

Cài đặt Apache Pig

1. Download Apache Pig

wget <https://dlcdn.apache.org/pig/pig-0.17.0/pig-0.17.0.tar.gz>

2. Giải nén và đổi tên thư mục pig

```
$ tar -xzf pig-0.17.0.tar.gz
```

```
$ mv pig-0.17.0 pig
```

```
hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop-3.3.2.tar.gz      output_dir      sample.txt  units.jar
db-derby-10.15.2.0-bin.tar.gz  hadoop-core-1.2.1.jar    pig             test.sh
derby                        hadoop-core-1.2.1.jar.zip  pig-0.17.0.tar.gz  tmp
hadoop                      hive                     ProcessUnits.java  units
hadoopminhchau@minhchau-master:~$
```

3. Khai báo biến môi trường

```
$ export PIG_HOME=/home/hadoopminhchau/pig
```

```
$ export PATH=$PATH:$PIG_HOME/bin
```

```
$ export PIG_CLASSPATH=$HADOOP_CONF_DIR
```

4. Check Pig

```
$ pig -help
```

```

-h, -help - Display this message. You can specify topic to get help for that topic.
properties is the only topic currently supported: -h properties.
-i, -version - Display version information
-l, -logfile - Path to client side log file; default is current working directory.
-m, -param_file - Path to the parameter file
-p, -param - Key value pair of the form param=val
-r, -dryrun - Produces script with substituted parameters. Script is not executed.
-t, -optimizer_off - Turn optimizations off. The following values are supported:
    ConstantCalculator - Calculate constants at compile time
    SplitFilter - Split filter conditions
    PushUpFilter - Filter as early as possible
    MergeFilter - Merge filter conditions
    PushDownForeachFlatten - Join or explode as late as possible
    LimitOptimizer - Limit as early as possible
    ColumnMapKeyPrune - Remove unused data
    AddForEach - Add ForEach to remove unneeded columns
    MergeForEach - Merge adjacent ForEach
    GroupByConstParallelSetter - Force parallel 1 for "group all" statement
    PartitionFilterOptimizer - Pushdown partition filter conditions to loader implementing LoadMetadata
    PredicatePushdownOptimizer - Pushdown filter predicates to loader implementing LoadPredicatePushDown
    All - Disable all optimizations
All optimizations listed here are enabled by default. Optimization values are case insensitive.
-v, -verbose - Print all error messages to screen
-w, -warning - Turn warning logging on; also turns warning aggregation off
-x, -exectype - Set execution mode: local|mapreduce|tez, default is mapreduce.
-F, -stop_on_failure - Aborts execution on the first failed job; default is off
-M, -no_multiquery - Turn multiquery optimization off; default is on
-N, -no_fetch - Turn fetch optimization off; default is on
-P, -propertyFile - Path to property file
-printCmdDebug - Overrides anything else and prints the actual command used to run Pig, including any environment variables that are set by the pig command.
2022-04-05 04:02:57,683 INFO pig.Main: Pig script completed in 75 milliseconds (75 ms)
hadoopminhchau@minhchau-master:~$ _

```

To direct input to this VM, click inside or press Ctrl+G.

5. Thực thi Pig

	Local Mode	Tez Local Mode	Spark Local Mode	MapReduce Mode	Tez Mode	Spark Mode
Interactive Mode	Yes	Experimental	Yes	Yes		
Batch Mode	Yes	Experimental	Yes	Yes		

Pig có 06 chế độ thực thi

- **Local Mode**

- Chạy trên máy đơn. Tất cả các tập tin được cài đặt và chạy trên hệ thống file và máy chủ cục bộ.

```
$ pig -x local
```

- **Tez Local Mode**

- Tương tự chế độ cục bộ, ngoại trừ bên trong Pig sẽ gọi công cụ thực thi Tez.

```
$ pig -x tez_local
```

- **Spark Local Mode**

- Tương tự chế độ cục bộ, ngoại trừ bên trong Pig sẽ gọi công cụ thực thi Spark.

```
$ pig -x spark local
```

- **Map Reduce Mode**

- Chạy ở chế độ mapreduce, cần truy cập cụm Hadoop và HDFS. Đây là chế độ chạy mặc định, có thể không cần cung cấp thông số -x

```
$ pig
```

Hoặc

```
$ pig -x mapreduce
```

- **Tez Mode**

- Cần truy cập cụm Hadoop và hệ thống HDFS

```
$ pig -x tez
```

- **Spark Mode**

- Cần truy cập Spark, cụm Yarn hoặc Mesos và hệ thống file HDFS. Pig Script chạy trên Spark có thể tận dụng tính năng cấp phát động (dynamic allocation). Tính năng này có thể được kích hoạt bằng cách bật `spark.dynamicAllocation.enabled`.

```
$ pig -x spark
```

6. Interactive Mode

- Có thể chạy Pig ở chế độ Interactive sử dụng Grunt Shell. Gọi Grunt Shell sử dụng lệnh Pig và nhập các câu lệnh Pig Latin và thực thi.

Ví dụ sau lấy tất cả các ID người dùng từ file `/etc/passwd`

[Copy file passwd ra thư mục làm việc cục bộ](#)

```
$ cp /etc/passwd /home/hadoopminhchau
```

```
hadoopminhchau@minhchau-master:~$ cp /etc/passwd /home/hadoopminhchau/
hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop-core-1.2.1.jar  pig  sample.txt
db-derby-10.15.2.0-bin.tar.gz  hadoop-core-1.2.1.jar.zip  pig-0.17.0.tar.gz  test.sh
derby  hive  pig_1650382433801.log  tmp
hadoop  output_dir  pig_1650382485427.log  units
hadoop-3.3.2.tar.gz  passwd  ProcessUnits.java  units.jar
hadoopminhchau@minhchau-master:~$
```

Khởi động Grunt Shell (trong cục bộ hoặc hadoop) và nhập các câu lệnh Pig trực tiếp

```
$ grunt> A = load '/home/hadoopminhchau/passwd' using
```

```
PigStorage(':');
```

```
$ grunt> B = foreach A generate $0 as id;
```

```
$ grunt> dump B;
```

```
hadoopminhchau@minhchau-master:~$ pig -x local
2022-04-19 16:26:00,596 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2022-04-19 16:26:00,596 INFO pig.ExecTypeProvider: Picked LOCAL as the ExecType
2022-04-19 16:26:00,638 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) c
iled Jun 02 2017, 15:41:58
2022-04-19 16:26:00,638 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoo
nhchau/pig_1650385560636.log
2022-04-19 16:26:00,655 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/h
opminhchau/.pigbootup not found
2022-04-19 16:26:00,735 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.job
racker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-19 16:26:00,737 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngi
- Connecting to hadoop file system at: file:///
2022-04-19 16:26:00,808 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2022-04-19 16:26:00,822 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PI
efault-97cd2578-2b38-46e3-91b3-f03b9abf487f
2022-04-19 16:26:00,822 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeli
service.enabled set to false
grunt> A = load '/home/hadoopminhchau/passwd' using PigStorage(':');
2022-04-19 16:26:17,649 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - io.bytes.
.checksum is deprecated. Instead, use dfs.bytes-per-checksum
grunt> B = foreach A generate $0 as id;
grunt> dump B;
```

```

Total input paths to process : 1
(root)
(daemon)
(bin)
(sys)
(sync)
(games)
(man)
(lp)
(mail)
(news)
(uucp)
(proxy)
(www-data)
(backup)
(list)
(irc)
(gnats)
(nobody)
(_apt)
(systemd-timesync)
(systemd-network)
(systemd-resolve)
(messagebus)
(pollinate)
(sshd)
(syslog)
(uuid)
(tcpdump)
(tss)
(landscape)
(usbmux)
(systemd-coredump)
(minhchau)
(lxd)
(hadoopminhchau)
grunt>

```

7. Batch Mode

Có thể chạy Pig ở chế độ batch mode sử dụng Pig Script và chạy pig ở chế độ cục bộ hoặc Hadoop

Ví dụ: chạy lại ví dụ trên ở chế độ Batch Mode

Tạo file Pig Script id.pig

\$ vim id.pig

```
hadoopminhchau@minhchau-master:~$ vim id.pig
```

```

/* id.pig */

A = load 'passwd' using PigStorage(':'); -- load the passwd file
B = foreach A generate $0 as id; -- extract the user IDs
store B into 'id.out'; -- write the results to a file name id.out

```

Thực thi file script

\$ pig -x local id.pig

```

Success!

Job Stats (time in seconds):
JobId      Maps      Reduces  MaxMapTime    MinMapTime    AvgMapTime    MedianMapTime  MaxReduce
e          MinReduceTime  AvgReduceTime  MedianReductime  Alias        Feature  Outputs
job_local1540184028_0001      1      0      n/a      n/a      n/a      n/a      0      0
A,B        MAP_ONLY      file:///home/hadoopminhchau/id.out,

Input(s):
Successfully read 35 records from: "file:///home/hadoopminhchau/passwd"

Output(s):
Successfully stored 35 records in: "file:///home/hadoopminhchau/id.out"

Counters:
Total records written : 35
Total bytes written : 0
Spillable Memory Manager spill count : 0
Total bags proactively spilled: 0
Total records proactively spilled: 0

Job DAG:
job_local1540184028_0001

2022-04-19 16:38:24,852 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTrack
metrics system already initialized!
2022-04-19 16:38:24,855 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTrack
metrics system already initialized!
2022-04-19 16:38:24,857 [main] WARN org.apache.hadoop.metrics2.impl.MetricsSystemImpl - JobTrack
metrics system already initialized!
2022-04-19 16:38:24,870 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
pReduceLauncher - Success!
2022-04-19 16:38:24,888 [main] INFO org.apache.pig.Main - Pig script completed in 2 seconds and
milliseconds (2258 ms)
hadoopminhchau@minhchau-master:~$ _

```

Xuất kết quả

```

hadoopminhchau@minhchau-master:~$ ls
apache-hive-3.1.3-bin.tar.gz  hadoop-core-1.2.1.jar.zip  pig          test.sh
db-derby-10.15.2.0-bin.tar.gz  hive                      pig-0.17.0.tar.gz  tmp
derby                          id.out                   pig_1650382433801.log  units
hadoop                         id.pig                  pig_1650382485427.log  units.jar
hadoop-3.3.2.tar.gz           output_dir               ProcessUnits.java
hadoop-core-1.2.1.jar         passwd                   sample.txt
hadoopminhchau@minhchau-master:~$ cd id.out/
hadoopminhchau@minhchau-master:~/id.out$ ls
part-m-00000  _SUCCESS
hadoopminhchau@minhchau-master:~/id.out$ cd
hadoopminhchau@minhchau-master:~$ cat id.out/part-m-00000

```

```
hadoopminhchau@minhchau-master:~$ cat id.out/part-m-00000
root
daemon
bin
sys
sync
games
man
lp
mail
news
uucp
proxy
www-data
backup
list
irc
gnats
nobody
_apt
systemd-timesync
systemd-network
systemd-resolve
messagebus
pollinate
sshd
syslog
uidd
tcpdump
tss
landscape
usbmux
systemd-coredump
minhchau
lxd
hadoopminhchau
hadoopminhchau@minhchau-master:~$
```

8. Chạy Pig trên mapreduce

```
$ hadoop fs -mkdir pigdata
```

```
$ hadoop fs -put /home/hadoopminhchau/passwd pigdata
```

```
$ pig
```

```
grunt> A = load './pigdata/passwd' using PigStorage (':');
```

```
grunt> B = foreach A generate $0 as id;
```

```
grunt> dump B;
```



```

hadoopminhchau@minhchau-master:~$ hadoop fs -mkdir pigdata
hadoopminhchau@minhchau-master:~$ hadoop fs -put /home/hadoopminhchau/passwd pigdata
hadoopminhchau@minhchau-master:~$ pig
2022-04-19 16:45:28,814 INFO pig.ExecTypeProvider: Trying ExecType : LOCAL
2022-04-19 16:45:28,815 INFO pig.ExecTypeProvider: Trying ExecType : MAPREDUCE
2022-04-19 16:45:28,815 INFO pig.ExecTypeProvider: Picked MAPREDUCE as the ExecType
2022-04-19 16:45:28,862 [main] INFO org.apache.pig.Main - Apache Pig version 0.17.0 (r1797386) compiled Jun 02 2017, 15:41:58
2022-04-19 16:45:28,862 [main] INFO org.apache.pig.Main - Logging error messages to: /home/hadoopminhchau/pig_1650386728851.log
2022-04-19 16:45:28,882 [main] INFO org.apache.pig.impl.util.Utils - Default bootup file /home/hadoopminhchau/.pigbootup not found
2022-04-19 16:45:29,134 [main] INFO org.apache.hadoop.conf.Configuration.deprecation - mapred.jobtracker is deprecated. Instead, use mapreduce.jobtracker.address
2022-04-19 16:45:29,134 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to hadoop file system at: hdfs://minhchau-master:9000
2022-04-19 16:45:29,507 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngine - Connecting to map-reduce job tracker at: minhchau-master:9001
2022-04-19 16:45:29,521 [main] INFO org.apache.pig.PigServer - Pig Script ID for the session: PIG-default-f4b14b41-a01e-45a1-82a0-89ea8d1f8093
2022-04-19 16:45:29,521 [main] WARN org.apache.pig.PigServer - ATS is disabled since yarn.timeline-service.enabled set to false
grunt> A = load './pigdata/passwd' using PigStorage(':');
grunt> B = foreach A generate $0 as id;
grunt> dump B;

```

```

Total input paths to process : 1
(root)
(daemon)
(bin)
(sys)
(sync)
(games)
(man)
(lp)
(mail)
(news)
(uucp)
(proxy)
(www-data)
(backup)
(list)
(irc)
(gnats)
(nobody)
(_apt)
(systemd-timesync)
(systemd-network)
(systemd-resolve)
(messagebus)
(pollinate)
(sshd)
(syslog)
(uuid)
(tcpcdump)
(tss)
(landscape)
(usbmux)
(systemd-coredump)
(minhchau)
(lxd)
(hadoopminhchau)
grunt> _

```