TRENDYTECH

UPLIFT YOUR CAREER!

APACHE HBASE

Assignment Solution

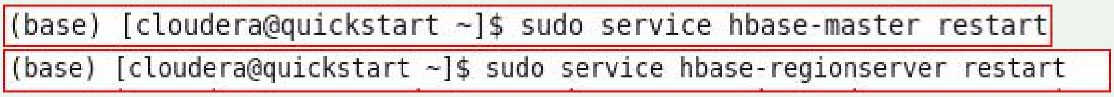
Week7: Introduction to NoSQL - Hbase

Solution :Hive HBase Integration

Creation of Hive Managed HBase Table:

Start the HBase Services-

In Another Terminal:

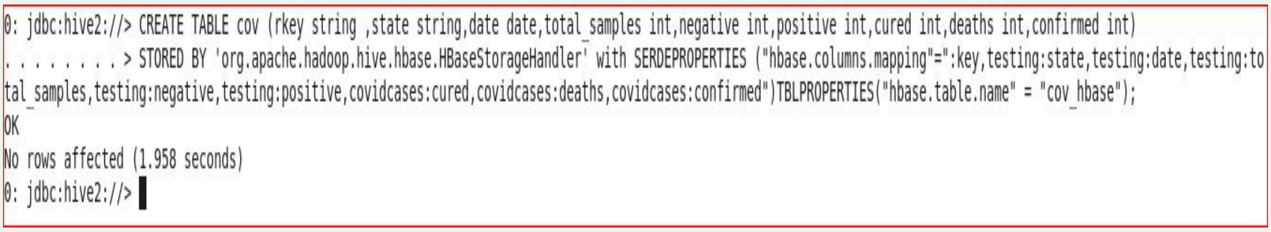


```
(base) [cloudera@quickstart ~]$ sudo service hbase-master restart
(base) [cloudera@quickstart ~]$ sudo service hbase-regionserver restart
```

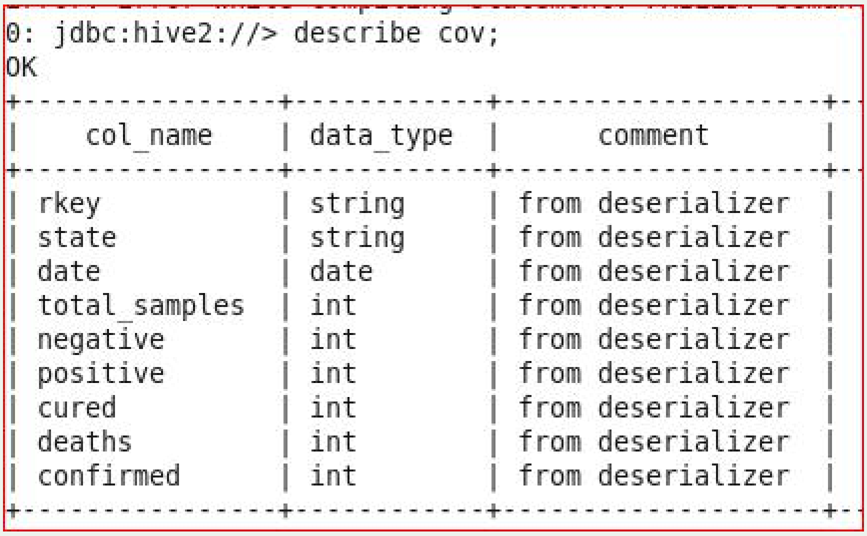
HBASE table creation in hive:

state + date will be a **Rowkey** in HBase table.Because state and date combination will uniquely identify each record,and Rowkeys in HBase must be unique.We specify **rkey** as the **Rowkey** for HBase table. rkey column is derived using - **concat(T.state,'_',T.date)as rkey**

CREATE TABLE cov (rkey string ,state string,date date,total_samples int,negative int,positive int,cured int,deaths int,confirmed int)
STORED BY
'org.apache.hadoop.hive.hbase.HBaseStorageHandler' with SERDEPROPERTIES
("hbase.columns.mapping"=":key,testing:state,testing:date,testing:total_samples,testing:negative,testing:positive,covidcases:cured,covidcases:deaths,covidcases:confirmed")TBLPROPERTIES("hbase.table.name" = "cov_hbase");

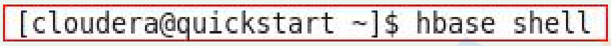


```
0: jdbc:hive2://> CREATE TABLE cov (rkey string ,state string,date date,total_samples int,negative int,positive int,cured int,deaths int,confirmed int)
. . . . . . . . > STORED BY 'org.apache.hadoop.hive.hbase.HBaseStorageHandler' with SERDEPROPERTIES ("hbase.columns.mapping"=":key,testing:state,testing:date,testing:to
tal_samples,testing:negative,testing:positive,covidcases:cured,covidcases:deaths,covidcases:confirmed")TBLPROPERTIES("hbase.table.name" = "cov_hbase");
OK
No rows affected (1.958 seconds)
0: jdbc:hive2://>
```

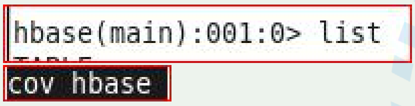


```
0: jdbc:hive2://> describe cov;
OK
+----------------+-----------+---------------------+--
|    col_name    | data_type |       comment       |
+----------------+-----------+---------------------+--
| rkey           | string    | from deserializer   |
| state          | string    | from deserializer   |
| date           | date      | from deserializer   |
| total_samples  | int       | from deserializer   |
| negative       | int       | from deserializer   |
| positive       | int       | from deserializer   |
| cured          | int       | from deserializer   |
| deaths         | int       | from deserializer   |
| confirmed      | int       | from deserializer   |
+----------------+-----------+---------------------+--
```

Verify the table created in HBase:

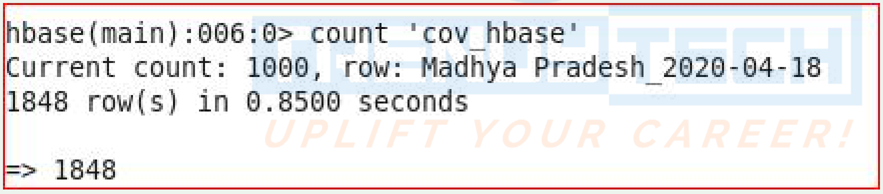


```
[cloudera@quickstart ~]$ hbase shell
```



```
hbase(main):001:0> list
cov_hbase
```

Count Number of records in Hive Managed HBase Table:



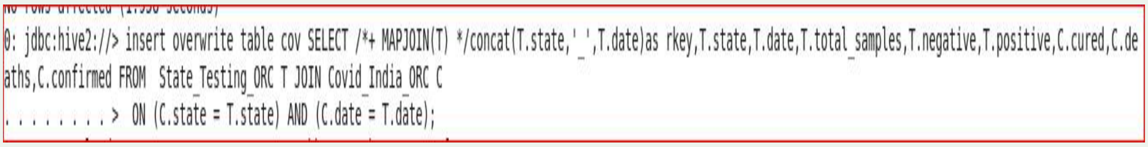
```
hbase(main):006:0> count 'cov_hbase'
Current count: 1000, row: Madhya Pradesh_2020-04-18
1848 row(s) in 0.8500 seconds

=> 1848
```

Populate the Hive-Managed-HBase Table-

Insert data into HBase table Managed by Hive by loading the output of mapside join operation into the hbase table.

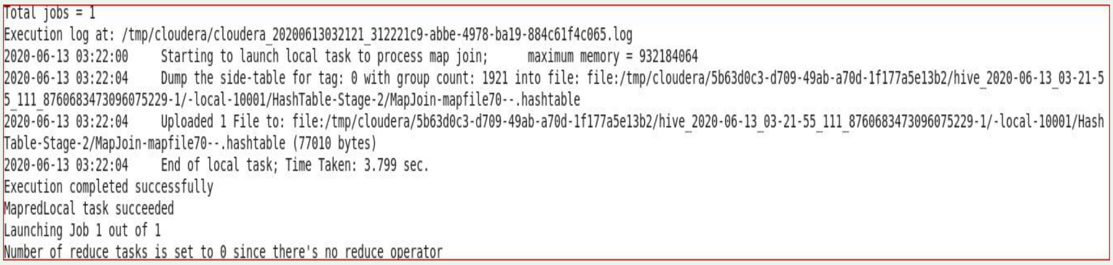
insert overwrite table cov SELECT /*+ MAPJOIN(T) */concat(T.state,'_',T.date)as rkey,T.state,T.date,T.total_samples,T.negative,T.positive,C.cured,C.deaths,C.confirmed FROM State_Testing_ORC T JOIN Covid_India_ORC C ON (C.state = T.state) AND (C.date = T.date);

Note:

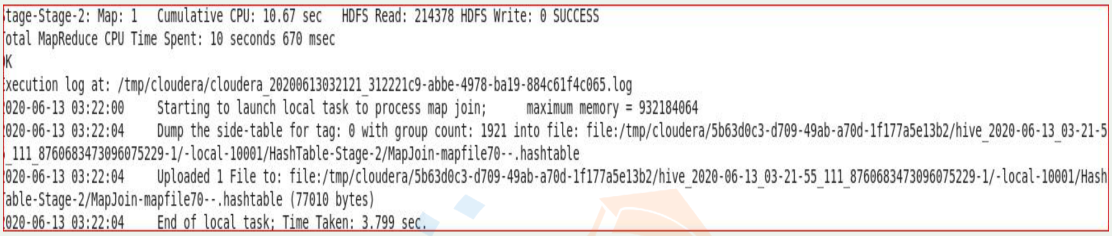


```
No rows affected (1.950 seconds)
0: jdbc:hive2://> insert overwrite table cov SELECT /*+ MAPJOIN(T) */concat(T.state,'_',T.date)as rkey,T.state,T.date,T.total_samples,T.negative,T.positive,C.cured,C.de
aths,C.confirmed FROM  State_Testing_ORC T JOIN Covid_India_ORC C
. . . . . . . . >  ON (C.state = T.state) AND (C.date = T.date);
```

We can see that Map-side join works and reducer is 0.



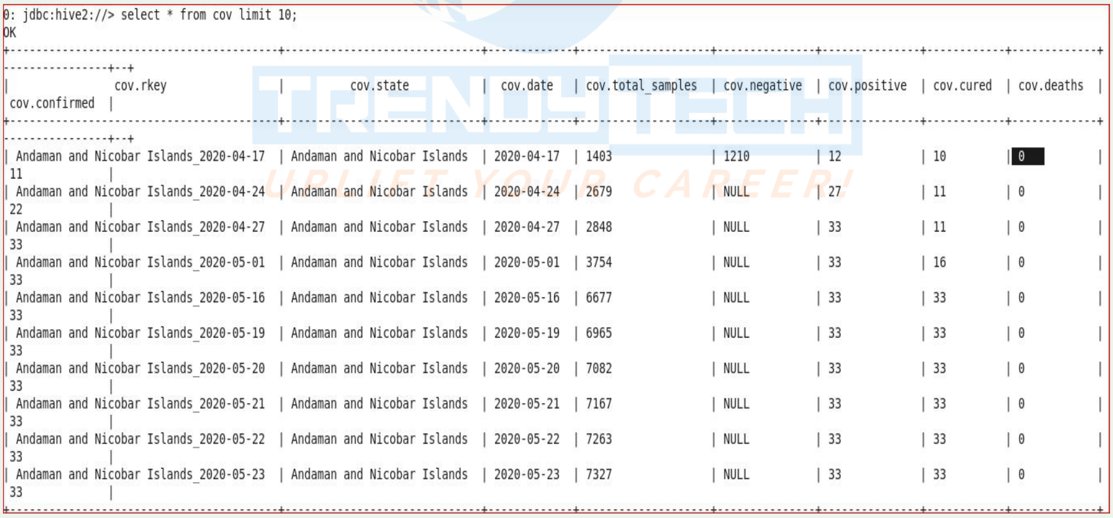
```
Total jobs = 1
Execution log at: /tmp/cloudera/cloudera_20200613032121_312221c9-abbe-4978-ba19-884c61f4c065.log
2020-06-13 03:22:00     Starting to launch local task to process map join;      maximum memory = 932184064
2020-06-13 03:22:04     Dump the side-table for tag: 0 with group count: 1921 into file: file:/tmp/cloudera/5b63d0c3-d709-49ab-a70d-1f177a5e13b2/hive_2020-06-13_03-21-5
5_111_8760683473096075229-1/-local-10001/HashTable-Stage-2/MapJoin-mapfile70--.hashtable
2020-06-13 03:22:04     Uploaded 1 File to: file:/tmp/cloudera/5b63d0c3-d709-49ab-a70d-1f177a5e13b2/hive_2020-06-13_03-21-55_111_8760683473096075229-1/-local-10001/Hash
Table-Stage-2/MapJoin-mapfile70--.hashtable (77010 bytes)
2020-06-13 03:22:04     End of local task; Time Taken: 3.799 sec.
Execution completed successfully
MapredLocal task succeeded
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
```



```
tage-Stage-2: Map: 1   Cumulative CPU: 10.67 sec   HDFS Read: 214378 HDFS Write: 0 SUCCESS
otal MapReduce CPU Time Spent: 10 seconds 670 msec
K
xecution log at: /tmp/cloudera/cloudera_20200613032121_312221c9-abbe-4978-ba19-884c61f4c065.log
020-06-13 03:22:00     Starting to launch local task to process map join;      maximum memory = 932184064
020-06-13 03:22:04     Dump the side-table for tag: 0 with group count: 1921 into file: file:/tmp/cloudera/5b63d0c3-d709-49ab-a70d-1f177a5e13b2/hive_2020-06-13_03-21-5
 111_8760683473096075229-1/-local-10001/HashTable-Stage-2/MapJoin-mapfile70--.hashtable
020-06-13 03:22:04     Uploaded 1 File to: file:/tmp/cloudera/5b63d0c3-d709-49ab-a70d-1f177a5e13b2/hive_2020-06-13_03-21-55_111_8760683473096075229-1/-local-10001/Hash
able-Stage-2/MapJoin-mapfile70--.hashtable (77010 bytes)
020-06-13 03:22:04     End of local task; Time Taken: 3.799 sec.
```

Display few records from the cov table in hive:

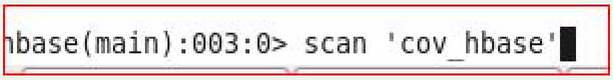
select * from cov limit 10;



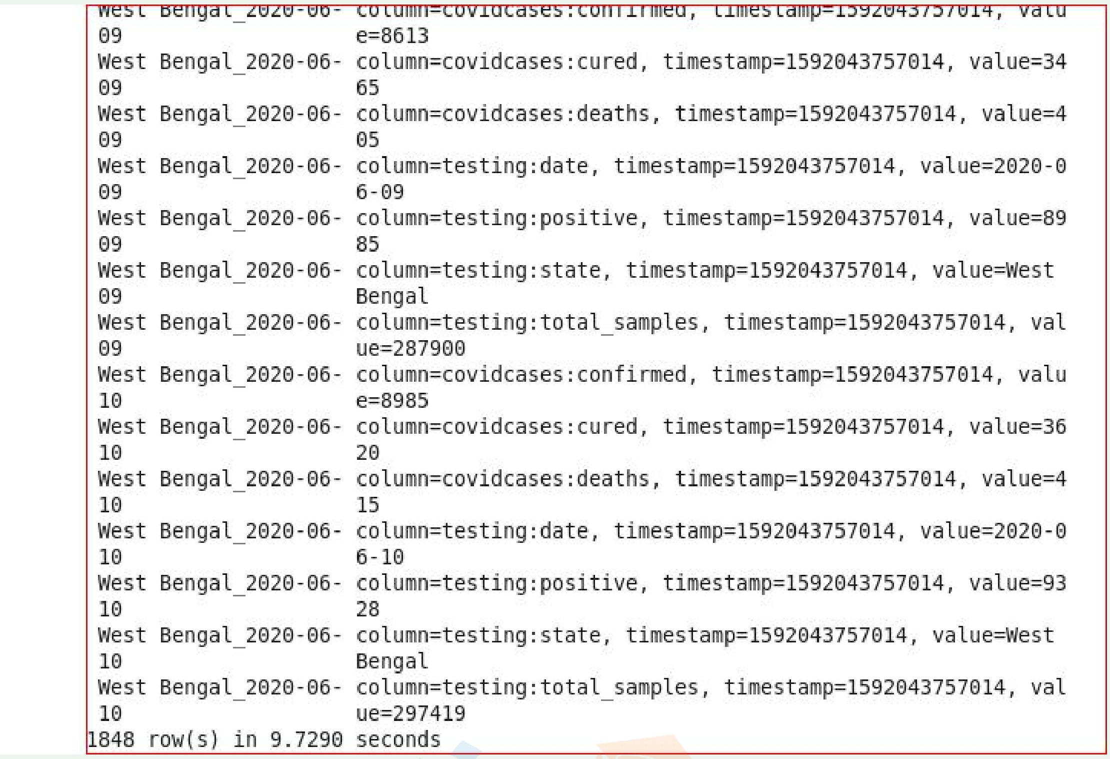
```
0: jdbc:hive2://> select * from cov limit 10;
OK
+----------------------------------------+-------------------------------+-------------+-----------------------+--------------+--------------+-------------+-------------+
----------------+--+
|                cov.rkey                |           cov.state           |  cov.date   | cov.total_samples  | cov.negative  | cov.positive  | cov.cured  | cov.deaths  |
 cov.confirmed  |
+----------------------------------------+-------------------------------+-------------+-----------------------+--------------+--------------+-------------+-------------+
----------------+--+
| Andaman and Nicobar Islands_2020-04-17  | Andaman and Nicobar Islands  | 2020-04-17  | 1403                | 1210         | 12           | 10          | 0           |
 11             |
| Andaman and Nicobar Islands_2020-04-24  | Andaman and Nicobar Islands  | 2020-04-24  | 2679                | NULL         | 27           | 11          | 0           |
 22             |
| Andaman and Nicobar Islands_2020-04-27  | Andaman and Nicobar Islands  | 2020-04-27  | 2848                | NULL         | 33           | 11          | 0           |
 33             |
| Andaman and Nicobar Islands_2020-05-01  | Andaman and Nicobar Islands  | 2020-05-01  | 3754                | NULL         | 33           | 16          | 0           |
 33             |
| Andaman and Nicobar Islands_2020-05-16  | Andaman and Nicobar Islands  | 2020-05-16  | 6677                | NULL         | 33           | 33          | 0           |
 33             |
| Andaman and Nicobar Islands_2020-05-19  | Andaman and Nicobar Islands  | 2020-05-19  | 6965                | NULL         | 33           | 33          | 0           |
 33             |
| Andaman and Nicobar Islands_2020-05-20  | Andaman and Nicobar Islands  | 2020-05-20  | 7082                | NULL         | 33           | 33          | 0           |
 33             |
| Andaman and Nicobar Islands_2020-05-21  | Andaman and Nicobar Islands  | 2020-05-21  | 7167                | NULL         | 33           | 33          | 0           |
 33             |
| Andaman and Nicobar Islands_2020-05-22  | Andaman and Nicobar Islands  | 2020-05-22  | 7263                | NULL         | 33           | 33          | 0           |
 33             |
| Andaman and Nicobar Islands_2020-05-23  | Andaman and Nicobar Islands  | 2020-05-23  | 7327                | NULL         | 33           | 33          | 0           |
 33             |
+----------------------------------------+-------------------------------+-------------+-----------------------+--------------+--------------+-------------+-------------+
```

Check the cov_hbase table in hbase:

scan 'cov_hbase'



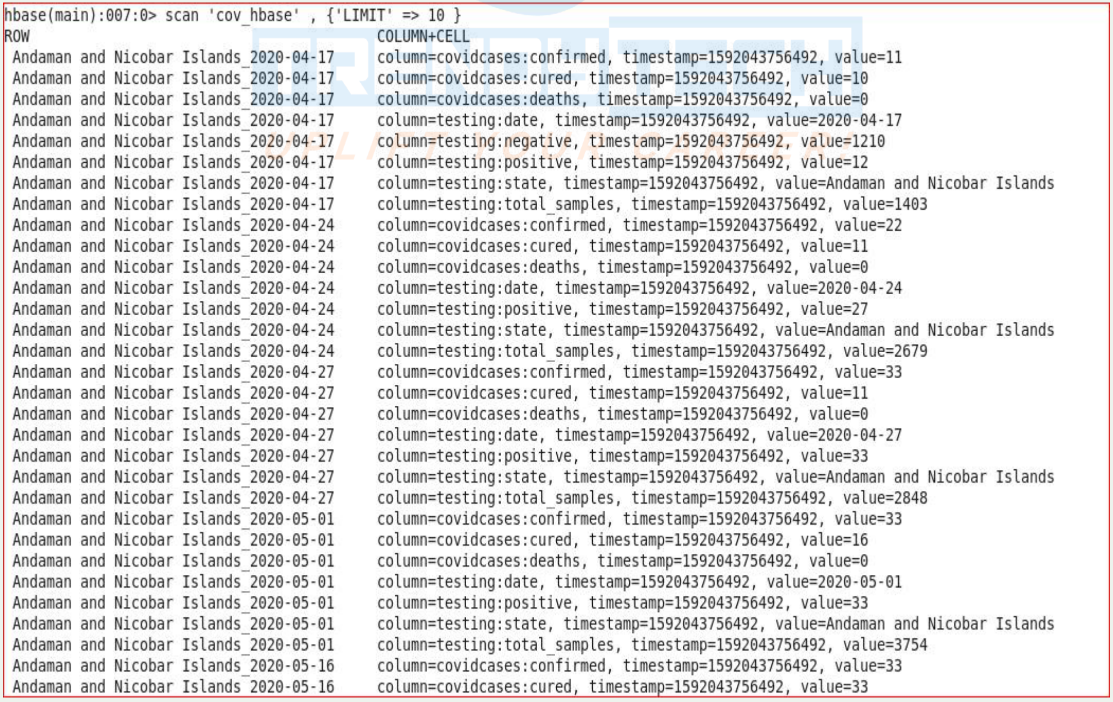
```
hbase(main):003:0> scan 'cov_hbase'
```



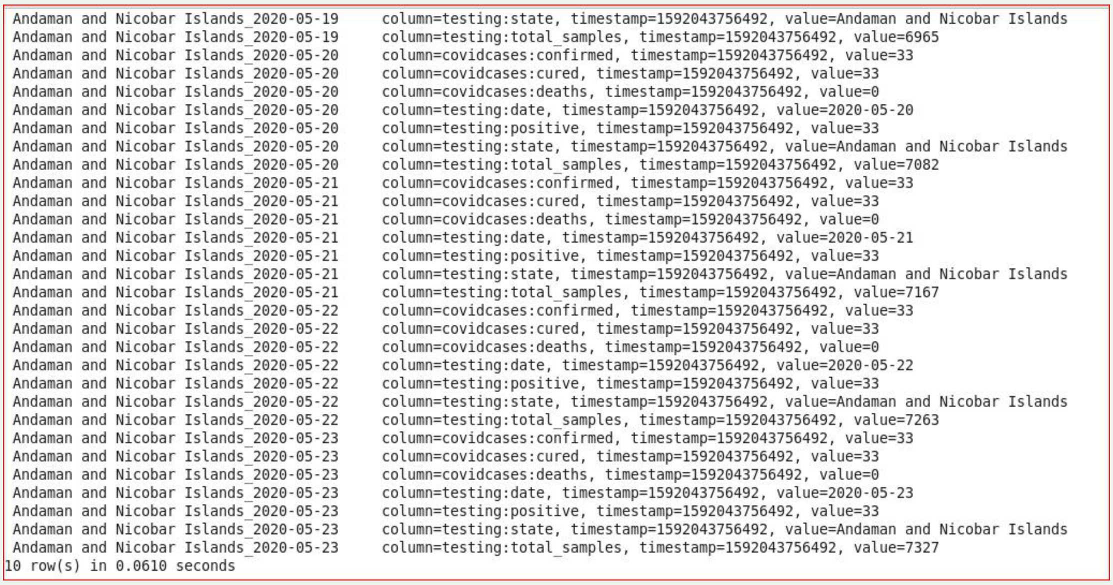
```
West Bengal_2020-06- column=covidcases:confirmed, timestamp=1592043757014, valu
09                   e=8613
West Bengal_2020-06- column=covidcases:cured, timestamp=1592043757014, value=34
09                   65
West Bengal_2020-06- column=covidcases:deaths, timestamp=1592043757014, value=4
09                   05
West Bengal_2020-06- column=testing:date, timestamp=1592043757014, value=2020-0
09                   6-09
West Bengal_2020-06- column=testing:positive, timestamp=1592043757014, value=89
09                   85
West Bengal_2020-06- column=testing:state, timestamp=1592043757014, value=West
09                   Bengal
West Bengal_2020-06- column=testing:total_samples, timestamp=1592043757014, val
09                   ue=287900
West Bengal_2020-06- column=covidcases:confirmed, timestamp=1592043757014, valu
10                   e=8985
West Bengal_2020-06- column=covidcases:cured, timestamp=1592043757014, value=36
10                   20
West Bengal_2020-06- column=covidcases:deaths, timestamp=1592043757014, value=4
10                   15
West Bengal_2020-06- column=testing:date, timestamp=1592043757014, value=2020-0
10                   6-10
West Bengal_2020-06- column=testing:positive, timestamp=1592043757014, value=93
10                   28
West Bengal_2020-06- column=testing:state, timestamp=1592043757014, value=West
10                   Bengal
West Bengal_2020-06- column=testing:total_samples, timestamp=1592043757014, val
10                   ue=297419
1848 row(s) in 9.7290 seconds
```

Scan only a few records in HBase:

scan 'cov_hbase' , {'LIMIT' => 10 }



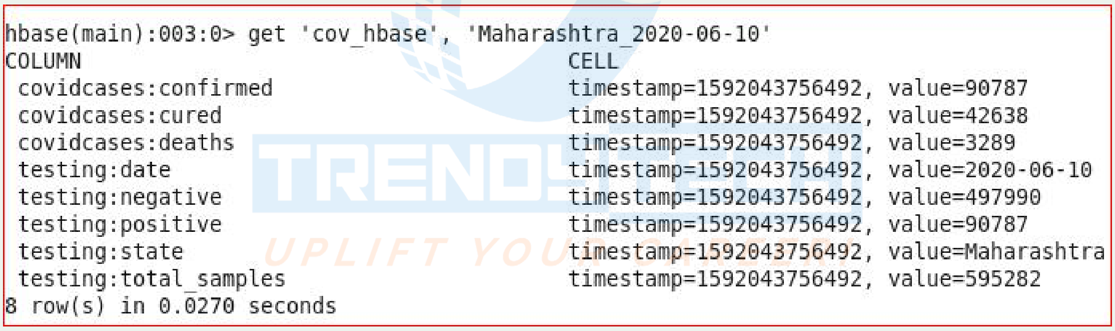
```
hbase(main):007:0> scan 'cov_hbase' , {'LIMIT' => 10 }
ROW                                          COLUMN+CELL
 Andaman and Nicobar Islands_2020-04-17      column=covidcases:confirmed, timestamp=1592043756492, value=11
 Andaman and Nicobar Islands_2020-04-17      column=covidcases:cured, timestamp=1592043756492, value=10
 Andaman and Nicobar Islands_2020-04-17      column=covidcases:deaths, timestamp=1592043756492, value=0
 Andaman and Nicobar Islands_2020-04-17      column=testing:date, timestamp=1592043756492, value=2020-04-17
 Andaman and Nicobar Islands_2020-04-17      column=testing:negative, timestamp=1592043756492, value=1210
 Andaman and Nicobar Islands_2020-04-17      column=testing:positive, timestamp=1592043756492, value=12
 Andaman and Nicobar Islands_2020-04-17      column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-04-17      column=testing:total_samples, timestamp=1592043756492, value=1403
 Andaman and Nicobar Islands_2020-04-24      column=covidcases:confirmed, timestamp=1592043756492, value=22
 Andaman and Nicobar Islands_2020-04-24      column=covidcases:cured, timestamp=1592043756492, value=11
 Andaman and Nicobar Islands_2020-04-24      column=covidcases:deaths, timestamp=1592043756492, value=0
 Andaman and Nicobar Islands_2020-04-24      column=testing:date, timestamp=1592043756492, value=2020-04-24
 Andaman and Nicobar Islands_2020-04-24      column=testing:positive, timestamp=1592043756492, value=27
 Andaman and Nicobar Islands_2020-04-24      column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-04-24      column=testing:total_samples, timestamp=1592043756492, value=2679
 Andaman and Nicobar Islands_2020-04-27      column=covidcases:confirmed, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-04-27      column=covidcases:cured, timestamp=1592043756492, value=11
 Andaman and Nicobar Islands_2020-04-27      column=covidcases:deaths, timestamp=1592043756492, value=0
 Andaman and Nicobar Islands_2020-04-27      column=testing:date, timestamp=1592043756492, value=2020-04-27
 Andaman and Nicobar Islands_2020-04-27      column=testing:positive, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-04-27      column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-04-27      column=testing:total_samples, timestamp=1592043756492, value=2848
 Andaman and Nicobar Islands_2020-05-01      column=covidcases:confirmed, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-01      column=covidcases:cured, timestamp=1592043756492, value=16
 Andaman and Nicobar Islands_2020-05-01      column=covidcases:deaths, timestamp=1592043756492, value=0
 Andaman and Nicobar Islands_2020-05-01      column=testing:date, timestamp=1592043756492, value=2020-05-01
 Andaman and Nicobar Islands_2020-05-01      column=testing:positive, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-01      column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-05-01      column=testing:total_samples, timestamp=1592043756492, value=3754
 Andaman and Nicobar Islands_2020-05-16      column=covidcases:confirmed, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-16      column=covidcases:cured, timestamp=1592043756492, value=33
```



```
 Andaman and Nicobar Islands_2020-05-19     column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-05-19     column=testing:total_samples, timestamp=1592043756492, value=6965
 Andaman and Nicobar Islands_2020-05-20     column=covidcases:confirmed, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-20     column=covidcases:cured, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-20     column=covidcases:deaths, timestamp=1592043756492, value=0
 Andaman and Nicobar Islands_2020-05-20     column=testing:date, timestamp=1592043756492, value=2020-05-20
 Andaman and Nicobar Islands_2020-05-20     column=testing:positive, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-20     column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-05-20     column=testing:total_samples, timestamp=1592043756492, value=7082
 Andaman and Nicobar Islands_2020-05-21     column=covidcases:confirmed, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-21     column=covidcases:cured, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-21     column=covidcases:deaths, timestamp=1592043756492, value=0
 Andaman and Nicobar Islands_2020-05-21     column=testing:date, timestamp=1592043756492, value=2020-05-21
 Andaman and Nicobar Islands_2020-05-21     column=testing:positive, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-21     column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-05-21     column=testing:total_samples, timestamp=1592043756492, value=7167
 Andaman and Nicobar Islands_2020-05-22     column=covidcases:confirmed, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-22     column=covidcases:cured, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-22     column=covidcases:deaths, timestamp=1592043756492, value=0
 Andaman and Nicobar Islands_2020-05-22     column=testing:date, timestamp=1592043756492, value=2020-05-22
 Andaman and Nicobar Islands_2020-05-22     column=testing:positive, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-22     column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-05-22     column=testing:total_samples, timestamp=1592043756492, value=7263
 Andaman and Nicobar Islands_2020-05-23     column=covidcases:confirmed, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-23     column=covidcases:cured, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-23     column=covidcases:deaths, timestamp=1592043756492, value=0
 Andaman and Nicobar Islands_2020-05-23     column=testing:date, timestamp=1592043756492, value=2020-05-23
 Andaman and Nicobar Islands_2020-05-23     column=testing:positive, timestamp=1592043756492, value=33
 Andaman and Nicobar Islands_2020-05-23     column=testing:state, timestamp=1592043756492, value=Andaman and Nicobar Islands
 Andaman and Nicobar Islands_2020-05-23     column=testing:total_samples, timestamp=1592043756492, value=7327
10 row(s) in 0.0610 seconds
```

Search a record in HBase:

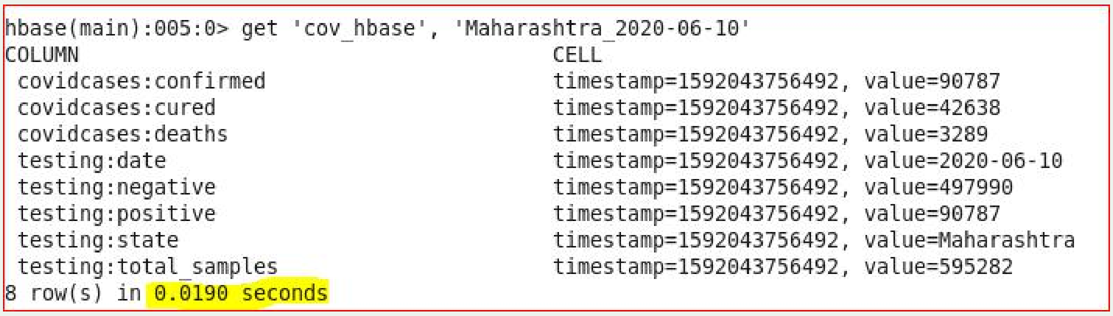
get 'cov_hbase', 'Maharashtra_2020-06-10'



```
hbase(main):003:0> get 'cov_hbase', 'Maharashtra_2020-06-10'
COLUMN                                      CELL
 covidcases:confirmed                       timestamp=1592043756492, value=90787
 covidcases:cured                           timestamp=1592043756492, value=42638
 covidcases:deaths                          timestamp=1592043756492, value=3289
 testing:date                               timestamp=1592043756492, value=2020-06-10
 testing:negative                           timestamp=1592043756492, value=497990
 testing:positive                           timestamp=1592043756492, value=90787
 testing:state                              timestamp=1592043756492, value=Maharashtra
 testing:total_samples                      timestamp=1592043756492, value=595282
8 row(s) in 0.0270 seconds
```

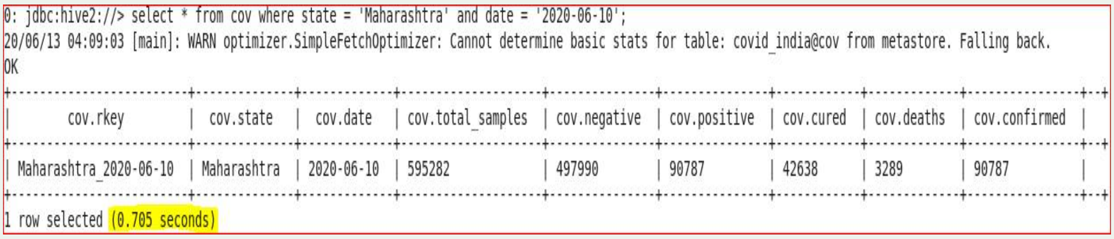
Time - 0.0270 sec

Next read of the same record takes lesser time as its cached in block cache:



```
hbase(main):005:0> get 'cov_hbase', 'Maharashtra_2020-06-10'
COLUMN                                      CELL
 covidcases:confirmed                       timestamp=1592043756492, value=90787
 covidcases:cured                           timestamp=1592043756492, value=42638
 covidcases:deaths                          timestamp=1592043756492, value=3289
 testing:date                               timestamp=1592043756492, value=2020-06-10
 testing:negative                           timestamp=1592043756492, value=497990
 testing:positive                           timestamp=1592043756492, value=90787
 testing:state                              timestamp=1592043756492, value=Maharashtra
 testing:total_samples                      timestamp=1592043756492, value=595282
8 row(s) in 0.0190 seconds
```

Same record when searched in hive:



```
0: jdbc:hive2://> select * from cov where state = 'Maharashtra' and date = '2020-06-10';
20/06/13 04:09:03 [main]: WARN optimizer.SimpleFetchOptimizer: Cannot determine basic stats for table: covid_india@cov from metastore. Falling back.
OK
+-------------------------+--------------+-------------+--------------------+---------------+---------------+------------+-------------+----------------+--+
|        cov.rkey         |  cov.state   |  cov.date   | cov.total_samples  | cov.negative  | cov.positive  | cov.cured  | cov.deaths  | cov.confirmed  |
+-------------------------+--------------+-------------+--------------------+---------------+---------------+------------+-------------+----------------+--+
| Maharashtra_2020-06-10  | Maharashtra  | 2020-06-10  | 595282             | 497990        | 90787         | 42638      | 3289        | 90787          |
+-------------------------+--------------+-------------+--------------------+---------------+---------------+------------+-------------+----------------+--+
1 row selected (0.705 seconds)
```

So hive takes more time to search the same record compared to hbase.

Conclusion: HBase performs quick searches (takes very less time) as compared to Hive.

****