP_address', 'hits:c2', 'v5'
0 row(s) in 0.0130 seconds

hbase(main):032:0> scan 'clicks'
ROW                                      COLUMN+CELL
 IP_address                              column=hits:c1, timestamp=1537921334697, value=v1
 IP_address                              column=hits:c2, timestamp=1537921408302, value=v5
 IP_address                              column=hits:c3, timestamp=1537921382253, value=y
1 row(s) in 0.0470 seconds

hbase(main):033:0>
```

get 'clicks', 'IP_address', {COLUMN=>'hits:c2',VERSIONS=>5}



```
hbase(main):032:0> scan 'clicks'
ROW                                      COLUMN+CELL
 IP_address                              column=hits:c1, timestamp=1537921334697, value=v1
 IP_address                              column=hits:c2, timestamp=1537921408302, value=v5
 IP_address                              column=hits:c3, timestamp=1537921382253, value=y
1 row(s) in 0.0470 seconds

hbase(main):033:0> get 'clicks', 'IP_address', {COLUMN=>'hits:c2',VERSIONS=>5}
COLUMN                                   CELL
 hits:c2                                 timestamp=1537921408302, value=v5
 hits:c2                                 timestamp=1537921406847, value=v4
 hits:c2                                 timestamp=1537921406785, value=v3
 hits:c2                                 timestamp=1537921406745, value=v2
4 row(s) in 0.0880 seconds

hbase(main):034:0>
```