**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HCM KHOA CÔNG NGHỆ THÔNG TIN**



**ỨNG DỤNG HADOOP THỰC HIỆN XÂY DỰNG HỆ THỐNG PHÂN TÍCH CẠNH TRANH THỊ TRƯỜNG DỰA TRÊN DỮ LIỆU GIÁ VÀ KHUYẾN MÃI SẢN PHẨM LAPTOP**

**SVTH: Phan Trọng Phú - 23133056**

**GVHD: Trần Quang Khải**

**HK1, NĂM 2025**

* 1. PHẦN MỞ ĐẦU

1. Cài đặt các file cấu hình
2. Cài hadoop: wget <https://dlcdn.apache.org/hadoop/common/hadoop-3.3.6/hadoop-3.3.6.tar.gz>
3. Cài jdk8: sudo apt install openjdk-8-jdk -y
4. Cài ssh : sudo apt install openssh-server -y
5. Cài đặt mysql : sudo apt install mysql-server -y
6. Cài đặt JDBC: wget <https://dev.mysql.com/get/Downloads/Connector-J/mysql-connector-j-8.0.33.tar.gz>
7. Minh chứng các phần cài đặt :
8. Hadoop

1.1 Lệnh cài đặt

A screen shot of a computer screen

AI-generated content may be incorrect.

* 1. Minh chứng hoàn thành

A computer screen shot of a program

AI-generated content may be incorrect.

Khi giải nén xong ta sẽ thấy được folder của hadoop

A screenshot of a computer

AI-generated content may be incorrect.

* 1. Kiểm tra version của hadoop

A computer screen with white text

AI-generated content may be incorrect.

1. Jdk 8
   1. Lệnh cài đặt

A screen shot of a computer

AI-generated content may be incorrect.

* 1. Minh chứng hoàn thành

A screen shot of a computer

AI-generated content may be incorrect.

* 1. Kiểm tra version của java

A black background with white text

AI-generated content may be incorrect.

1. SSH

3.1 Lệnh cài ssh

A screenshot of a computer

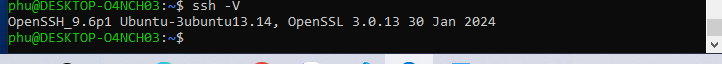
AI-generated content may be incorrect.

* 1. Minh chứng hoàn thành

A screenshot of a computer screen

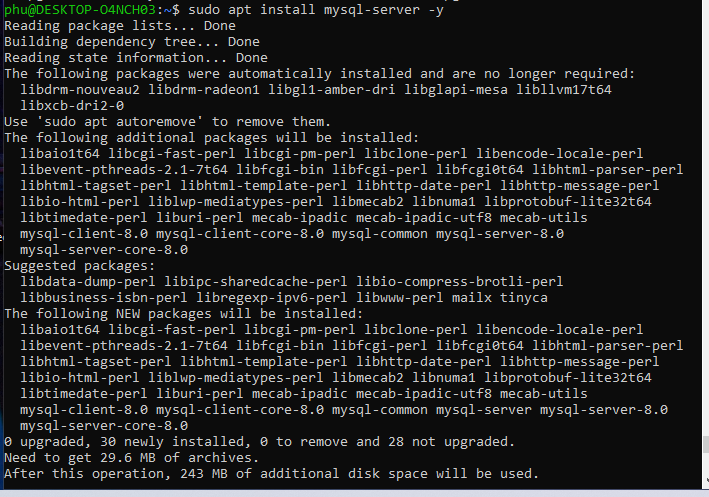
AI-generated content may be incorrect.

* 1. Kiểm tra version của ssh



1. Mysql

4.1 Lệnh cài đặt



* 1. Minh chứng hoàn thành

A screenshot of a computer program

AI-generated content may be incorrect.

* 1. Kiểm tra version

A computer screen with white text

AI-generated content may be incorrect.

1. JDBC
   1. Lệnh cài đặt

A computer screen shot of a program

AI-generated content may be incorrect.

* 1. Minh chứng hoàn thành

A computer screen shot of a program

AI-generated content may be incorrect.

1. SQOOP

71. Cài đặt và minh chứng thành công

A computer screen with white text

AI-generated content may be incorrect.

7.2 Kiểm tra version

Phần cấu hình hadoop

Chỉnh lại các file hệ thống core-site

A screenshot of a computer screen

AI-generated content may be incorrect.

File hdfs-site

A screenshot of a computer program

AI-generated content may be incorrect.

Chỉnh lại file mapred-stie

A screenshot of a computer

AI-generated content may be incorrect.

Chỉnh lại file yarn-site

A screenshot of a computer

AI-generated content may be incorrect.

Chỉnh lại export JAVA\_HOME trong file hadoop-env.sh

A white background with black text

AI-generated content may be incorrect.

Cấu hình ssh

1. Tạo keygen không mật khẩu

A computer screen with white text

AI-generated content may be incorrect.

1. Thêm khóa công khai vừa tạo



1. Đặt quyền truy cập



1. Kiểm tra kết nối

A screenshot of a computer screen

AI-generated content may be incorrect.

Định dạng hdfs format và tiến hành cài đặt

A screen shot of a computer screen

AI-generated content may be incorrect.

Kiểm tra cấu hình hệ thống

A screenshot of a computer

AI-generated content may be incorrect.

Phần cấu hình SQOOP

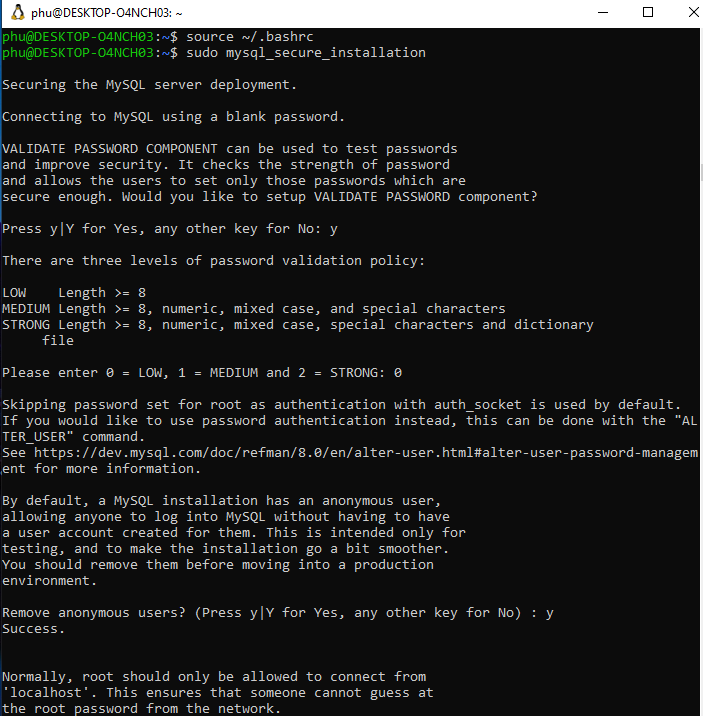
Cài đặt như phần trên

Đặt cấu hình trên bashrc

A screenshot of a computer screen

AI-generated content may be incorrect.

Chỉnh sửa các cấu hình trong mysql



Chỉnh lại lắng nghe mọi port

A screenshot of a computer

AI-generated content may be incorrect.

Di chuyển thư mục đó vào trong lib

A screen shot of a computer

AI-generated content may be incorrect.

Chỉnh file cấu hình SQOOP

A screenshot of a computer screen

AI-generated content may be incorrect.

Copy cấu hình JDBC vào liv của SQOOP

A computer code on a black background

AI-generated content may be incorrect.

Tải coomon lang

A computer screen with white text

AI-generated content may be incorrect.

Copy jarr vào lib

A computer screen with text

AI-generated content may be incorrect.

PHẦN VỀ THỰC HIỆN CẤU HÌNH PIG

Thực hiện cài đặt

A computer screen with white text

AI-generated content may be incorrect.

Di chuyển và thực hiện việc cấp quyền



Chỉnh lại file bashrc

A screenshot of a computer

AI-generated content may be incorrect.

Sao chép file cấu hỉnh mẫu sang file cấu hình hiện tại

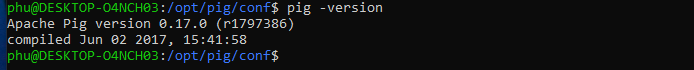
A computer screen with yellow and black text

AI-generated content may be incorrect.

Cấp quyền thực thi



Kiểm tra version của pig



PHẦN VỀ CẤU HÌNH HIVE

1 Tải hive : wget <https://archive.apache.org/dist/hive/hive-3.1.3/apache-hive-3.1.3-bin.tar.gz>

A computer screen shot of a black screen

AI-generated content may be incorrect.

1. Giải nén và thực hiện di chuyển

A screen shot of a computer program

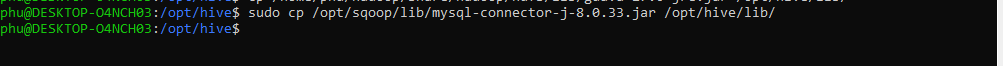
AI-generated content may be incorrect.

Thực hiện xóa gua cũ và thay guava mới

A computer screen with white text

AI-generated content may be incorrect.

1. Cấu hình metastore



Tạo cấu hình metastore



Tạo file cấu hình hive-site.xml

A screenshot of a computer program

AI-generated content may be incorrect.

Nội dung :

<configuration>

<property>

<name>javax.jdo.option.ConnectionURL</name>

<value>jdbc:mysql://localhost/hive\_metastore?createDatabaseIfNotExist=true&amp;useSSL=false&amp;autoReconnect=true</value>

<description>JDBC connect string for a JDBC metastore</description>

</property>

<property>

<name>javax.jdo.option.ConnectionDriverName</name>

<value>com.mysql.cj.jdbc.Driver</value>

<description>Driver class name for a JDBC metastore</description>

</property>

<property>

<name>javax.jdo.option.ConnectionUserName</name>

<value>root</value>

<description>Username for the database</description>

</property>

<property>

<name>javax.jdo.option.ConnectionPassword</name>

<value>Nhocta2207@</value>

<description>Password for the database</description>

</property>

<property>

<name>hive.metastore.warehouse.dir</name>

<value>/user/hive/warehouse</value>

<description>location of default database for the warehouse</description>

</property>

</configuration>

Khởi tạo metastore

A screenshot of a computer screen

AI-generated content may be incorrect.

Kết quả

A black background with white text

AI-generated content may be incorrect.

Tạo thư mục warehouse trên hdfs

A computer screen shot of a black screen

AI-generated content may be incorrect.

Đảm bảo trên thưu mục

A black background with white text

AI-generated content may be incorrect.

A computer screen shot of a black screen

AI-generated content may be incorrect.

Kiểm tra cấu hình hive

A computer screen shot of a computer code

AI-generated content may be incorrect.

Và kết quả khi show databse

A computer screen shot of a black screen

AI-generated content may be incorrect.

PHẦN VỀ THỰC HIỆN APAHCE DRILL

A screen shot of a computer code

AI-generated content may be incorrect.

Di chuyển vào file thư mục

A black background with white text

AI-generated content may be incorrect.

Cấp quyền



Chỉnh thêm trong file cấu hình bashrc

A screenshot of a computer

AI-generated content may be incorrect.

Cấp quyền và cấu hình tích hợp các dịch vụ hadoop vào trong drill

A screen shot of a computer code

AI-generated content may be incorrect.

Cài đặt zookeeper

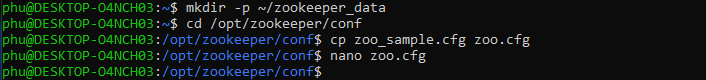
A screen shot of a computer

AI-generated content may be incorrect.

Cấp quyền



Tạo thư mục để tiến hành lưu dữ liệu



Code cấu hình file đó, chỉnh lại dataDir

A screenshot of a computer

AI-generated content may be incorrect.

Áp dụng các cấu hình bashrc

A screenshot of a computer

AI-generated content may be incorrect.

Khởi động dịch vụ zookeeper

A black screen with white text

AI-generated content may be incorrect.

Kiểm tra kết qua : /opt/drill/bin/drill-embedded

A computer screen with text on it

AI-generated content may be incorrect.

PHẦN TẢI VÀ CÀI ĐẶT VỀ HBASE

1. Tải và cài đặt

A computer screen with white text

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.

Thực hiện giải nén và đổi tên dễ nhớ

A screen shot of a computer

AI-generated content may be incorrect.

Cấu hình file bashrc

A screenshot of a computer

AI-generated content may be incorrect.

Cấu hình hbase

Chỉnh sửa hbase-env

A screenshot of a computer screen

AI-generated content may be incorrect.

Chỉnh hbase-site

A screenshot of a computer screen

AI-generated content may be incorrect.

Bỏ bình luận dòng này để tránh trùng classplath

A screenshot of a computer

AI-generated content may be incorrect.

Khởi chạy dịch vụ và kiểm tra jps

A screen shot of a computer program

AI-generated content may be incorrect.

Kiểm tra hbase shell

A screen shot of a computer screen

AI-generated content may be incorrect.

PHẦN VỀ CẤU HÌNH SPARK

1. Lệnh cài đặt

A screenshot of a computer program

AI-generated content may be incorrect.

1. Giải nén đổi tên và di chuyển vào thư mục

A screen shot of a computer screen

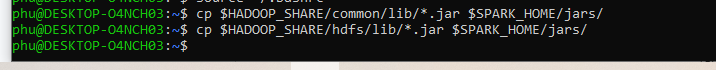
AI-generated content may be incorrect.

Chỉnh lại cấu hình bashrc

A screenshot of a computer

AI-generated content may be incorrect.

Sao chép các JAR của hadoop vào lib của SPARK



Sửa spark.env

A screenshot of a computer

AI-generated content may be incorrect.

Kiểm tra cấu hình cuối cùng

A screenshot of a computer program

AI-generated content may be incorrect.

PHẦN VỀ CẤU HÌNH PHOENIX

1. Các lệnh cài đặt

A screen shot of a computer code

AI-generated content may be incorrect.

1. Giải nén thư mục và tiến hành đổi tên, di chuyển

A computer screen with colorful text

AI-generated content may be incorrect.

1. Tích hợp Phoenix Server JAR vào HBASE

A computer screen with white text

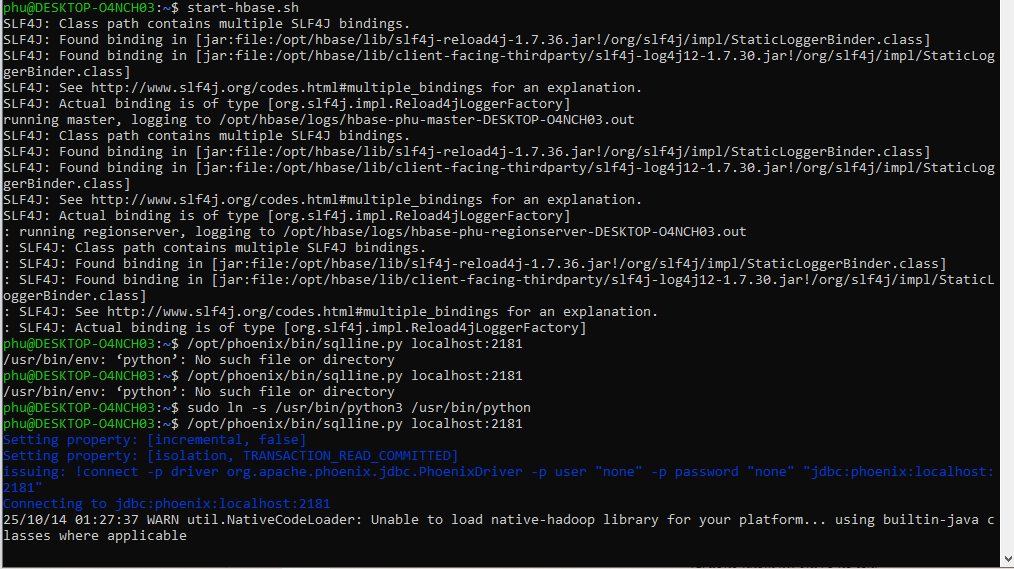
AI-generated content may be incorrect.

1. Cấu hình biến môi trường bashrc

A screenshot of a computer program

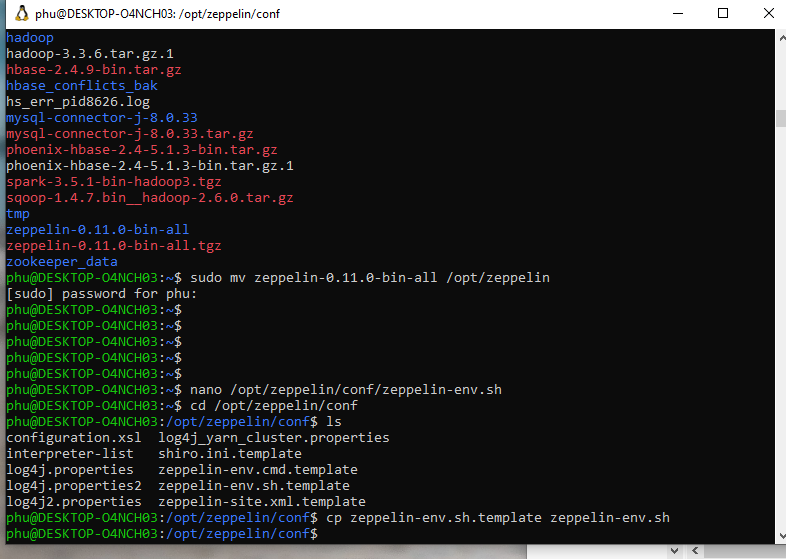
AI-generated content may be incorrect.

Khởi động lại hbase và kiểm tra hoạt động của phoenix



THỰC HIỆN PHẦN CÀI ZEPPELIN

1. Cài đặt và giải nén



1. Chỉnh sửa lại file zeppelin-env

A screenshot of a computer

AI-generated content may be incorrect.

1. Thêm các cấu hình bashrc

A computer screen shot of a computer screen

AI-generated content may be incorrect.

1. Khởi động thử các dịch vụ

A computer screen shot of a computer screen

AI-generated content may be incorrect.

1. Thêm các JAR cuẩ JDBC vào

A screenshot of a computer

AI-generated content may be incorrect.

Tiến hành kiểm tra trên webUI

A screenshot of a computer

AI-generated content may be incorrect.

Sau khi cài xong tất cả tôi sẽ có các dịch vụ chạy webUI ở các port sau :

HDFS: 9870

YARN:8088

HBASE:16010

ZEPPELIN : 8080

HIVE: 10002

DRILL: 8047

Các lệnh để khởi chạy

1. HDFS và YARN : start-all.sh
2. HBASE VÀ Phoenix: /opt/hbase/bin/start-hbase.sh
3. DRILL : cd /opt/drill và ./bin/drill-embedded
4. HIVE : cd /opt/hive và ./bin/hive --service metastore & và ./bin/hive --service hiveserver2 &

Truy cập hive : ./bin/hive

5: PIG : cd /opt/pig và ./bin/pig ten\_script\_cua\_ban.pig và mở sell : ./bin/pig

THỰC HIỆN PHẦN VỀ CÀO DỮ LIỆU

THỰC HIỆN VIỆC ĐƯA DỮ LIỆU LÊN HDFS

1. Tạo 1 thưu mục gốc trên hdfs

1.1 Tạo thư mục cha là raw\_data



* 1. Tạo thư mục chứa các sản phẩm



* 1. Tạo thư mục chứa các khuyến mãi



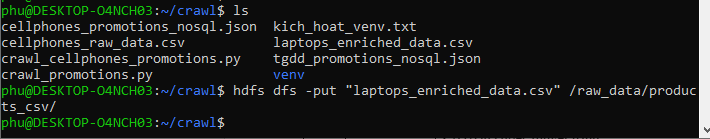
1. Đưa các file vào các thư mục đã có trên hdfs

2.1 Put dữ liệu của cellphoneS lên trước

A screen shot of a computer

AI-generated content may be incorrect.

* 1. Put dữ liệu của thegioididong



1. Đưa các file JSON lên hdfs

3.1 Dữ liệu khuyến mãi của CellPhoneS



3.2 Dữ liệu khuyến mãi của TheGioiDiDong



1. Kiểm tra xác nhận lại

A computer screen shot of a code

AI-generated content may be incorrect.

Kiểm tra xác nhận trên web

A screenshot of a computer

AI-generated content may be incorrect.

Và 2 file JSON

A screenshot of a computer

AI-generated content may be incorrect.

PHẦN VỀ DÙNG APACHE PIG ĐỂ TIẾN HÀNH XỬ LÝ DỮ LIỆU TRƯỚC KHI DÙNG SQOOP ĐẨY VÀO MYSQL

1. Chuẩn bị Môi trường và File
   1. Tạo file pig\_script để tạo môi trường làm việc

A computer screen shot of a program code

AI-generated content may be incorrect.

* 1. Tạo file mysql\_brand\_etl.pig

A computer screen with white text

AI-generated content may be incorrect.

Code của toàn bộ phần này : 'output/mysql\_export\_temp\_fixed'

|  |
| --- |
| -- Cấu hình: Thiết lập số lượng job song song  SET default\_parallel 5;  -- Thiết lập encoding là UTF-8 để xử lý ký tự Tiếng Việt  SET default\_charset utf-8;  -- BẮT BUỘC: Đăng ký và định nghĩa Piggybank UDFs  REGISTER /opt/pig/lib/piggybank.jar;  -- Hàm này sẽ được dùng để tách chuỗi số hợp lệ từ chuỗi giá Cellphones  DEFINE REGEX\_EXTRACT org.apache.pig.piggybank.evaluation.string.RegexExtract;  -----------------------------------------------------------------  -- 1. Xử lý Dữ liệu CellphoneS (LÀM SẠCH GIÁ VÀ CHUYỂN ĐỔI KIỂU DỮ LIỆU)  -----------------------------------------------------------------  -- LOAD từ Local File System (Đường dẫn cần tồn tại trên máy local)  CELLPHONES\_DATA = LOAD 'file:///home/phu/raw\_data/products\_csv/cellphones\_raw\_data.csv'  USING org.apache.pig.piggybank.storage.CSVExcelStorage()  AS (id:chararray, product\_name:chararray, current\_price\_raw:chararray, list\_price\_raw:chararray, raw\_specs:chararray, url:chararray);  -- Lọc dòng Header  CELLPHONES\_FILTERED = FILTER CELLPHONES\_DATA BY NOT (id == 'id' AND product\_name == 'product\_name');  -- LÀM SẠCH VÀ CHUYỂN ĐỔI KIỂU DỮ LIỆU  CELLPHONES\_CLEANED = FOREACH CELLPHONES\_FILTERED {  -- 1. Tách chuỗi số (bao gồm dấu chấm) khỏi đơn vị tiền tệ và chữ cái (ví dụ: '22.990.000đ' -> '22.990.000')  -- Biểu thức chính quy: '([0-9\.]+)' lấy ra chuỗi số và dấu chấm.  price\_cur\_str = REGEX\_EXTRACT(current\_price\_raw, '([0-9\\.]+)', 1);  price\_list\_str = REGEX\_EXTRACT(list\_price\_raw, '([0-9\\.]+)', 1);  -- 2. Loại bỏ dấu chấm phân cách hàng nghìn (ví dụ: '22.990.000' -> '22990000')  price\_cur\_clean = REPLACE(price\_cur\_str, '\\.', '');  price\_list\_clean = REPLACE(price\_list\_str, '\\.', '');  GENERATE  (long)id AS id,  product\_name,  -- Chuyển đổi sang double. Nếu chuỗi không hợp lệ, Pig sẽ đặt là NULL.  (double)price\_cur\_clean AS current\_price,  (double)price\_list\_clean AS list\_price,  'cellphones' AS source\_brand;  };  -----------------------------------------------------------------  -- 2. Xử lý Dữ liệu TGDD (LOAD từ LOCAL)  -----------------------------------------------------------------  -- LOAD từ Local File System (Giá đã là số hợp lệ)  TGDD\_DATA = LOAD 'file:///home/phu/raw\_data/products\_csv/laptops\_enriched\_data.csv'  USING org.apache.pig.piggybank.storage.CSVExcelStorage()  AS (id:chararray, product\_name:chararray, current\_price:double, list\_price:double, brand:chararray, category:chararray, cpu:chararray, ram:chararray, storage:chararray, screen\_size:chararray, screen\_resolution:chararray, os:chararray, software:chararray, average\_rating:float, product\_url:chararray);  -- Lọc dòng Header  TGDD\_FILTERED = FILTER TGDD\_DATA BY NOT (id == 'id' AND product\_name == 'product\_name');  -- Chuẩn hóa: giữ nguyên các cột cần thiết (Giá đã là double)  TGDD\_CLEANED = FOREACH TGDD\_FILTERED GENERATE  (long)id AS id,  product\_name,  current\_price,  list\_price,  'thegioididong' AS source\_brand;  -----------------------------------------------------------------  -- 3. Hợp nhất Dữ liệu và Lưu Tạm  -----------------------------------------------------------------  -- Hợp nhất hai bộ dữ liệu (UNION) với Schema thống nhất: (long, chararray, double, double, chararray)  ALL\_DATA\_FOR\_MYSQL = UNION CELLPHONES\_CLEANED, TGDD\_CLEANED;  -- Lưu dữ liệu. NullStorage('\N') sẽ xử lý các giá trị NULL thành '\N', phù hợp cho MySQL.  STORE ALL\_DATA\_FOR\_MYSQL  INTO 'output/mysql\_export\_temp\_fixed'  USING PigStorage(',', 'NullStorage(\'\\\N\')'); |

1. Chuẩn bị bảng Đích trên MYSQL
   1. Đăng nhập vào MySQL : mysql -u root

A computer screen shot of a black screen

AI-generated content may be incorrect.

1. Thực thi ETL : Pig (HDFS 🡪 HDFS tạm) để làm sạch dữ liệu trước

A computer screen with white text

AI-generated content may be incorrect.

Nhưu vậy ta thấy 1164 record đã được lưu

Kiểm tra sơ qua các file

A screen shot of a computer

AI-generated content may be incorrect.

Tiến hành sqoop dữ liệu lên mysql