

## 1.load the given textfile in HDFS.

```
[cloudera@quickstart ~]$ ls
mouni          eclipse          kp
rakeshdata
avro            emp.java        kpi_hadoop
rakeshdata1
cloudera-manager enterprise-deployment.json lib
sample.txt
cm_api.py       express-deployment.json Music
sparkjars_exec
data1           external_jars    parcels          table.csv
Desktop         external-unified parquet_write     Templates
devices.json    HiveDirectory    part_dir         Videos
Documents       input.txt        Pictures         workspace
Downloads       kerberos         Public
zeyo_tab.java
[cloudera@quickstart ~]$ hdfs dfs -ls
22/01/20 22:20:53 WARN ipc.Client: Failed to connect to server:
quickstart.cloudera/127.0.0.1:8020: try once and fail.
java.net.ConnectException: Connection refused
    at sun.nio.ch.SocketChannelImpl.checkConnect(Native Method)
    at
sun.nio.ch.SocketChannelImpl.finishConnect(SocketChannelImpl.java:714)
    at
org.apache.hadoop.net.SocketIOWithTimeout.connect(SocketIOWithTimeout.
java:206)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:530)
    at org.apache.hadoop.net.NetUtils.connect(NetUtils.java:494)
    at
org.apache.hadoop.ipc.Client$Connection.setupConnection(Client.java:64
8)
    at
org.apache.hadoop.ipc.Client$Connection.setupIOstreams(Client.java:744
)
    at
org.apache.hadoop.ipc.Client$Connection.access$3000(Client.java:396)
    at org.apache.hadoop.ipc.Client.getConnection(Client.java:1557)
    at org.apache.hadoop.ipc.Client.call(Client.java:1480)
    at org.apache.hadoop.ipc.Client.call(Client.java:1441)
    at
org.apache.hadoop.ipc.ProtobufRpcEngine$Invoker.invoke(ProtobufRpcEngi
ne.java:230)
    at com.sun.proxy.$Proxy10.getFileInfo(Unknown Source)
    at
org.apache.hadoop.hdfs.protocolPB.ClientNamenodeProtocolTranslatorPB.g
etFileInfo(ClientNamenodeProtocolTranslatorPB.java:786)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at
sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.j
ava:62)
    at
```

```

sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccess
orImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at
org.apache.hadoop.io.retry.RetryInvocationHandler.invokeMethod(RetryIn
vocationHandler.java:260)
    at
org.apache.hadoop.io.retry.RetryInvocationHandler.invoke(RetryInvocati
onHandler.java:104)
    at com.sun.proxy.$Proxy11.getFileInfo(Unknown Source)
    at
org.apache.hadoop.hdfs.DFSClient.getFileInfo(DFSClient.java:2131)
    at
org.apache.hadoop.hdfs.DistributedFileSystem$20.doCall(DistributedFile
System.java:1265)
    at
org.apache.hadoop.hdfs.DistributedFileSystem$20.doCall(DistributedFile
System.java:1261)
    at
org.apache.hadoop.fs.FileSystemLinkResolver.resolve(FileSystemLinkReso
lver.java:81)
    at
org.apache.hadoop.hdfs.DistributedFileSystem.getFileStatus(Distributed
FileSystem.java:1261)
    at org.apache.hadoop.fs.Globber.getFileStatus(Globber.java:64)
    at org.apache.hadoop.fs.Globber.doGlob(Globber.java:272)
    at org.apache.hadoop.fs.Globber.glob(Globber.java:151)
    at
org.apache.hadoop.fs.FileSystem.globStatus(FileSystem.java:1715)
    at
org.apache.hadoop.fs.shell.PathData.expandAsGlob(PathData.java:326)
    at
org.apache.hadoop.fs.shell.Command.expandArgument(Command.java:235)
    at
org.apache.hadoop.fs.shell.Command.expandArguments(Command.java:218)
    at
org.apache.hadoop.fs.shell.FsCommand.processRawArguments(FsCommand.jav
a:102)
    at org.apache.hadoop.fs.shell.Command.run(Command.java:165)
    at org.apache.hadoop.fs.FsShell.run(FsShell.java:315)
    at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:70)
    at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
    at org.apache.hadoop.fs.FsShell.main(FsShell.java:372)
ls: Call From quickstart.cloudera/127.0.0.1 to quickstart.cloudera:8020
failed on connection exception: java.net.ConnectException: Connection
refused; For more details see:
http://wiki.apache.org/hadoop/ConnectionRefused
[cloudera@quickstart ~]$ list
bash: list: command not found
[cloudera@quickstart ~]$ ls
mouni          cm_api.py      devices.json    eclipse
express-deployment.json HiveDirectory kp Music

```

```

part_dir  rakeshdata  sparkjars_exec  Videos
avro      data1      Documents  emp.java
external_jars  input.txt  kpi_hadoop  parcels
Pictures  rakeshdata1  table.csv  workspace
cloudera-manager  DesktopDownloads  enterprise-deployment.json
external-unified  kerberos  lib  parquet_write
Public  sample.txt  Templates  zeyo_tab.java
[cloudera@quickstart ~]$ dir
mouni      cm_api.py  devices.json  eclipse
express-deployment.json  HiveDirectory  kp  Music  part_dir
rakeshdata  sparkjars_exec  Videos
avro      data1      Documents  emp.java
external_jars  input.txt  kpi_hadoop  parcels  Pictures
rakeshdata1  table.csv  workspace
cloudera-manager  Desktop  Downloads  enterprise-deployment.json
external-unified  kerberos  lib  parquet_write  Public
sample.txt  Templates  zeyo_tab.java

```

## 2.Perform WordCount on the text file using mapreduce

```

[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera
Found 19 items
drwx----- - cloudera cloudera 0 2022-01-12 09:03
/user/cloudera/.staging
drwxr-xr-x - cloudera cloudera 0 2020-05-23 23:09
/user/cloudera/avro_json_write
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:15
/user/cloudera/csv_dir
drwxr-xr-x - cloudera cloudera 0 2022-01-12 09:03
/user/cloudera/emp
drwxr-xr-x - cloudera cloudera 0 2020-06-04 08:36
/user/cloudera/import_avro
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:56
/user/cloudera/json_avro_1
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:13
/user/cloudera/json_dir
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:39
/user/cloudera/json_orc
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:56
/user/cloudera/json_orc_1
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:38
/user/cloudera/json_parquet
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:56
/user/cloudera/json_parquet_1
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:11
/user/cloudera/orc_dir
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:14
/user/cloudera/parquet_dir
drwxr-xr-x - cloudera cloudera 0 2020-05-23 23:11
/user/cloudera/parquet_json_write

```

```

drwxr-xr-x   - cloudera cloudera          0 2020-05-22 13:40
/user/cloudera/part_dir
drwxr-xr-x   - cloudera cloudera          0 2020-05-22 14:01
/user/cloudera/part_dir2
-rw-r--r--    1 cloudera cloudera          81 2022-01-11 02:29
/user/cloudera/table.csv
-rw-r--r--    1 cloudera cloudera        1173 2022-01-20 22:48
/user/cloudera/words.txt
drwxr-xr-x   - cloudera cloudera          0 2020-06-04 09:04
/user/cloudera/zeyo_dir
[cloudera@quickstart ~]$ cat words.txt
cat: words.txt: No such file or directory
[cloudera@quickstart ~]$ cd user/cloudera
bash: cd: user/cloudera: No such file or directory
[cloudera@quickstart ~]$ cd user
bash: cd: user: No such file or directory
[cloudera@quickstart ~]$ hdfs dfs -cat /user/cloudera
cat: `/user/cloudera': Is a directory
[cloudera@quickstart ~]$ hdfs dfs -cat words.txt /user/cloudera
It's a truly pleasant experience to read this book, actually I should
confess that I laughed A LOT in the reading. The book is hilarious.

```

Besides the fun part, I was inspired by this book too. This book went through the early history of Personal Computer industry, gave the vivid silhouettes of the people, the companies and Silicon Valley in this industry. Mr.Cringely examined why today's Information Technology industry is what it is now, and how it became like this.

The book provided the facts and opinion about how the high tech companies succeeded, and how many more failed. Why Bill Gates is the richest person in the world, and how Steve Jobs and Steve Wozniak created the most beloved high tech company in the world.

It used to say that reading history can make people understand the rise and fall of things. We can learn the lessons from it, and get new ideas or patterns from the past success. Today Personal Computer is declining, and the focus is shifting to Smart Phone and Tablet. Although product is changing, the similar struggles, fights, winning and loss are still happening lively everyday in this industry, just like what it did in the old days.

```

cat: `/user/cloudera': Is a directory
[cloudera@quickstart ~]$ ^C
[cloudera@quickstart ~]$ hadoop jar
/usr/lib/hadoop-mapreduce/hadoop-map-reduce-examples.jar
Not a valid JAR:
/usr/lib/hadoop-mapreduce/hadoop-map-reduce-examples.jar
[cloudera@quickstart ~]$ hadoop jar
/usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount
/user/cloudera/sample_kpi.txt /user/cloudera/output
22/01/20 23:12:10 INFO client.RMPProxy: Connecting to ResourceManager at

```

```
quickstart.cloudera/127.0.0.1:8032
22/01/20 23:12:13 INFO mapreduce.JobSubmitter: Cleaning up the staging
area /user/cloudera/.staging/job_1642747289338_0001
22/01/20 23:12:13 WARN security.UserGroupInformation:
PrivilegedActionException as:cloudera (auth:SIMPLE)
cause:org.apache.hadoop.mapreduce.lib.input.InvalidInputException:
Input path does not exist:
hdfs://quickstart.cloudera:8020/user/cloudera/sample_kpi.txt
org.apache.hadoop.mapreduce.lib.input.InvalidInputException: Input path
does not exist:
hdfs://quickstart.cloudera:8020/user/cloudera/sample_kpi.txt
    at
org.apache.hadoop.mapreduce.lib.input.FileInputFormat.singleThreadedLi
stStatus(FileInputFormat.java:323)
    at
org.apache.hadoop.mapreduce.lib.input.FileInputFormat.listStatus(FileI
nputFormat.java:265)
    at
org.apache.hadoop.mapreduce.lib.input.FileInputFormat.getSplits(FileIn
putFormat.java:387)
    at
org.apache.hadoop.mapreduce.JobSubmitter.writeNewSplits(JobSubmitter.j
ava:305)
    at
org.apache.hadoop.mapreduce.JobSubmitter.writeSplits(JobSubmitter.java
:322)
    at
org.apache.hadoop.mapreduce.JobSubmitter.submitJobInternal(JobSubmitte
r.java:200)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1307)
        at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1304)
        at java.security.AccessController.doPrivileged(Native Method)
        at javax.security.auth.Subject.doAs(Subject.java:422)
        at
org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformat
ion.java:1917)
            at org.apache.hadoop.mapreduce.Job.submit(Job.java:1304)
            at
org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1325)
                at org.apache.hadoop.examples.WordCount.main(WordCount.java:87)
                at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
                at
sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.j
ava:62)
                at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccess
orImpl.java:43)
                    at java.lang.reflect.Method.invoke(Method.java:498)
                    at
org.apache.hadoop.util.ProgramDriver$ProgramDescription.invoke(Program
Driver.java:71)
                        at
```

```
org.apache.hadoop.util.ProgramDriver.run(ProgramDriver.java:144)
    at
org.apache.hadoop.examples.ExampleDriver.main(ExampleDriver.java:74)
    at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
    at
sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:62)
    at
sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccess
orImpl.java:43)
    at java.lang.reflect.Method.invoke(Method.java:498)
    at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
    at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
[cloudera@quickstart ~]$ hadoop jar
/usr/lib/hadoop-mapreduce/hadoop-mapreduce-examples.jar wordcount
/user/cloudera/words.txt /user/cloudera/output
22/01/20 23:12:45 INFO client.RMPProxy: Connecting to ResourceManager at
quickstart.cloudera/127.0.0.1:8032
22/01/20 23:12:49 INFO input.FileInputFormat: Total input paths to process
: 1
22/01/20 23:12:50 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1252)
    at java.lang.Thread.join(Thread.java:1326)
    at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFS
OutputStream.java:967)
    at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutput
Stream.java:705)
    at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStrea
m.java:894)
22/01/20 23:12:50 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1252)
    at java.lang.Thread.join(Thread.java:1326)
    at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFS
OutputStream.java:967)
    at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutput
Stream.java:705)
    at
org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStrea
m.java:894)
22/01/20 23:12:50 INFO mapreduce.JobSubmitter: number of splits:1
22/01/20 23:12:50 INFO mapreduce.JobSubmitter: Submitting tokens for job:
job_1642747289338_0002
22/01/20 23:12:52 INFO impl.YarnClientImpl: Submitted application
```

application\_1642747289338\_0002

22/01/20 23:12:53 INFO mapreduce.Job: The url to track the job:

[http://quickstart.cloudera:8088/proxy/application\\_1642747289338\\_0002/](http://quickstart.cloudera:8088/proxy/application_1642747289338_0002/)

22/01/20 23:12:53 INFO mapreduce.Job: Running job: job\_1642747289338\_0002

22/01/20 23:13:33 INFO mapreduce.Job: Job job\_1642747289338\_0002 running in uber mode : false

22/01/20 23:13:33 INFO mapreduce.Job: map 0% reduce 0%

22/01/20 23:14:05 INFO mapreduce.Job: map 100% reduce 0%

22/01/20 23:14:35 INFO mapreduce.Job: map 100% reduce 100%

22/01/20 23:14:37 INFO mapreduce.Job: Job job\_1642747289338\_0002

completed successfully

22/01/20 23:14:38 INFO mapreduce.Job: Counters: 49

#### File System Counters

FILE: Number of bytes read=1261

FILE: Number of bytes written=297411

FILE: Number of read operations=0

FILE: Number of large read operations=0

FILE: Number of write operations=0

HDFS: Number of bytes read=1293

HDFS: Number of bytes written=1143

HDFS: Number of read operations=6

HDFS: Number of large read operations=0

HDFS: Number of write operations=2

#### Job Counters

Launched map tasks=1

Launched reduce tasks=1

Data-local map tasks=1

Total time spent by all maps in occupied slots (ms)=14978560

Total time spent by all reduces in occupied slots (ms)=14747648

Total time spent by all map tasks (ms)=29255

Total time spent by all reduce tasks (ms)=28804

Total vcore-milliseconds taken by all map tasks=29255

Total vcore-milliseconds taken by all reduce tasks=28804

Total megabyte-milliseconds taken by all map tasks=14978560

Total megabyte-milliseconds taken by all reduce tasks=14747648

#### Map-Reduce Framework

Map input records=9

Map output records=203

Map output bytes=1980

Map output materialized bytes=1257

Input split bytes=120

Combine input records=203

Combine output records=134

Reduce input groups=134

Reduce shuffle bytes=1257

Reduce input records=134

Reduce output records=134

Spilled Records=268

Shuffled Maps =1

Failed Shuffles=0

Merged Map outputs=1

```

GC time elapsed (ms)=1352
CPU time spent (ms)=2520
Physical memory (bytes) snapshot=251547648
Virtual memory (bytes) snapshot=3890036736
Total committed heap usage (bytes)=101449728

Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0

File Input Format Counters
  Bytes Read=1173

File Output Format Counters
  Bytes Written=1143

[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hadoop dfs -ls /user/cloudera/output
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

```

```

Found 2 items
-rw-r--r--    1 cloudera cloudera          0 2022-01-20 23:14
/user/cloudera/output/_SUCCESS
-rw-r--r--    1 cloudera cloudera      1143 2022-01-20 23:14
/user/cloudera/output/part-r-00000
[cloudera@quickstart ~]$ hadoop dfs -cat
/user/cloudera/output/part-r-00000
DEPRECATED: Use of this script to execute hdfs command is deprecated.
Instead use the hdfs command for it.

```

```

A      1
Although  1
Besides  1
Bill  1
Computer  2
Gates  1
I      3
Information  1
It      1
It's    1
Jobs    1
LOT     1
Mr.Cringely  1
Personal  2
Phone  1
Silicon  1
Smart  1
Steve  2
Tablet.  1
Technology 1
The    2

```



This 1  
Today 1  
Valley 1  
We 1  
Why 1  
Wozniak 1  
a 1  
about 1  
actually 1  
and 11  
are 1  
became 1  
beloved 1  
book 4  
book, 1  
by 1  
can 2  
changing, 1  
companies 2  
company 1  
confess 1  
created 1  
days. 1  
declining, 1  
did 1  
early 1  
everyday 1  
examined 1  
experience 1  
facts 1  
failed. 1  
fall 1  
fights, 1  
focus 1  
from 2  
fun 1  
gave 1  
get 1  
happening 1  
high 2  
hilarious. 1  
history 2  
how 4  
ideas 1  
in 6  
industry 1  
industry, 2  
industry. 1  
inspired 1  
is 7  
it 3  
it, 1

just	1	
laughed	1	
learn	1	
lessons	1	
like	2	
lively	1	
loss	1	
make	1	
many	1	
more	1	
most	1	
new	1	
now,	1	
of	3	
old	1	
opinion	1	
or	1	
part,	1	
past	1	
patterns	1	
people	1	
people,	1	
person	1	
pleasant	1	
product	1	
provided	1	
read	1	
reading	1	
reading.	1	
richest	1	
rise	1	
say	1	
shifting	1	
should	1	
silhouettes		1
similar	1	
still	1	
struggles,	1	
succeeded,	1	
success.	1	
tech	2	
that	2	
the	18	
things.	1	
this	4	
this.	1	
through	1	
to	3	
today's	1	
too.	1	
truly	1	
understand	1	

```

used 1
vivid 1
was 1
went 1
what 2
why 1
winning 1
world, 1
world. 1
[cloudera@quickstart ~]$

```

### 3. Create a HBase table 'Census' using java with Column Family as 'Personal', 'Professional'.

```

[cloudera@quickstart ~]$ hbase shell
OpenJDK 64-Bit Server VM warning: Using incremental CMS is deprecated and
will likely be removed in a future release
OpenJDK 64-Bit Server VM warning: If the number of processors is expected
to increase from one, then you should configure the number of parallel GC
threads appropriately using -XX:ParallelGCThreads=N
22/01/21 01:29:53 INFO Configuration.deprecation: hadoop.native.lib 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-cdh5.13.0, rUnknown, Wed Oct 4 11:16:18 PDT 2017

hbase(main):001:0> create table 'census', 'personal', 'professional'
NoMethodError: undefined method `table' for #<Object:0x5809fa26>

hbase(main):002:0> create 'census', 'personal', 'professional'
0 row(s) in 3.3680 seconds

=> Hbase::Table - census
hbase(main):003:0> describe census
NameError: undefined local variable or method `census' for
#<Object:0x5809fa26>

hbase(main):004:0> describe 'census'
Table census is ENABLED
census
COLUMN FAMILIES DESCRIPTION
{NAME => 'personal', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY =>
'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL
=> 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKC
ACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'professional', BLOOMFILTER => 'ROW', VERSIONS => '1', 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'}
2 row(s) in 0.5510 seconds
```

#### 4.Put 2 rows in the Census table each having columns name and gender in personal and occupation in professional and display data using HBase shell.

```
hbase(main):006:0> scan 'census'
ROW
0 row(s) in 0.1280 seconds
```

COLUMN+CELL

```
hbase(main):007:0> put 'census', '1', 'personal:name,gender',
'mouni,female'
0 row(s) in 0.1800 seconds
```

```
hbase(main):008:0> put 'census', '1', 'professional:occupation', 'design'
0 row(s) in 0.0260 seconds
```

```
hbase(main):009:0> scan 'census'
ROW
```

COLUMN+CELL

```
1
column=personal:name,gender, timestamp=1642758230179,
value=mouni,female
1
column=professional:occupation, timestamp=1642758281887, value=design
1 row(s) in 0.0360 seconds
```

```
hbase(main):010:0> put 'census', '2', 'personal:name,gender',
'yugesh,male'
0 row(s) in 0.0150 seconds
```

```
hbase(main):011:0> put 'census', '1', 'professional:occupation',
'cricket'
0 row(s) in 0.0060 seconds
```

```
hbase(main):012:0> scan 'census'
ROW
```

COLUMN+CELL

```
1
column=personal:name,gender, timestamp=1642758230179,
value=mouni,female
1
column=professional:occupation, timestamp=1642758449677, value=cricket
2
column=personal:name,gender, timestamp=1642758439344, value=yugesh,male
2 row(s) in 0.0290 seconds
```

```
hbase(main):013:0> truncate 'census'
```

```

Truncating 'census' table (it may take a while):
- Disabling table...
- Truncating table...
0 row(s) in 4.0800 seconds

hbase(main):014:0> put 'census', '1', 'personal:name,gender',
'mouni,female'
0 row(s) in 0.1700 seconds

hbase(main):015:0> put 'census', '1', 'professional:occupation', 'design'
0 row(s) in 0.0150 seconds

hbase(main):016:0> put 'census', '2', 'personal:name,gender',
'yugesh,male'
0 row(s) in 0.0780 seconds

hbase(main):017:0> put 'census', '2', 'professional:occupation',
'cricket'
0 row(s) in 0.0380 seconds

hbase(main):018:0> scan 'census'
ROW                                                    COLUMN+CELL
1
column=personal:name,gender, timestamp=1642758509217,
value=mouni,female
1
column=professional:occupation, timestamp=1642758519896, value=design
2
column=personal:name,gender, timestamp=1642758533100, value=yugesh,male
2
column=professional:occupation, timestamp=1642758543828, value=cricket
2 row(s) in 0.0950 seconds

```

## 5. Load the groceries data file using Hdfs, Hbase, Sqoop with a schema and describe and display the data.

```

[cloudera@quickstart ~]$ hadoop fs -put groceries.csv /user/cloudera
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 21 items
drwx-----   - cloudera cloudera          0 2022-01-20 23:14 .staging
drwxr-xr-x   - cloudera cloudera          0 2020-05-23 23:09
avro_json_write
drwxr-xr-x   - cloudera cloudera          0 2020-05-22 22:15 csv_dir
drwxr-xr-x   - cloudera cloudera          0 2022-01-12 09:03 emp
-rw-r--r--   1 cloudera cloudera        456 2022-01-21 02:33
groceries.csv
drwxr-xr-x   - cloudera cloudera          0 2020-06-04 08:36 import_avro
drwxr-xr-x   - cloudera cloudera          0 2020-05-23 22:56 json_avro_1

```

```

drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:13 json_dir
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:39 json_orc
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:56 json_orc_1
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:38 json_parquet
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:56
json_parquet_1
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:11 orc_dir
drwxr-xr-x - cloudera cloudera 0 2022-01-20 23:14 output
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:14 parquet_dir
drwxr-xr-x - cloudera cloudera 0 2020-05-23 23:11
parquet_json_write
drwxr-xr-x - cloudera cloudera 0 2020-05-22 13:40 part_dir
drwxr-xr-x - cloudera cloudera 0 2020-05-22 14:01 part_dir2
-rw-r--r-- 1 cloudera cloudera 81 2022-01-11 02:29 table.csv
-rw-r--r-- 1 cloudera cloudera 1173 2022-01-20 22:48 words.txt
drwxr-xr-x - cloudera cloudera 0 2020-06-04 09:04 zeyo_dir
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera
Found 21 items
drwx----- - cloudera cloudera 0 2022-01-20 23:14
/user/cloudera/.staging
drwxr-xr-x - cloudera cloudera 0 2020-05-23 23:09
/user/cloudera/avro_json_write
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:15
/user/cloudera/csv_dir
drwxr-xr-x - cloudera cloudera 0 2022-01-12 09:03
/user/cloudera/emp
-rw-r--r-- 1 cloudera cloudera 456 2022-01-21 02:33
/user/cloudera/groceries.csv
drwxr-xr-x - cloudera cloudera 0 2020-06-04 08:36
/user/cloudera/import_avro
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:56
/user/cloudera/json_avro_1
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:13
/user/cloudera/json_dir
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:39
/user/cloudera/json_orc
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:56
/user/cloudera/json_orc_1
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:38
/user/cloudera/json_parquet
drwxr-xr-x - cloudera cloudera 0 2020-05-23 22:56
/user/cloudera/json_parquet_1
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:11
/user/cloudera/orc_dir
drwxr-xr-x - cloudera cloudera 0 2022-01-20 23:14
/user/cloudera/output
drwxr-xr-x - cloudera cloudera 0 2020-05-22 22:14
/user/cloudera/parquet_dir
drwxr-xr-x - cloudera cloudera 0 2020-05-23 23:11
/user/cloudera/parquet_json_write
drwxr-xr-x - cloudera cloudera 0 2020-05-22 13:40
/user/cloudera/part_dir

```

```

drwxr-xr-x - cloudera cloudera 0 2020-05-22 14:01
/user/cloudera/part_dir2
-rw-r--r-- 1 cloudera cloudera 81 2022-01-11 02:29
/user/cloudera/table.csv
-rw-r--r-- 1 cloudera cloudera 1173 2022-01-20 22:48
/user/cloudera/words.txt
drwxr-xr-x - cloudera cloudera 0 2020-06-04 09:04
/user/cloudera/zeyo_dir
[cloudera@quickstart ~]$ hbase shell

```

## HBASE

```

cloudera@quickstart ~]$ hbase shell
OpenJDK 64-Bit Server VM warning: Using incremental CMS is deprecated and
will likely be removed in a future release
OpenJDK 64-Bit Server VM warning: If the number of processors is expected
to increase from one, then you should configure the number of parallel GC
threads appropriately using -XX:ParallelGCThreads=N
22/01/21 01:29:53 INFO Configuration.deprecation: hadoop.native.lib 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-cdh5.13.0, rUnknown, Wed Oct 4 11:16:18 PDT 2017

hbase(main):001:0> create table 'census', 'personal', 'professional'
NoMethodError: undefined method `table' for #<Object:0x5809fa26>

hbase(main):002:0> create 'census', 'personal', 'professional'
0 row(s) in 3.3680 seconds

=> Hbase::Table - census
hbase(main):003:0> describe census
NameError: undefined local variable or method `census' for
#<Object:0x5809fa26>

hbase(main):004:0> describe 'census'
Table census is ENABLED
census
COLUMN FAMILIES DESCRIPTION
{NAME => 'personal', BLOOMFILTER => 'ROW', VERSIONS => '1', IN_MEMORY =>
'false', KEEP_DELETED_CELLS => 'FALSE', DATA_BLOCK_ENCODING => 'NONE', TTL
=> 'FOREVER', COMPRESSION => 'NONE', MIN_VERSIONS => '0', BLOCKC
ACHE => 'true', BLOCKSIZE => '65536', REPLICATION_SCOPE => '0'}
{NAME => 'professional', BLOOMFILTER => 'ROW', VERSIONS => '1', 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'}
2 row(s) in 0.5510 seconds
```

```
hbase(main):005:0> put 'census', '1', 'personal:name,gender',
'mrspy,male', 'professional:occupation', 'spy'
```

```
ERROR: no method 'add' for arguments
(org.jruby.java.proxies.ArrayJavaProxy,org.jruby.java.proxies.ArrayJavaProxy,org.jruby.RubyString,org.jruby.java.proxies.ArrayJavaProxy) on
Java::OrgApacheHadoopHbaseClient::Put
  available overloads:
    (byte[],byte[],long,byte[])
    (byte[],java.nio.ByteBuffer,long,java.nio.ByteBuffer)
```

Put a cell 'value' at specified table/row/column and optionally timestamp coordinates. To put a cell value into table 'ns1:t1' or 't1' at row 'r1' under column 'c1' marked with the time 'ts1', do:

```
hbase> put 'ns1:t1', 'r1', 'c1', 'value'
hbase> put 't1', 'r1', 'c1', 'value'
hbase> put 't1', 'r1', 'c1', 'value', ts1
hbase> put 't1', 'r1', 'c1', 'value',
{ATTRIBUTES=>{'mykey'=>'myvalue'}}
hbase> put 't1', 'r1', 'c1', 'value', ts1,
{ATTRIBUTES=>{'mykey'=>'myvalue'}}
hbase> put 't1', 'r1', 'c1', 'value', ts1,
{VISIBILITY=>'PRIVATE|SECRET'}
```

The same commands also can be run on a table reference. Suppose you had a reference t to table 't1', the corresponding command would be:

```
hbase> t.put 'r1', 'c1', 'value', ts1,
{ATTRIBUTES=>{'mykey'=>'myvalue'}}}
```

```
hbase(main):006:0> scan 'census'
ROW
0 row(s) in 0.1280 seconds
```

```
hbase(main):007:0> put 'census', '1', 'personal:name,gender',
'mouni,female'
0 row(s) in 0.1800 seconds
```

```
hbase(main):008:0> put 'census', '1', 'professional:occupation',
'design'
0 row(s) in 0.0260 seconds
```

```
hbase(main):009:0> scan 'census'
ROW
1
column=personal:name,gender, timestamp=1642758230179,
value=mouni,female
```



```

1
column=professional:occupation, timestamp=1642758281887, value=design
1 row(s) in 0.0360 seconds

hbase(main):010:0> put 'census', '2', 'personal:name,gender', 'abhi,male'
0 row(s) in 0.0150 seconds

hbase(main):011:0> put 'census', '1', 'professional:occupation',
'cricket'
0 row(s) in 0.0060 seconds

hbase(main):012:0> scan 'census'
ROW                                COLUMN+CELL
1
column=personal:name,gender, timestamp=1642758230179,
value=mouni,female 1
column=professional:occupation, timestamp=1642758449677, value=cricket
2
column=personal:name,gender, timestamp=1642758439344, value=abhi,male
2 row(s) in 0.0290 seconds

hbase(main):013:0> truncate 'census'
Truncating 'census' table (it may take a while):
- Disabling table...
- Truncating table...
0 row(s) in 4.0800 seconds

hbase(main):014:0> put 'census', '1', 'personal:name,gender',
'mouni,female'
0 row(s) in 0.1700 seconds

hbase(main):015:0> put 'census', '1', 'professional:occupation',
'design'
0 row(s) in 0.0150 seconds

hbase(main):016:0> put 'census', '2', 'personal:name,gender', 'abhi,male'
0 row(s) in 0.0780 seconds

hbase(main):017:0> put 'census', '2', 'professional:occupation',
'cricket'
0 row(s) in 0.0380 seconds

hbase(main):018:0> scan 'census'
ROW                                COLUMN+CELL
1
column=personal:name,gender, timestamp=1642758509217,
value=mouni,female 1
column=professional:occupation, timestamp=1642758519896,
value=design 2
column=personal:name,gender, timestamp=1642758533100, value=abhi,male
2
column=professional:occupation, timestamp=1642758543828, value=cricket
2 row(s) in 0.0950 seconds

```

```

hbase(main):019:0> [cloudera@quickstart ~]$ ^C
[cloudera@quickstart ~]$ hadoop fs -put groceries.csv /user/cloudera
[cloudera@quickstart ~]$ hdfs dfs -ls
Found 21 items
drwx-----   - cloudera cloudera          0 2022-01-20 23:14 .staging
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 23:09
avro_json_write
drwxr-xr-x    - cloudera cloudera          0 2020-05-22 22:15 csv_dir
drwxr-xr-x    - cloudera cloudera          0 2022-01-12 09:03 emp
-rw-r--r--    1 cloudera cloudera        456 2022-01-21 02:33
groceries.csv
drwxr-xr-x    - cloudera cloudera          0 2020-06-04 08:36 import_avro
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 22:56 json_avro_1
drwxr-xr-x    - cloudera cloudera          0 2020-05-22 22:13 json_dir
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 22:39 json_orc
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 22:56 json_orc_1
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 22:38 json_parquet
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 22:56
json_parquet_1
drwxr-xr-x    - cloudera cloudera          0 2020-05-22 22:11 orc_dir
drwxr-xr-x    - cloudera cloudera          0 2022-01-20 23:14 output
drwxr-xr-x    - cloudera cloudera          0 2020-05-22 22:14 parquet_dir
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 23:11
parquet_json_write
drwxr-xr-x    - cloudera cloudera          0 2020-05-22 13:40 part_dir
drwxr-xr-x    - cloudera cloudera          0 2020-05-22 14:01 part_dir2
-rw-r--r--    1 cloudera cloudera          81 2022-01-11 02:29 table.csv
-rw-r--r--    1 cloudera cloudera       1173 2022-01-20 22:48 words.txt
drwxr-xr-x    - cloudera cloudera          0 2020-06-04 09:04 zeyo_dir
[cloudera@quickstart ~]$ hdfs dfs -ls /user/cloudera
Found 21 items
drwx-----   - cloudera cloudera          0 2022-01-20 23:14
/user/cloudera/.staging
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 23:09
/user/cloudera/avro_json_write
drwxr-xr-x    - cloudera cloudera          0 2020-05-22 22:15
/user/cloudera/csv_dir
drwxr-xr-x    - cloudera cloudera          0 2022-01-12 09:03
/user/cloudera/emp
-rw-r--r--    1 cloudera cloudera        456 2022-01-21 02:33
/user/cloudera/groceries.csv
drwxr-xr-x    - cloudera cloudera          0 2020-06-04 08:36
/user/cloudera/import_avro
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 22:56
/user/cloudera/json_avro_1
drwxr-xr-x    - cloudera cloudera          0 2020-05-22 22:13
/user/cloudera/json_dir
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 22:39
/user/cloudera/json_orc
drwxr-xr-x    - cloudera cloudera          0 2020-05-23 22:56
/user/cloudera/json_orc_1

```

```

drwxr-xr-x   - cloudera cloudera          0 2020-05-23 22:38
/user/cloudera/json_parquet
drwxr-xr-x   - cloudera cloudera          0 2020-05-23 22:56
/user/cloudera/json_parquet_1
drwxr-xr-x   - cloudera cloudera          0 2020-05-22 22:11
/user/cloudera/orc_dir
drwxr-xr-x   - cloudera cloudera          0 2022-01-20 23:14
/user/cloudera/output
drwxr-xr-x   - cloudera cloudera          0 2020-05-22 22:14
/user/cloudera/parquet_dir
drwxr-xr-x   - cloudera cloudera          0 2020-05-23 23:11
/user/cloudera/parquet_json_write
drwxr-xr-x   - cloudera cloudera          0 2020-05-22 13:40
/user/cloudera/part_dir
drwxr-xr-x   - cloudera cloudera          0 2020-05-22 14:01
/user/cloudera/part_dir2
-rw-r--r--    1 cloudera cloudera          81 2022-01-11 02:29
/user/cloudera/table.csv
-rw-r--r--    1 cloudera cloudera        1173 2022-01-20 22:48
/user/cloudera/words.txt
drwxr-xr-x   - cloudera cloudera          0 2020-06-04 09:04
/user/cloudera/zeyo_dir
[cloudera@quickstart ~]$ hbase shell
OpenJDK 64-Bit Server VM warning: Using incremental CMS is deprecated and
will likely be removed in a future release
OpenJDK 64-Bit Server VM warning: If the number of processors is expected
to increase from one, then you should configure the number of parallel GC
threads appropriately using -XX:ParallelGCThreads=N
22/01/21 02:37:15 INFO Configuration.deprecation: hadoop.native.lib 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-cdh5.13.0, rUnknown, Wed Oct 4 11:16:18 PDT 2017

hbase(main):001:0> [cloudera@quickstart ~]$ ^C
[cloudera@quickstart ~]$ hbase shell
OpenJDK 64-Bit Server VM warning: Using incremental CMS is deprecated and
will likely be removed in a future release
OpenJDK 64-Bit Server VM warning: If the number of processors is expected
to increase from one, then you should configure the number of parallel GC
threads appropriately using -XX:ParallelGCThreads=N
22/01/21 02:46:42 INFO Configuration.deprecation: hadoop.native.lib 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-cdh5.13.0, rUnknown, Wed Oct 4 11:16:18 PDT 2017

hbase(main):001:0> create 'groceries','info'
0 row(s) in 2.5750 seconds

=> Hbase::Table - groceries
hbase(main):002:0> put 'groceries','info:itmno,city,item,date,quantity'
```

ERROR: wrong number of arguments (2 for 4)

Put a cell 'value' at specified table/row/column and optionally timestamp coordinates. To put a cell value into table 'ns1:t1' or 't1' at row 'r1' under column 'c1' marked with the time 'ts1', do:

```
hbase> put 'ns1:t1', 'r1', 'c1', 'value'
hbase> put 't1', 'r1', 'c1', 'value'
hbase> put 't1', 'r1', 'c1', 'value', ts1
hbase> put 't1', 'r1', 'c1', 'value',
{ATTRIBUTES=>{'mykey'=>'myvalue'}}
hbase> put 't1', 'r1', 'c1', 'value', ts1,
{ATTRIBUTES=>{'mykey'=>'myvalue'}}
hbase> put 't1', 'r1', 'c1', 'value', ts1,
{VISIBILITY=>'PRIVATE|SECRET'}
```

The same commands also can be run on a table reference. Suppose you had a reference

t to table 't1', the corresponding command would be:

```
hbase> t.put 'r1', 'c1', 'value', ts1,
{ATTRIBUTES=>{'mykey'=>'myvalue'}}
```

```
hbase(main):003:0> put
'groceries','info:itmno,city,item,date,quantity',' , , ,'
```

ERROR: wrong number of arguments (3 for 4)

Put a cell 'value' at specified table/row/column and optionally timestamp coordinates. To put a cell value into table 'ns1:t1' or 't1' at row 'r1' under column 'c1' marked with the time 'ts1', do:

```
hbase> put 'ns1:t1', 'r1', 'c1', 'value'
hbase> put 't1', 'r1', 'c1', 'value'
hbase> put 't1', 'r1', 'c1', 'value', ts1
hbase> put 't1', 'r1', 'c1', 'value',
{ATTRIBUTES=>{'mykey'=>'myvalue'}}
hbase> put 't1', 'r1', 'c1', 'value', ts1,
{ATTRIBUTES=>{'mykey'=>'myvalue'}}
hbase> put 't1', 'r1', 'c1', 'value', ts1,
{VISIBILITY=>'PRIVATE|SECRET'}
```

The same commands also can be run on a table reference. Suppose you had a reference

t to table 't1', the corresponding command would be:

```
hbase> t.put 'r1', 'c1', 'value', ts1,
{ATTRIBUTES=>{'mykey'=>'myvalue'}}
```

COLUMN+CELL

COLUMN+CELL  
column=info:city,  
column=info:date,  
column=info:item,  
column=info:itemno,  
column=info:city,  
column=info:date,  
column=info:item,  
column=info:itemno,  
column=info:city,  
column=info:date,  
column=info:item,  
column=info:itemno,  
column=info:city,  
column=info:date,  
column=info:item,  
column=info:itemno,  
column=info:city,  
column=info:date,  
column=info:item,  
column=info:itemno,

o14	timestamp=1642764326730, value=Issaquah	column=info:itemno,
o2	timestamp=1642764326730, value=Apples	column=info:city,
o2	timestamp=1642764326730, value=20	column=info:date,
o2	timestamp=1642764326730, value=02-01-2017	column=info:item,
o2	timestamp=1642764326730, value=Kent	column=info:itemno,
o3	timestamp=1642764326730, value=Flowers	column=info:city,
o3	timestamp=1642764326730, value=10	column=info:date,
o3	timestamp=1642764326730, value=02-01-2017	column=info:item,
o3	timestamp=1642764326730, value=Bellevue	column=info:itemno,
o4	timestamp=1642764326730, value=Meat	column=info:city,
o4	timestamp=1642764326730, value=40	column=info:date,
o4	timestamp=1642764326730, value=03-01-2017	column=info:item,
o4	timestamp=1642764326730, value=Redmond	column=info:itemno,
o5	timestamp=1642764326730, value=Potatoes	column=info:city,
o5	timestamp=1642764326730, value=9	column=info:date,
o5	timestamp=1642764326730, value=04-01-2017	column=info:item,
o5	timestamp=1642764326730, value=Seattle	column=info:itemno,
o6	timestamp=1642764326730, value=Bread	column=info:city,
o6	timestamp=1642764326730, value=5	column=info:date,
o6	timestamp=1642764326730, value=04-01-2017	column=info:item,
o6	timestamp=1642764326730, value=Bellevue	column=info:itemno,
o7	timestamp=1642764326730, value=Bread	column=info:city,
o7	timestamp=1642764326730, value=5	column=info:date,
o7	timestamp=1642764326730, value=05-01-2017	column=info:item,
o7	timestamp=1642764326730, value=Redmond	column=info:itemno,
o8	timestamp=1642764326730, value=Onion	column=info:city,

```
o8
timestamp=1642764326730, value=4
o8
timestamp=1642764326730, value=05-01-2017
o8
timestamp=1642764326730, value=Issaquah
o9
timestamp=1642764326730, value=Cheese
o9
timestamp=1642764326730, value=15
o9
timestamp=1642764326730, value=05-01-2017
o9
timestamp=1642764326730, value=Redmond
14 row(s) in 1.4130 seconds
```

```
column=info:date,
column=info:item,
column=info:itemno,
column=info:city,
column=info:date,
column=info:item,
column=info:itemno,
```