Cài đặt Apache Pig

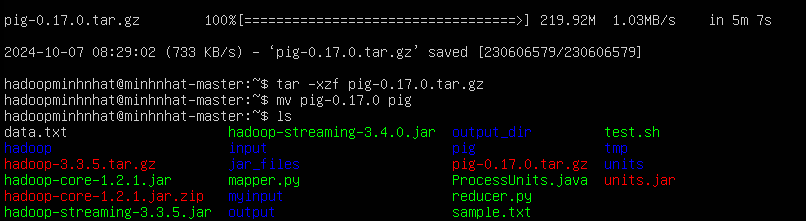
1. Download Apache Pig

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

1. Giải nén và đổi tên thư mục pig

$ tar -xzf pig-0.17.0.tar.gz

$ mv pig-0.17.0 pig



1. Khai báo biến môi trường

$ export PIG\_HOME=/home/hadoopminhnhat/pig

$ export PATH=$PATH:$PIG\_HOME/bin

$ export PIG\_CLASSPATH=$HADOOP\_CONF\_DIR



1. Check Pig

$ pig -help



1. 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.

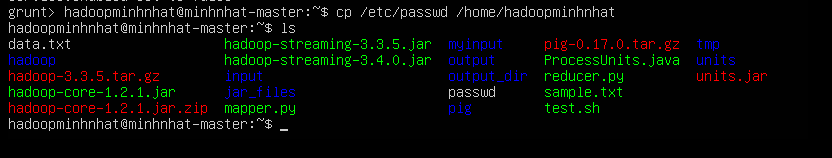
$ pig -x spark

1. 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/hadoopminhnhat

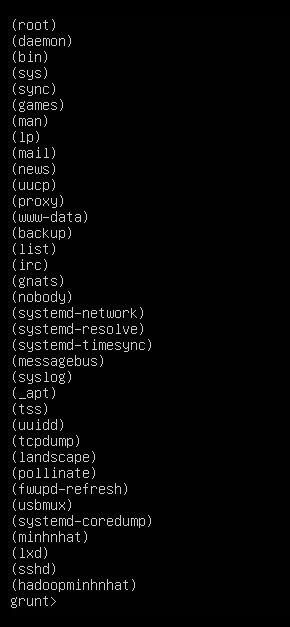


Khởi động Grunt Shell (trong cục bộ hoặc hadoop)và nhập các câu lệnh Pig trực tiếp

$ grunt> A = load '/home/hadoopminhnhat/passwd' using PigStorage(':');

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

$ grunt> dump B;



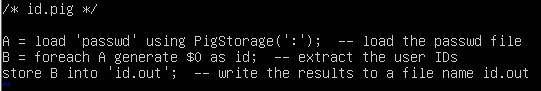
1. Batch Mode

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

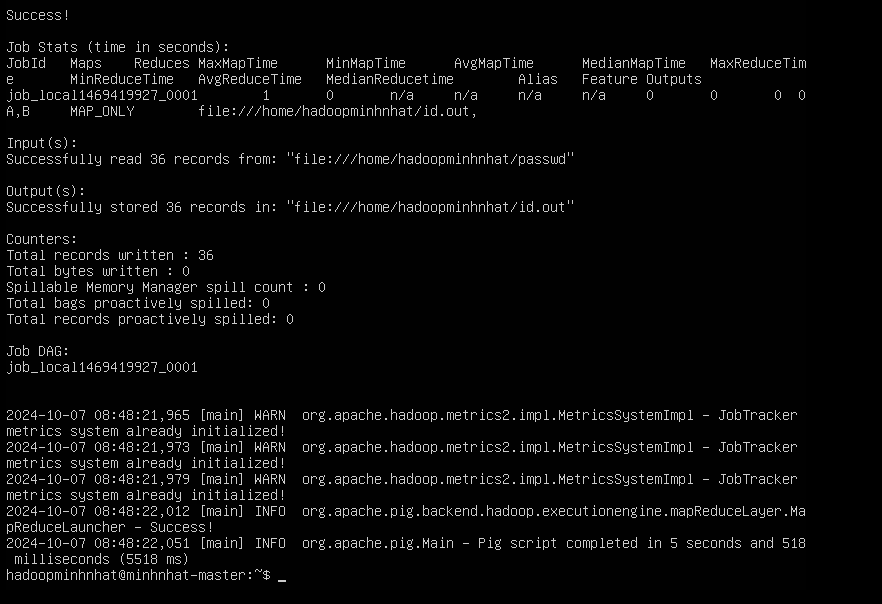
# Tạo file Pig Script id.pig

$ vim id.pig



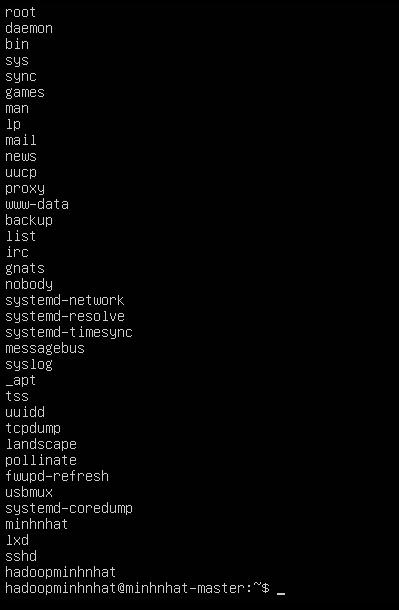
# Thực thi file script

$ pig -x local id.pig



# Xuất kết quả

# 



1. Chạy Pig trên mapreduce

$ hadoop fs -mkdir pigdata

$ hadoop fs -put /home/hadoopminhnhat/passwd pigdata

$ pig

grunt> A = load './pigdata/passwd' using PigStorage (':'); grunt> B = foreach A generate $0 as id;

grunt> dump B;

