Kalyan Big Data Projects – Project 14
How To Stream Text Data Into Hbase Using Apache Flume

Pre-Requisites of Flume Project:

hadoop-2.6.0, flume-1.6.0
hbase-0.98.4, java-1.7

Project Compatibility :
1. hadoop-2.6.0 + hbase-0.98.4 + flume-1.6.0
2. hadoop-2.7.2 + hbase-1.1.2 + flume-1.7.0

NOTE: Make sure that install all the above components

Flume Project Download Links:

`hadoop-2.6.0.tar.gz` ==> link
(https://archive.apache.org/dist/hadoop/core/hadoop-2.6.0/hadoop-2.6.0.tar.gz)

`apache-flume-1.6.0-bin.tar.gz` ==> link
(https://archive.apache.org/dist/flume/1.6.0/apache-flume-1.6.0-bin.tar.gz)

`kalyan-flume-project-0.1.jar` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/kalyan/kalyan-flume-project-0.1.jar)

`kalyan-text-hbase-agent.conf` ==> link
(https://github.com/kalyanhadooptraining/kalyan-bigdata-realtime-projects/blob/master/flume/project14-hbase-text/kalyan-text-hbase-agent.conf)

Learnings of this Project:

- We will learn Flume Configurations and Commands
- Flume Agent
  1. Source (Netcat Source)
  2. Channel (Memory Channel)
  3. Sink (Hbase Sink)
- Major project in Real Time `Chat Applications`
  1. We are extracting the data from Chat Applications
  2. This data will be useful to do analysis on Sentiment on Tweets
  3. Complex Data is the output format then REGEX is best solution
- We can use Hbase to analyze this data

----

1. create "**kalyan-text-hbase-agent.conf**" file with below content

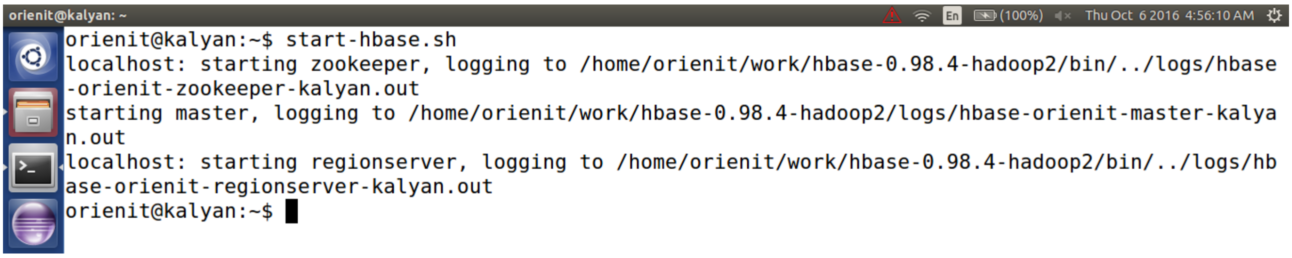
```
agent.sources = NETCAT
agent.channels = MemChannel
agent.sinks = HBASE

agent.sources.NETCAT.type = netcat
agent.sources.NETCAT.bind = localhost
agent.sources.NETCAT.port = 3000
agent.sources.NETCAT.channels = MemChannel

agent.sinks.HBASE.type = hbase
agent.sinks.HBASE.table = sample1
agent.sinks.HBASE.columnFamily = cf
agent.sinks.HBASE.serializer = org.apache.flume.sink.hbase.SimpleHbaseEventSerializer
agent.sinks.HBASE.serializer.payloadColumn=message
agent.sinks.HBASE.serializer.incrementColumn=inc
agent.sinks.HBASE.channel = MemChannel

agent.channels.MemChannel.type = memory
agent.channels.MemChannel.capacity = 1000
agent.channels.MemChannel.transactionCapacity = 100
```

2. Copy "**kalyan-text-hbase-agent.conf**" file into "**$FUME_HOME/conf**" folder

3. Copy "**kalyan-flume-project-0.1.jar**" file into "**$FLUME_HOME/lib**" folder

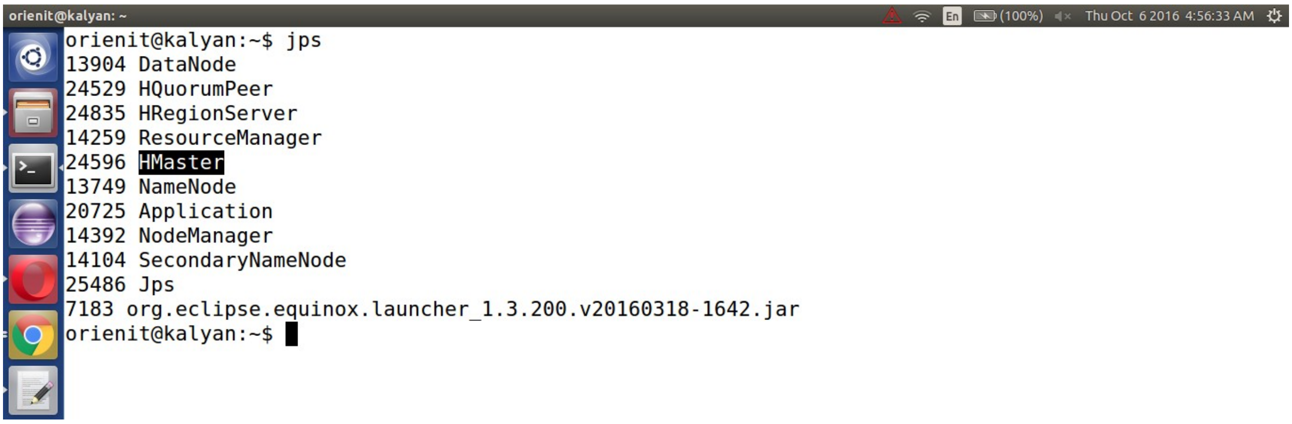
4. To work with **Flume + Hbase Integration,** Follow the below steps

i) Start the hbase using below '**start-hbase.sh**' command.



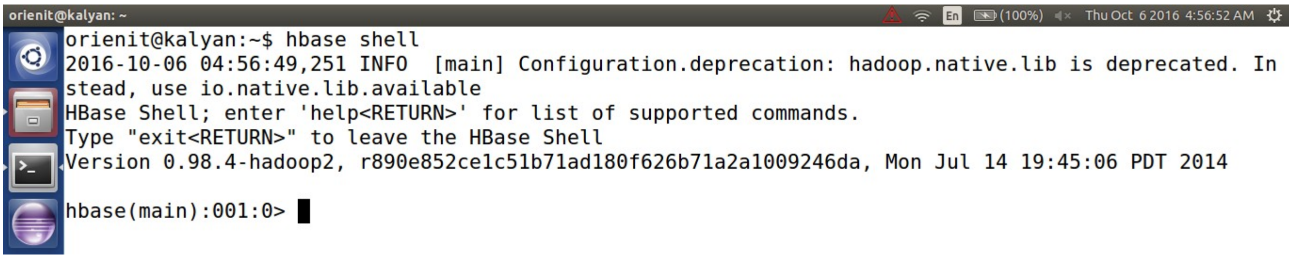
```
orienit@kalyan:~$ start-hbase.sh
localhost: starting zookeeper, logging to /home/orienit/work/hbase-0.98.4-hadoop2/bin/../logs/hbase
-orienit-zookeeper-kalyan.out
starting master, logging to /home/orienit/work/hbase-0.98.4-hadoop2/logs/hbase-orienit-master-kalya
n.out
localhost: starting regionserver, logging to /home/orienit/work/hbase-0.98.4-hadoop2/bin/../logs/hb
ase-orienit-regionserver-kalyan.out
orienit@kalyan:~$ 
```

ii. verify the hbase is running or not with "**jps**" command



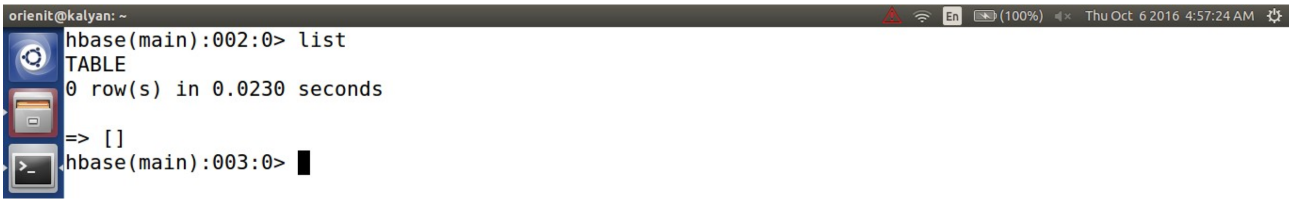
```
orienit@kalyan:~$ jps
13904 DataNode
24529 HQuorumPeer
24835 HRegionServer
14259 ResourceManager
24596 HMaster
13749 NameNode
20725 Application
14392 NodeManager
14104 SecondaryNameNode
25486 Jps
7183 org.eclipse.equinox.launcher_1.3.200.v20160318-1642.jar
orienit@kalyan:~$
```

iii. connect to hbase using '**hbase shell**' command



```
orienit@kalyan:~$ hbase shell
2016-10-06 04:56:49,251 INFO  [main] Configuration.deprecation: hadoop.native.lib is deprecated. In
stead, use io.native.lib.available
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 0.98.4-hadoop2, r890e852ce1c51b71ad180f626b71a2a1009246da, Mon Jul 14 19:45:06 PDT 2014

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

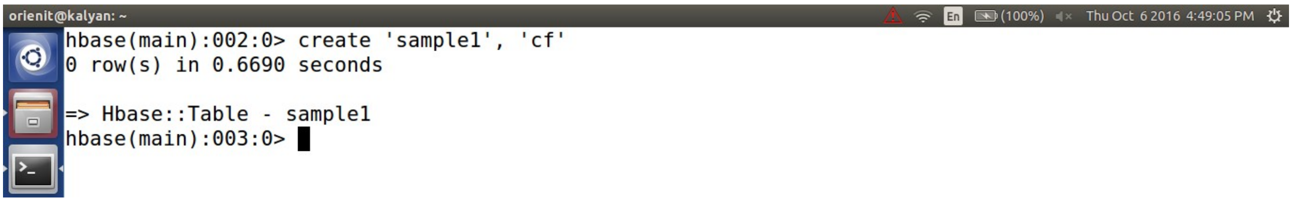
iv. list out all the tables in hbase using '**list**' command



```
hbase(main):002:0> list
TABLE
0 row(s) in 0.0230 seconds

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

v. create the hbase table name is '**sample1**' with column family name is '**cf**' using below command.

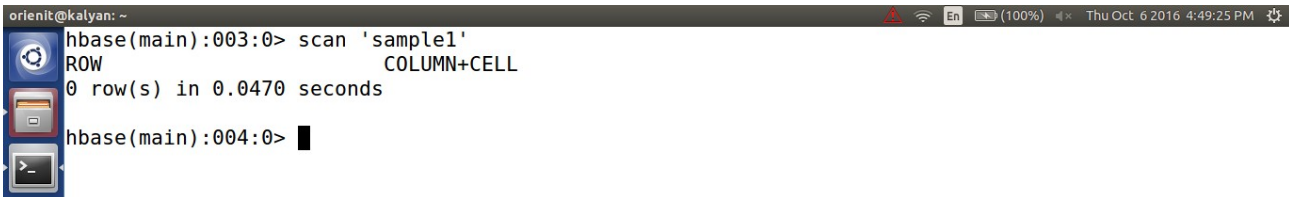
create 'sample1', 'cf'



```
hbase(main):002:0> create 'sample1', 'cf'
0 row(s) in 0.6690 seconds

=> Hbase::Table - sample1
hbase(main):003:0>
```

vi. read the data from hbase table 'sample1' using below command.

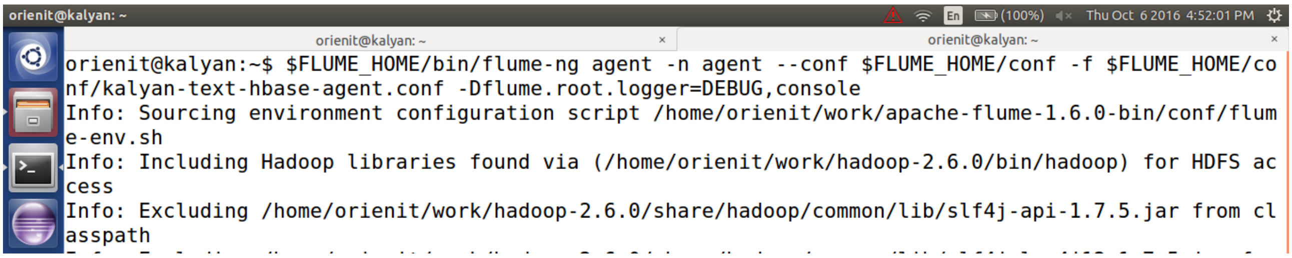
scan 'sample1'



```
hbase(main):003:0> scan 'sample1'
ROW                         COLUMN+CELL
0 row(s) in 0.0470 seconds

hbase(main):004:0>
```

5. Execute the below command to `**Extract data from Text data into HBase using Flume**`

$FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f $FLUME_HOME/conf/kalyan-text-hbase-agent.conf -Dflume.root.logger=DEBUG,console

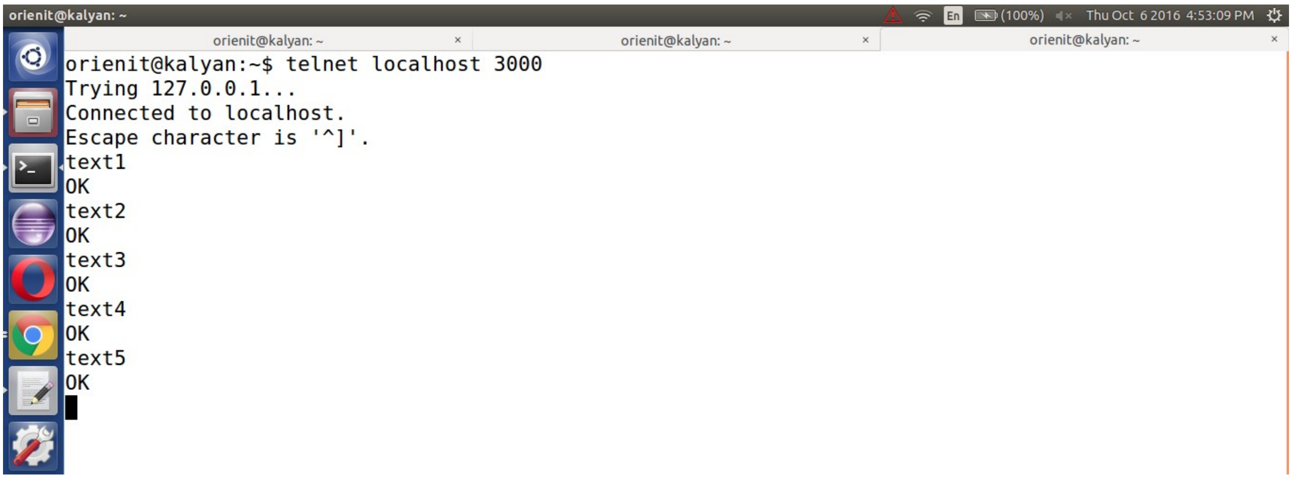


```
orienit@kalyan:~$ $FLUME_HOME/bin/flume-ng agent -n agent --conf $FLUME_HOME/conf -f $FLUME_HOME/co
nf/kalyan-text-hbase-agent.conf -Dflume.root.logger=DEBUG,console
Info: Sourcing environment configuration script /home/orienit/work/apache-flume-1.6.0-bin/conf/flum
e-env.sh
Info: Including Hadoop libraries found via (/home/orienit/work/hadoop-2.6.0/bin/hadoop) for HDFS ac
cess
Info: Excluding /home/orienit/work/hadoop-2.6.0/share/hadoop/common/lib/slf4j-api-1.7.5.jar from cl
asspath
```

6. Connect to Socket Server using below command

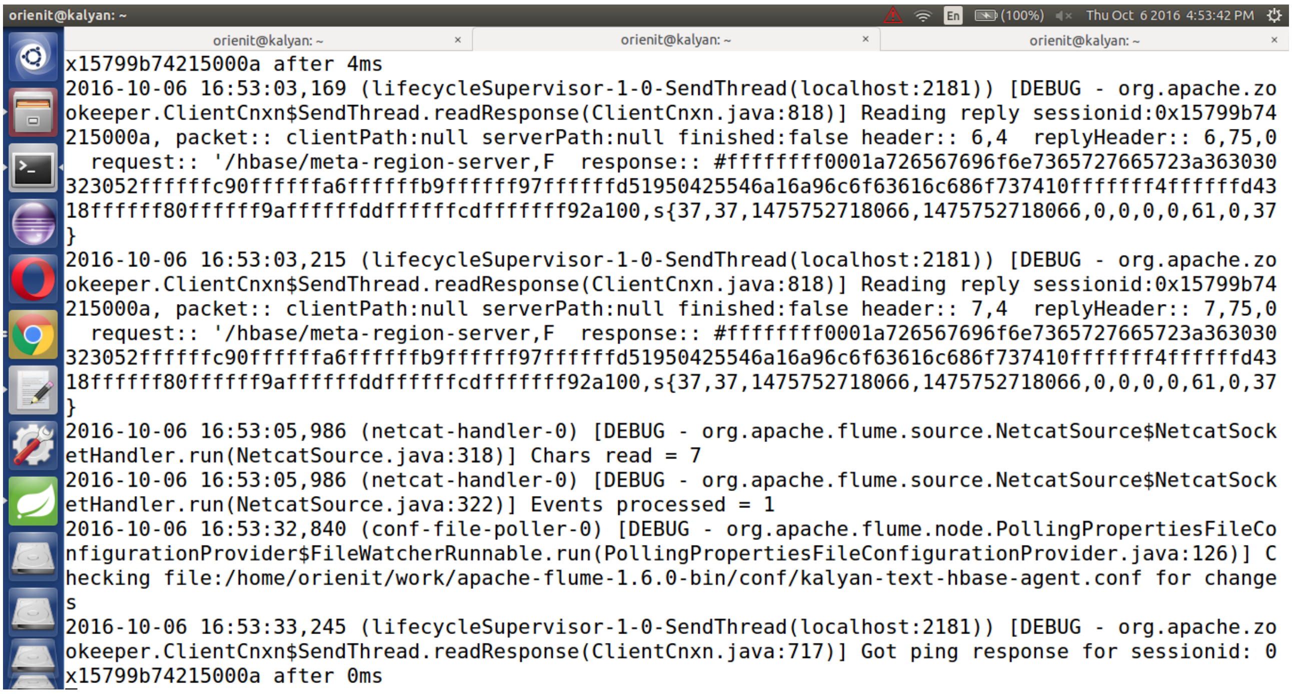
telnet localhost 3000

NOTE: send the sample text to flume like below screen



```
orienit@kalyan:~$ telnet localhost 3000
Trying 127.0.0.1...
Connected to localhost.
Escape character is '^]'.
text1
OK
text2
OK
text3
OK
text4
OK
text5
OK
```

7. Verify the data in console



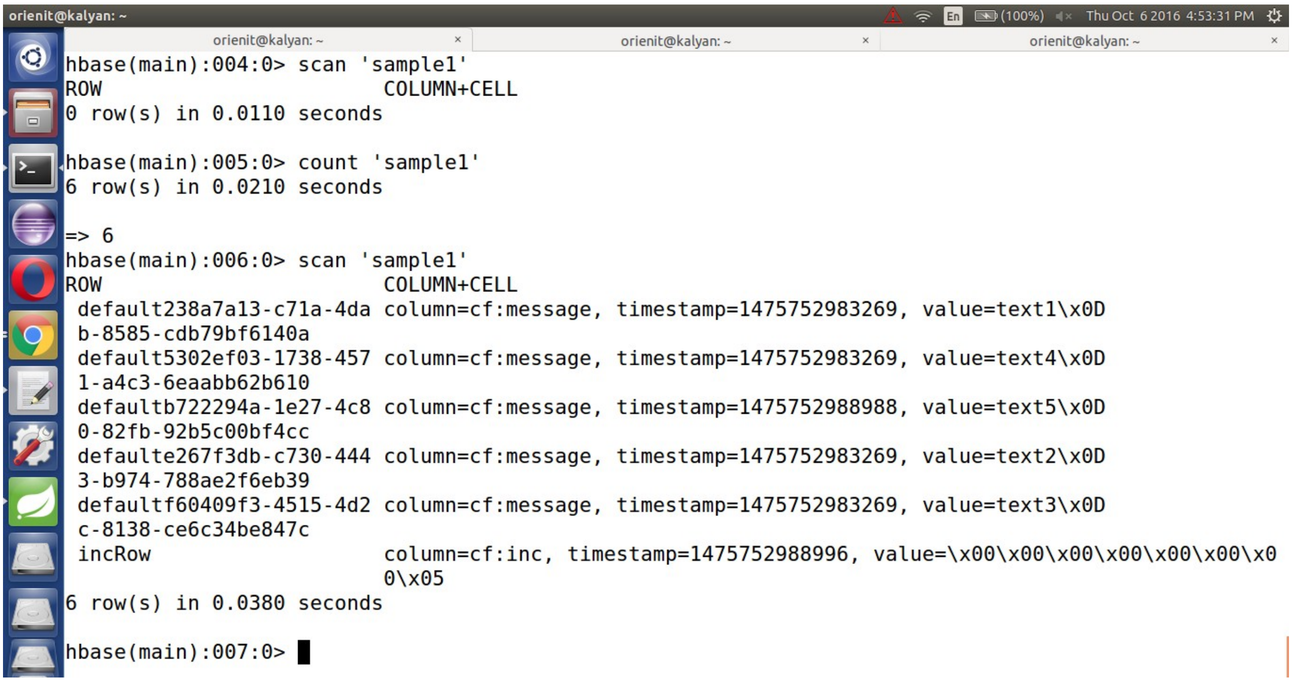
```
x15799b74215000a after 4ms
2016-10-06 16:53:03,169 (lifecycleSupervisor-1-0-SendThread(localhost:2181)) [DEBUG - org.apache.zo
okeeper.ClientCnxn$SendThread.readResponse(ClientCnxn.java:818)] Reading reply sessionid:0x15799b74
215000a, packet:: clientPath:null serverPath:null finished:false header:: 6,4  replyHeader:: 6,75,0
  request:: '/hbase/meta-region-server,F  response:: #ffffffff0001a726567696f6e7365727665723a363030
323052ffffffc90ffffffa6ffffffb9ffffff97ffffffd51950425546a16a96c6f63616c686f737410ffffff4ffffffd43
18ffffff80ffffff9affffffddffffffcdffffff92a100,s{37,37,1475752718066,1475752718066,0,0,0,0,61,0,37
}
2016-10-06 16:53:03,215 (lifecycleSupervisor-1-0-SendThread(localhost:2181)) [DEBUG - org.apache.zo
okeeper.ClientCnxn$SendThread.readResponse(ClientCnxn.java:818)] Reading reply sessionid:0x15799b74
215000a, packet:: clientPath:null serverPath:null finished:false header:: 7,4  replyHeader:: 7,75,0
  request:: '/hbase/meta-region-server,F  response:: #ffffffff0001a726567696f6e7365727665723a363030
323052ffffffc90ffffffa6ffffffb9ffffff97ffffffd51950425546a16a96c6f63616c686f737410ffffff4ffffffd43
18ffffff80ffffff9affffffddffffffcdffffff92a100,s{37,37,1475752718066,1475752718066,0,0,0,0,61,0,37
}
2016-10-06 16:53:05,986 (netcat-handler-0) [DEBUG - org.apache.flume.source.NetcatSource$NetcatSock
etHandler.run(NetcatSource.java:318)] Chars read = 7
2016-10-06 16:53:05,986 (netcat-handler-0) [DEBUG - org.apache.flume.source.NetcatSource$NetcatSock
etHandler.run(NetcatSource.java:322)] Events processed = 1
2016-10-06 16:53:32,840 (conf-file-poller-0) [DEBUG - org.apache.flume.node.PollingPropertiesFileCo
nfigurationProvider$FileWatcherRunnable.run(PollingPropertiesFileConfigurationProvider.java:126)] C
hecking file:/home/orienit/work/apache-flume-1.6.0-bin/conf/kalyan-text-hbase-agent.conf for change
s
2016-10-06 16:53:33,245 (lifecycleSupervisor-1-0-SendThread(localhost:2181)) [DEBUG - org.apache.zo
okeeper.ClientCnxn$SendThread.readResponse(ClientCnxn.java:717)] Got ping response for sessionid: 0
x15799b74215000a after 0ms
```

8. Verify the data in HBase

Execute below command to get the data from hbase table '**sample1**'

count 'sample1'

scan 'sample1'



```
hbase(main):004:0> scan 'sample1'
ROW                       COLUMN+CELL
0 row(s) in 0.0110 seconds

hbase(main):005:0> count 'sample1'
6 row(s) in 0.0210 seconds

=> 6
hbase(main):006:0> scan 'sample1'
ROW                       COLUMN+CELL
 default238a7a13-c71a-4da column=cf:message, timestamp=1475752983269, value=text1\x0D
 b-8585-cdb79bf6140a
 default5302ef03-1738-457 column=cf:message, timestamp=1475752983269, value=text4\x0D
 1-a4c3-6eaabb62b610
 defaultb722294a-1e27-4c8 column=cf:message, timestamp=1475752988988, value=text5\x0D
 0-82fb-92b5c00bf4cc
 defaulte267f3db-c730-444 column=cf:message, timestamp=1475752983269, value=text2\x0D
 3-b974-788ae2f6eb39
 defaultf60409f3-4515-4d2 column=cf:message, timestamp=1475752983269, value=text3\x0D
 c-8138-ce6c34be847c
 incRow                   column=cf:inc, timestamp=1475752988996, value=\x00\x00\x00\x00\x00\x00\x0
                          0\x05
6 row(s) in 0.0380 seconds

hbase(main):007:0> 
```