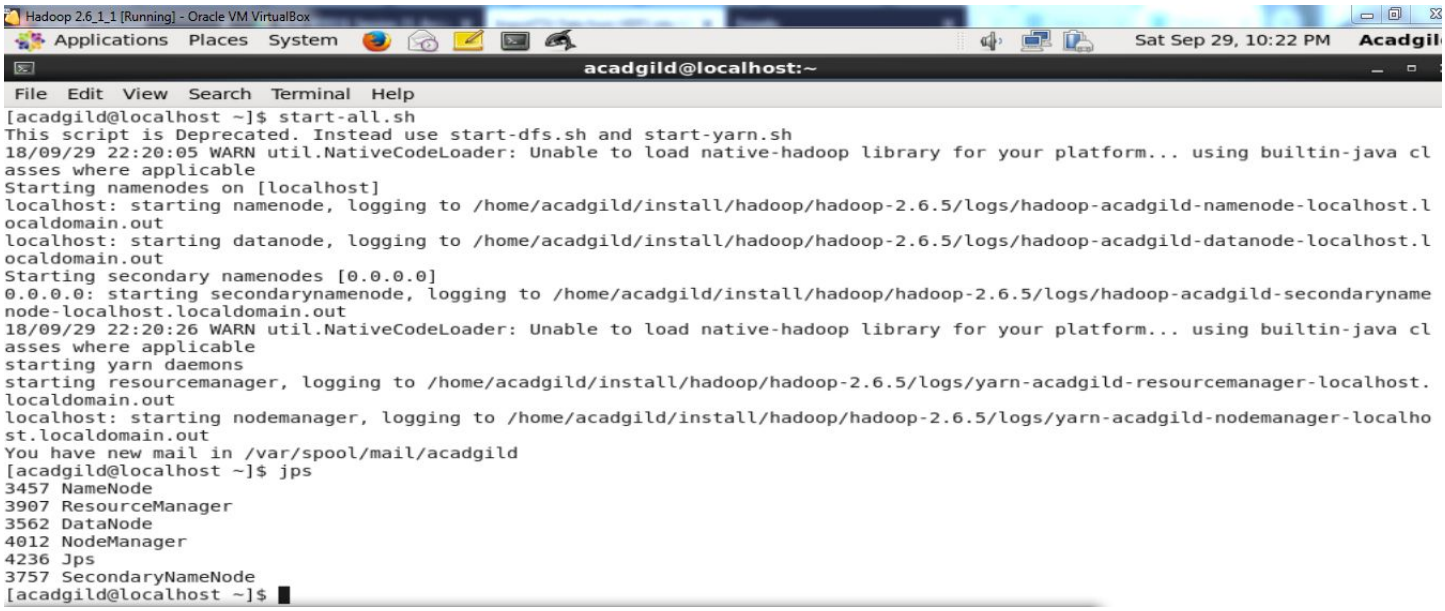
Import TSV Data from HDFS into HBase

ImportTSV is a utility that loads data in the TSV format into HBase. ImportTSV takes data from HDFS into HBase via puts. Find below the syntax used to load data via puts (i.e., non-bulk loading):

$bin/hbase org.apache.hadoop.hbase.mapreduce.ImportTsv –Dimporttsv.columns=a,b,c <tablename> <hdfs-inputdir>

Before starting practice on TSV import, it is compulsory to start all the Hadoop and HBase daemons.

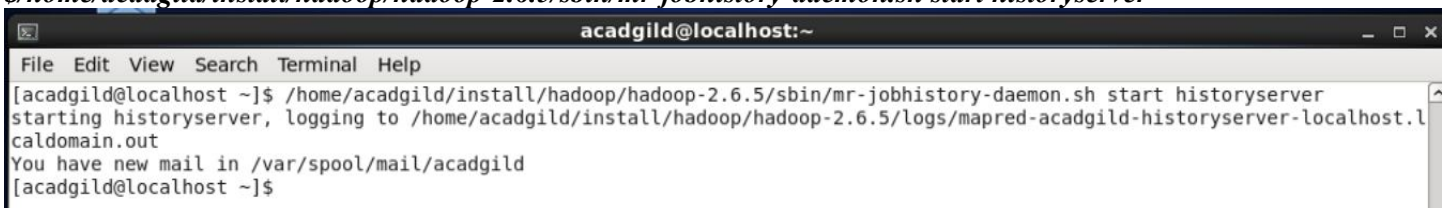
1. Start hadoop in VM using the following:
$start-all.sh
2. Now check all the daemons are started or not:
$jps



3. If HBase is not running, start the Hbase as follows:

4. Now HBase is started, we can observe this by giving ***jps*** command which produces HMaster.

5. Now start the job history server as follows:
$/home/acadgild/install/hadoop/hadoop-2.6.5/sbin/mr-jobhistory-daemon.sh start historyserver



Step 1:

1. Inside HBase shell give the following command to create table along with 2 column family:
hbase(main):001:0> **create 'bulktable', 'cf1', 'cf2'**

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
[acadgild@localhost ~]$ hbase shell
2018-09-29 22:35:15,694 WARN  [main] util.NativeCodeLoader: Unable to load nativ
builtin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/s
cLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/sha
!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
HBase Shell; enter 'help<RETURN>' for list of supported commands.
Type "exit<RETURN>" to leave the HBase Shell
Version 1.2.6, rUnknown, Mon May 29 02:25:32 CDT 2017

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

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Help
hbase(main):001:0> create 'bulktable', 'cf1', 'cf2'
0 row(s) in 1.7130 seconds

=> Hbase::Table - bulktable
hbase(main):002:0> list
TABLE
bulktable
clicks
employee
htest
4 row(s) in 0.0400 seconds

=> ["bulktable", "clicks", "employee", "htest"]
hbase(main):003:0>
```

2. We can see that the table ***bulktable*** is created.

Step 2:

1. Take a new terminal and also make a directory called ***hbase*** in the local drive; so, since you have your own path you can use it.
   $mkdir hbase

2. Now change current directory into the ***hbase*** directory as follows:
   $cd hbase

```
acadgild@localhost:~/hbase
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~   acadgild@localhost:~/hbase
[acadgild@localhost ~]$ mkdir -p hbase
[acadgild@localhost ~]$ cd hbase/
[acadgild@localhost hbase]$
```

3. Now open a file ***bulk_daa.tsv*** the directory ***hbase*** as follows:
   $vi bulk_data.tsv
4. Now add data into that file as foolws:

```
acadgild@localhost:~/hbase
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~   acadgild@localhost:~/hbase
1       Amit    4
2       Girija  3
3       Jatin   5
4       Swati   3
~
~
~
~
~
~
~
~
:wq!
```

5. Once create the file save the file with ***esc + :wq!***+enter.

Step 4:

1. Now create a directory called ***hbase*** inside the HDFS. And copy the file ***hdfs/bulk_data.tsv*** into the directory ***hbase*** which is created inside the HDFS as follows:

$hadoop fs –mkdir /hbase

```
acadgild@localhost:~/hbase
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~   acadgild@localhost:~/hbase
[acadgild@localhost hbase]$ hadoop fs -mkdir /hbase
18/09/29 22:55:26 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
```

```
acadgild@localhost:~/hbase
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~   acadgild@localhost:~/hbase
[acadgild@localhost hbase]$ hadoop fs -put bulk_data.tsv /hbase/
18/09/29 22:56:31 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
[acadgild@localhost hbase]$ hadoop fs -cat /hbase/bulk_data.tsv
18/09/29 22:56:56 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
1       Amit    4
2       Girija  3
3       Jatin   5
4       Swati   3
[acadgild@localhost hbase]$
```

2. Now use ***–cat*** option for printing the data as follows:
   $hadoop fs –cat /hbase/bulk_data.tsv

3. We can see that the data present in side the ***bulk_data.tsv.***

Step 5:

1. Now in terminal we give the following command along with arguments as <table name> and <path of the bulk_data.tsv file in HDFS> as follows:

$hbase org.apache.hadoop.hbase.mapreduce.ImportTsv –Dimporttsv.columns=HBASE_ROW_KEY,cf1:name,cf2:exp bulktable /hbase/bulk_data.tsv

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~   acadgild@localhost:~/hbase   acadgild@localhost:~
[acadgild@localhost ~]$ hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.columns=HBASE_ROW_KEY,cf1:name,cf2:exp
bulktable /hbase/bulk_data.tsv
2018-09-29 23:07:17,995 WARN  [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using b
uiltin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/Static
LoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!
/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-09-29 23:07:18,720 INFO  [main] zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x6025e1b6 connecting to
ZooKeeper ensemble=localhost:2181
2018-09-29 23:07:18,733 INFO  [main] zookeeper.ZooKeeper: Client environment:zookeeper.version=3.4.6-1569965, built on 02/20/
2014 09:09 GMT
2018-09-29 23:07:18,733 INFO  [main] zookeeper.ZooKeeper: Client environment:host.name=localhost
2018-09-29 23:07:18,733 INFO  [main] zookeeper.ZooKeeper: Client environment:java.version=1.8.0_151
2018-09-29 23:07:18,733 INFO  [main] zookeeper.ZooKeeper: Client environment:java.vendor=Oracle Corporation
2018-09-29 23:07:18,733 INFO  [main] zookeeper.ZooKeeper: Client environment:java.home=/usr/java/jdk1.8.0_151/jre
2018-09-29 23:07:18,733 INFO  [main] zookeeper.ZooKeeper: Client environment:java.class.path=/home/acadgild/install/hbase/hba
se-1.2.6/conf:/usr/java/jdk1.8.0_151/lib/tools.jar:/home/acadgild/install/hbase/hbase-1.2.6:/home/acadgild/install/hbase/hbas
e-1.2.6/lib/activation-1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/aopalliance-1.0.jar:/home/acadgild/install/hbase/
hbase-1.2.6/lib/apacheds-i18n-2.0.0-M15.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/apacheds-kerberos-codec-2.0.0-M15.ja
r:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-asn1-api-1.0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-ut
il-1.0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/asm-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/avro-1
.7.4.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-beanutils-1.7.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/li
b/commons-beanutils-core-1.8.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-cli-1.2.jar:/home/acadgild/install/hb
ase/hbase-1.2.6/lib/commons-codec-1.9.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-collections-3.2.2.jar:/home/ac
adgild/install/hbase/hbase-1.2.6/lib/commons-compress-1.4.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-configur
ation-1.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-daemon-1.0.13.jar:/home/acadgild/install/hbase/hbase-1.2.6
/lib/commons-digester-1.8.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-el-1.0.jar:/home/acadgild/install/hbase/hb
ase-1.2.6/lib/commons-httpclient-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-io-2.4.jar:/home/acadgild/insta
ll/hbase/hbase-1.2.6/lib/commons-lang-2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-logging-1.2.jar:/home/acad
gild/install/hbase/hbase-1.2.6/lib/commons-math-2.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-math3-3.1.1.jar:
/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-net-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/disruptor-3.3.0
.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/findbugs-annotations-1.3.9-1.jar:/home/acadgild/install/hbase/hbase-1.2.6/l
ib/guava-12.0.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guice-3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/g
```

```
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~   acadgild@localhost:~/hbase   acadgild@localhost:~
2018-09-29 23:07:18,737 INFO  [main] zookeeper.ZooKeeper: Client environment:java.library.path=/usr/java/packages/lib/amd64:/
usr/lib64:/lib64:/lib:/usr/lib
2018-09-29 23:07:18,738 INFO  [main] zookeeper.ZooKeeper: Client environment:java.io.tmpdir=/tmp
2018-09-29 23:07:18,738 INFO  [main] zookeeper.ZooKeeper: Client environment:java.compiler=<NA>
2018-09-29 23:07:18,738 INFO  [main] zookeeper.ZooKeeper: Client environment:os.name=Linux
2018-09-29 23:07:18,738 INFO  [main] zookeeper.ZooKeeper: Client environment:os.arch=amd64
2018-09-29 23:07:18,738 INFO  [main] zookeeper.ZooKeeper: Client environment:os.version=2.6.32-696.18.7.el6.x86_64
2018-09-29 23:07:18,738 INFO  [main] zookeeper.ZooKeeper: Client environment:user.name=acadgild
2018-09-29 23:07:18,738 INFO  [main] zookeeper.ZooKeeper: Client environment:user.home=/home/acadgild
2018-09-29 23:07:18,738 INFO  [main] zookeeper.ZooKeeper: Client environment:user.dir=/home/acadgild
2018-09-29 23:07:18,740 INFO  [main] zookeeper.ZooKeeper: Initiating client connection, connectString=localhost:2181 sessionT
imeout=90000 watcher=hconnection-0x6025e1b60x0, quorum=localhost:2181, baseZNode=/hbase
2018-09-29 23:07:18,792 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Opening socket connection to server loc
alhost/127.0.0.1:2181. Will not attempt to authenticate using SASL (unknown error)
2018-09-29 23:07:18,827 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Socket connection established to localh
ost/127.0.0.1:2181, initiating session
2018-09-29 23:07:18,854 INFO  [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Session establishment complete on serve
r localhost/127.0.0.1:2181, sessionid = 0x166264034530007, negotiated timeout = 90000
2018-09-29 23:07:20,626 INFO  [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2018-09-29 23:07:20,757 INFO  [main] client.ConnectionManager$HConnectionImplementation: Closing zookeeper sessionid=0x166264
034530007
2018-09-29 23:07:20,760 INFO  [main] zookeeper.ZooKeeper: Session: 0x166264034530007 closed
2018-09-29 23:07:20,760 INFO  [main-EventThread] zookeeper.ClientCnxn: EventThread shut down
2018-09-29 23:07:20,895 INFO  [main] client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-09-29 23:07:21,290 INFO  [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2018-09-29 23:07:23,532 INFO  [main] input.FileInputFormat: Total input paths to process : 1
2018-09-29 23:07:23,639 INFO  [main] mapreduce.JobSubmitter: number of splits:1
2018-09-29 23:07:23,660 INFO  [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-p
er-checksum
2018-09-29 23:07:24,085 INFO  [main] mapreduce.JobSubmitter: Submitting tokens for job: job_1538239831446_0001
2018-09-29 23:07:24,851 INFO  [main] impl.YarnClientImpl: Submitted application application_1538239831446_0001
2018-09-29 23:07:24,949 INFO  [main] mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1538239
831446_0001/
2018-09-29 23:07:24,951 INFO  [main] mapreduce.Job: Running job: job_1538239831446_0001
2018-09-29 23:07:38,482 INFO  [main] mapreduce.Job: Job job_1538239831446_0001 running in uber mode : false
```

```
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~    acadgild@localhost:~/hbase    acadgild@localhost:~
                FILE: Number of bytes written=139463
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=146
                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)=6068
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=6068
                Total vcore-seconds taken by all map tasks=6068
                Total megabyte-seconds taken by all map tasks=6213632
        Map-Reduce Framework
                Map input records=4
                Map output records=4
                Input split bytes=106
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=134
                CPU time spent (ms)=2450
                Physical memory (bytes) snapshot=183951360
                Virtual memory (bytes) snapshot=2087329792
                Total committed heap usage (bytes)=100663296
        ImportTsv
                Bad Lines=0
        File Input Format Counters
                Bytes Read=40
        File Output Format Counters
                Bytes Written=0
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```

2. Now lets check the data which imported into the table that we created in the ***hbase shell*** called ***bulktable.***
   hbase(main):003:0> scan 'bulktable'

```
acadgild@localhost:~
File  Edit  View  Search  Terminal  Tabs  Help
acadgild@localhost:~    acadgild@localhost:~/hbase    acadgild@localhost:~
hbase(main):003:0> scan 'bulktable'
ROW                                    COLUMN+CELL
 1                                     column=cf1:name, timestamp=1538242637908, value=Amit
 1                                     column=cf2:exp, timestamp=1538242637908, value=4
 2                                     column=cf1:name, timestamp=1538242637908, value=Girija
 2                                     column=cf2:exp, timestamp=1538242637908, value=3
 3                                     column=cf1:name, timestamp=1538242637908, value=Jatin
 3                                     column=cf2:exp, timestamp=1538242637908, value=5
 4                                     column=cf1:name, timestamp=1538242637908, value=Swati
 4                                     column=cf2:exp, timestamp=1538242637908, value=3
4 row(s) in 0.2350 seconds

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

We see all the data are present in the table, thus configuring our mapping successful for tab separated values.

Notes:

Running ***ImportTsv*** with no arguments prints brief usage information:

Usage: importtsv –Dimporttsv.columns=a,b,c <tablename> <inputdir>

Imports the given input directory of TSV data into the specified table. The column names of the TSV data must be specified using the –Dimporttsv.columns options. This option takes the form of comma-separated column names, where each column name is either a simple column family or a columnfamily:qualifier. Also, the special column name HBASE_ROW_KEY is used to designate that this column should be used as the row key for each imported record. You must specify exactly one column to be the row key and consequently you must specify a column name for every column that exists in the input data.

Especially relevant, this importtsv will load data directly into HBase. To instead generate HFiles of data to prepare for bulk data load, pass the option: ***-Dimporttsv.bulk.output=/path/for/output.***

Note: The target table will be created with default column family descriptors if it does not already exist.

Other options that may be specified with –D include:

-Dimporttsv.skip.bad.lines=false **–** fail if encountering an invalid line.
'-Dimportsv.separator=|' eg separate on pipes instead of tabs.
-Dimporttsv.timestamp=currentTimeAsLong **–** use the specified timestamp for the import
-Dimporttsv.mapper.class=my.Mapper – A user-defined Mapper to use instead of org.apache.hadoop.hbase.mapreduce.TsvImportMapper.