

# IBM Inter-University Programming Contest 2012 Chapter 7: BigInsight

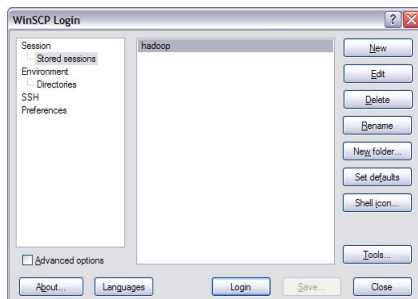
## Objectives

In this exercise, we will learn:

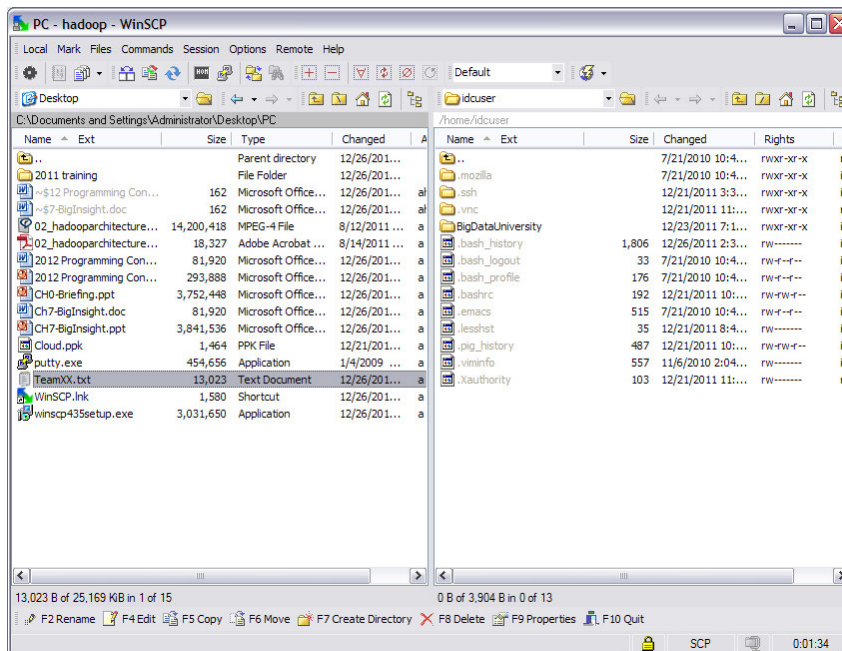
- Basic command on the HDFS system
- Transfer files from local system to HDFS system
- Run Word Count command which is store in a Jar file
- Use the HIVE shell to consolidate the result to table and query on it
- Load HIVE query result into text file

## Exercises

1. Rename the TeamXX.txt to your team number (eg Team01.txt)
2. Open WinSCP and highlight “hadoop” and click “Login”

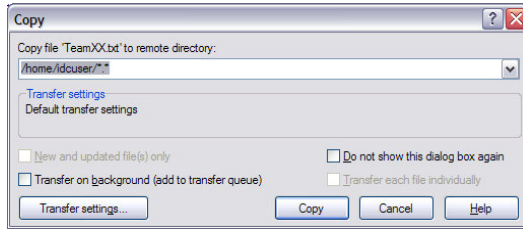


3. Locate your TeamXX.txt and press F5

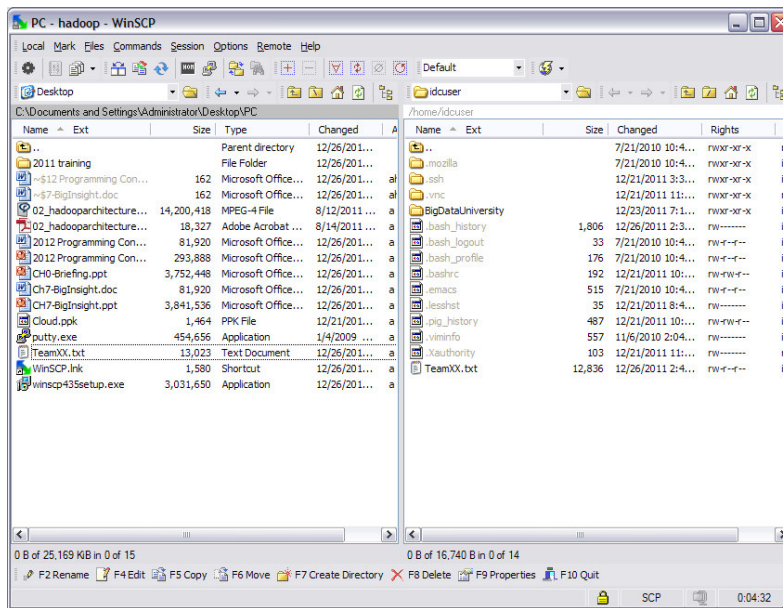


# IBM Inter-University Programming Contest 2012

## 4. Press Copy

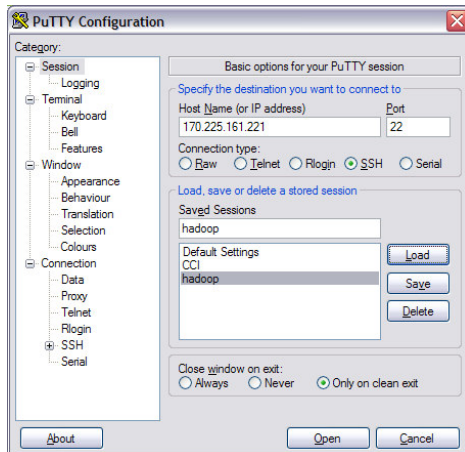


## 5. You will see the TeamXX.txt file at “/home/idcuser”



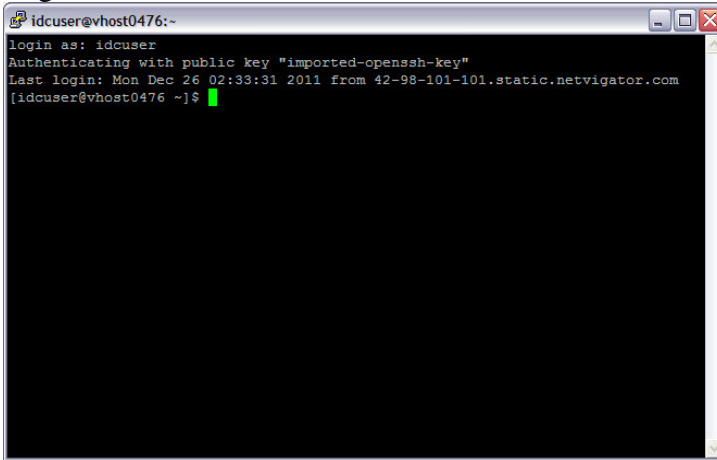
## 6. Open Putty on your desktop.

## 7. Double click on hadoop at the Saved Sessions



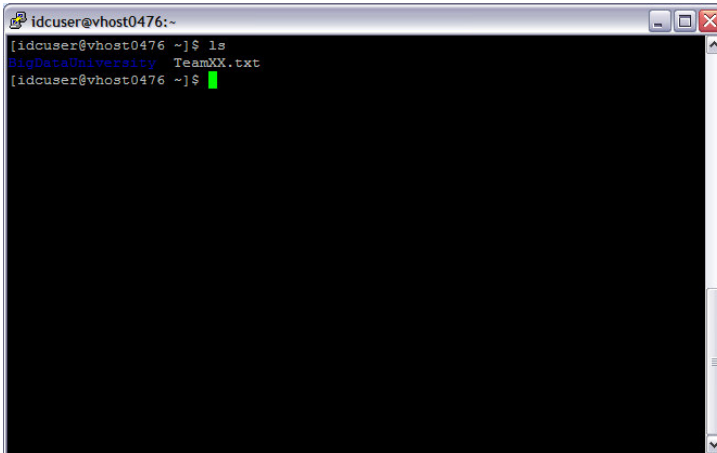
# IBM Inter-University Programming Contest 2012

## 8. Login as “idcuser”



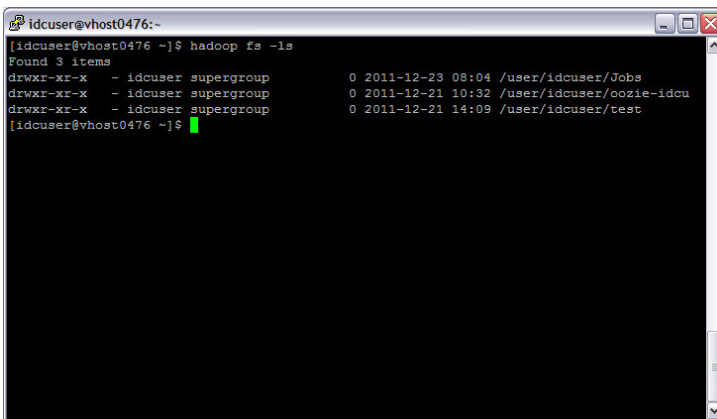
```
idcuser@vhost0476:~  
login as: idcuser  
Authenticating with public key "imported-openssh-key"  
Last login: Mon Dec 26 02:33:31 2011 from 42-98-101-101.static.netvigator.com  
[idcuser@vhost0476 ~]$
```

## 9. Type “ls” to see the file structure and confirm the TeamXX.txt file



```
[idcuser@vhost0476 ~]$ ls  
BigDataUniversity TeamXX.txt  
[idcuser@vhost0476 ~]$
```

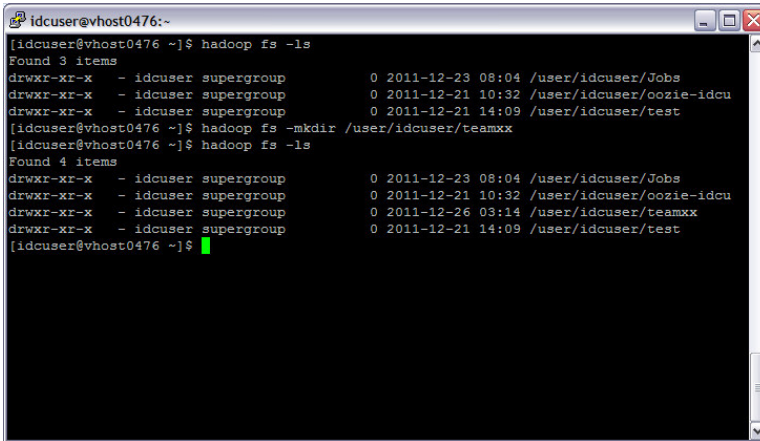
## 10. To access the HDFS system you will need to type the following “hadoop fs -ls”, you will notice the difference between the two file system



```
[idcuser@vhost0476 ~]$ hadoop fs -ls  
Found 3 items  
drwxr-xr-x - idcuser supergroup 0 2011-12-23 08:04 /user/idcuser/Jobs  
drwxr-xr-x - idcuser supergroup 0 2011-12-21 10:32 /user/idcuser/oozie-idcu  
drwxr-xr-x - idcuser supergroup 0 2011-12-21 14:09 /user/idcuser/test  
[idcuser@vhost0476 ~]$
```

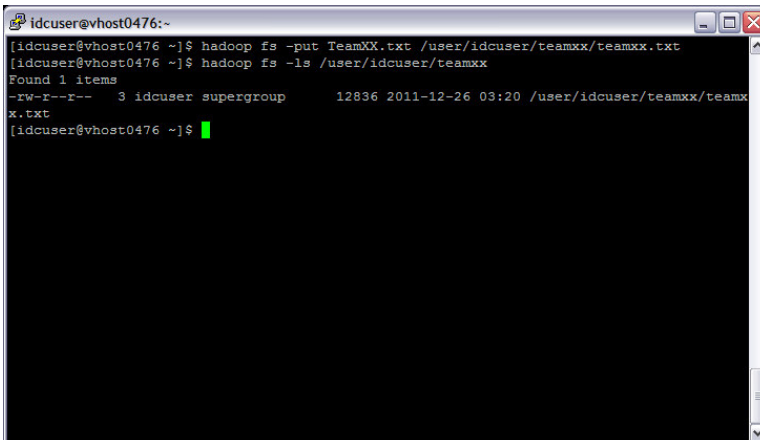
## IBM Inter-University Programming Contest 2012

11. Create a folder to store your test file, type “`hadoop fs -mkdir /user/idcuser/teamxx`”  
xx is your team number. Use the `ls` command to confirm the creation of your team  
folder



```
idcuser@vhost0476:~$ hadoop fs -ls
Found 3 items
drwxr-xr-x - idcuser supergroup      0 2011-12-23 08:04 /user/idcuser/Jobs
drwxr-xr-x - idcuser supergroup      0 2011-12-21 10:32 /user/idcuser/oozie-idcu
drwxr-xr-x - idcuser supergroup      0 2011-12-21 14:09 /user/idcuser/test
[idcuser@vhost0476 ~]$ hadoop fs -mkdir /user/idcuser/teamxx
[idcuser@vhost0476 ~]$ hadoop fs -ls
Found 4 items
drwxr-xr-x - idcuser supergroup      0 2011-12-23 08:04 /user/idcuser/Jobs
drwxr-xr-x - idcuser supergroup      0 2011-12-21 10:32 /user/idcuser/oozie-idcu
drwxr-xr-x - idcuser supergroup      0 2011-12-26 03:14 /user/idcuser/teamxx
drwxr-xr-x - idcuser supergroup      0 2011-12-21 14:09 /user/idcuser/test
[idcuser@vhost0476 ~]$
```

12. To transfer the file from local system to HDFS system, type the following “`hadoop fs -put TeamXX.txt /user/idcuser/teamxx/teamxx.txt`”. Then type “`hadoop fs -ls /user/idcuser/teamxx`” to see the file is transferred

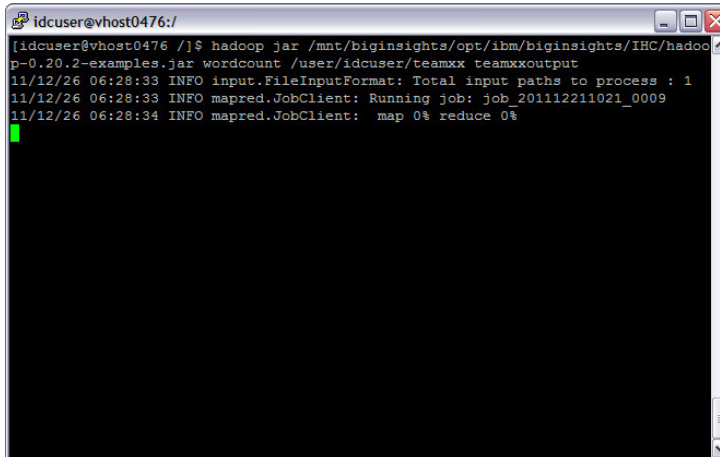


```
idcuser@vhost0476:~$ hadoop fs -put TeamXX.txt /user/idcuser/teamxx/teamxx.txt
[idcuser@vhost0476 ~]$ hadoop fs -ls /user/idcuser/teamxx
Found 1 items
-rw-r--r--  3 idcuser supergroup      12836 2011-12-26 03:20 /user/idcuser/teamxx/teamxx.txt
[idcuser@vhost0476 ~]$
```

## IBM Inter-University Programming Contest 2012

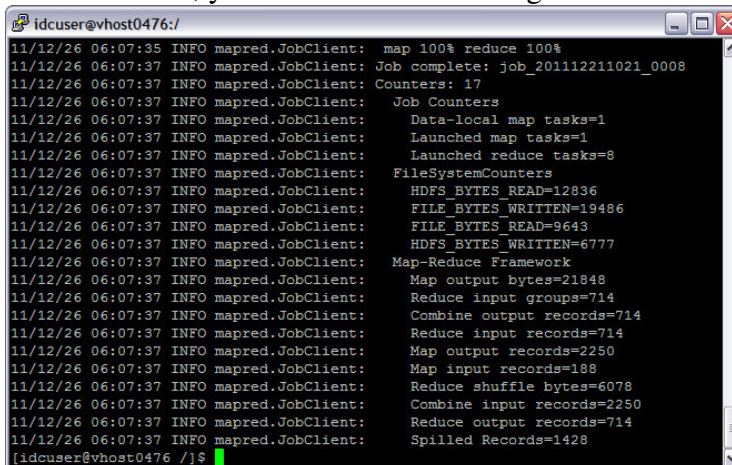
13. In order to count the word in the text file that is uploaded, a sample jar file which content the word count command will be deployed. The command line will like the following “`hadoop jar /mnt/biginsights/opt/ibm/biginsights/IHC/hadoop-0.20.2-examples.jar wordcount /user/idcuser/teamxx teamxxoutput`”

Where the “`mnt.../hadoop-0.20.2-examples.jar`” is the path of the jar, “`/user/idcuser/teamxx`” is the folder which store the text file (multiple files can count at the same time) and “`output`” is the path the output files will be stored



```
idcuser@vhost0476:/$ hadoop jar /mnt/biginsights/opt/ibm/biginsights/IHC/hadoop-0.20.2-examples.jar wordcount /user/idcuser/teamxx teamxxoutput
11/12/26 06:28:33 INFO input.FileInputFormat: Total input paths to process : 1
11/12/26 06:28:33 INFO mapred.JobClient: Running job: job_201112211021_0009
11/12/26 06:28:34 INFO mapred.JobClient: map 0% reduce 0%
```

When finished, you will see the following

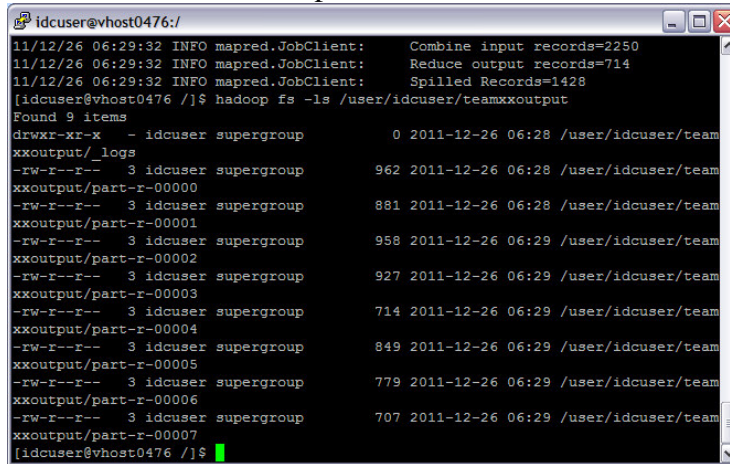


```
11/12/26 06:07:35 INFO mapred.JobClient: map 100% reduce 100%
11/12/26 06:07:37 INFO mapred.JobClient: Job complete: job_201112211021_0008
11/12/26 06:07:37 INFO mapred.JobClient: Counters: 17
11/12/26 06:07:37 INFO mapred.JobClient: Job Counters
11/12/26 06:07:37 INFO mapred.JobClient: Data-local map tasks=1
11/12/26 06:07:37 INFO mapred.JobClient: Launched map tasks=1
11/12/26 06:07:37 INFO mapred.JobClient: Launched reduce tasks=8
11/12/26 06:07:37 INFO mapred.JobClient: FileSystemCounters
11/12/26 06:07:37 INFO mapred.JobClient: HDFS_BYTES_READ=12836
11/12/26 06:07:37 INFO mapred.JobClient: FILE_BYTES_WRITTEN=19486
11/12/26 06:07:37 INFO mapred.JobClient: FILE_BYTES_READ=9643
11/12/26 06:07:37 INFO mapred.JobClient: HDFS_BYTES_WRITTEN=6777
11/12/26 06:07:37 INFO mapred.JobClient: Map-Reduce Framework
11/12/26 06:07:37 INFO mapred.JobClient: Map output bytes=21848
11/12/26 06:07:37 INFO mapred.JobClient: Reduce input groups=714
11/12/26 06:07:37 INFO mapred.JobClient: Combine output records=714
11/12/26 06:07:37 INFO mapred.JobClient: Reduce input records=714
11/12/26 06:07:37 INFO mapred.JobClient: Map output records=2250
11/12/26 06:07:37 INFO mapred.JobClient: Map input records=188
11/12/26 06:07:37 INFO mapred.JobClient: Reduce shuffle bytes=6078
11/12/26 06:07:37 INFO mapred.JobClient: Combine input records=2250
11/12/26 06:07:37 INFO mapred.JobClient: Reduce output records=714
11/12/26 06:07:37 INFO mapred.JobClient: Spilled Records=1428
[idcuser@vhost0476 /]$
```

Please make notes of the job number after “Job complete” = `job_201112211021_0008`

## IBM Inter-University Programming Contest 2012

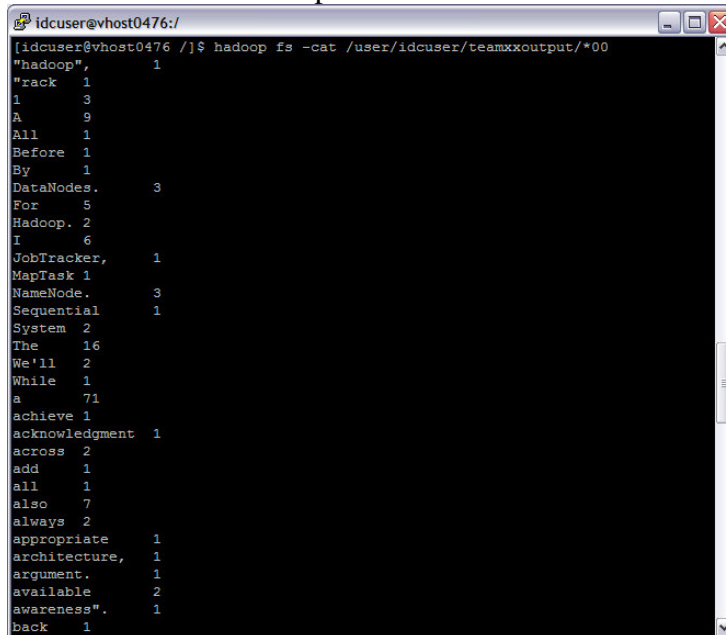
14. We can see the output files, but using the following command “`hadoop fs -ls /user/idcuser/teamxxoutput`”



```
idcuser@vhost0476:/
11/12/26 06:29:32 INFO mapred.JobClient: Combine input records=2250
11/12/26 06:29:32 INFO mapred.JobClient: Reduce output records=714
11/12/26 06:29:32 INFO mapred.JobClient: Spilled Records=1428
[idcuser@vhost0476 /]$ hadoop fs -ls /user/idcuser/teamxxoutput
Found 9 items
drwxr-xr-x - idcuser supergroup 0 2011-12-26 06:28 /user/idcuser/team
xxoutput/_logs
-rw-r--r-- 3 idcuser supergroup 962 2011-12-26 06:28 /user/idcuser/team
xxoutput/part-r-00000
-rw-r--r-- 3 idcuser supergroup 881 2011-12-26 06:28 /user/idcuser/team
xxoutput/part-r-00001
-rw-r--r-- 3 idcuser supergroup 958 2011-12-26 06:29 /user/idcuser/team
xxoutput/part-r-00002
-rw-r--r-- 3 idcuser supergroup 927 2011-12-26 06:29 /user/idcuser/team
xxoutput/part-r-00003
-rw-r--r-- 3 idcuser supergroup 714 2011-12-26 06:29 /user/idcuser/team
xxoutput/part-r-00004
-rw-r--r-- 3 idcuser supergroup 849 2011-12-26 06:29 /user/idcuser/team
xxoutput/part-r-00005
-rw-r--r-- 3 idcuser supergroup 779 2011-12-26 06:29 /user/idcuser/team
xxoutput/part-r-00006
-rw-r--r-- 3 idcuser supergroup 707 2011-12-26 06:29 /user/idcuser/team
xxoutput/part-r-00007
[idcuser@vhost0476 /]$
```

You will notice there are 8 files starting from “part-r-00000” to “part-r-00007”. That is because 8 parallel tasks are used to perform this action

15. To view the output file, use the following command “`hadoop fs -cat /user/idcuser/teamxxoutput/*00`”



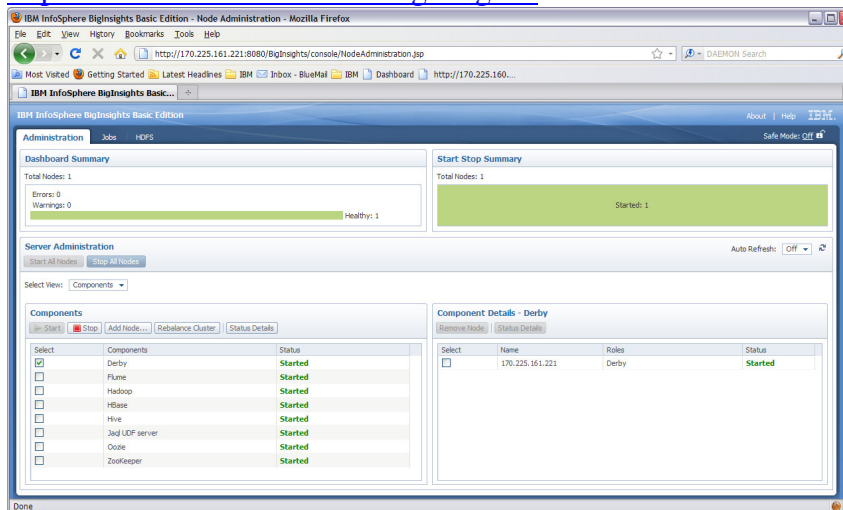
```
[idcuser@vhost0476 /]$ hadoop fs -cat /user/idcuser/teamxxoutput/*00
"hadoop", 1
"rack 1
1 3
A 9
All 1
Before 1
By 1
DataNodes. 3
For 5
Hadoop. 2
I 6
JobTracker, 1
MapTask 1
NameNode. 3
Sequential 1
System 2
The 16
We'll 2
While 1
a 71
achieve 1
acknowledgment 1
across 2
add 1
all 1
also 7
always 2
appropriate 1
architecture, 1
argument. 1
available 2
awareness". 1
back 1
```

The list of word counted will be shown

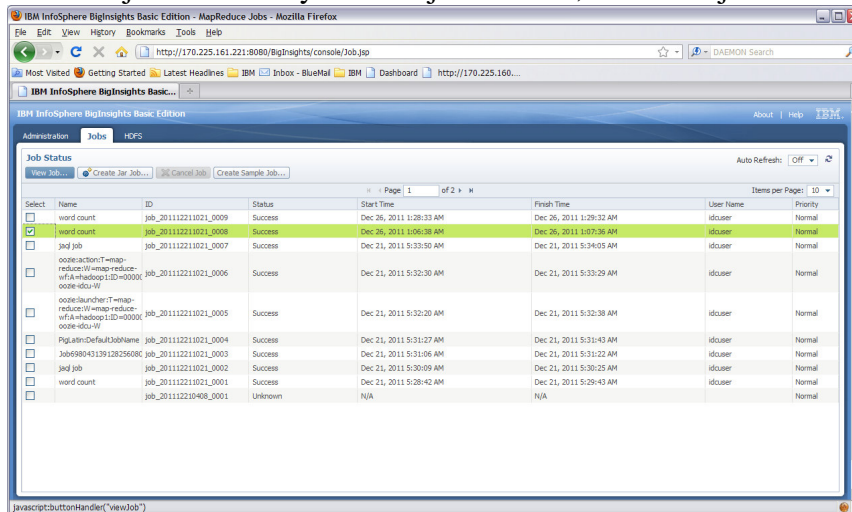
# IBM Inter-University Programming Contest 2012

16. We can review the job over the web as well, open a browser and go to the following address

<http://170.225.161.221:8080/BigInsights/>



17. Click on jobs and find your own job number, select the job and click on “View job”



# IBM Inter-University Programming Contest 2012

18. The details of the tasks is show here

The screenshot shows the IBM InfoSphere BigInsights Basic Edition interface. The main content area displays 'Job Details' for a job named 'word count'. The job status is 'Success'. Below this, there is a 'Task Summary' section with a table showing task details.

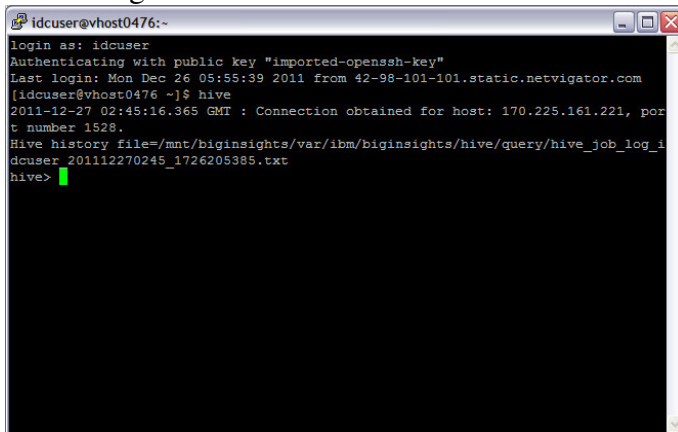
Select	Type	Total Tasks	Successful Tasks	Failed Tasks	Killed Tasks	Running Tasks	Pending Tasks	Unknown	Start Time	Finish Time
<input type="radio"/>	Setup	1	1	0	0	0	0	0	Dec 26, 2011 1:06:39 AM	Dec 26, 2011 1:06:42 AM
<input type="radio"/>	Map	1	1	0	0	0	0	0	Dec 26, 2011 1:06:45 AM	Dec 26, 2011 1:06:48 AM
<input type="radio"/>	Reduce	8	8	0	0	0	0	0	Dec 26, 2011 1:06:51 AM	Dec 26, 2011 1:07:33 AM
<input type="radio"/>	Cleanup	1	1	0	0	0	0	0	Dec 26, 2011 1:07:33 AM	Dec 26, 2011 1:07:36 AM

You will notice the 8 tasks is run on the Reduce type



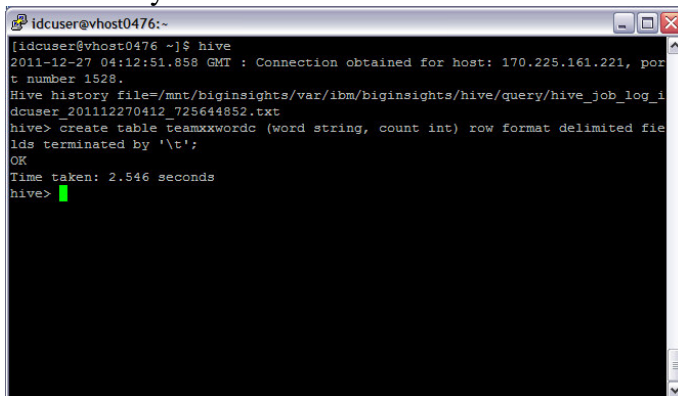
## IBM Inter-University Programming Contest 2012

19. Now we have to word count result, we will leverage one of the language HIVE to store the result and select the top words. Return to the Putty Windows and type “hive” to get into hive interface



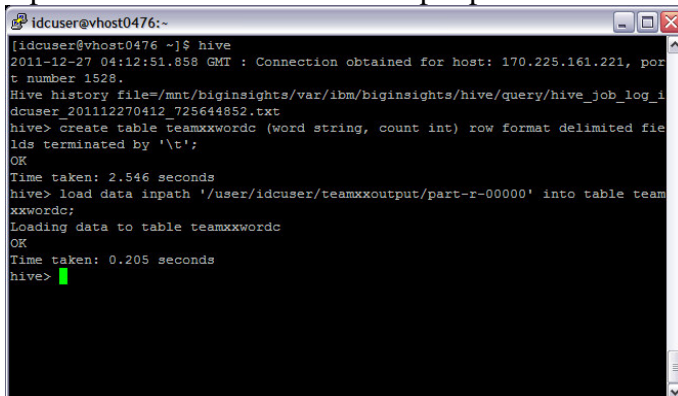
```
idcuser@vhost0476:~  
login as: idcuser  
Authenticating with public key "imported-openssh-key"  
Last login: Mon Dec 26 05:55:39 2011 from 42-98-101-101.static.netvigator.com  
[idcuser@vhost0476 ~]$ hive  
2011-12-27 02:45:16.365 GMT : Connection obtained for host: 170.225.161.221, port number 1528.  
Hive history file=/mnt/biginsights/var/ibm/biginsights/hive/query/hive_job_log_idcuser_201112270245_1726205385.txt  
hive>
```

20. You will create a table to store the result, it will have 2 column, one is the word, other is the count. Use the following command “create table teamxxwordc (word string, count int) row format delimited fields terminated by ‘\t’;” where the create table with name teamxxwordc with two column word of string and count of interger and field delimited by tab



```
[idcuser@vhost0476 ~]$ hive  
2011-12-27 04:12:51.858 GMT : Connection obtained for host: 170.225.161.221, port number 1528.  
Hive history file=/mnt/biginsights/var/ibm/biginsights/hive/query/hive_job_log_idcuser_201112270412_725644852.txt  
hive> create table teamxxwordc (word string, count int) row format delimited fields terminated by '\t';  
OK  
Time taken: 2.546 seconds  
hive>
```

21. You will load the first data file to the table by the following command “load data inpath ‘/user/idcuser/teamxxoutput/part-r-00000’ into table teamxxwordc;”



```
[idcuser@vhost0476 ~]$ hive  
2011-12-27 04:12:51.858 GMT : Connection obtained for host: 170.225.161.221, port number 1528.  
Hive history file=/mnt/biginsights/var/ibm/biginsights/hive/query/hive_job_log_idcuser_201112270412_725644852.txt  
hive> create table teamxxwordc (word string, count int) row format delimited fields terminated by '\t';  
OK  
Time taken: 2.546 seconds  
hive> load data inpath '/user/idcuser/teamxxoutput/part-r-00000' into table teamxxwordc;  
Loading data to table teamxxwordc  
OK  
Time taken: 0.205 seconds  
hive>
```

## IBM Inter-University Programming Contest 2012

22. Repeat the same loading for the other 7 files

```
idcuser@vhost0476:~$
Time taken: 0.134 seconds
hive> load data inpath '/user/idcuser/teamxxoutput/part-r-00004' into table team
xxwords;
Loading data to table teamxxwords
OK
Time taken: 0.118 seconds
hive> load data inpath '/user/idcuser/teamxxoutput/part-r-00005' into table team
xxwords;
Loading data to table teamxxwords
OK
Time taken: 0.118 seconds
hive> load data inpath '/user/idcuser/teamxxoutput/part-r-00006' into table team
xxwords;
Loading data to table teamxxwords
OK
Time taken: 0.13 seconds
hive> load data inpath '/user/idcuser/teamxxoutput/part-r-00007' into table team
xxwords;
Loading data to table teamxxwords
OK
Time taken: 0.116 seconds
hive>
```

23. You can now use SQL like state to query the table, for example “Select word, sum(count) from teamxxwords group by word;”

```
idcuser@vhost0476:~$
OK
Time taken: 0.116 seconds
hive> select word sum(count) from teamxxwords group by word;
FAILED: Parse Error: line 1:15 mismatched input '(' expecting FROM in from claus
e

hive> select word,sum(count) from teamxxwords group by word;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201112211021_0020, Tracking URL = http://vhost0476.dcl.sg.ap.
compute.ihost.com:50030/jobdetails.jsp?jobid=job_201112211021_0020
Kill Command = /mnt/biginsights/opt/ibm/biginsights/IHC/bin/./bin/hadoop job -
Dmapred.job.tracker=170.225.161.221:9001 -kill job_201112211021_0020
2011-12-27 04:21:29,847 Stage-1 map = 0%, reduce = 0%
```

The result will be show like this

```
idcuser@vhost0476:~$
sequential      1
software        1
specify 1
state 1
say 1
seeks, 1
sends 1
size 1
size, 1
specifying      1
start.sh        1
submits 1
submitted      1
succeeds,      1
schedules      1
several 1
stores 1
stores. 1
subset 1
successful      1
Time taken: 46.856 seconds
hive>
```

## IBM Inter-University Programming Contest 2012

24. You can output the result to a test file in the local system, the command is as follow  
“insert overwrite local directory ‘/home/idcuser/result.txt’ select \* from teamxxwordc  
where word like ‘s%’ sort by count desc;”

```
idcuser@vhost0476:~$
Time taken: 46.856 seconds
hive> insert overwrite local directory '/home/idcuser/result.txt' select * from
teamxxwordc where word like 's%' sort by count desc;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks not specified. Estimated from input data size: 1
In order to change the average load for a reducer (in bytes):
  set hive.exec.reducers.bytes.per.reducer=<number>
In order to limit the maximum number of reducers:
  set hive.exec.reducers.max=<number>
In order to set a constant number of reducers:
  set mapred.reduce.tasks=<number>
Starting Job = job_201112211021_0024, Tracking URL = http://vhost0476.dcl.sg.ap.
compute.ihost.com:50030/jobdetails.jsp?jobid=job_201112211021_0024
Kill Command = /mnt/biginsights/opt/ibm/biginsights/IHC/bin/./bin/hadoop job -
Dmapred.job.tracker=170.225.161.221:9001 -kill job_201112211021_0024
2011-12-27 04:43:49,010 Stage-1 map = 0%, reduce = 0%
2011-12-27 04:43:59,176 Stage-1 map = 13%, reduce = 0%
2011-12-27 04:44:02,190 Stage-1 map = 25%, reduce = 0%
2011-12-27 04:44:05,207 Stage-1 map = 38%, reduce = 0%
2011-12-27 04:44:08,241 Stage-1 map = 50%, reduce = 0%
```

Exit hive by “exit;” and list the directory, you will find the result.txt file

```
idcuser@vhost0476:~$
Dmapred.job.tracker=170.225.161.221:9001 -kill job_201112211021_0024
2011-12-27 04:43:49,010 Stage-1 map = 0%, reduce = 0%
2011-12-27 04:43:59,176 Stage-1 map = 13%, reduce = 0%
2011-12-27 04:44:02,190 Stage-1 map = 25%, reduce = 0%
2011-12-27 04:44:05,207 Stage-1 map = 38%, reduce = 0%
2011-12-27 04:44:08,241 Stage-1 map = 50%, reduce = 0%
2011-12-27 04:44:11,259 Stage-1 map = 63%, reduce = 0%
2011-12-27 04:44:14,283 Stage-1 map = 75%, reduce = 0%
2011-12-27 04:44:17,301 Stage-1 map = 88%, reduce = 0%
2011-12-27 04:44:20,321 Stage-1 map = 100%, reduce = 0%
2011-12-27 04:44:31,400 Stage-1 map = 100%, reduce = 100%
Ended Job = job_201112211021_0024
Copying data to local directory /home/idcuser/result.txt
Copying data to local directory /home/idcuser/result.txt
78 Rows loaded to /home/idcuser/result.txt
OK
Time taken: 47.415 seconds
hive> exit
> ;
[idcuser@vhost0476 ~]$ ls
BigDataUniversity result.txt TeamXX.txt
[idcuser@vhost0476 ~]$
```

## IBM Inter-University Programming Contest 2012

25. Transfer the file from result folder to local machine and open the file in word pad, and the result will be shown as follow



The screenshot shows a WordPad window titled "attempt\_201112211021\_0024\_r\_000000\_0 - WordPad". The window contains a list of words with suffixes, such as "systemr7", "secondr7", "same r6", "shownr5", "suchr4", "seer4", "shouldr4", "storedr4", "store r3", "supportsr3", "single r3", "schemer3", "systemr3", "specificr2", "some r2", "system. r2", "seeksr2", "supported. r2", "seekingr1", "sent r1", "set r1", "since r1", "sizetr1", "space. r1", "spawnsr1", and "statr1". The window also shows a menu bar with "File", "Edit", "View", "Insert", "Format", and "Help", and a status bar at the bottom that says "For Help, press F1".

```
systemr7
secondr7
same r6
shownr5
suchr4
seer4
shouldr4
storedr4
store r3
supportsr3
single r3
schemer3
systemr3
specificr2
some r2
system. r2
seeksr2
supported. r2
seekingr1
sent r1
set r1
since r1
sizetr1
space. r1
spawnsr1
statr1
```

26. This is the end of the lab for BigInsight