LAB 4:

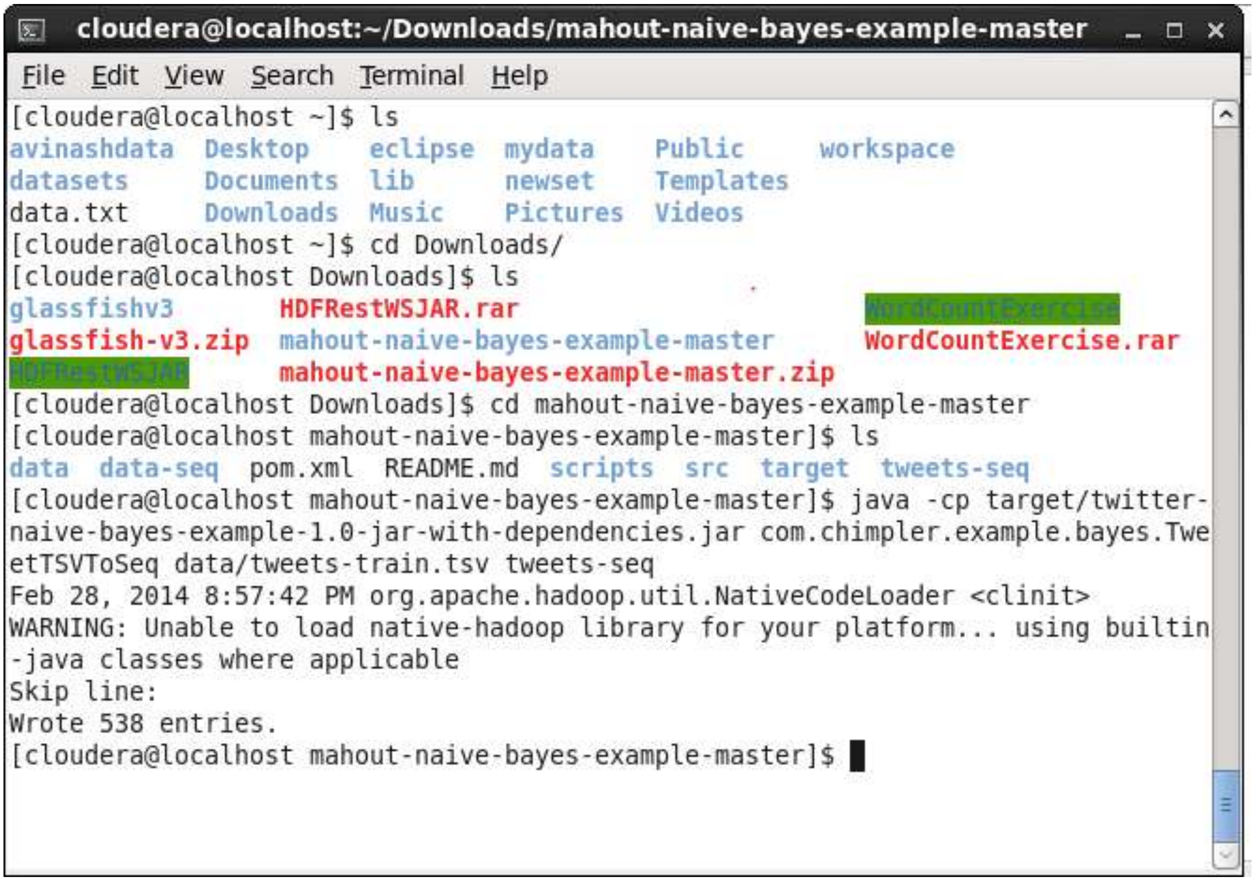
Submitted by

Mahesh Vemula

Task 1: Create a Mahout application

Data is taken from twitter tweets and stored in target folder under main directory

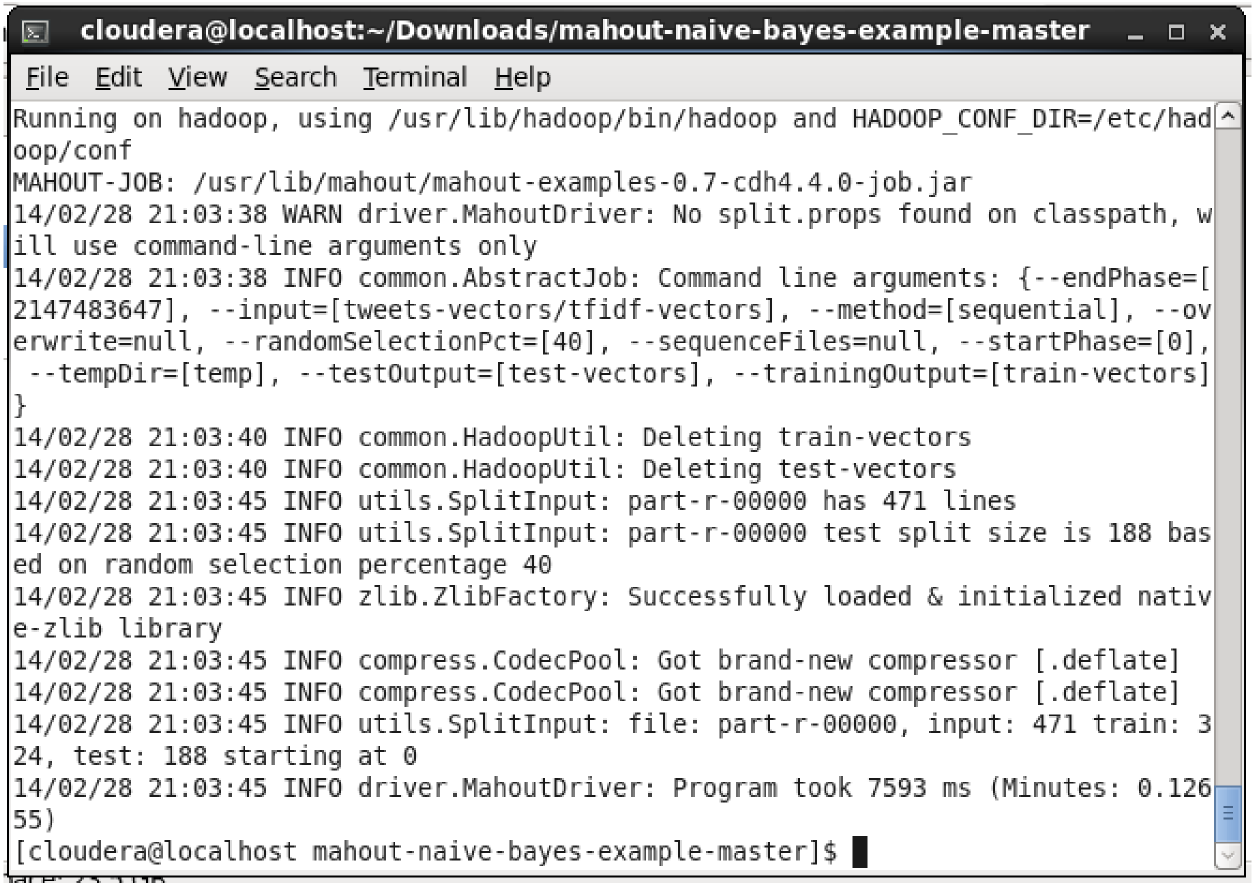
Step 1: Storing training data using HDFS



cloudera@localhost:~/Downloads/mahout-naive-bayes-example-master
File Edit View Search Terminal Help

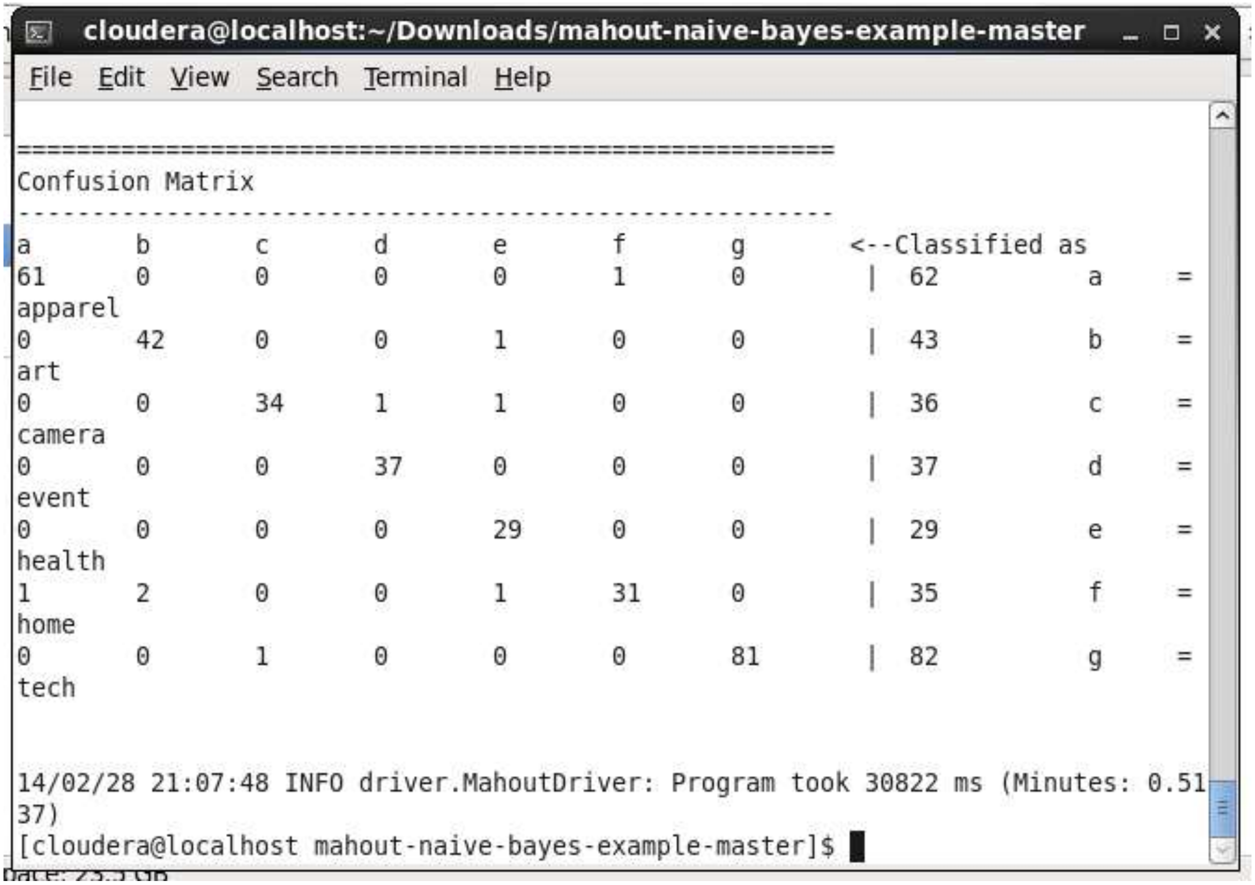
```
[cloudera@localhost ~]$ ls
avinashdata  Desktop    eclipse  mydata    Public     workspace
datasets     Documents  lib      newset    Templates
data.txt     Downloads  Music    Pictures  Videos
[cloudera@localhost ~]$ cd Downloads/
[cloudera@localhost Downloads]$ ls
glassfishv3       HDFRestWSJAR.rar                           WordCountExercise
glassfish-v3.zip  mahout-naive-bayes-example-master         WordCountExercise.rar
HDFRestWSJAR      mahout-naive-bayes-example-master.zip
[cloudera@localhost Downloads]$ cd mahout-naive-bayes-example-master
[cloudera@localhost mahout-naive-bayes-example-master]$ ls
data  data-seq  pom.xml  README.md  scripts  src  target  tweets-seq
[cloudera@localhost mahout-naive-bayes-example-master]$ java -cp target/twitter-
naive-bayes-example-1.0-jar-with-dependencies.jar com.chimpler.example.bayes.Twe
etTSVToSeq data/tweets-train.tsv tweets-seq
Feb 28, 2014 8:57:42 PM org.apache.hadoop.util.NativeCodeLoader <clinit>
WARNING: Unable to load native-hadoop library for your platform... using builtin
-java classes where applicable
Skip line:
Wrote 538 entries.
[cloudera@localhost mahout-naive-bayes-example-master]$ 
```

Step 2: Copy file into Hadoop directory and run mahout naïve bayes algorithms on training data


cloudera@localhost:~/Downloads/mahout-naive-bayes-example-master
File
Edit
View
Search
Terminal
Help

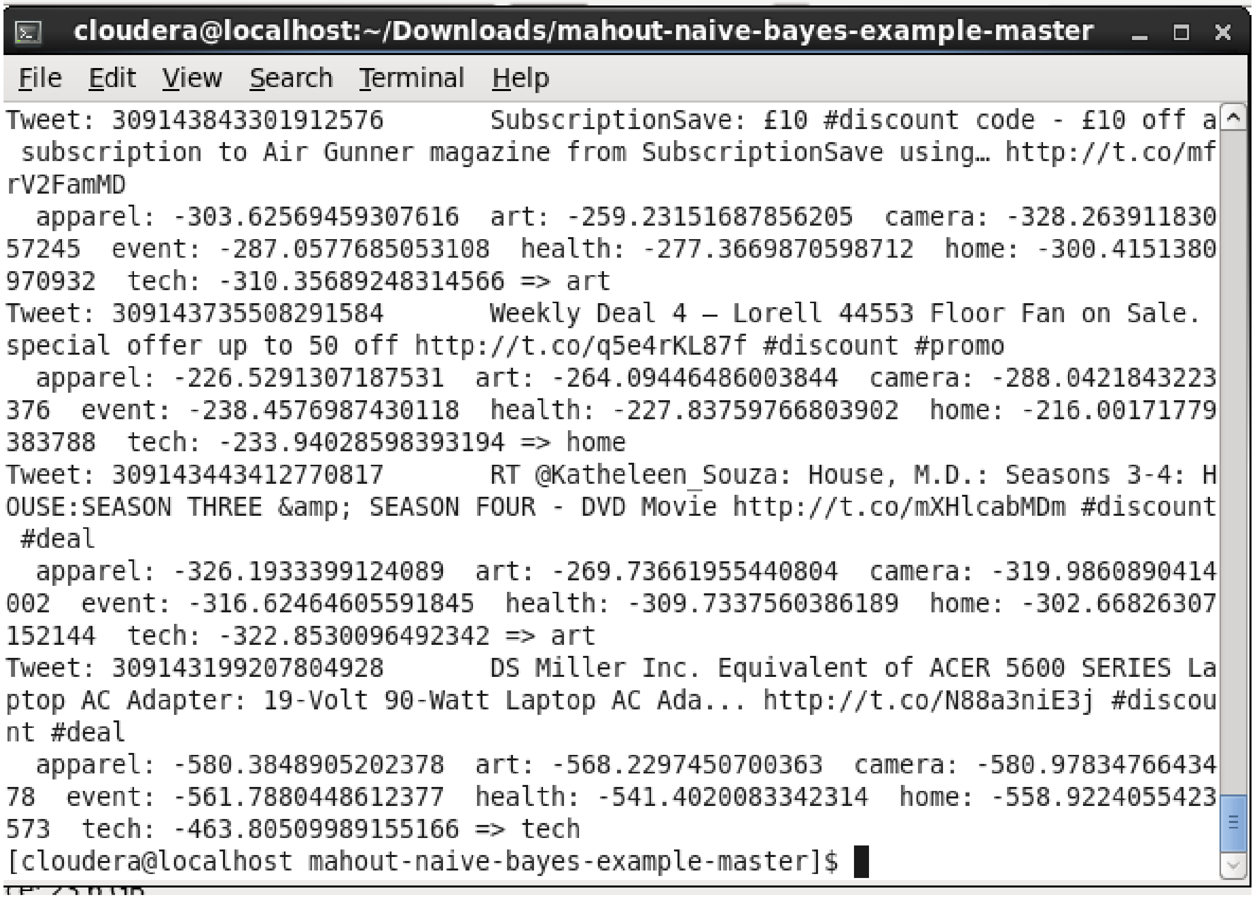
```
Running on hadoop, using /usr/lib/hadoop/bin/hadoop and HADOOP_CONF_DIR=/etc/had
oop/conf
MAHOUT-JOB: /usr/lib/mahout/mahout-examples-0.7-cdh4.4.0-job.jar
14/02/28 21:03:38 WARN driver.MahoutDriver: No split.props found on classpath, w
ill use command-line arguments only
14/02/28 21:03:38 INFO common.AbstractJob: Command line arguments: {--endPhase=[
2147483647], --input=[tweets-vectors/tfidf-vectors], --method=[sequential], --ov
erwrite=null, --randomSelectionPct=[40], --sequenceFiles=null, --startPhase=[0],
 --tempDir=[temp], --testOutput=[test-vectors], --trainingOutput=[train-vectors]
}
14/02/28 21:03:40 INFO common.HadoopUtil: Deleting train-vectors
14/02/28 21:03:40 INFO common.HadoopUtil: Deleting test-vectors
14/02/28 21:03:45 INFO utils.SplitInput: part-r-00000 has 471 lines
14/02/28 21:03:45 INFO utils.SplitInput: part-r-00000 test split size is 188 bas
ed on random selection percentage 40
14/02/28 21:03:45 INFO zlib.ZlibFactory: Successfully loaded & initialized nativ
e-zlib library
14/02/28 21:03:45 INFO compress.CodecPool: Got brand-new compressor [.deflate]
14/02/28 21:03:45 INFO compress.CodecPool: Got brand-new compressor [.deflate]
14/02/28 21:03:45 INFO utils.SplitInput: file: part-r-00000, input: 471 train: 3
24, test: 188 starting at 0
14/02/28 21:03:45 INFO driver.MahoutDriver: Program took 7593 ms (Minutes: 0.126
55)
[cloudera@localhost mahout-naive-bayes-example-master]$ 
```

Step 3: Test algorithm on training data and then on test data to make a classifier


cloudera@localhost:~/Downloads/mahout-naive-bayes-example-master
File Edit View Search Terminal Help

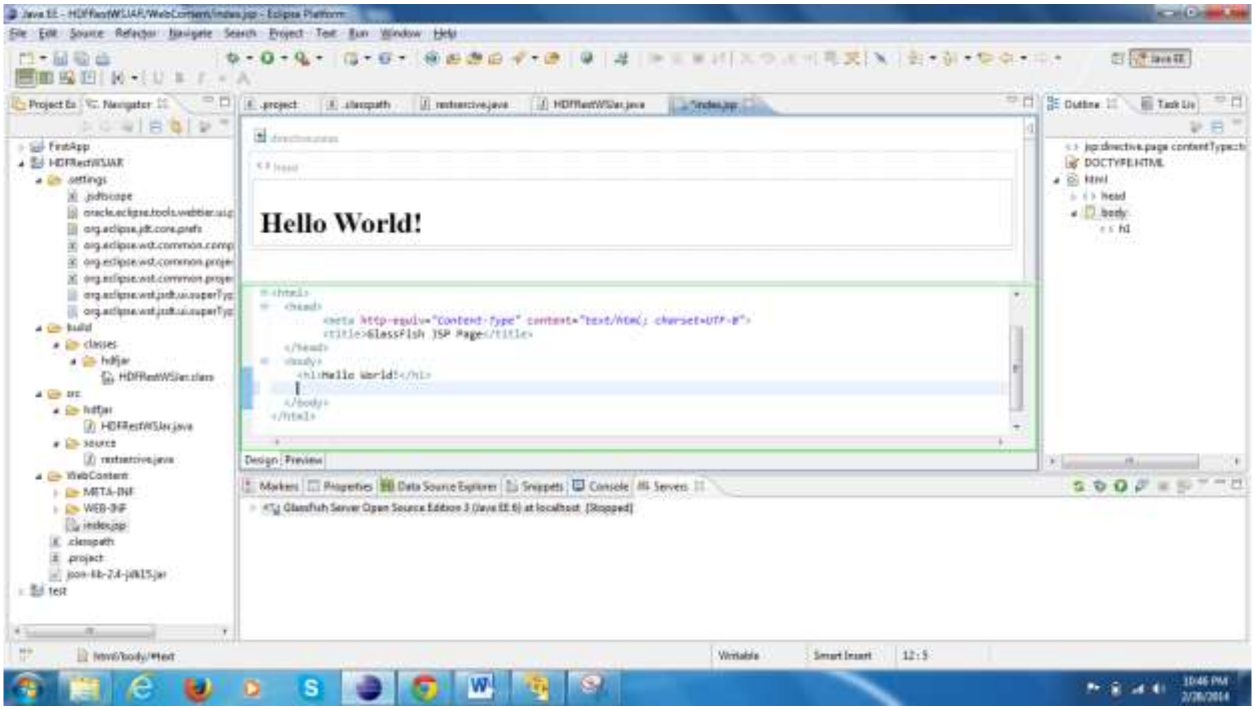
```

=======================================================
Confusion Matrix
-------------------------------------------------------
a       b       c       d       e       f       g       <--Classified as
61      0       0       0       0       1       0        |  62          a     =
apparel
0       42      0       0       1       0       0        |  43          b     =
art
0       0       34      1       1       0       0        |  36          c     =
camera
0       0       0       37      0       0       0        |  37          d     =
event
0       0       0       0       29      0       0        |  29          e     =
health
1       2       0       0       1       31      0        |  35          f     =
home
0       0       1       0       0       0       81       |  82          g     =
tech


14/02/28 21:07:48 INFO driver.MahoutDriver: Program took 30822 ms (Minutes: 0.51
37)
[cloudera@localhost mahout-naive-bayes-example-master]$ 
```

Step 4: Finally classification of tweets


cloudera@localhost:~/Downloads/mahout-naive-bayes-example-master
File Edit View Search Terminal Help

```
Tweet: 309143843301912576       SubscriptionSave: £10 #discount code - £10 off a
 subscription to Air Gunner magazine from SubscriptionSave using… http://t.co/mf
rV2FamMD
  apparel: -303.62569459307616  art: -259.23151687856205  camera: -328.263911830
57245  event: -287.0577685053108  health: -277.3669870598712  home: -300.4151380
970932  tech: -310.35689248314566 => art
Tweet: 309143735508291584       Weekly Deal 4 – Lorell 44553 Floor Fan on Sale.
special offer up to 50 off http://t.co/q5e4rKL87f #discount #promo
  apparel: -226.5291307187531  art: -264.09446486003844  camera: -288.0421843223
376  event: -238.4576987430118  health: -227.83759766803902  home: -216.00171779
383788  tech: -233.94028598393194 => home
Tweet: 309143443412770817       RT @Katheleen_Souza: House, M.D.: Seasons 3-4: H
OUSE:SEASON THREE &amp; SEASON FOUR - DVD Movie http://t.co/mXHlcabMDm #discount
 #deal
  apparel: -326.1933399124089  art: -269.73661955440804  camera: -319.9860890414
002  event: -316.62464605591845  health: -309.7337560386189  home: -302.66826307
152144  tech: -322.8530096492342 => art
Tweet: 309143199207804928       DS Miller Inc. Equivalent of ACER 5600 SERIES La
ptop AC Adapter: 19-Volt 90-Watt Laptop AC Ada... http://t.co/N88a3niE3j #discou
nt #deal
  apparel: -580.3848905202378  art: -568.2297450700363  camera: -580.97834766434
78  event: -561.7880448612377  health: -541.4020083342314  home: -558.9224055423
573  tech: -463.80509989155166 => tech
[cloudera@localhost mahout-naive-bayes-example-master]$ 
```

Now creating a rest server: Install Glassfish and add requite libraries and start the server

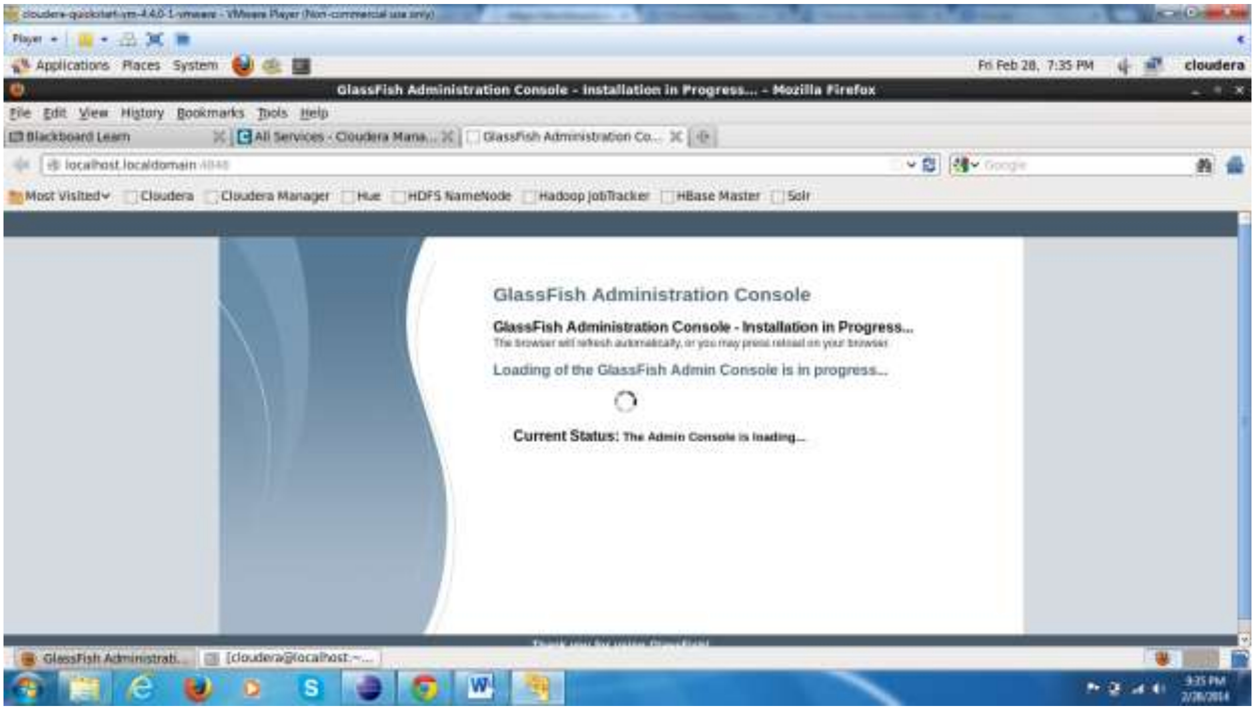


Hello World!

```
<html>
  <head>
        <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
        <title>GlassFish JSP Page</title>
  </head>
  <body>
    <h1>Hello World!</h1>
    
  </body>
</html>
```

Glassfish Server Open Source Edition 3 (Java EE 6) at localhost [Stopped]

Download glassfish server in cloudera and start it


cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Fri Feb 28, 7:35 PM
cloudera
GlassFish Administration Console - Installation in Progress... - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Blackboard Learn
All Services - Cloudera Mana...
GlassFish Administration Co...
localhost.localdomain:4848
Most Visited
Cloudera
Cloudera Manager
Hue
HDFS NameNode
Hadoop JobTracker
HBase Master
Solr
GlassFish Administrati...
[cloudera@localhost:~...

GlassFish Administration Console

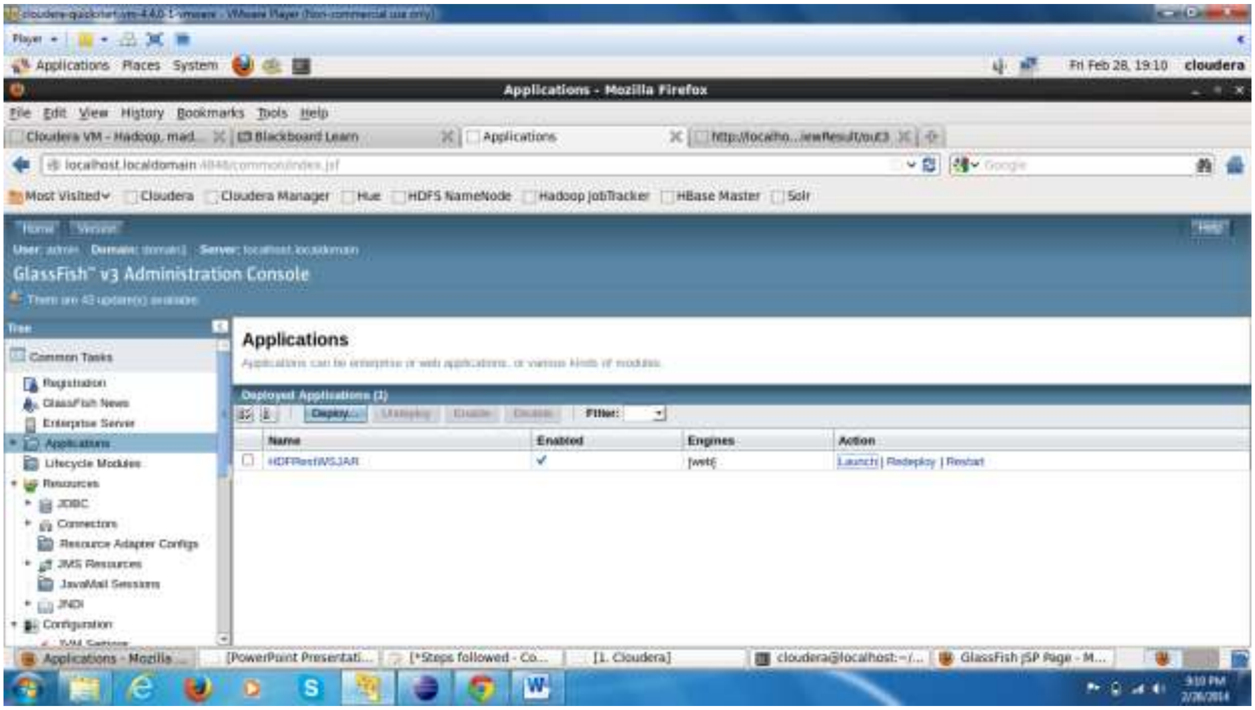
GlassFish Administration Console - Installation in Progress...

The browser will refresh automatically, or you may press reload on your browser

Loading of the GlassFish Admin Console is in progress...

Current Status: The Admin Console is loading...

Load the rest service created into the server



User: admin Domain: domain1 Server: localhost.localdomain

GlassFish™ v3 Administration Console

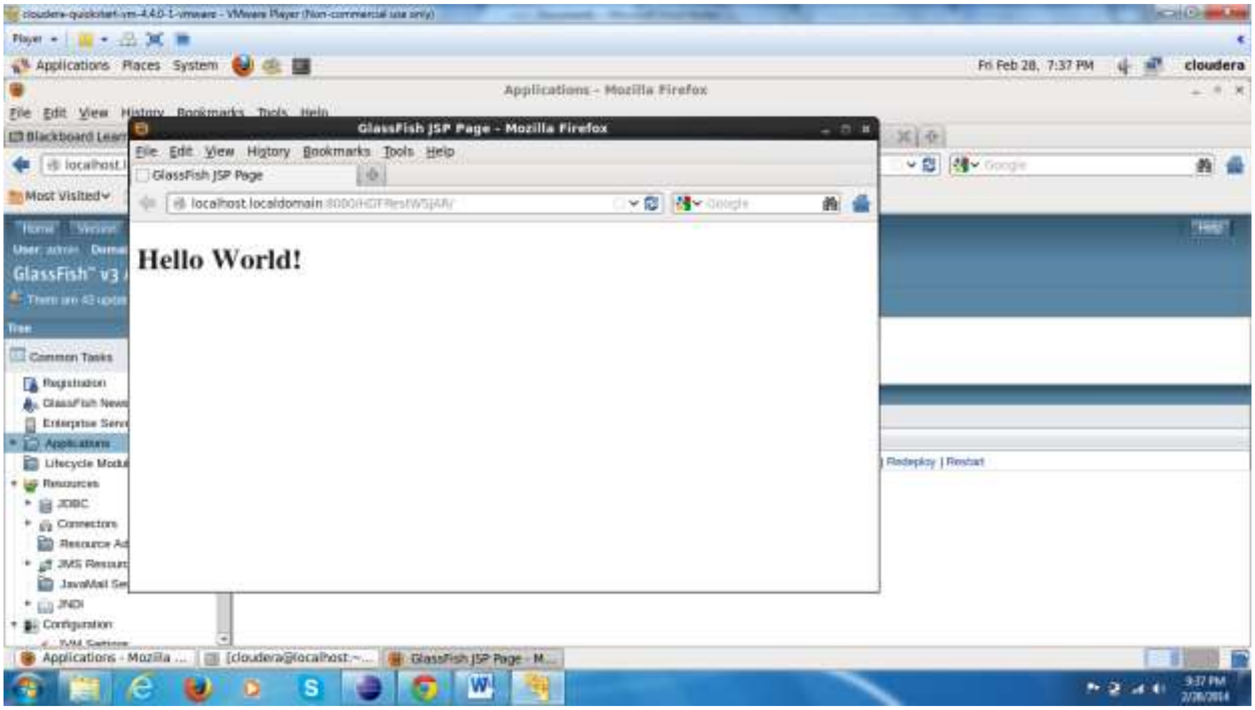
Applications

Applications can be enterprise or web applications, or various kinds of modules.

Deployed Applications (1)

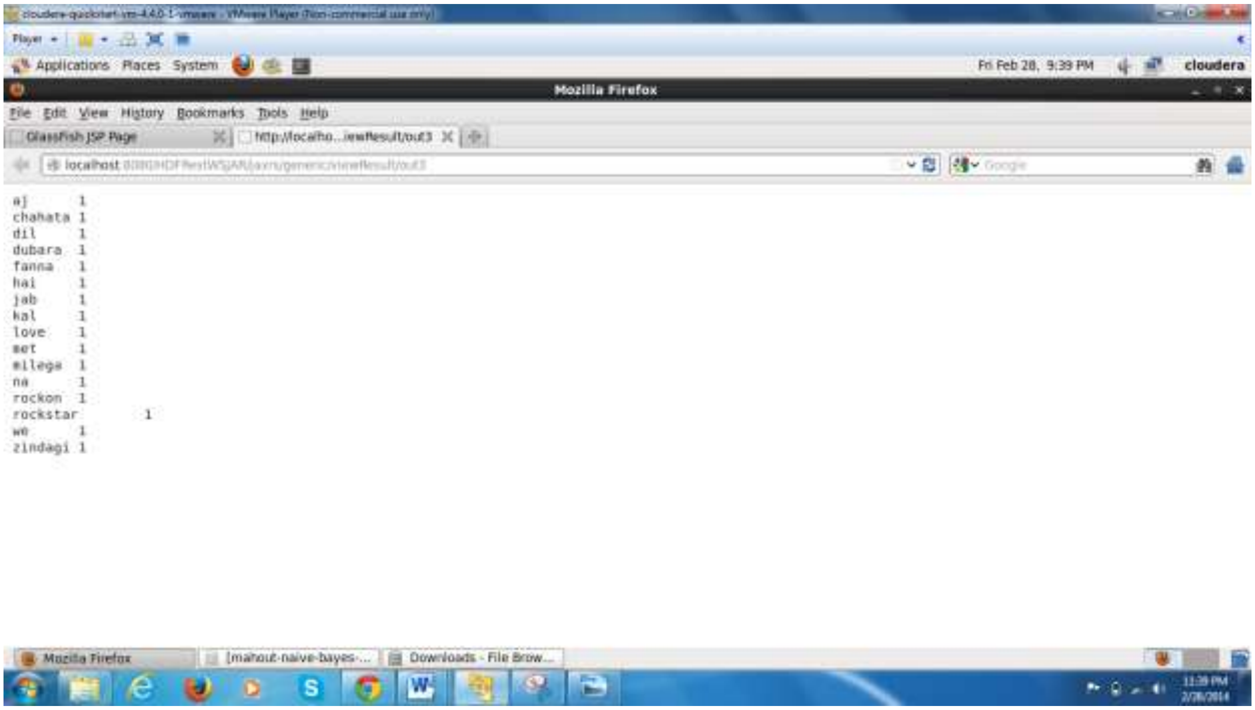
	Name	Enabled	Engines	Action
□	HDFRestWSJAR	✓	[web]	Launch | Redeploy | Restart

Launch the server


Applications - Mozilla Firefox
GlassFish JSP Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help
GlassFish JSP Page
localhost.localdomain:8080/HDFRestWSJAR/
Fri Feb 28, 7:37 PM
cloudera

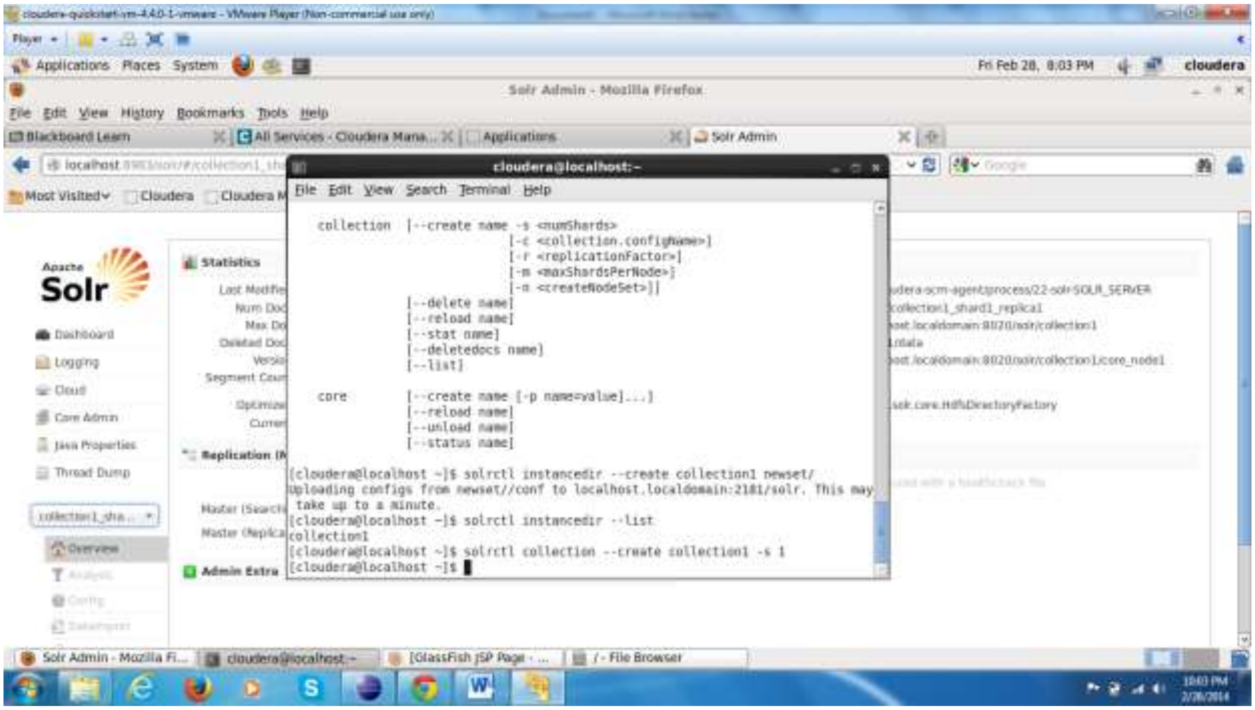
Hello World!

Viewing data from Hadoop file system using server



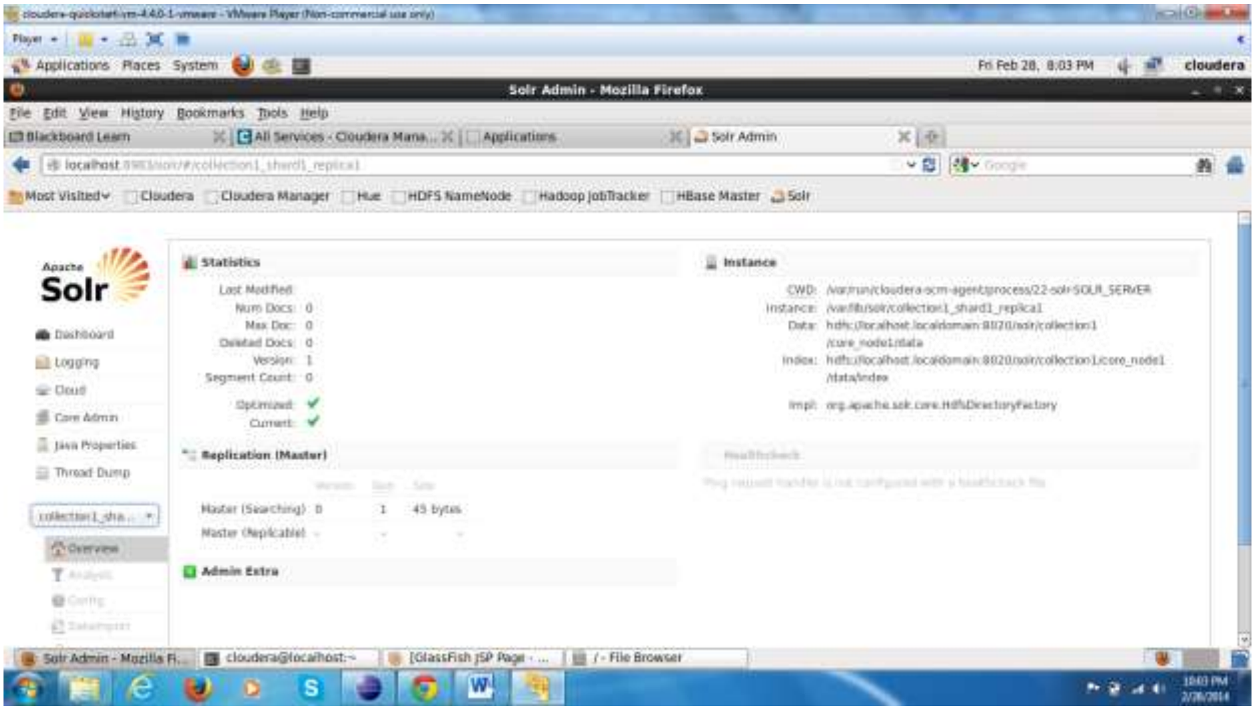
```
aj      1
chahata 1
dil     1
dubara  1
fanna   1
hai     1
jab     1
kal     1
love    1
met     1
milega  1
na      1
rockon  1
rockstar        1
wo      1
zindagi 1
```

Now initializing solar and create a directory name collection1



```
   collection  [--create name -s <numShards>
                             [-c <collection.configName>]
                             [-r <replicationFactor>]
                             [-m <maxShardsPerNode>]
                             [-n <createNodeSet>]]
               [--delete name]
               [--reload name]
               [--stat name]
               [--deletedocs name]
               [--list]

   core        [--create name [-p name=value]...]
               [--reload name]
               [--unload name]
               [--status name]

[cloudera@localhost ~]$ solrctl instancedir --create collection1 newset/
Uploading configs from newset//conf to localhost.localdomain:2181/solr. This may
 take up to a minute.
[cloudera@localhost ~]$ solrctl instancedir --list
collection1
[cloudera@localhost ~]$ solrctl collection --create collection1 -s 1
[cloudera@localhost ~]$ 
```

Details of collection1 directory


Solr Admin - Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost:8983/solr/#/collection1_shard1_replica1
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
Dashboard
Logging
Cloud
Core Admin
Java Properties
Thread Dump
collection1_sha...
Overview
Analysis
Config
DataImport

Statistics

Last Modified:
Num Docs: 0
Max Doc: 0
Deleted Docs: 0
Version: 1
Segment Count: 0
Optimized: ✓
Current: ✓

Instance

CWD: /var/run/cloudera-scm-agent/process/22-solr-SOLR_SERVER
Instance: /var/lib/solr/collection1_shard1_replica1
Data: hdfs://localhost.localdomain:8020/solr/collection1/core_node1/data
Index: hdfs://localhost.localdomain:8020/solr/collection1/core_node1/data/index
Impl: org.apache.solr.core.HdfsDirectoryFactory

Replication (Master)

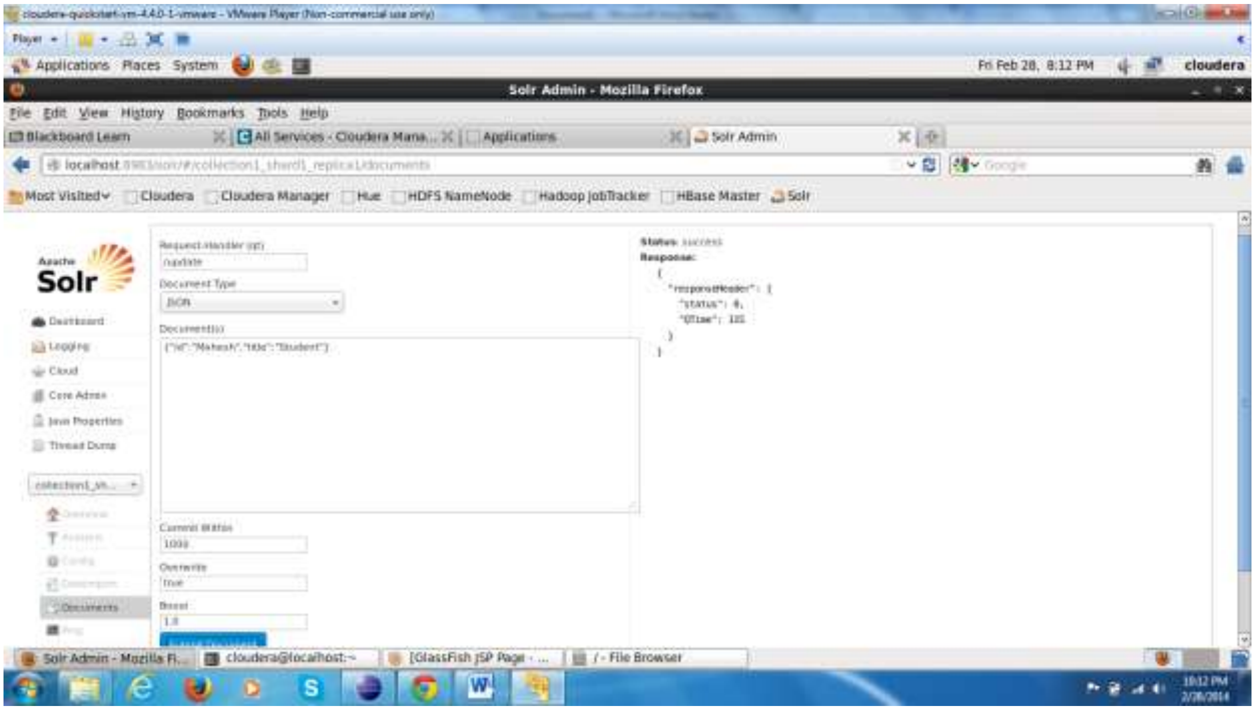
	Version	Gen	Size
Master (Searching)	0	1	45 bytes
Master (Replicable)	-	-	-

Admin Extra

Healthcheck

Ping request handler is not configured with a healthcheck file.

Insert the json data into solr


Solr Admin - Mozilla Firefox
File Edit View History Bookmarks Tools Help
localhost:8983/solr/#/collection1_shard1_replica1/documents
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
Dashboard
Logging
Cloud
Core Admin
Java Properties
Thread Dump
collection1_sh...
Overview
Analysis
Config
DataImport
Documents
Ping

Request-Handler (qt): /update

Document Type: JSON

Document(s):

```
{"id":"Mahesh","title":"Student"}
```

Commit Within: 1000

Overwrite: true

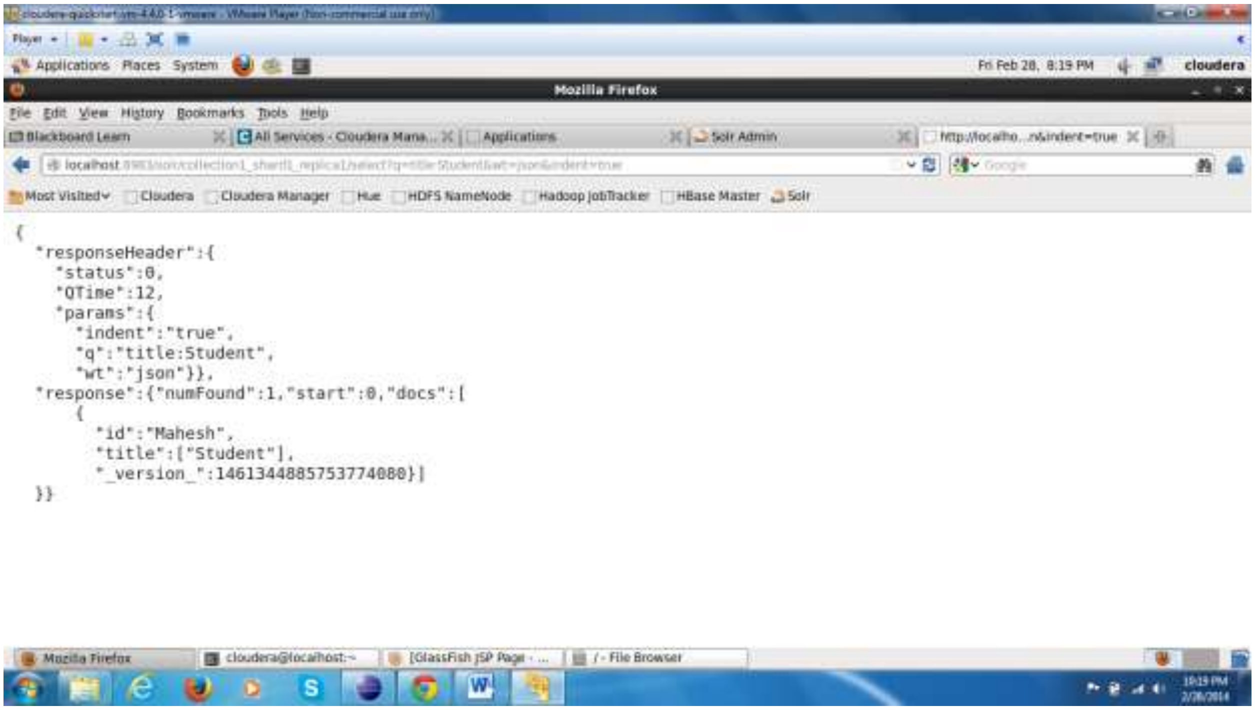
Boost: 1.0

Status: success

Response:

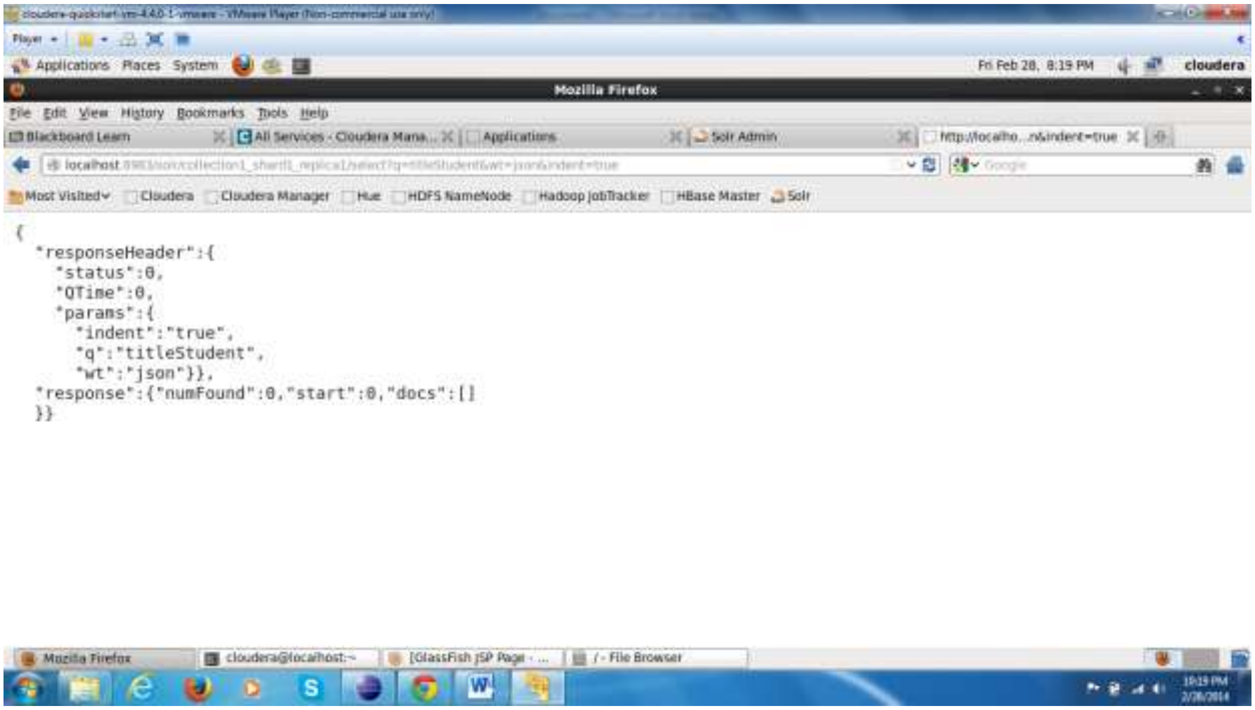
```
{
  "responseHeader": {
    "status": 0,
    "QTime": 131
  }
}
```

Output using query url


cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Fri Feb 28, 8:19 PM
cloudera
Mozilla Firefox
File Edit View History Bookmarks Tools Help
Blackboard Learn
All Services - Cloudera Mana...
Applications
Solr Admin
http://localho...n&indent=true
localhost:8983/solr/collection1_shard1_replica1/select?q=title:Student&wt=json&indent=true
Google
Most Visited
Cloudera
Cloudera Manager
Hue
HDFS NameNode
Hadoop JobTracker
HBase Master
Solr
Mozilla Firefox
cloudera@localhost:~
[GlassFish JSP Page - ...
/ - File Browser
10:19 PM
2/28/2014

```
{
  "responseHeader":{
    "status":0,
    "QTime":12,
    "params":{
      "indent":"true",
      "q":"title:Student",
      "wt":"json"}},
  "response":{"numFound":1,"start":0,"docs":[
      {
        "id":"Mahesh",
        "title":["Student"],
        "_version_":1461344885753774080}]
  }}
```

Output 2 using query url from solr


cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Fri Feb 28, 8:19 PM
cloudera
Mozilla Firefox
File Edit View History Bookmarks Tools Help
Blackboard Learn
All Services - Cloudera Mana...
Applications
Solr Admin
http://localho...n&indent=true
localhost:8983/solr/collection1_shard1_replica1/select?q=titleStudent&wt=json&indent=true
Google
Most Visited
Cloudera
Cloudera Manager
Hue
HDFS NameNode
Hadoop JobTracker
HBase Master
Solr
Mozilla Firefox
cloudera@localhost:~
[GlassFish JSP Page - ...
/ - File Browser
10:19 PM
2/28/2014

```
{
  "responseHeader":{
    "status":0,
    "QTime":0,
    "params":{
      "indent":"true",
      "q":"titleStudent",
      "wt":"json"}},
  "response":{"numFound":0,"start":0,"docs":[]
  }}
```