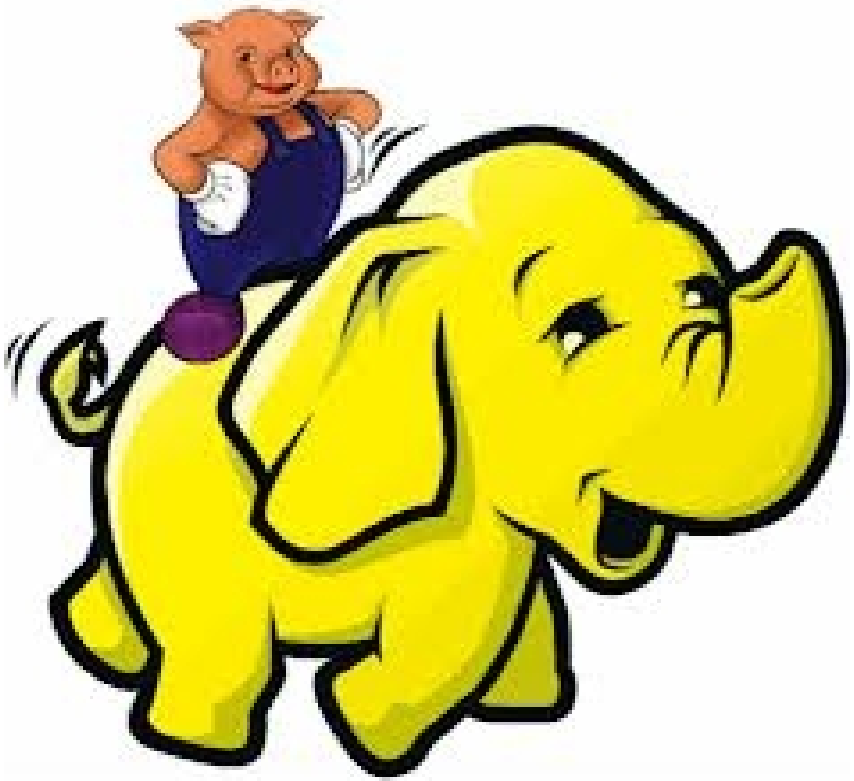
Pig



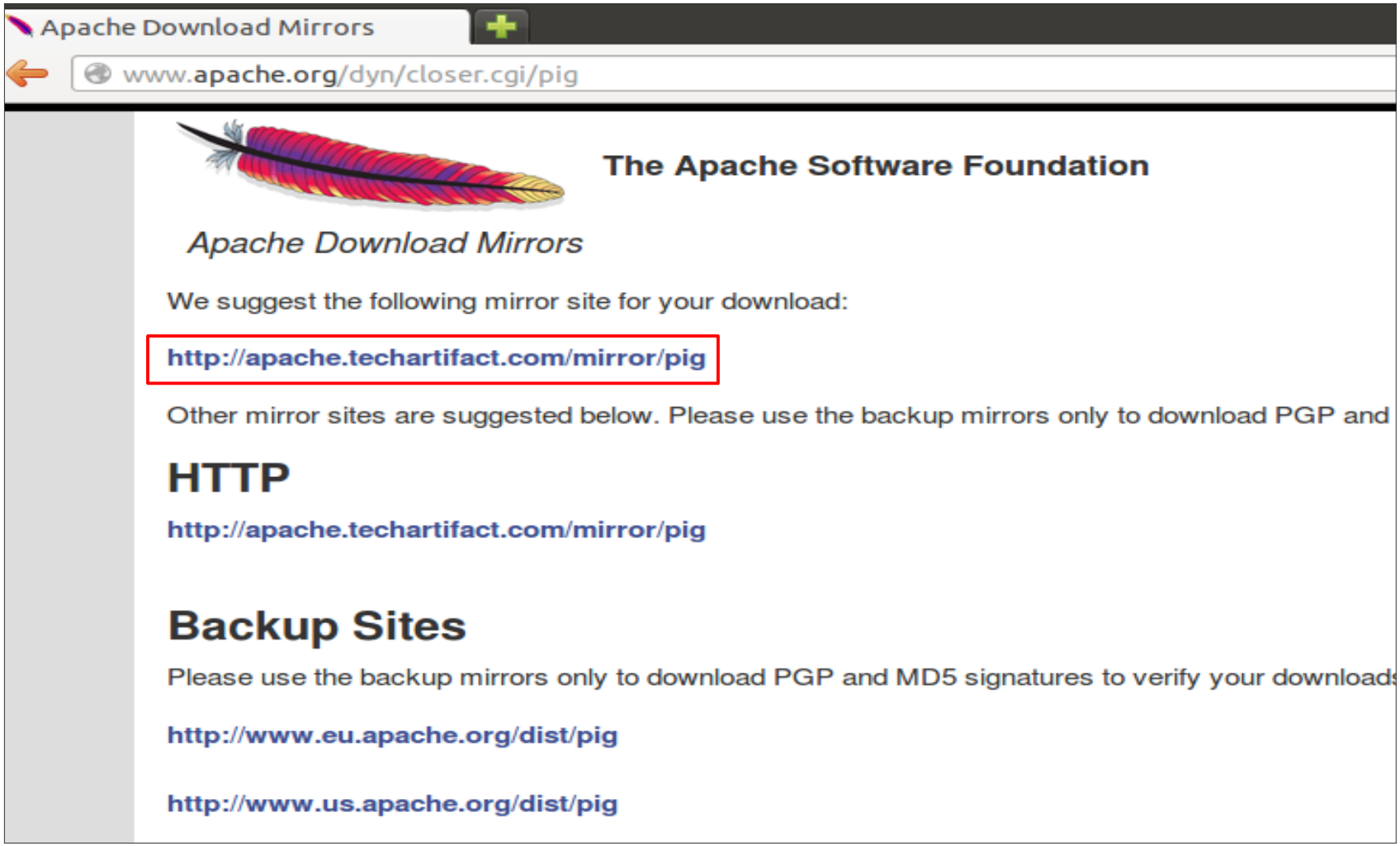
For online Hadoop training, send mail to neeraj.ymca.2k6@gmail.com

Agenda

Download Pig tar.gz file
Extract the content of Pig tar.gz
Configure pig-env.sh file
Configure pig.properties file
Start your Hadoop
Start Pig shell
Input file for Pig query
Access HDFS from Pig shell
Execute Pig commands
Store Pig query's output into HDFS
Check the output
Comparison of HBase/Hive/Pig

Download Pig from Apache website

www.apache.org/dyn/closer.cgi/pig



Select a stable version of Pig

Index of /mirror/pig

apache.techartifact.com/mirror/pig/

Pig Releases

Please make sure you're downloading from a nearby mirror site, not from www.apache.org

We suggest downloading the current stable release.

Older releases are available from the archives.

Name	Last modified	Size	Description
Parent Directory		-	
pig-0.10.0/	25-Apr-2012 12:25	-	
pig-0.10.1/	05-Jan-2013 13:43	-	
pig-0.11.0/	15-Feb-2013 05:00	-	
pig-0.9.2/	22-Jan-2012 04:32	-	
stable/	15-Feb-2013 05:00	-	
HEADER.html	11-Nov-2011 04:08	397	
KEYS	09-Dec-2010 03:23	4.5K	

Apache/2.2.23 (Unix) mod_ssl/2.2.23 OpenSSL/1.0.0-fips DAV/2 mod_auth_passthrough/2.1 r
Port 80

Click on pig-0.11.0-tar.gz

Index of /mirror/pig/stable

apache.techartifact.com/mirror/pig/stable/

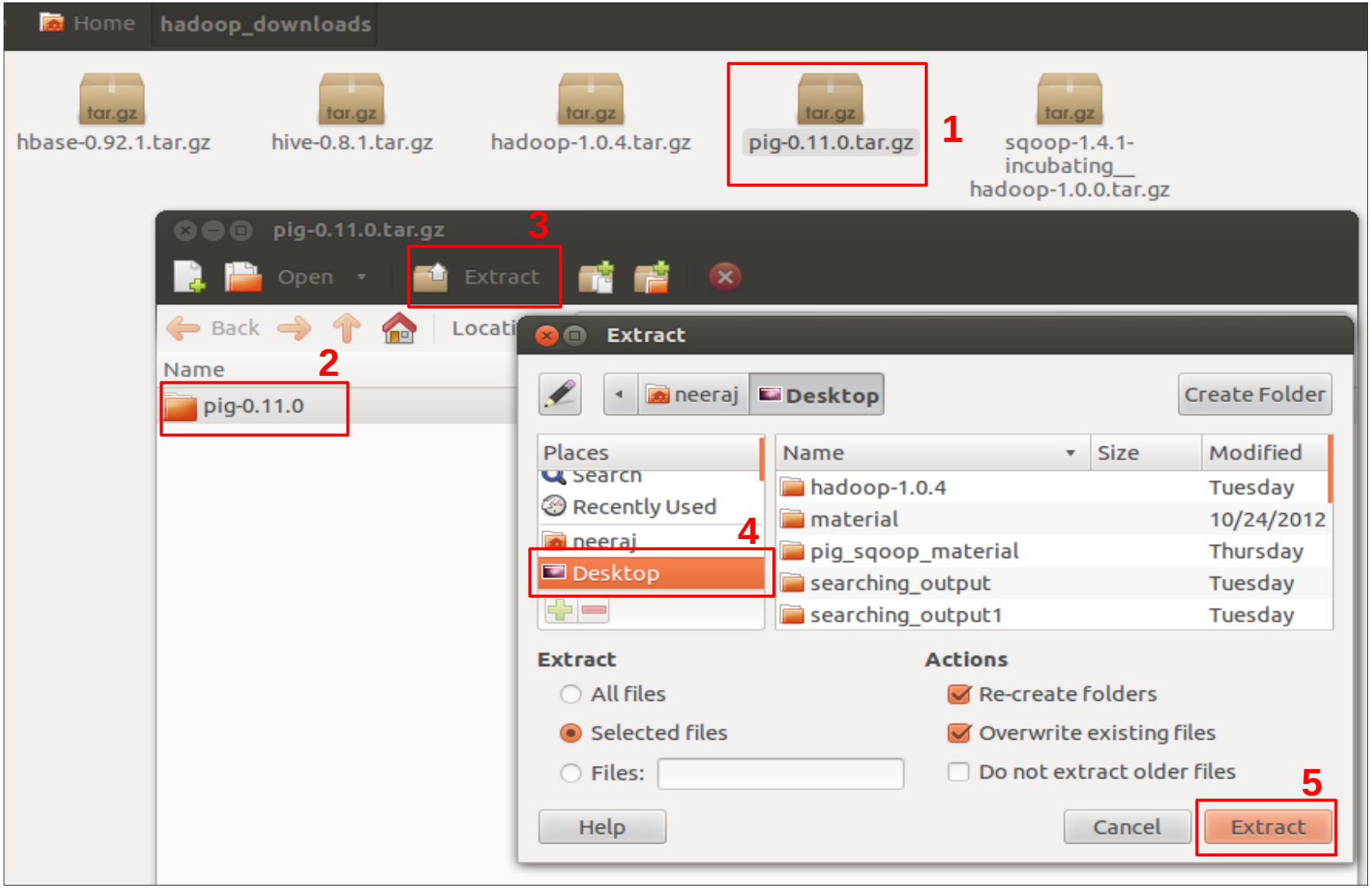
Index of /mirror/pig/stable

Name	Last modified	Size	Description
Parent Directory		-	
RELEASE_NOTES.txt	15-Feb-2013 04:22	2.9K	
pig-0.11.0-1.i386.rpm	15-Feb-2013 04:24	39M	
pig-0.11.0-1.i386.rpm.asc	15-Feb-2013 04:46	535	
pig-0.11.0-src.tar.gz	15-Feb-2013 03:45	14M	
pig-0.11.0-src.tar.gz.asc	15-Feb-2013 04:46	535	
pig-0.11.0.tar.gz	15-Feb-2013 03:46	54M	
pig-0.11.0.tar.gz.asc	15-Feb-2013 04:46	535	
pig_0.11.0-1_i386.deb	15-Feb-2013 03:29	38M	
pig_0.11.0-1_i386.deb.asc	15-Feb-2013 04:46	535	
rat_report.txt	15-Feb-2013 04:51	352K	

Apache/2.2.23 (Unix) mod_ssl/2.2.23 OpenSSL/1.0.0-fips DAV/2 mod_auth_passthrou
Port 80

Save pig-0.11.0-tar.gz file

Untar pig-0.11.0-tar.gz file



Configure pig-env.sh file

Create pig-env.sh file in PIG_HOME/conf

Add the following entries in PIG_HOME/conf/pig-env.sh file

```
export JAVA_HOME=/usr
export PIG_HOME=/home/neeraj/local_cluster_home/pig-0.11.0
export HADOOP_HOME=/home/neeraj/local_cluster_home/hadoop-1.0.3
export PIG_CLASSPATH=$HADOOP_HOME/conf/
```

Configure pig.properties file

Add the following entries in PIG_HOME/conf/pig.properties file

```
fs.default.name=hdfs://localhost:9000
mapred.job.tracker=localhost:9001
```

Copy core-site.xml, hdfs-site.xml & mapred-site.xml file from HADOOP_HOME/conf to PIG_HOME/conf

Start your Hadoop

```
Terminal
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ pwd
/home/neeraj/local_cluster_home/hadoop-1.0.3/bin
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ ./start-all.sh
starting namenode, logging to /home/neeraj/local_cluster_home/hadoop-1.0.3/logs/hadoop-neeraj-namenode-
ubuntu.out
myubuntu: starting datanode, logging to /home/neeraj/local_cluster_home/hadoop-1.0.3/logs/hadoop-neeraj
-datanode-ubuntu.out
myubuntu: starting secondarynamenode, logging to /home/neeraj/local_cluster_home/hadoop-1.0.3/logs/hado
op-neeraj-secondarynamenode-ubuntu.out
starting jobtracker, logging to /home/neeraj/local_cluster_home/hadoop-1.0.3/logs/hadoop-neeraj-jobtrac
ker-ubuntu.out
myubuntu: starting tasktracker, logging to /home/neeraj/local_cluster_home/hadoop-1.0.3/logs/hadoop-nee
raj-tasktracker-ubuntu.out
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$
```

Check Hadoop processes
&
Safemode

Make sure that safe mode is off before you start Pig

```
Terminal
neeraj@ubuntu:~/local_cluster_home/hadoop-1.2.1/bin$ pwd
/home/neeraj/local_cluster_home/hadoop-1.2.1/bin
neeraj@ubuntu:~/local_cluster_home/hadoop-1.2.1/bin$ jps
3915 TaskTracker
4178 Jps
3610 SecondaryNameNode
3399 DataNode
3159 NameNode
3701 JobTracker
neeraj@ubuntu:~/local_cluster_home/hadoop-1.2.1/bin$ ./hadoop dfsadmin -safemode get
Safe mode is OFF
neeraj@ubuntu:~/local_cluster_home/hadoop-1.2.1/bin$
```

Start Pig shell

```
Terminal
neeraj@ubuntu:~/local_cluster_home/pig-0.11.0/bin$ pwd
/home/neeraj/local_cluster_home/pig-0.11.0/bin
neeraj@ubuntu:~/local_cluster_home/pig-0.11.0/bin$ ./pig
2013-03-12 19:22:30,454 [main] INFO  org.apache.pig.Main - Apache Pig version 0.11.0 (r1446324) compil
ed Feb 14 2013, 16:40:57
2013-03-12 19:22:30,455 [main] INFO  org.apache.pig.Main - Logging error messages to: /home/neeraj/loc
al_cluster_home/pig-0.11.0/bin/pig_1363096350450.log
2013-03-12 19:22:30,530 [main] INFO  org.apache.pig.impl.util.Utils - Default bootup file /home/neeraj
/.pigbootup not found
2013-03-12 19:22:30,785 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine -
Connecting to hadoop file system at: hdfs://localhost:9000
2013-03-12 19:22:31,340 [main] INFO  org.apache.pig.backend.hadoop.executionengine.HExecutionEngine -
Connecting to map-reduce job tracker at: localhost:9001
grunt>
grunt>
grunt>
grunt>
```

Input file for Pig

```
Terminal
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ pwd
/home/neeraj/local_cluster_home/hadoop-1.0.3/bin
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ ./hadoop fs -ls /
Found 3 items
drwxr-xr-x   - neeraj supergroup          0 2013-07-25 22:36 /hbase
drwxr-xr-x   - neeraj supergroup          0 2013-07-24 21:59 /home
drwxr-xr-x   - neeraj supergroup          0 2013-08-11 11:55 /pig_input_files
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ ./hadoop fs -ls /pig_input_files
Found 1 items
-rw-r--r--   1 neeraj supergroup         86 2013-08-11 11:55 /pig_input_files/temprature.txt
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ ./hadoop fs -cat /pig_input_files/temprature.txt
2001    43
2001    42
2001    9999
2002    47
2002    45
2002    9999
2002    49
2003    9999
2003    32
2003    35
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$
```

Access HDFS from Pig shell

```
Terminal
File  Edit  View  Search  Terminal  Help
grunt> fs -ls /
Found 3 items
drwxr-xr-x   - neeraj supergroup          0 2014-04-25 08:18 /home
drwxr-xr-x   - neeraj supergroup          0 2014-04-27 17:55 /pig_input_files
drwxr-xr-x   - neeraj supergroup          0 2014-04-25 08:42 /test_dir
grunt> fs -copyFromLocal /home/neeraj/PDF/Pig/temprature.txt /pig_input_files/
grunt> fs -ls /pig_input_files
Found 1 items
-rw-r--r--   1 neeraj supergroup         86 2014-04-27 17:56 /pig_input_files/temprature.txt
grunt> fs -cat /pig_input_files/temprature.txt
2001    43
2001    42
2001    9999
2002    47
2002    45
2002    9999
2002    49
2003    9999
2003    32
2003    35
grunt>
```

Execute Pig query

```
records = LOAD '/pig_input_files/temprature.txt' AS (year:chararray, temperature:int);

filtered_records = FILTER records BY temperature != 9999;

grouped_records = GROUP filtered_records BY year;

max_temp = FOREACH grouped_records GENERATE group,MAX(filtered_records.temperature);

DUMP max_temp;
```

Terminal

```
grunt> records = LOAD '/pig_input_files/temprature.txt' AS (year:chararray, temperature:int);
grunt>
grunt> filtered_records = FILTER records BY temperature != 9999;
grunt>
grunt> grouped_records = GROUP filtered_records BY year;
grunt>
grunt> max_temp = FOREACH grouped_records GENERATE group, MAX(filtered_records.temperature);
grunt>
grunt> DUMP max_temp;
```

Execute Pig query

records = LOAD '/pig_input_files/temprature.txt' AS (year:chararray, temperature:int);

filtered_records = FILTER records BY temperature != 9999;

grouped_records = GROUP filtered_records BY year;

max_temp = FOREACH grouped_records GENERATE group,MAX(filtered_records.temperature);

STORE max_temp INTO '/pig_output_files';

```
Terminal
grunt> records = LOAD '/pig_input_files/temprature.txt' AS (year:chararray, temperature:int);
grunt>
grunt> filtered_records = FILTER records BY temperature != 9999;
grunt>
grunt> grouped_records = GROUP filtered_records BY year;
grunt>
grunt> max_temp = FOREACH grouped_records GENERATE group, MAX(filtered_records.temperature);
grunt>
grunt> STORE max_temp INTO '/pig_output_files';
```

Pig job details

```
Terminal

HadoopVersion   PigVersion      UserId  StartedAt       FinishedAt      Features
1.0.3   0.11.0  neeraj  2013-08-11 12:21:43     2013-08-11 12:22:30     GROUP_BY,FILTER

Success!

Job Stats (time in seconds):
JobId   Maps    Reduces MaxMapTime      MinMapTIme      AvgMapTime      MedianMapTime   MaxReduceTime
ReduceTime      MedianReducetime        Alias   Feature Outputs
job_201308111151_0003   1       1       6       6       6       6       15      15      15      15
ouped_records,max_temp,records  GROUP_BY,COMBINER       /pig_output_files,

Input(s):
Successfully read 10 records (457 bytes) from: "/pig_input_files/temprature.txt"

Output(s):
Successfully stored 3 records (24 bytes) in: "/pig_output_files"

Counters:
Total records written : 3
Total bytes written : 24
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0
```

Output of Pig query

```
Terminal
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ pwd
/home/neeraj/local_cluster_home/hadoop-1.0.3/bin
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ ./hadoop fs -ls /
Found 5 items
drwxr-xr-x   - neeraj supergroup          0 2013-07-25 22:36 /hbase
drwxr-xr-x   - neeraj supergroup          0 2013-07-24 21:59 /home
drwxr-xr-x   - neeraj supergroup          0 2013-08-11 11:55 /pig_input_files
drwxr-xr-x   - neeraj supergroup          0 2013-08-11 12:22 /pig_output_files
drwxr-xr-x   - neeraj supergroup          0 2013-08-11 12:10 /tmp
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ ./hadoop fs -ls /pig_output_files
Found 3 items
-rw-r--r--   1 neeraj supergroup          0 2013-08-11 12:22 /pig_output_files/_SUCCESS
drwxr-xr-x   - neeraj supergroup          0 2013-08-11 12:21 /pig_output_files/_logs
-rw-r--r--   1 neeraj supergroup         24 2013-08-11 12:22 /pig_output_files/part-r-00000
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$ ./hadoop fs -cat /pig_output_files/part-r-00000
2001    43
2002    49
2003    35
neeraj@ubuntu:~/local_cluster_home/hadoop-1.0.3/bin$
```

Exit from Pig shell

```
Terminal
File  Edit  View  Search  Terminal  Help
grunt> fs -ls /
Found 3 items
drwxr-xr-x   - neeraj supergroup          0 2014-04-25 08:18 /home
drwxr-xr-x   - neeraj supergroup          0 2014-04-27 17:56 /pig_input_files
drwxr-xr-x   - neeraj supergroup          0 2014-04-25 08:42 /test_dir
grunt> quit
neeraj@myubuntu:~/local_cluster_home/pig-0.11.0/bin$
```

HBase/Hive/Pig

Features	HBase	Hive	Pig
Unstructured data	Yes	No	No
Data editing	Yes	No	No
Versioned data	Yes	No	No
Key-Value concept	Yes	No	No
Column-family/qualifier	Yes	No	No
Tables	Yes	Yes	No
Indexes	No	Yes	No
Order by/Group by	No	Yes	Yes
Join	No	Yes	Yes
UDF	No	Yes	Yes

HBase/Hive/Pig suitability

HBase is suitable when...

When you need to handle unstructured data
When you need to edit the data
When you need versioned data

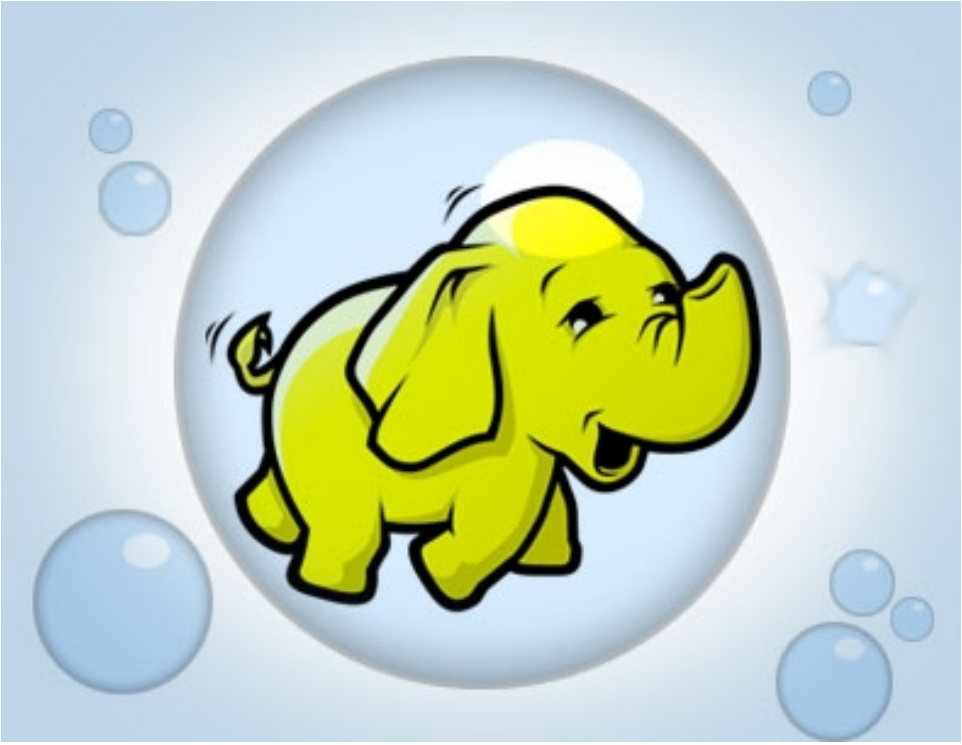
Hive is suitable when...

When you need to handle structured data
When you don't need to edit the data
When you comfortable in SQL syntax

Pig is suitable when...

When you need to handle structured data
When you don't need to edit the data
When you are comfortable in scripting

...Thanks...



For online Hadoop training, send mail to neeraj.ymca.2k6@gmail.com