HBASE TUTORIAL

Hands-on Session

by Suchitra Jayaprakash
suchitra@cmi.ac.in

HBASE

- Distributed column-oriented database built on top of the Hadoop file system
- Designed for quick random access.
- Columns are grouped into column families, which must be defined during table creation.
- Basic structure
  - Column – single field in table
  - Column-family – is group of columns
  - Row-key – it is a mandatory field which serves as the unique identifier for every record.

Run HBASE

- **Start Cloudera server**

docker run --hostname=quickstart.cloudera --privileged=true -t -i --publish-all=true -p 8888:8888 -p 8080:80 -p 50070:50070 -p 8088:8088 -p 50075:50075 -p 8032:8032 -p 8042:8042 -p 19888:19888 cloudera/quickstart /usr/bin/docker-quickstart

- **To get HBASE command prompt**
- type "hbase shell" and press enter

HBASE COMMANDS

- **Create table**

create <'tablename'>, <'columnfamilyname'>

create 'Student' , 'personal_data','academic_data','other_data'

```
[root@quickstart /]# hbase shell
2020-03-03 14:35:40,686 INFO  [main] Configuration.deprecation: hadoop.native.li
b is deprecated. Instead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.0-cdh5.7.0, rUnknown, Wed Mar 23 11:39:14 PDT 2016

hbase(main):001:0> create 'Student' , 'personal_data','academic_data','other_dat
a'
0 row(s) in 4.3130 seconds

=> Hbase::Table - Student
hbase(main):002:0>
```

- **Display All tables**

list

displays all the tables that are present or created in HBase

```
hbase(main):002:0> list
TABLE
Student
1 row(s) in 0.1070 seconds

=> ["Student"]
hbase(main):003:0>
```

HBASE COMMANDS

- **Add cell value**

put <'tablename'>,<'rowname'>,<'colfamily:colname'>,<'value'>

put 'Student','S101','personal_data:name','John'
put 'Student','S101','personal_data:address','#145, New Road,Chennai'
put 'Student','S101','academic_data:class','Course A'
put 'Student','S101','academic_data:year','second'

Inserting data into HBase

```
=> ["Student"]
hbase(main):003:0> put 'Student','S101','personal_data:name','John'
0 row(s) in 0.8380 seconds

hbase(main):004:0> put 'Student','S101','personal_data:address','#145, New Road,
Chennai'
0 row(s) in 0.0580 seconds

hbase(main):005:0> put 'Student','S101','academic_data:class','Course A'
0 row(s) in 0.0350 seconds

hbase(main):006:0> put 'Student','S101','academic_data:year','second'
0 row(s) in 0.0650 seconds
```

- **Get a value**

get '<table-name>','<row-key>','<column-family>'

get 'Student','S101‘

retrieve single record from HBase table

```
hbase(main):012:0> get 'Student','S101'
COLUMN                CELL
 academic_data:class  timestamp=1583246362745, value=Course A
 academic_data:year   timestamp=1583246363123, value=second
 personal_data:addres timestamp=1583246362285, value=#145, New Road,Chenna
 s
 personal_data:name   timestamp=1583246430789, value=John
4 row(s) in 0.1480 seconds

hbase(main):013:0>
```

HBASE COMMANDS

- **Diplay row count of table**

count <'tablename'>

count 'Student'

```
hbase(main):013:0> count 'Student'
1 row(s) in 0.3770 seconds

=> 1
hbase(main):014:0>
```

- **Display table data**

scan '<table -name>'

scan 'Student'

get the all the records

```
hbase(main):020:0> scan 'Student'
ROW                   COLUMN+CELL
 S101                 column=academic_data:class, timestamp=1583246362745, value
                      =Course A
 S101                 column=academic_data:year, timestamp=1583246363123, value=
                      second
 S101                 column=personal_data:address, timestamp=1583246362285, val
                      ue=#145, New Road,Chennai
 S101                 column=personal_data:name, timestamp=1583246430789, value=
                      John
1 row(s) in 0.1530 seconds
```

HBASE COMMANDS

- **Deleting records**
- delete '<table-name>','<rowkey>‘,'<column-family>'

- delete 'Student','S101','academic_data:class'

```
hbase(main):021:0> delete 'Student','S101','academic_data:class'
0 row(s) in 0.1480 seconds

hbase(main):022:0> scan 'Student'
ROW                     COLUMN+CELL
 S101                   column=academic_data:year, timestamp=1583246363123, value=
                        second
 S101                   column=personal_data:address, timestamp=1583246362285, val
                        ue=#145, New Road,Chennai
 S101                   column=personal_data:name, timestamp=1583246430789, value=
                        John
1 row(s) in 0.2120 seconds
```

- **Delete table**

disable 'Employee'

drop 'Employee'

HBASE COMMANDS

- **loading HDFS file as table**

- 1)Using ImportTsv to load txt to Hbase

```
docker cp E:/MSc_Datascience/BigDataHadoop/Slides/SampleName.txt
c1ce02333912:/tmp/SampleName.txt
    hadoop fs -copyFromLocal /tmp/SampleName.txt SampleName.txt
```

- 2) create table

```
create 'Person' , 'Name'
```

- 3) import data (execute it outside hbase shell)

```
hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -
Dimporttsv.separator=',' -Dimporttsv.columns='HBASE_ROW_KEY,Name'
Person hdfs://quickstart.cloudera/user/root/SampleName.txt
```

OUTPUT

```
2020-03-03 15:11:54,141 INFO  [main] mapreduce.Job: Job job_1583245414405_0001 c
ompleted successfully
2020-03-03 15:11:54,936 INFO  [main] mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=146345
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=155
                HDFS: Number of bytes written=0
                HDFS: Number of read operations=2
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=0
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=34621
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=34621
                Total vcore-seconds taken by all map tasks=34621
                Total megabyte-seconds taken by all map tasks=35451904
        Map-Reduce Framework
                Map input records=5
                Map output records=5
                Input split bytes=116
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=578
                CPU time spent (ms)=6810
                Physical memory (bytes) snapshot=142856192
                Virtual memory (bytes) snapshot=1307754496
                Total committed heap usage (bytes)=62652416
        ImportTsv
                Bad Lines=0
        File Input Format Counters
                Bytes Read=39
        File Output Format Counters
                Bytes Written=0
[root@quickstart /]#
```

```
hbase(main):001:0> scan 'Person'
ROW                     COLUMN+CELL
 1                      column=Name:, timestamp=1583248180951, value=John
 2                      column=Name:, timestamp=1583248180951, value=Mary
 3                      column=Name:, timestamp=1583248180951, value=Sam
 4                      column=Name:, timestamp=1583248180951, value=Jerry
 5                      column=Name:, timestamp=1583248180951, value=Tom
5 row(s) in 1.5610 seconds

hbase(main):002:0>
```

THANK YOU