TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN ĐIỆN TOÁN ĐÁM MÂY

-----📖-----

PHẠM TRUNG KIÊN - 19110384

BÙI ANH ĐỨC - 19110348

TRẦN VIỆT ANH - 19110325

Đề Tài:

TÌM HIỂU APACHE HIVE VÀ VIẾT ỨNG DỤNG DEMO

GIẢNG VIÊN HƯỚNG DẪN

TS. HUỲNH XUÂN PHỤNG

KHÓA 2019 - 2023

ĐH SƯ PHẠM KỸ THUẬT TP. HCM
KHOA CNTT
****

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh Phúc
****

PHIẾU NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

Họ và tên Sinh viên 1: **Phạm Trung Kiên** MSSV 1: **19110384**

Họ và tên Sinh viên 2: **Bùi Anh Đức** MSSV 2: **19110348**

Họ và tên Sinh viên 2: **Trần Việt Anh** MSSV 2: **19110325**

Ngành: **Công nghệ Thông tin**

Tên đề tài: **Tìm hiểu Apache Hive và viết ứng dụng demo**

Họ và tên Giáo viên hướng dẫn: **TS. Huỳnh Xuân Phụng**

NHẬN XÉT

1. Về nội dung đề tài & khối lượng thực hiện:

.....

.....

.....

2. Ưu điểm:

.....

.....

.....

3. Khuyết điểm:

.....

.....

.....

4. Đề nghị cho bảo vệ hay không?

.....

5. Đánh giá loại:

.....

6. Điểm:

.....

Tp. Hồ Chí Minh, ngày tháng năm 2021

Giáo viên hướng dẫn
(Ký & ghi rõ họ tên)

ĐH SƯ PHẠM KỸ THUẬT TP. HCM
KHOA CNTT
****

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh Phúc
****

PHIẾU NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

Họ và tên Sinh viên 1: **Phạm Trung Kiên** MSSV 1: **19110384**

Họ và tên Sinh viên 2: **Bùi Anh Đức** MSSV 2: **19110348**

Họ và tên Sinh viên 2: **Trần Việt Anh** MSSV 2: **19110325**

Ngành: **Công nghệ Thông tin**

Tên đề tài: **Tìm hiểu Apache Hive và viết ứng dụng demo**

Họ và tên Giáo viên phản biện: **TS. Huỳnh Xuân Phụng**

NHẬN XÉT

1. Về nội dung đề tài & khối lượng thực hiện:

.....

.....

.....

2. Ưu điểm:

.....

.....

.....

3. Khuyết điểm:

.....

.....

4. Đề nghị cho bảo vệ hay không?

.....

5. Đánh giá loại:

.....

6. Điểm:

.....

Tp. Hồ Chí Minh, ngày tháng năm 2021

Giáo viên phản biện

(Ký & ghi rõ họ tên)

LỜI CẢM ƠN

Lời đầu tiên nhóm xin phép được gửi lời cảm ơn chân thành và sâu sắc nhất đến với Khoa Công Nghệ Thông Tin – Trường Đại Học Sư Phạm Kỹ Thuật Thành Phố Hồ Chí Minh đã tạo điều kiện cho nhóm chúng em được học tập, phát triển nền tảng kiến thức sâu sắc và thực hiện đề tài này.

Bên cạnh đó nhóm chúng em xin gửi đến thầy Huỳnh Xuân Phụng lời cảm ơn sâu sắc nhất. Trải qua một quá trình dài học tập và thực hiện đề tài trong thời gian qua. Thầy đã tận tâm chỉ bảo nhiệt tình nhóm chúng em trong suốt quá trình từ lúc bắt đầu cũng như kết thúc đề tài này.

Với sự hướng dẫn nhiệt tình, giảng dạy tận tình đầy đủ kiến thức của thầy Huỳnh Xuân Phụng, chúng em đã học tập và hiểu được những kiến thức cơ bản về Apache Hive – một dạng kho dữ liệu được sử dụng để quản lý và phân tích khối lượng dữ liệu lớn. Qua đó tụi em biết cách cài đặt và sử dụng Apache Hive.

Tuy nhiên lượng kiến thức là vô tận và với khả năng hạn hẹp chúng em đã rất cố gắng để hoàn thành một cách tốt nhất. Chính vì vậy việc xảy ra những thiếu sót là điều khó có thể tránh khỏi. Chúng em hi vọng nhận được sự góp ý tận tình của quý thầy (cô) qua đó chúng em có thể rút ra được bài học kinh nghiệm và hoàn thiện và cải thiện nâng cấp lại sản phẩm của mình một cách tốt nhất có thể.

Chúng em xin chân thành cảm ơn!

Nhóm thực hiện

Phạm Trung Kiên – 19110384

Bùi Anh Đức – 19110348

Trần Việt Anh – 19110325

MỤC LỤC

DANH MỤC CÁC HÌNH

PHẦN MỞ ĐẦU

1. Tính cấp thiết của đề tài

Hiện nay, thuật ngữ Big Data được sử dụng cho các bộ tập dữ liệu khổng lồ bao gồm khối lượng lớn, tốc độ cao và nhiều loại dữ liệu đang tăng lên từng ngày. Sử dụng các hệ thống quản lý dữ liệu truyền thống, rất khó để xử lý Big Data. Do đó, Quỹ phần mềm Apache (Apache Software Foundation) đã giới thiệu một framework tên là Hadoop và trong đó là Apache Hive để giải quyết các thách thức quản lý và xử lý Big Data.

Theo.

2. Đối tượng nghiên cứu

Đối với đề tài này, đối tượng nghiên cứu là Big Data. Đồng thời kèm theo đó là các công nghệ áp dụng để xây dựng warehouse đơn giản, cụ thể như:

- JDK và JRE: Bộ công cụ phát triển Java.
- Hadoop: Apache framework mã nguồn mở cho phép phát triển các ứng dụng phân tán (distributed processing) để lưu trữ và quản lý các tập dữ liệu lớn.
- Apache Hive: Công cụ cơ sở hạ tầng kho dữ liệu để xử lý dữ liệu có cấu trúc trong Hadoop.
- Apache Derby: External database để cấu hình Metastore.

3. Phạm vi nghiên cứu

Đề tài này chủ yếu tập trung vào việc xử lý và phân tích các câu truy vấn đối với dữ liệu lớn khi sử dụng Hive.

4. Kết quả dự kiến đạt được

- Cài đặt được Hadoop và Apache Hive
- Xây dựng được một data warehouse đơn giản bằng các câu truy vấn.
- Phân tích, thống kê các dữ liệu từ các câu truy vấn tùy vào mục đích của người dùng.
- Cài đặt giao diện Hive Web Interface.

PHẦN NỘI DUNG

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT VỀ APACHE HIVE

1.1 Tổng quan về Hive

Apache Hive là 1 kho dữ liệu (data warehouse) hỗ trợ người sử dụng có thể dễ dàng hơn trong việc quản lý và truy vấn đối với các tập dữ liệu lớn được lưu trữ trên các hệ thống lưu trữ phân tán (distributed storage). Hive được xây dựng dựa trên cơ sở của Apache Hadoop, nó cung cấp các tính năng chính sau:

- Công cụ cho phép dễ dàng thực hiện tác vụ như trích xuất, vận chuyển và lưu trữ dữ liệu.
- Cơ chế để xử lý cho nhiều định dạng dữ liệu khác nhau.
- Truy cập tới dữ liệu dạng files được lưu trữ trực tiếp ở trong Apache HDFS hoặc đối với nhiều hệ thống lưu trữ dữ liệu khác như Apache HBase.
- Thực hiện query thông qua MapReduce.

Hive định nghĩa ra một ngôn ngữ truy vấn đơn giản có cú pháp gần giống với SQL (SQL-like query language) được gọi là HiveQL, nó cho phép người sử dụng đã quen thuộc với các truy vấn SQL thực hiện việc truy vấn dữ liệu. Ngoài ra ngôn ngữ này còn cho phép các lập trình viên người đã quen thuộc với MapReduce framework có thể nhúng các mappers và reducers cho chính họ viết ra để thực thi nhiều hơn nữa các phân tích phức tập mà không được hỗ trợ bởi các hàm đã có sẵn trong ngôn ngữ HiveQL. HiveQL cung có thể được mở rộng với các custom scalar functions (UDF's), aggregations (UDAF's) và các table funtions (UDTF's)

Hive không yêu cầu dữ liệu phải được đọc và ghi dưới một định dạng của riêng Hive (Hive format). Hive hoạt động tốt trên Thrift và các định dạng dữ liệu riêng của người sử dụng.

Hive không được thiết kế để cho các giao dịch online (OLTP workloads) và không nên dùng cho các real-time queries và các cập nhật trên từng dòng trong 1 table (row-level). Hive hoạt động tốt nhất cho các batch jobs trên các tập dữ liệu lớn, mà ở đó dữ liệu được thêm vào liên tục (append-only data) ví dụ như web logs. Hive có khả năng mở rộng theo chiều ngang tốt (thực thi tốt trên 1 hadoop cluster có số tượng máy biến

đổi), có khả năng tích hợp với MapReduce framework và UDF, UDAF, UDTF; có khả năng chống chịu lỗi và mềm dẻo đối với các dữ liệu đầu vào của chính nó.

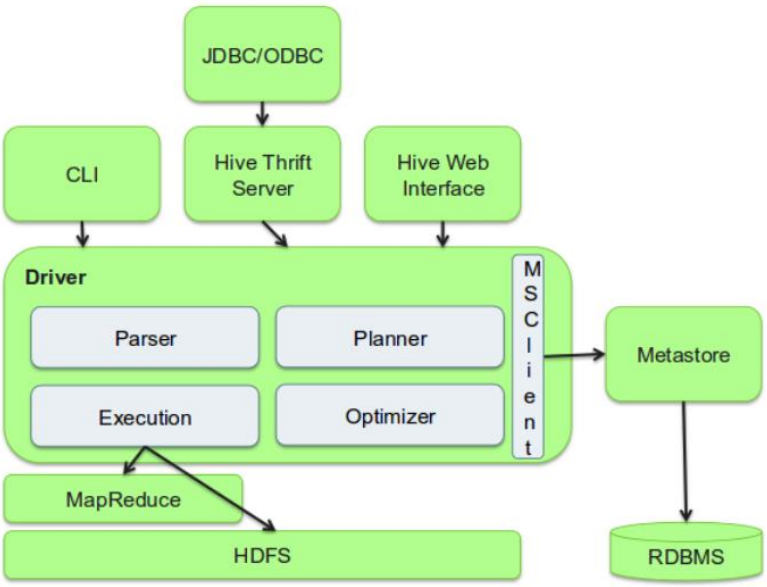
Các thành phần cấu hình Hive bao gồm HCatalog và WebHCat. HCatalog là một thành phần của Hive. Đây là lớp quản lý lưu trữ cho Hadoop (table and management layer), nó cho phép người dùng với các công cụ xử lý dữ liệu khác nhau bao gồm cả Pig và MapReduce thực thi hoạt động đọc, ghi một cách dễ dàng hơn. WebHCat cung cấp một dịch vụ cho phép bạn có thể thực thi Hadoop MapReduce (hoặc YARN), Pig, Hive.

1.2 Kiến trúc của Hive

Hive có các thành phần chính là:

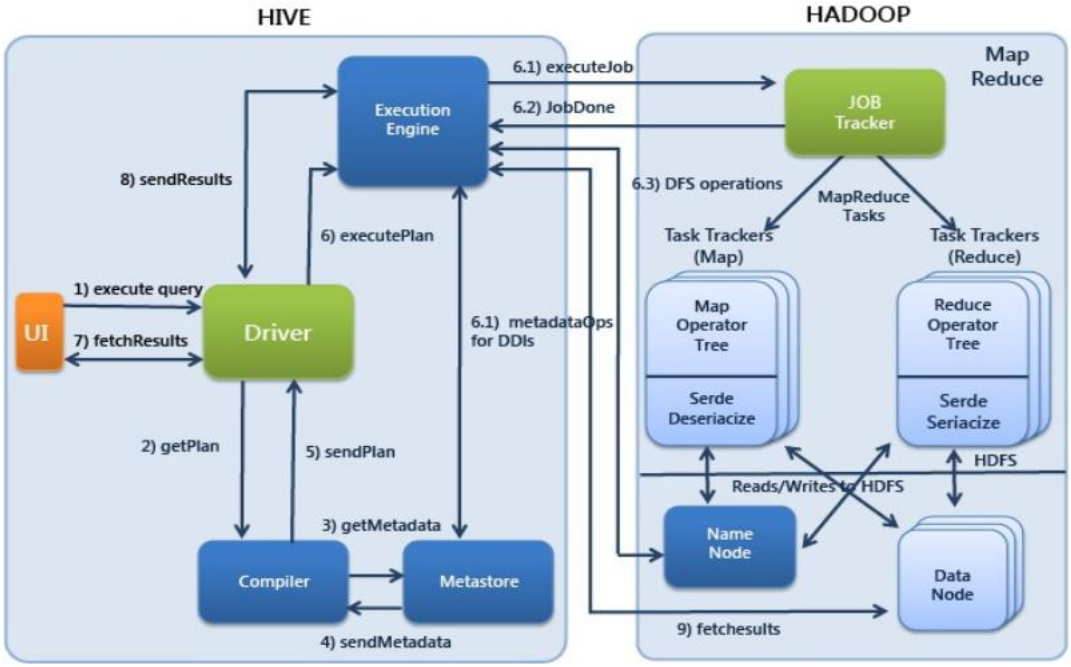
- Hive UI: cung cấp giao diện cho phép người sử dụng tương tác với hệ thống Hive. Hive cung cấp nhiều phương thức khác nhau cho phép người sử dụng tương tác với Hive:
  - CLI: giao diện dạng shell cho phép người sử dụng tương tác trực tiếp qua command line.
  - Hive Web Interface: giao diện Web cho phép người sử dụng thực hiện các truy vấn thông qua giao diện Web.
  - Hive Thrift Server: cho phép các client từ nhiều ngôn ngữ lập trình khác nhau có thể thực hiện tương tác với Hive.
- Hive Driver: thành phần nhận các truy vấn và chuyển các truy vấn này thành các MapReduce Jobs để tiến hành xử lý yêu cầu của người sử dụng.
  - Driver: nhận các truy vấn, thành phần này thực hiện việc quản lý các sessions và cung cấp các API để thực thi và lấy dữ liệu trên JDBC/ODBC interfaces.
  - Compiler: thành phần hiện việc phân tích ngữ nghĩa đối với các query, lấy các thông tin metadata cần thiết về table và partion từ metastore để sinh ra các excution plan.
  - Execute engine: thành phần thực thi các execution plan được tạo bởi compiler (submit các job tới MapReduce). Ngoài ra thành phần execution enginen này thực hiện việc quản lý các dependencies của các bước trong mỗi execution plan, thực thi từng bước này.

- Hive Metastore: thành phần lưu trữ các metadata của Hive: table, partion, buckets bao gồm cả thông tin về các column trong mỗi table, các serializers và desrializers cần thiết để thực hiện việc đọc và ghi dữ liệu. Metastore sử dụng một cơ sở dữ liệu quan hệ để lưu trữ dữ liệu của chính mình.



Hình 1: Kiến trúc của Hive

1.3 Hoạt động của Hive



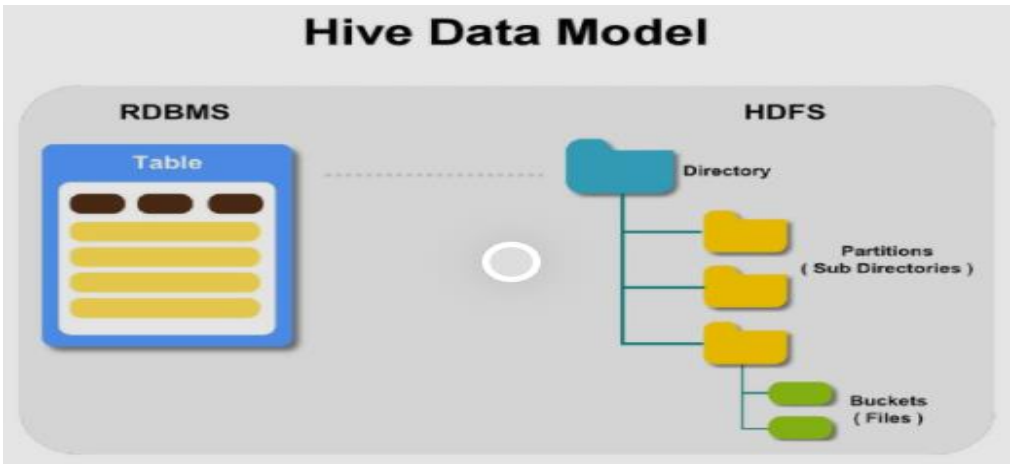
Hình 2: Sơ đồ luồng hoạt động của Hive

Quy trình hoạt động của Hive có thể được mô tả theo các bước sau:

1. Các truy vấn tới từ User Interface (CLI, Hive Web Interface, Thirft Server) được gửi tới thành phần Driver (Bước 1 hình 2)
2. Driver tạo ra mới 1 session cho truy vấn này và gửi query tới compiler để nhận lấy Execution Plan (Bước 2 hình 2)
3. Compilter nhận các metadata cần thiết từ Metastore (Bước 3, 4 hình 2). Các metadata này sẽ được sử dụng để kiểm tra các biểu thức bên trong query mà Compiler nhận được.
4. Plan được sinh ra bởi Compiler (thông tin về các job (map-reduce) cần thiết để thực thi query sẽ được gửi lại tới thành phần thực thi (Bước 5 hình 2)
5. Execution engine nhận yêu cầu thực thi và lấy các metadata cần thiết và yêu cầu mapreduce thực thi công việc (Bước 6.1, 6.2, 6.3 hình 2)
6. Khi output được sinh ra, nó sẽ được ghi dưới dạng 1 temporary file, temorary file này sẽ cung cấp các thông tin cần thiết cho các stages khác của plan. Nội dung của các temporary file này được execution đọc trực tiếp từ HDFS như là 1 phần của các lời gọi từ Driver (bước 7, 8, 9 hình 2)

1.4 Mô hình dữ liệu trong Hive

1.4.1. Tổ chức dữ liệu



Hình 3: Tổ chức dữ liệu trong Hive

Dữ liệu trong Hive được tổ chức thành các kiểu sau:

- Databases: là namespace cho các tables, dùng để nhóm và quản lý các nhóm tables khác nhau.
- Tables: tương tự như table trong các hệ cơ sở dữ liệu quan hệ. Trong Hive table có thể thực hiện các phép toán filter, join và union… Mặc định thì dữ liệu của

Hive sẽ được lưu bên trong thư mục warehouse trên HDFS. Tuy nhiên Hive cũng cung cấp kiểu external table cho phép ta tạo ra và quản lý các table mà dữ liệu của nó đã tồn tại từ trước khi ta tạo ra table này hoặc nó được lưu trữ ở 1 thư mục khác bên trong hệ thống HDFS. Tổ chức row và column bên trong Hive có nhiều điểm tương đồng với tổ chức Row và Column trong các hệ cơ sở dữ liệu quan hệ. Hive có 2 kiểu table đó là: Managed Table và External tables.

- Partions: Mỗi table có thể có 1 hoặc nhiều các khóa mà từ đó xác định dữ liệu sẽ được lưu trữ ở đâu. Ví dụ table web_log có thể phân chia dữ liệu của mình theo từng ngày là lưu dữ liệu của mỗi ngày trong 1 thư mục khác nhau bên dưới đường dẫn warehouse.

  Ví dụ: /warehouse/web_log/date=”01-01-2014″

- Buckets: Dữ liệu trong mỗi partion có thể được phân chia thành nhiều buckets khác nhau dựa trên 1 hash của 1 colume bên trong table. Mỗi bucket lưu trữ dữ liệu của nó bên dưới 1 thư mục riêng. Việc phân chia các partion thành các bucket giúp việc thực thi các query dễ dàng hơn.

1.4.2. Kiểu dữ liệu

Kiểu dữ liệu nguyên thủy:

Mỗi columns có 1 kiểu dữ liệu cố định. Các kiểu dữ liệu nguyên thủy sau sẽ được hỗ trợ đối với Hive:

- Integers:
  - TINYINT – 1 byte integer
  - SMALLINT – 2 bytes integer
  - INT – 4 bytes integer
  - BIGINT – 8 bytes integer
- Boolean type
  - BOOLEAN – TRUE/FALSE
- Floating point numbers
  - FLOAT – single precision
  - DOUBLE – Double precision
  - String type
- STRING – sequence of characters in a specified character set

Các kiểu dữ liệu khác:

- Structs: là kiểu dữ liệu mà mỗi phần tử bên trong đó có thể được truy cập thông qua việc sử dụng ký hiệu (.)

  Ví dụ, với kiểu dữ liệu STRUCT {a INT; b INT} ví dụ trường a của nó có thể truy cập thông qua c.a

- Maps (key-value tuples): là kiểu dữ liệu mà các phần tử sẽ được truy cập thông qua ký hiệu [‘element name’]. Đối với map M thực hiện việc map dữ liệu đối với khóa ‘group’ -> thì dữ liệu sẽ được sử dụng bởi trường M[‘group’]
- Arrays (indexable lists): Kiểu mảng.

1.5 HiveSQL (HQL)

Ngôn ngữ truy vấn Hive cung cấp các toán tử cơ bản giống SQL. Đây là một số tác vụ mà HQL có thể làm dễ dàng.

- Tạo và quản lý tables và partitions.
- Hỗ trợ các toán tử Relational, Arithmetic và Logical khác nhau.
- Evaluate functions
- Tải về nội dung 1 table từ thư mục cục bộ hoặc kết quả của câu truy vấn đến thư mục HDFS.

Đây là ví dụ truy vấn HQL:

```
SELECT upper(name), salesprice

FROM sales;

SELECT category, count(1)

FROM products

GROUP BY category;
```

CHƯƠNG 2: CÀI ĐẶT, THIẾT KẾ VÀ XÂY DỰNG DATA WAREHOUSE

2.1 Cài đặt

2.1.1 Cài đặt môi trường Hadoop

Cài đặt OpenJDK

```
root@ip-172-31-28-250:/home/ubuntu# sudo apt install openjdk-8-jdk -y
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  adwaita-icon-theme at-spi2-core fontconfig fonts-dejavu-extra gtk-update-icon-cache hicolor-icon-theme
  humanity-icon-theme libasound2 libasound2-data libasyncns0 libatk-bridge2.0-0 libatk-wrapper-java
  libatk-wrapper-java-jni libatk1.0-0 libatk1.0-data libatspi2.0-0 libcairo2 libcroco3 libdatrie1 libdrm-amdgpu1
  libdrm-intel1 libdrm-nouveau2 libdrm-radeon1 libflac8 libfontenc1 libgail-common libgail18 libgdk-pixbuf2.0-0
  libgdk-pixbuf2.0-bin libgdk-pixbuf2.0-common libgif7 libgl1 libgl1-mesa-dri libgl1-mesa-glx libglapi-mesa libglvnd0
  libglx-mesa0 libglx0 libgraphite2-3 libgtk2.0-0 libgtk2.0-bin libgtk2.0-common libharfbuzz0b libice-dev libice6
  libjbig0 libllvm10 libogg0 libpango-1.0-0 libpangocairo-1.0-0 libpangoft2-1.0-0 libpciaccess0 libpixman-1-0
  libpthread-stubs0-dev libpulse0 librsvg2-2 librsvg2-common libsensors4 libsm-dev libsm6 libsndfile1 libthai-data
  libthai0 libtiff5 libvorbis0a libvorbisenc2 libx11-6 libx11-dev libx11-doc libx11-xcb1 libxau-dev libxaw7
  libxcb-dri2-0 libxcb-dri3-0 libxcb-glx0 libxcb-present0 libxcb-render0 libxcb-shape0 libxcb-shm0 libxcb-sync1
  libxcb1-dev libxcomposite1 libxcursor1 libxdamage1 libxdmcp-dev libxfixes3 libxft2 libxinerama1 libxmu6 libxpm4
  libxrandr2 libxshmfence1 libxt-dev libxt6 libxv1 libxxf86dga1 libxxf86vm1 openjdk-8-jdk-headless openjdk-8-jre
  openjdk-8-jre-headless ubuntu-mono x11-utils x11proto-core-dev x11proto-dev xorg-sgml-doctools xtrans-dev
```

Kiểm tra phiên bản java

```
root@ip-172-31-28-250:/home/ubuntu# java -version
openjdk version "1.8.0_292"
OpenJDK Runtime Environment (build 1.8.0_292-8u292-b10-0ubuntu1~18.04-b10)
OpenJDK 64-Bit Server VM (build 25.292-b10, mixed mode)
root@ip-172-31-28-250:/home/ubuntu# javac -version
javac 1.8.0_292
```

Cài đặt OpenSSH

```
root@ip-172-31-28-250:/home/ubuntu# sudo apt install openssh-server openssh-client -y
Reading package lists... Done
Building dependency tree
Reading state information... Done
The following additional packages will be installed:
  openssh-sftp-server
Suggested packages:
  keychain libpam-ssh monkeysphere ssh-askpass molly-guard rssh
The following packages will be upgraded:
  openssh-client openssh-server openssh-sftp-server
3 upgraded, 0 newly installed, 0 to remove and 216 not upgraded.
Need to get 990 kB of archives.
After this operation, 5120 B of additional disk space will be used.
Get:1 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-sftp-server amd64 1:7.6p1-4ubuntu0.5 [45.5 kB]
Get:2 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-server amd64 1:7.6p1-4ubuntu0.5 [332 kB]
Get:3 http://us-east-1.ec2.archive.ubuntu.com/ubuntu bionic-updates/main amd64 openssh-client amd64 1:7.6p1-4ubuntu0.5 [612 kB]
Fetched 990 kB in 0s (16.2 MB/s)
Preconfiguring packages ...
(Reading database ... 119300 files and directories currently installed.)
```

Sau khi cài đặt OpenJDK và OpenSSH, ta cần 1 người dùng để sử dụng để tiếp tục cài đặt Hadoop

Ở đây, tụi em tạo user tên là hdoop

```
root@ip-172-31-28-250:/home/ubuntu# sudo adduser hdoop
Adding user `hdoop' ...
Adding new group `hdoop' (1001) ...
Adding new user `hdoop' (1001) with group `hdoop' ...
Creating home directory `/home/hdoop' ...
Copying files from `/etc/skel' ...
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
Changing the user information for hdoop
Enter the new value, or press ENTER for the default
        Full Name []:
        Room Number []:
        Work Phone []:
        Home Phone []:
        Other []:
Is the information correct? [Y/n] y
root@ip-172-31-28-250:/home/ubuntu#
```

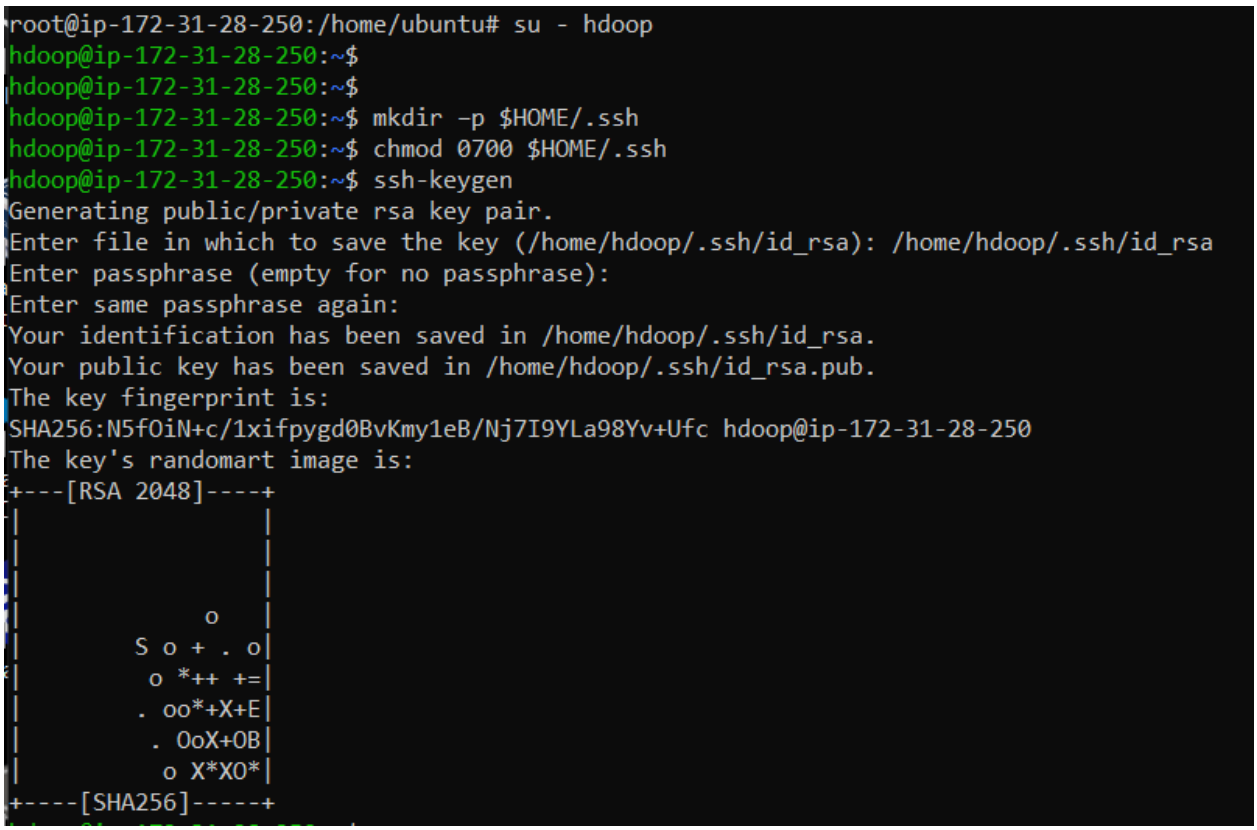
Sau khi tạo xong, ta cần thêm quyền root vào user vừa tạo

```
root@ip-172-31-28-250:/home/ubuntu# sudo passwd root
Enter new UNIX password:
Retype new UNIX password:
passwd: password updated successfully
root@ip-172-31-28-250:/home/ubuntu#
root@ip-172-31-28-250:/home/ubuntu# su root
root@ip-172-31-28-250:/home/ubuntu# visudo
```

Sử dụng lệnh visudo và thêm quyền root vào dưới root ALL= (ALL:ALL) ALL

```
# User privilege specification
root    ALL=(ALL:ALL) ALL
hdoop ALL=(ALL:ALL) ALL
```

Bật không mật khẩu cho người dùng hdoop để tránh phải nhập lại mật khẩu nhiều lần.



Sử dụng lệnh cat để lưu trữ khóa công khai dưới dạng allow_keys trong thư mục ssh.

Sau đó đặt quyền cho user bằng lệnh chmod.

Người dùng sau đó có thể ssh mà không cần dùng mật khẩu mỗi lần. Có thể kiểm tra mọi thứ đã thiết lập chính xác bằng cách sử dụng user hdoop đã tạo ở trên để ssh vào localhost.

```
hdoop@ip-172-31-28-250:~$ cat ~/.ssh/id_rsa.pub >> ~/.ssh/authorized_keys
hdoop@ip-172-31-28-250:~$ chmod 0600 ~/.ssh/authorized_keys
hdoop@ip-172-31-28-250:~$ ssh localhost
The authenticity of host 'localhost (127.0.0.1)' can't be established.
ECDSA key fingerprint is SHA256:lPgV11XIgKicaU2Cgg+HmEvrnhA6/uLhyd652X3lCxA.
Are you sure you want to continue connecting (yes/no)? yes
Warning: Permanently added 'localhost' (ECDSA) to the list of known hosts.
Welcome to Ubuntu 18.04.5 LTS (GNU/Linux 5.3.0-1033-aws x86_64)

 * Documentation:  https://help.ubuntu.com
 * Management:     https://landscape.canonical.com
 * Support:        https://ubuntu.com/advantage

 System information disabled due to load higher than 1.0

 * Ubuntu Pro delivers the most comprehensive open source security and
   compliance features.

   https://ubuntu.com/aws/pro

  Get cloud support with Ubuntu Advantage Cloud Guest:
    http://www.ubuntu.com/business/services/cloud

217 packages can be updated.
161 updates are security updates.
```

Tiếp theo là tải gói hadoop về, ở đây chúng em sử dụng phiên bản hadoop 3.2.1

Link tải: https://archive.apache.org/dist/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz

Sau khi tải xong, chúng ta giải nén bằng lệnh tar xzf

```
hdoop@ip-172-31-28-250:~$ wget https://archive.apache.org/dist/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz
--2021-11-11 07:28:34--  https://archive.apache.org/dist/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz
Resolving archive.apache.org (archive.apache.org)... 138.201.131.134, 2a01:4f8:172:2ec5::2
Connecting to archive.apache.org (archive.apache.org)|138.201.131.134|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 359196911 (343M) [application/x-gzip]
Saving to: ‘hadoop-3.2.1.tar.gz’

hadoop-3.2.1.tar.gz          100%[=============================================>] 342.56M  15.3MB/s    in 24s

2021-11-11 07:28:59 (14.2 MB/s) - ‘hadoop-3.2.1.tar.gz’ saved [359196911/359196911]

hdoop@ip-172-31-28-250:~$ tar xzf hadoop-3.2.1.tar.gz
```

Tiếp theo ta cần cấu hình các biến môi trường Hadoop(bashrc)

```
hdoop@ip-172-31-28-250:~$ sudo nano .bashrc
[sudo] password for hdoop:
```

Xác định nội dung biến môi trường bằng các lệnh sau:

```
#Hadoop Related Options
export HADOOP_HOME=/home/hdoop/hadoop-3.2.1
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib/native"
```

Sau khi xác định xong, ta lưu lại bằng lệnh sau

```
hdoop@ip-172-31-28-250:~$ source ~/.bashrc
```

Chỉnh sửa tệp hadoop.env.sh

```
hdoop@ip-172-31-28-250:~$ sudo nano $HADOOP_HOME/etc/hadoop/hadoop-env.sh
```

Bỏ ghi chú biến $JAVA_HOME (bỏ dấu '#') và thêm đường dẫn đầy đủ đến cài đặt OpenJDK trên hệ thống

```
# The java implementation to use. By default, this environment
# variable is REQUIRED on ALL platforms except OS X!
export JAVA_HOME=/usr/lib/jvm/java-8-openjdk-amd64

# Location of Hadoop.  By default, Hadoop will attempt to determine
# this location based upon its execution path.
```

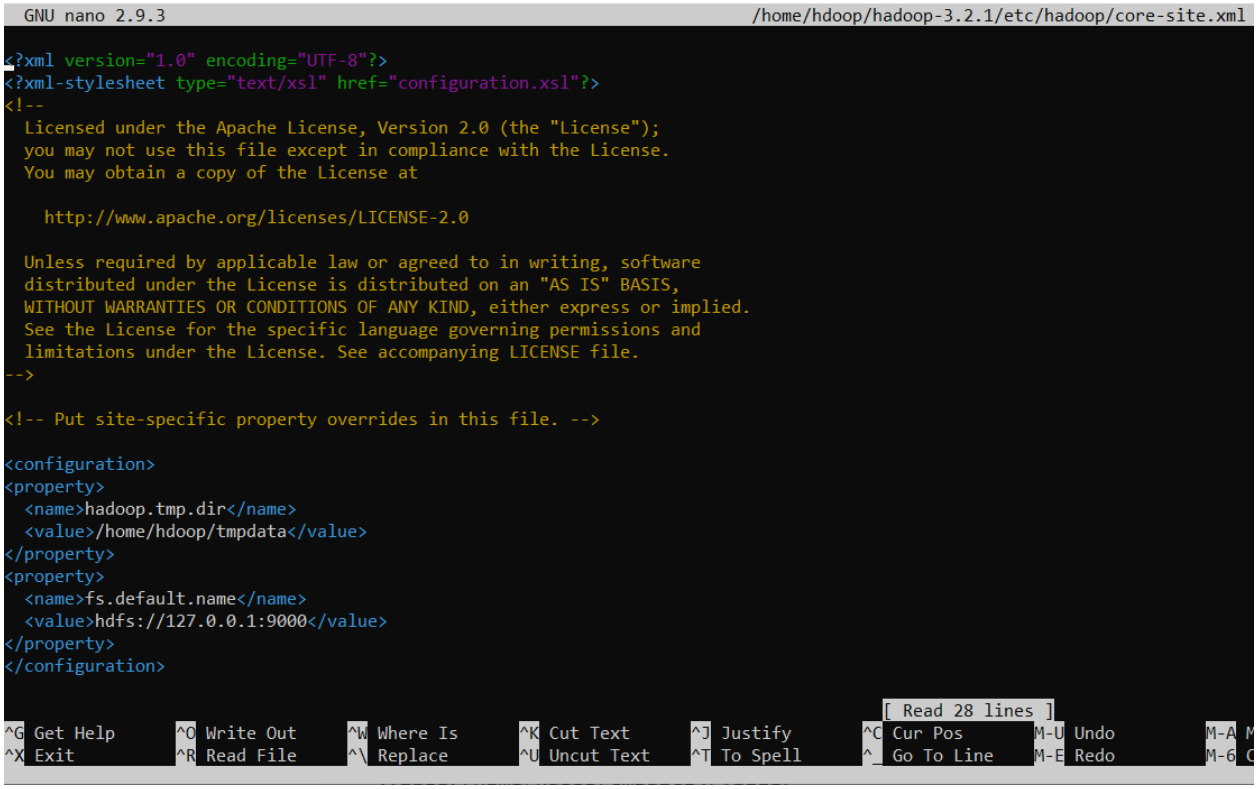
Nếu như không biết đường dẫn Java chính xác, hãy chạy lệnh sau:

```
hdoop@ip-172-31-28-250:~$ which javac
/usr/bin/javac
hdoop@ip-172-31-28-250:~$ readlink -f /usr/bin/javac
/usr/lib/jvm/java-8-openjdk-amd64/bin/javac
hdoop@ip-172-31-28-250:~$
```

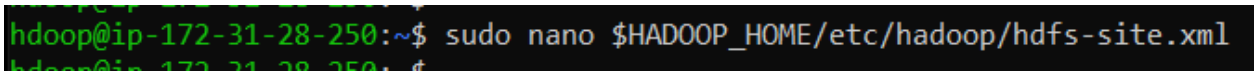
Chỉnh sửa tệp core-site.xml

```
hdoop@ip-172-31-28-250:~$ sudo nano $HADOOP_HOME/etc/hadoop/core-site.xml
```

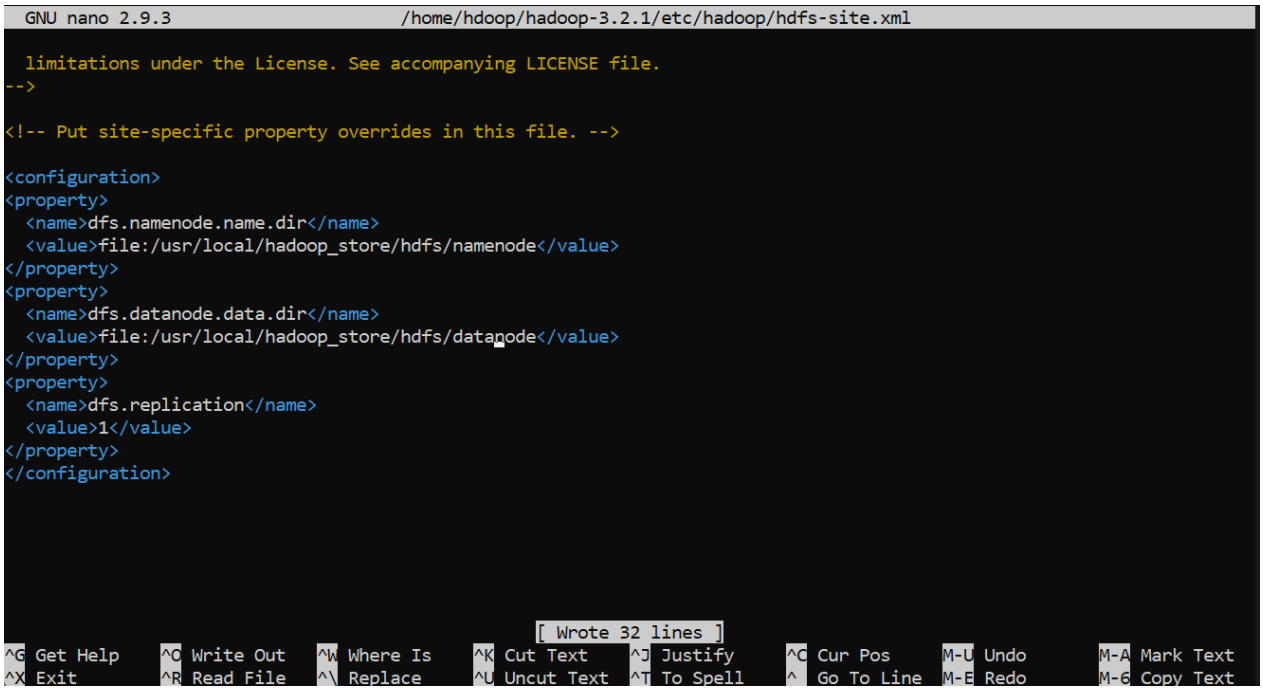
Thêm cấu hình sau để ghi đè các giá trị mặc định cho thư mục tạm thời và thêm URL HDFS để thay thế cài đặt hệ thống tệp cục bộ mặc định:



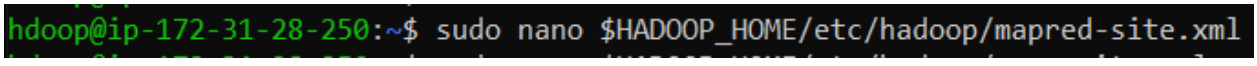
Chỉnh sửa tệp hdfs-site.xml



Thêm cấu hình sau vào tệp và nếu cần, hãy điều chỉnh các thư mục NameNode và DataNode cho các vị trí tùy chỉnh:



Chỉnh sửa tệp mapred-site.xml



Thêm cấu hình sau để thay đổi giá trị MapReduce framework mặc định thành yarn:

```
  GNU nano 2.9.3                                  /home/hdoop/hadoop-3.2.1/etc/hadoop/mapred-site.xml

<?xml version="1.0"?>
<?xml-stylesheet type="text/xsl" href="configuration.xsl"?>
<!--
  Licensed under the Apache License, Version 2.0 (the "License");
  you may not use this file except in compliance with the License.
  You may obtain a copy of the License at

    http://www.apache.org/licenses/LICENSE-2.0

  Unless required by applicable law or agreed to in writing, software
  distributed under the License is distributed on an "AS IS" BASIS,
  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
  See the License for the specific language governing permissions and
  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
</configuration>

                                                                [ Read 24 lines ]
^G Get Help  ^O Write Out  ^W Where Is  ^K Cut Text  ^J Justify  ^C Cur Pos  M-U Undo  M-A Ma
^X Exit      ^R Read File  ^\ Replace   ^U Uncut Text ^T To Spell ^_ Go To Line M-E Redo  M-6 Co
```

Chỉnh sửa tập tin yarn-site.xml

```
hdoop@ip-172-31-28-250:~$ sudo nano $HADOOP_HOME/etc/hadoop/yarn-site.xml
hdoop@ip-172-31-28-250:~$
```

Thêm các cấu hình sau vào file

```
  GNU nano 2.9.3                                  /home/hdoop/hadoop-3.2.1/etc/hadoop/yarn-site.xml

<!-- Site specific YARN configuration properties -->
<property>
  <name>yarn.nodemanager.aux-services</name>
  <value>mapreduce_shuffle</value>
</property>
<property>
  <name>yarn.nodemanager.aux-services.mapreduce.shuffle.class</name>
  <value>org.apache.hadoop.mapred.ShuffleHandler</value>
</property>
<property>
  <name>yarn.resourcemanager.hostname</name>
  <value>127.0.0.1</value>
</property>
<property>
  <name>yarn.acl.enable</name>
  <value>0</value>
</property>
<property>
  <name>yarn.nodemanager.env-whitelist</name>
  <value>JAVA_HOME,HADOOP_COMMON_HOME,HADOOP_HDFS_HOME,HADOOP_CONF_DIR,CLASSPATH_PERPEND_DISTCACHE,HADOOP_YARN_HOME,HADOOP_MAPRED_HOME</value>
</property>
</configuration>
```

Định dạng HDFS NameNode:

```
hdoop@ip-172-31-28-250:~$ hdfs namenode -format
WARNING: /home/hdoop/hadoop-3.2.1/logs does not exist. Creating.
2021-11-11 08:12:51,734 INFO namenode.NameNode: STARTUP_MSG:
/************************************************************
STARTUP_MSG: Starting NameNode
STARTUP_MSG:   host = ip-172-31-28-250/172.31.28.250
STARTUP_MSG:   args = [-format]
STARTUP_MSG:   version = 3.2.1
STARTUP_MSG:   classpath = /home/hdoop/hadoop-3.2.1/etc/hadoop:/hom
9.3.24.v20180605.jar:/home/hdoop/hadoop-3.2.1/share/hadoop/common/l
```

```
2021-11-11 08:12:54,254 INFO namenode.FSImage: FSImageSaver clean checkpoint: txid=0 when meet shutdown.
2021-11-11 08:12:54,255 INFO namenode.NameNode: SHUTDOWN_MSG:
/************************************************************
SHUTDOWN_MSG: Shutting down NameNode at ip-172-31-28-250/172.31.28.250
************************************************************/
```

Khởi động Hadoop Cluster

```
hdoop@ip-172-31-28-250:~$ cd hadoop-3.2.1
hdoop@ip-172-31-28-250:~/hadoop-3.2.1$ cd sbin
hdoop@ip-172-31-28-250:~/hadoop-3.2.1/sbin$ ./start-dfs.sh
Starting namenodes on [localhost]
Starting datanodes
Starting secondary namenodes [ip-172-31-28-250]
ip-172-31-28-250: Warning: Permanently added 'ip-172-31-28-250' (ECDSA) to the list of known hosts.
```

Như với lệnh trước, đầu ra thông báo rằng các quá trình đang bắt đầu

Nhập lệnh đơn giản jps để kiểm tra xem tất cả các daemon có đang hoạt động và đang chạy như các quy trình Java hay không

```
hdoop@ip-172-31-28-250:~/hadoop-3.2.1/sbin$ ./start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hdoop@ip-172-31-28-250:~/hadoop-3.2.1/sbin$ jps
27969 NameNode
29108 Jps
28597 ResourceManager
28774 NodeManager
28383 SecondaryNameNode
```

2.1.2 Cài đặt Apache Hive

Tải gói apache hive về, chúng em sử dụng phiên bản hive 3.1.2

Link tải: https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz

Sau khi tải xong, giải nén bằng lệnh tar xzf

```
hdoop@ip-172-31-28-250:~$ wget https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz
--2021-11-11 08:26:03--  https://downloads.apache.org/hive/hive-3.1.2/apache-hive-3.1.2-bin.tar.gz
Resolving downloads.apache.org (downloads.apache.org)... 88.99.95.219, 135.181.214.104, 2a01:4f9:3a:2c57::2, ...
Connecting to downloads.apache.org (downloads.apache.org)|88.99.95.219|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 278813748 (266M) [application/x-gzip]
Saving to: ‘apache-hive-3.1.2-bin.tar.gz’

apache-hive-3.1.2-bin.tar 100%[==================================>] 265.90M  14.8MB/s    in 19s

2021-11-11 08:26:22 (14.3 MB/s) - ‘apache-hive-3.1.2-bin.tar.gz’ saved [278813748/278813748]

hdoop@ip-172-31-28-250:~$ tar xzf apache-hive-3.1.2-bin.tar.gz
```

Cấu hình các biến môi trường Hive (bashrc)

```
hdoop@ip-172-31-28-250:~$ sudo nano .bashrc
[sudo] password for hdoop:
```

Thêm các biến môi trường hive sau vào tệp .bashrc

```
#Hive Related Options
export HIVE_HOME=/home/hdoop/apache-hive-3.1.2-bin
export PATH=$PATH:$HIVE_HOME/bin
```

Sau khi cấu hình xong, chạy lệnh sau để lưu tệp

```
hdoop@ip-172-31-28-250:~$ source ~/.bashrc
```

Chỉnh sửa tập hive-config.sh

```
hdoop@ip-172-31-28-250:~$ sudo nano $HIVE_HOME/bin/hive-config.sh
```

Thêm biến HADOOP_HOME và đường dẫn đầy đủ đến thư mục Hadoop

```
 GNU nano 2.9.3                          /home/hdoop/apache-hive-3.1.2-bin/bin/hive-config.sh

        shift
        confdir=$1
        shift
        HIVE_CONF_DIR=$confdir
        ;;
    --auxpath)
        shift
        HIVE_AUX_JARS_PATH=$1
        shift
        ;;
    *)
        break;
        ;;
  esac
done


# Allow alternate conf dir location.
HIVE_CONF_DIR="${HIVE_CONF_DIR:-$HIVE_HOME/conf}"

export HIVE_CONF_DIR=$HIVE_CONF_DIR
export HADOOP_HOME=/home/hdoop/hadoop-3.2.1
export HIVE_AUX_JARS_PATH=$HIVE_AUX_JARS_PATH

# Default to use 256MB
export HADOOP_HEAPSIZE=${HADOOP_HEAPSIZE:-256}
```

Tạo đường dẫn thư mục Hive trong HDFS

Thêm quyền ghi và thực thi cho file tmp

Kiểm tra quyền bằng lệnh –ls/

```
hdoop@ip-172-31-28-250:~$ hdfs dfs -mkdir /tmp
hdoop@ip-172-31-28-250:~$ hdfs dfs -chmod g+w /tmp
hdoop@ip-172-31-28-250:~$ hdfs dfs -ls /
Found 1 items
drwxrwxr-x   - hdoop supergroup          0 2021-11-11 08:43 /tmp
```

Tạo thư mục warehouse

Thêm quyền ghi và thực thi cho warehouse

Kiểm tra bằng lệnh –ls/

```
hdoop@ip-172-31-28-250:~$ hdfs dfs -mkdir -p /user/hive/warehouse
hdoop@ip-172-31-28-250:~$ hdfs dfs -chmod g+w /user/hive/warehouse
hdoop@ip-172-31-28-250:~$ hdfs dfs -ls /user/hive
Found 1 items
drwxrwxr-x   - hdoop supergroup          0 2021-11-11 08:45 /user/hive/warehouse
```

Cấu hình tập hive-site.xml

```
hdoop@ip-172-31-28-250:~$ cd $HIVE_HOME/conf
hdoop@ip-172-31-28-250:~/apache-hive-3.1.2-bin/conf$ ls
beeline-log4j2.properties.template    ivysettings.xml
hive-default.xml.template             llap-cli-log4j2.properties.template
hive-env.sh.template                  llap-daemon-log4j2.properties.template
hive-exec-log4j2.properties.template  parquet-logging.properties
hive-log4j2.properties.template
```

Sử dụng lệnh copy để tạo tệp hive-site.xml từ hive-default.xml.template

```
hdoop@ip-172-31-28-250:~/apache-hive-3.1.2-bin/conf$ cp hive-default.xml.template hive-site.xml
```

Khởi tạo Derby Database

```
hdoop@ip-172-31-28-250:~$ $HIVE_HOME/bin/schematool -dbType derby -initSchema
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.2.1/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Metastore connection URL:        jdbc:derby:;databaseName=metastore_db;create=true
Metastore Connection Driver :    org.apache.derby.jdbc.EmbeddedDriver
Metastore connection User:       APP
Starting metastore schema initialization to 3.1.0
Initialization script hive-schema-3.1.0.derby.sql
```

Sau khi khởi tạo xong sẽ hiện ra schemaTool completed

```
Initialization script completed
schemaTool completed
```

Sau đó ta có thể sử dụng Hive

```
hdoop@ip-172-31-28-250:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-28-250:~/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.2.1/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = 490da63f-be2c-4e8a-8668-e3a581c35478

Logging initialized using configuration in jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Hive-on-MR is deprecated in Hive 2 and may not be available in the future versions. Consider using a different execution engine (i.e. spark, tez) or using Hive 1.X releases.
hive>
```

2.1.3 Lỗi phát sinh trong quá trình cài đặt

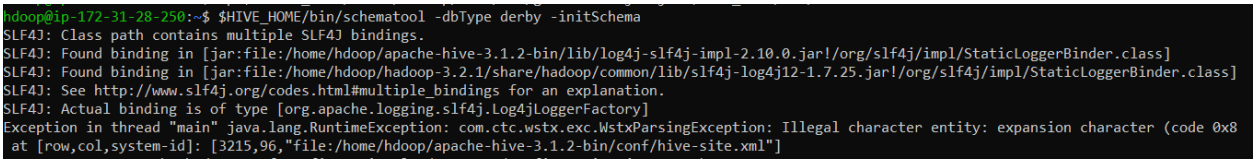
Lỗi 1:

```
hdoop@ip-172-31-28-250:~$ $HIVE_HOME/bin/schematool -dbType derby -initSchema
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.2.1/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Exception in thread "main" java.lang.NoSuchMethodError: com.google.common.base.Preconditions.checkArgument(ZLjava/lang/String;Ljava/lang/Object;)V
```

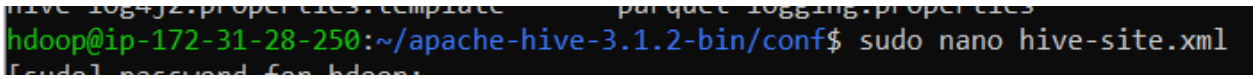
Ta cần xóa guave của hive, rồi copy guave của hadoop qua hive để đồng bộ bằng cách sau:

```
hdoop@ip-172-31-28-250:~$ rm $HIVE_HOME/lib/guava-19.0.jar
hdoop@ip-172-31-28-250:~$ cp $HADOOP_HOME/share/hadoop/hdfs/lib/guava-27.0-jre.jar $HIVE_HOME/lib/
```

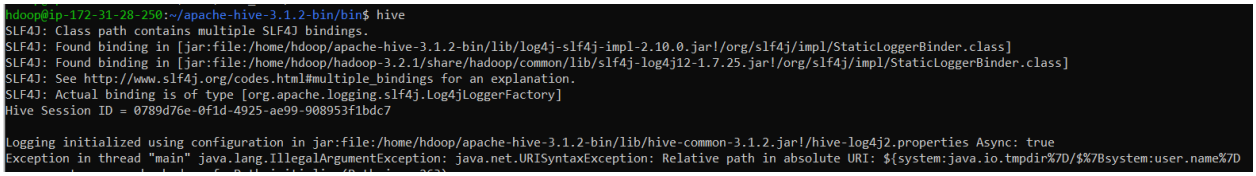
Lỗi 2:



Xóa những kí tự đặc biệt bị dư, vào hive-site.xml xóa những kí tự bị dư ở dòng 3215:



Lỗi 3:



Thêm các thẻ property sau vào đầu file hive-site.xml:

```
<property>
  <name>system:java.io.tmpdir</name>
  <value>/tmp/hive/java</value>
 </property>
 <property>
  <name>system:user.name</name>
  <value>${user.name}</value>
 </property>
```

```
 GNU nano 2.9.3                                                  hive-site.xml

   limitations under the License.
--><configuration>
  <!-- WARNING!!! This file is auto generated for documentation purposes ONLY! -->
  <!-- WARNING!!! Any changes you make to this file will be ignored by Hive.   -->
  <!-- WARNING!!! You must make your changes in hive-site.xml instead.         -->
  <!-- Hive Execution Parameters -->
<property>
    <name>system:java.io.tmpdir</name>
    <value>/tmp/hive/java</value>
  </property>
  <property>
    <name>system:user.name</name>
    <value>${user.name}</value>
  </property>
```

Lỗi 4:

```
hdoop@ip-172-31-28-250:~/apache-hive-3.1.2-bin/bin$ hive
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/log4j-slf4j-impl-2.10.0.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-3.2.1/share/hadoop/common/lib/slf4j-log4j12-1.7.25.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
Hive Session ID = bda137a9-604b-4c6e-bc27-dbf23856c787

Logging initialized using configuration in jar:file:/home/hdoop/apache-hive-3.1.2-bin/lib/hive-common-3.1.2.jar!/hive-log4j2.properties Async: true
Exception in thread "main" java.lang.RuntimeException: java.net.ConnectException: Call From ip-172-31-28-250/172.31.28.250 to localhost:9000 failed on connect
ion exception: java.net.ConnectException: Connection refused; For more details see:  http://wiki.apache.org/hadoop/ConnectionRefused
```

Exception in thread "main" java.lang.RuntimeException: java.net.ConnectException: Connection refused

Sửa bằng cách thực hiện các bước sau:

```
hdoop@ip-172-31-28-250:~$ stop-all.sh
```

```
hdoop@ip-172-31-28-250:~$ hadoop namenode -format
```

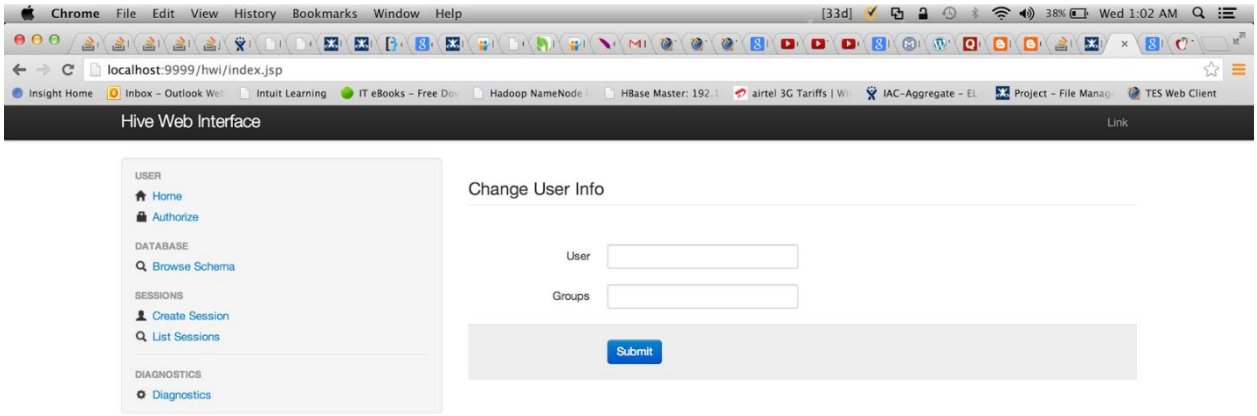
```
hdoop@ip-172-31-28-250:~$ start-all.sh
```

```
hdoop@ip-172-31-28-250:~$ cd $HIVE_HOME/bin
hdoop@ip-172-31-28-250:~/apache-hive-3.1.2-bin/bin$ hive
```

2.2 Triển khai Hive Web Interface

Hive Web Interface là dao diện dùng để sử dụng và truy vấn HQL trên một giao diện người dùng thay vì trên Command Line.

Việc sử dụng HWI sẽ đơn giản và thuận tiện hơn trong việc truy vấn



Do HWI chỉ hỗ trợ cho các phiên bản Apache Hive 2.2.0 trở xuống nên để có thể sử dụng được HWI, ta cần cài đặt lại phiên bản mà HWI hỗ trợ.

Cụ thể, tụi em sẽ sử dụng các phiên bản sau:

Hadoop: https://archive.apache.org/dist/hadoop/common/hadoop-2.7.3/hadoop-2.7.3.tar.gz

Hive: https://archive.apache.org/dist/hive/hive-2.1.0/apache-hive-2.1.0-bin.tar.gz

Hive-source: https://archive.apache.org/dist/hive/hive-2.1.0/apache-hive-2.1.0-src.tar.gz

Vì bản Apache Hive này không hỗ trợ Derby nên ta cần phải cài thêm

Derby: https://archive.apache.org/dist/db/derby/db-derby-10.13.1.1/db-derby-10.13.1.1-bin.tar.gz

Cuối cùng là sử dụng Apache Ant để có thể kết nối với HWI

Ant: https://archive.apache.org/dist/ant/source/apache-ant-1.9.7-src.tar.gz

Sau khi tải xong, ta cần giải nén bằng lệnh tar xzf (đã hướng dẫn trong bước cài đặt)

```
hdoop@ip-172-31-88-175:~$ ls
apache-ant-1.9.7          apache-hive-2.1.0-bin      apache-hive-2.1.0-src      db-derby-10.13.1.1-bin      derby.log  hadoop-2.7.3.tar.gz  tmpdata
apache-ant-1.9.7-bin.tar.gz  apache-hive-2.1.0-bin.tar.gz  apache-hive-2.1.0-src.tar.gz  db-derby-10.13.1.1-bin.tar.gz  hadoop-2.7.3  metastore_db
```

Tương tự theo các bước cài đặt ở trên, sau khi giải nén xong, ta vào file bashrc để cấu hình các đường dẫn

```
hdoop@ip-172-31-88-175:~$ nano .bashrc
```

```
#Hadoop Related Options
export HADOOP_HOME=/home/hdoop/hadoop-2.7.3
export HADOOP_INSTALL=$HADOOP_HOME
export HADOOP_MAPRED_HOME=$HADOOP_HOME
export HADOOP_COMMON_HOME=$HADOOP_HOME
export HADOOP_HDFS_HOME=$HADOOP_HOME
export YARN_HOME=$HADOOP_HOME
export HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_HOME/lib/native
export PATH=$PATH:$HADOOP_HOME/sbin:$HADOOP_HOME/bin
export HADOOP_OPTS="-Djava.library.path=$HADOOP_HOME/lib"
#Hive Related Option
export HIVE_HOME=/home/hdoop/apache-hive-2.1.0-bin
export HIVE_CONF_DIR=/home/hdoop/apache-hive-2.1.0-bin/conf
export PATH=$HIVE_HOME/bin:$PATH
export CLASSPATH=$CLASSPATH:/home/hdoop/hadoop/lib/*:.
export CLASSPATH=$CLASSPATH:/home/hdoop/apache-hive-2.1.0-bin/lib/*:.
#Derby Related Option
export DERBY_HOME=/home/hdoop/db-derby-10.13.1.1-bin
export PATH=$PATH:$DERBY_HOME/bin
export CLASSPATH=$CLASSPATH:$DERBY_HOME/lib/derby.jar:$DERBY_HOME/lib/derbytools.jar
#Ant
export ANT_HOME=/home/hdoop/apache-ant-1.9.7
export ANT_LIB=$ANT_HOME/lib
```

Tiếp theo thực hiện cấu hình các file (theo hướng dẫn ở trên)

hadoop.env.sh

core-site.xml

hdfs-site.xml

mapred-site.xml

yarn-site.xml

HDFS NameNode

Hadoop Cluster

Sau khi cấu hình xong, nhập lệnh jps để kiểm tra xem tất cả các daemon có đang hoạt động và đang chạy như các quy trình Java hay không

Nếu trong trường hợp thiếu Datanode hoặc Namenode:

```
hdoop@ip-172-31-28-250:~$ jps
3314 NameNode
3730 SecondaryNameNode
4131 NodeManager
4499 Jps
3947 ResourceManager
```

Ta thực hiện các bước sau để cấp quyền và khởi động lại hệ thống:

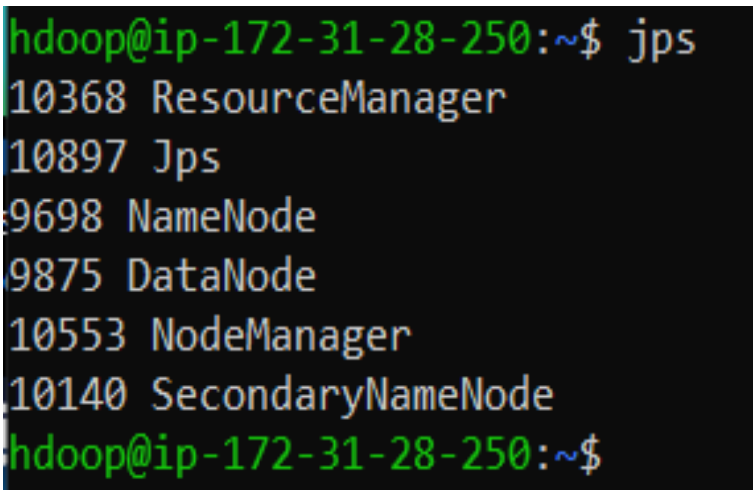
Tiến hành sửa đường dẫn của Namenode và Datanode:

```
  GNU nano 2.9.3                                    hdfs-site.xml

  limitations under the License. See accompanying LICENSE file.
-->

<!-- Put site-specific property overrides in this file. -->

<configuration>
<property>
  <name>dfs.namenode.name.dir</name>
  <value>file:/usr/local/hadoop_store/hdfs/namenode</value>
</property>
<property>
  <name>dfs.datanode.data.dir</name>
  <value>file:/usr/local/hadoop_store/hdfs/datanode</value>
</property>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
</configuration>
```

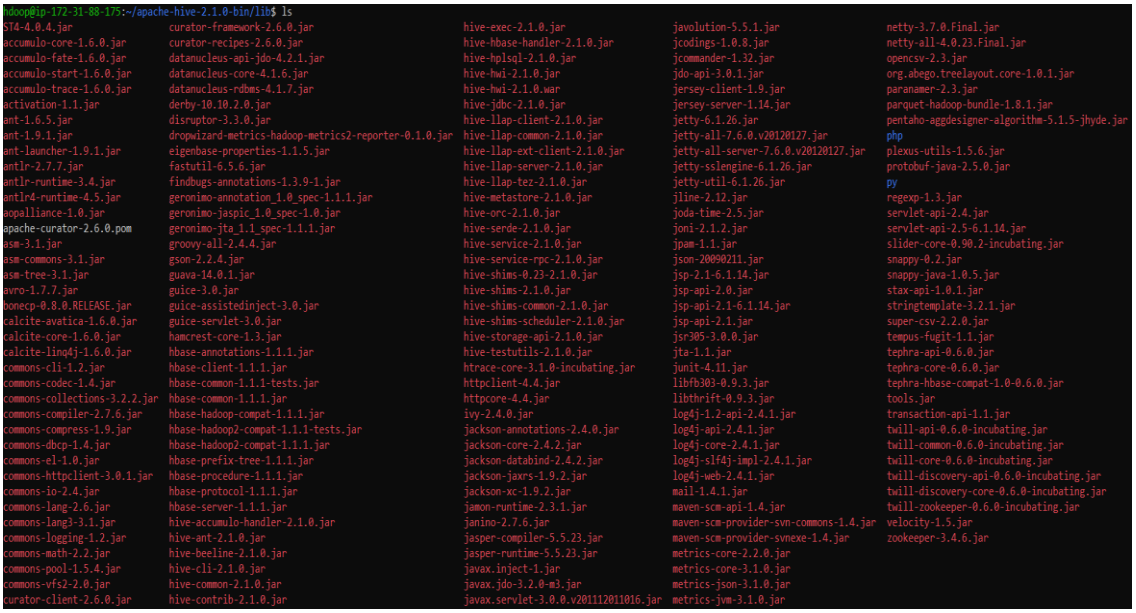
Cấp quyền cho user hdoop trên các tệp NameNode và DataNode (usr/local/hadoop_store/hdfs):

```
hdoop@ip-172-31-28-250:~$ sudo chown -R hdoop:hdoop /usr/local/hadoop_store/hdfs
hdoop@ip-172-31-28-250:~$
hdoop@ip-172-31-28-250:~$ sudo chmod -R 755 /usr/local/hadoop_store/hdfs
```

Stop .sh, chỉnh lại format NameNode và start .sh lại. Sau đó nhập jps để kiểm tra



Tiến hành copy file hive-hwi-2.1.0.war từ apache-hive-2.1.0-src và copy tools.jar từ thư viện của java vào thư viện lib của Apache Hive

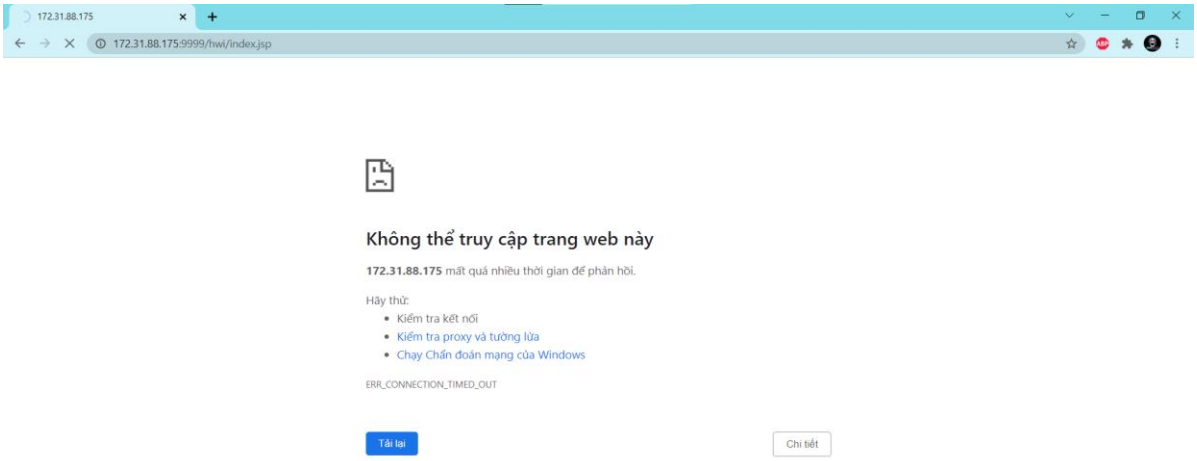


Chỉnh sửa file hive-site.xml, truy cập bằng lệnh nano hive-site.xml

```
<property>
  <name>hive.hwi.listen.host</name>
  <value>0.0.0.0</value>
  <description>This is the host address the Hive Web Interface will listen on</description>
</property>
<property>
  <name>hive.hwi.listen.port</name>
  <value>9999</value>
  <description>This is the port the Hive Web Interface will listen on</description>
</property>
<property>
  <name>hive.hwi.war.file</name>
  <value>lib/hive-hwi-2.1.0.war</value>
  <description>This sets the path to the HWI war file, relative to ${HIVE_HOME}. </description>
</property>
```

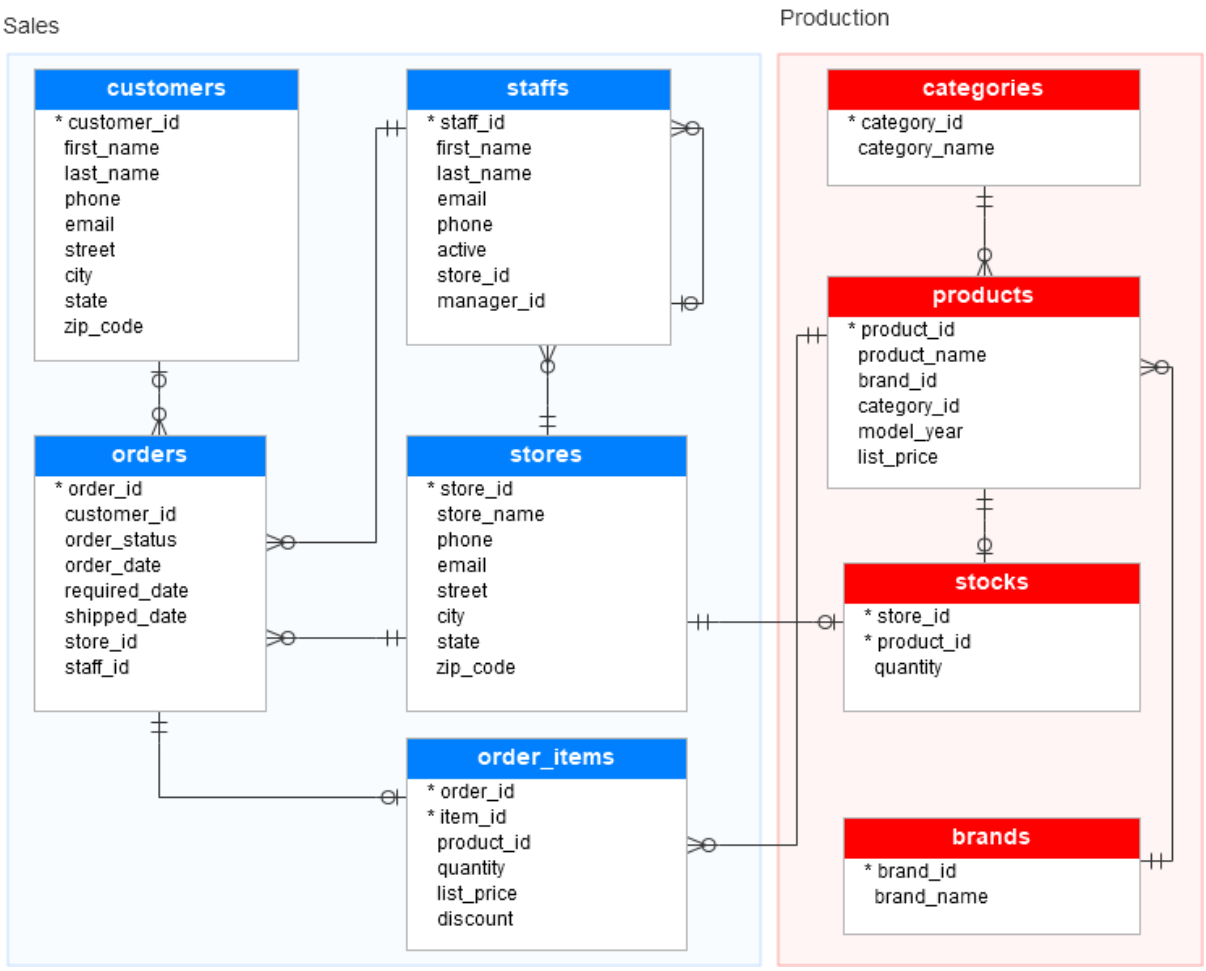
Chạy thử Hive Web Interface bằng lệnh: hive --service hwi

```
hdoop@ip-172-31-88-175:~$ hive --service hwi
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/hdoop/apache-hive-2.1.0-bin/lib/log4j-slf4j-impl-2.4.1.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/hdoop/hadoop-2.7.3/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.apache.logging.slf4j.Log4jLoggerFactory]
```



2.3 Thiết kế Data warehouse

Thiết kế lược đồ quan hệ:



2.4 Xây dựng Data warehouse

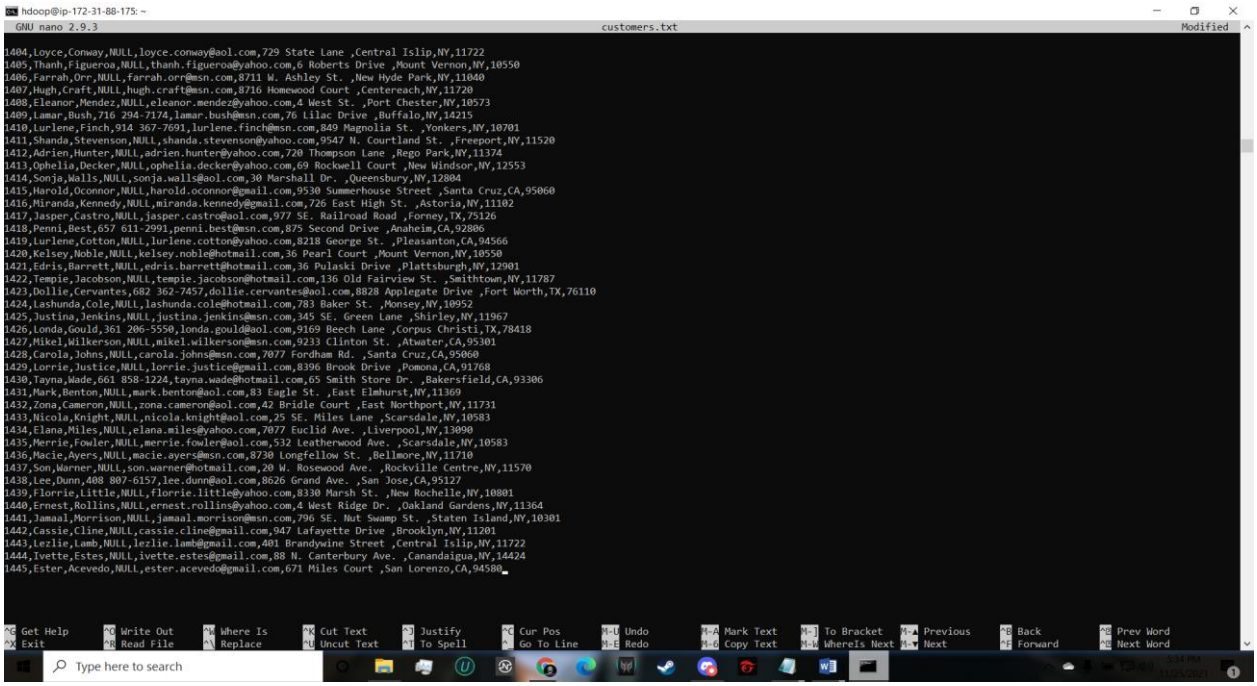
Tại đây tụi em thêm dữ liệu từ các file có sẵn nên làm theo các bước sau:

2.4.1 Tạo các file dữ liệu

Tạo các file chứa dữ liệu ở ngoài /local ứng với từng bảng đã thiết kế

```
hdoop@ip-172-31-88-175:~$ touch brands.txt
hdoop@ip-172-31-88-175:~$ touch categories.txt
hdoop@ip-172-31-88-175:~$ touch customers.txt
hdoop@ip-172-31-88-175:~$ touch order_items.txt
hdoop@ip-172-31-88-175:~$ touch orders.txt
hdoop@ip-172-31-88-175:~$ touch products.txt
hdoop@ip-172-31-88-175:~$ touch staffs.txt
hdoop@ip-172-31-88-175:~$ touch stocks.txt
hdoop@ip-172-31-88-175:~$ touch stores.txt
hdoop@ip-172-31-88-175:~$ ls
apache-ant-1.9.7            apache-hive-2.1.0-bin.tar.gz  brands.txt      db-derby-10.13.1.1-bin        hadoop-2.7.3        order_items.txt  staffs.txt  tmpdata
apache-ant-1.9.7-bin.tar.gz apache-hive-2.1.0-src         categories.txt  db-derby-10.13.1.1-bin.tar.gz hadoop-2.7.3.tar.gz orders.txt      stocks.txt
apache-hive-2.1.0-bin       apache-hive-2.1.0-src.tar.gz  customers.txt   derby.log                     metastore_db        products.txt     stores.txt
```

Dữ liệu trong file customers.txt



Đẩy dữ liệu các file vào /tmp/hive/hdoop

```
hdoop@ip-172-31-88-175:~$ hdfs dfs -ls -R /
21/11/25 10:42:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
drwx-wx-wx   - hdoop supergroup          0 2021-11-25 09:42 /tmp
drwx-wx-wx   - hdoop supergroup          0 2021-11-25 09:42 /tmp/hive
drwx------   - hdoop supergroup          0 2021-11-25 10:42 /tmp/hive/hdoop
-rw-r--r--   1 hdoop supergroup         90 2021-11-25 10:40 /tmp/hive/hdoop/brands.txt
-rw-r--r--   1 hdoop supergroup        128 2021-11-25 10:39 /tmp/hive/hdoop/categories.txt
-rw-r--r--   1 hdoop supergroup     125777 2021-11-25 10:41 /tmp/hive/hdoop/customers.txt
-rw-r--r--   1 hdoop supergroup     109322 2021-11-25 10:41 /tmp/hive/hdoop/order_items.txt
-rw-r--r--   1 hdoop supergroup      66230 2021-11-25 10:41 /tmp/hive/hdoop/orders.txt
-rw-r--r--   1 hdoop supergroup      16872 2021-11-25 10:39 /tmp/hive/hdoop/products.txt
-rw-r--r--   1 hdoop supergroup        360 2021-11-25 10:41 /tmp/hive/hdoop/staffs.txt
-rw-r--r--   1 hdoop supergroup       7804 2021-11-25 10:42 /tmp/hive/hdoop/stocks.txt
-rw-r--r--   1 hdoop supergroup        261 2021-11-25 10:41 /tmp/hive/hdoop/stores.txt
drwxr-xr-x   - hdoop supergroup          0 2021-11-25 10:29 /user
drwxr-xr-x   - hdoop supergroup          0 2021-11-25 10:29 /user/hive
drwxr-xr-x   - hdoop supergroup          0 2021-11-25 10:29 /user/hive/warehouse
drwxr-xr-x   - hdoop supergroup          0 2021-11-25 10:29 /user/hive/warehouse/bike.db
```

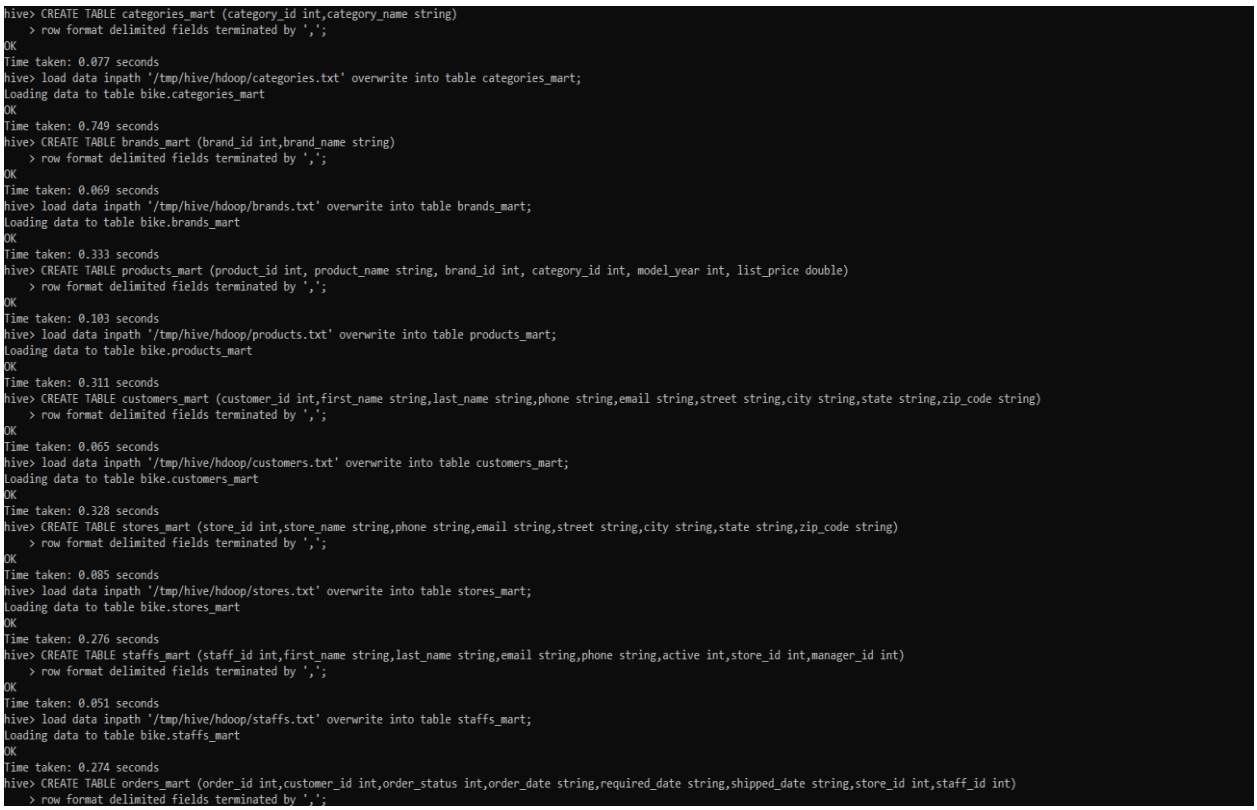
2.4.2 Tạo Database trong Hive

```
hive> show databases;
OK
default
Time taken: 1.419 seconds, Fetched: 1 row(s)
hive> create database bike;
OK
Time taken: 0.396 seconds
hive> show databases;
OK
bike
default
Time taken: 0.045 seconds, Fetched: 2 row(s)
hive>
```

2.4.3 Tạo Table và load dữ liệu

Chúng em đã thêm dữ liệu trong các file .txt ở ngoài /local trước, do đó khi tạo bảng, chúng em sẽ load dữ liệu từ các file .txt lên các table trong Hive.

Tạo bảng trong Hive để chứa dữ liệu và load dữ liệu từ tmp vào:



Sau khi load xong dữ liệu, ta thực hiện truy vấn.

- Bảng products

```
hive> select * from products;
OK
1       Trek 820 - 2016 9       6       2016    379.99
2       Ritchey Timberwolf Frameset - 2016      5       6       2016    749.99
3       Surly Wednesday Frameset - 2016 8       6       2016    999.99
4       Trek Fuel EX 8 29 - 2016        9       6       2016    2899.99
5       Heller Shagamaw Frame - 2016    3       6       2016    1320.99
6       Surly Ice Cream Truck Frameset - 2016   8       6       2016    469.99
7       Trek Slash 8 27.5 - 2016        9       6       2016    3999.99
8       Trek Remedy 29 Carbon Frameset - 2016   9       6       2016    1799.99
9       Trek Conduit+ - 2016    9       5       2016    2999.99
10      Surly Straggler - 2016  8       4       2016    1549.0
11      Surly Straggler 650b - 2016     8       4       2016    1680.99
12      Electra Townie Original 21D - 2016      1       3       2016    549.99
13      Electra Cruiser 1 (24-Inch) - 2016      1       3       2016    269.99
14      Electra Girls Hawaii 1 (16-inch) - 2015/2016    1       3       2016    269.99
15      Electra Moto 1 - 2016   1       3       2016    529.99
16      Electra Townie Original 7D EQ - 2016    1       3       2016    599.99
17      Pure Cycles Vine 8-Speed - 2016 4       3       2016    429.0
18      Pure Cycles Western 3-Speed - Womens - 2015/2016        4       3       2016    449.0
19      Pure Cycles William 3-Speed - 2016      4       3       2016    449.0
20      Electra Townie Original 7D EQ - Womens - 2016   1       3       2016    599.99
21      Electra Cruiser 1 (24-Inch) - 2016      1       1       2016    269.99
22      Electra Girls Hawaii 1 (16-inch) - 2015/2016    1       1       2016    269.99
23      Electra Girls Hawaii 1 (20-inch) - 2015/2016    1       1       2016    299.99
24      Electra Townie Original 21D - 2016      1       2       2016    549.99
25      Electra Townie Original 7D - 2015/2016  1       2       2016    499.99
26      Electra Townie Original 7D EQ - 2016    1       2       2016    599.99
27      Surly Big Dummy Frameset - 2017 8       6       2017    999.99
28      Surly Karate Monkey 27.5+ Frameset - 2017       8       6       2017    2499.99
29      Trek X-Caliber 8 - 2017 9       6       2017    999.99
30      Surly Ice Cream Truck Frameset - 2017   8       6       2017    999.99
31      Surly Wednesday - 2017  8       6       2017    1632.99
32      Trek Farley Alloy Frameset - 2017       9       6       2017    469.99
33      Surly Wednesday Frameset - 2017 8       6       2017    469.99
34      Trek Session DH 27.5 Carbon Frameset - 2017     9       6       2017    469.99
35      Sun Bicycles Spider 3i - 2017   7       6       2017    832.99
36      Surly Troll Frameset - 2017     8       6       2017    832.99
37      Haro Flightline One ST - 2017   2       6       2017    379.99
38      Haro Flightline Two 26 Plus - 2017      2       6       2017    549.99
39      Trek Stache 5 - 2017    9       6       2017    1499.99
40      Trek Fuel EX 9.8 29 - 2017      9       6       2017    4999.99
41      Haro Shift R3 - 2017    2       6       2017    1469.99
42      Trek Fuel EX 5 27.5 Plus - 2017 9       6       2017    2299.99
43      Trek Fuel EX 9.8 27.5 Plus - 2017       9       6       2017    5299.99
44      Haro SR 1.1 - 2017      2       6       2017    539.99
45      Haro SR 1.2 - 2017      2       6       2017    869.99
46      Haro SR 1.3 - 2017      2       6       2017    1409.99
```

- Bảng categories

```
hive> select * from categories;
OK
1       Children Bicycles
2       Comfort Bicycles
3       Cruisers Bicycles
4       Cyclocross Bicycles
5       Electric Bikes
6       Mountain Bikes
7       Road Bikes
Time taken: 0.191 seconds, Fetched: 7 row(s)
```

- Bảng brands

```
hive> select * from brands;
OK
1       Electra
2       Haro
3       Heller
4       Pure Cycles
5       Ritchey
6       Strider
7       Sun Bicycles
8       Surly
9       Trek
Time taken: 0.19 seconds, Fetched: 9 row(s)
```

- Bảng customers

```
hive> select * from customers;
OK
1       Debra   Burks   NULL    debra.burks@yahoo.com   9273 Thorne Ave.        Orchard Park    NY      14127
2       Kasha   Todd    NULL    kasha.todd@yahoo.com    910 Vine Street         Campbell        CA      95008
3       Tameka  Fisher  NULL    tameka.fisher@aol.com   769C Honey Creek St.    Redondo Beach   CA      90278
4       Daryl   Spence  NULL    daryl.spence@aol.com    988 Pearl Lane  Uniondale       NY      11553
5       Charolette      Rice    916 381-6003    charolette.rice@msn.com 107 River Dr.   Sacramento      CA      95820
6       Lyndsey Bean    NULL    lyndsey.bean@hotmail.com        769 West Road   Fairport        NY      14450
7       Latasha Hays    716 986-3359    latasha.hays@hotmail.com        7014 Manor Station Rd.  Buffalo NY      14215
8       Jacquline       Duncan  NULL    jacquline.duncan@yahoo.com      15 Brown St.    Jackson Heights NY      11372
9       Genoveva        Baldwin NULL    genoveva.baldwin@msn.com        8550 Spruce Drive       Port Washington NY      11050
10      Pamelia Newman  NULL    pamelia.newman@gmail.com        476 Chestnut Ave.       Monroe  NY      10950
11      Deshawn Mendoza NULL    deshawn.mendoza@yahoo.com       8790 Cobblestone Street         Monsey  NY      10952
12      Robby   Sykes   516 583-7761    robby.sykes@hotmail.com 486 Rock Maple Street   Hempstead       NY      11550
13      Lashawn Ortiz   NULL    lashawn.ortiz@msn.com   27 Washington Rd.       Longview        TX      75604
14      Garry   Espinoza        NULL    garry.espinoza@hotmail.com      7858 Rockaway Court     Forney  TX      75126
15      Linnie  Branch  NULL    linnie.branch@gmail.com 314 South Columbia Ave.         Plattsburgh     NY      12901
16      Emmitt  Sanchez 212 945-8823    emmitt.sanchez@hotmail.com      461 Squaw Creek Road    New York        NY      10002
17      Caren   Stephens        NULL    caren.stephens@msn.com  914 Brook St.   Scarsdale       NY      10583
18      Georgetta       Hardin  NULL    georgetta.hardin@aol.com        474 Chapel Dr.  Canandaigua     NY      14424
19      Lizzette        Stein   NULL    lizzette.stein@yahoo.com        19 Green Hill Lane      Orchard Park    NY      14127
20      Aleta   Shepard NULL    aleta.shepard@aol.com   684 Howard St.  Sugar Land      TX      77478
21      Tobie   Little  NULL    tobie.little@gmail.com  10 Silver Spear Dr.     Victoria        TX      77904
22      Adelle  Larsen  NULL    adelle.larsen@gmail.com 683 West Kirkland Dr.   East Northport  NY      11731
23      Kaylee  English NULL    kaylee.english@msn.com  8786 Fulton Rd.         Hollis  NY      11423
24      Corene  Wall    NULL    corene.wall@msn.com     9601 Ocean Rd.  Atwater CA      95301
25      Regenia Vaughan NULL    regenia.vaughan@gmail.com       44 Stonybrook Street    Mahopac NY      10541
26      Theo    Reese   562 215-2907    theo.reese@gmail.com    8755 W. Wild Horse St.  Long Beach      NY      11561
27      Santos  Valencia        NULL    santos.valencia@yahoo.com       7479 Carpenter Street   Sunnyside       NY      11104
28      Jeanice Frost   NULL    jeanice.frost@hotmail.com       76 Devon Lane   Ossining        NY      10562
29      Syreeta Hendricks       NULL    syreeta.hendricks@msn.com       193 Spruce Road         Mahopac NY      10541
30      Jamaal  Albert  NULL    jamaal.albert@gmail.com 853 Stonybrook Street   Torrance        CA      90505
31      Williemae       Holloway        510 246-8375    williemae.holloway@msn.com      69 Cypress St.  Oakland CA      94603
32      Araceli Golden  NULL    araceli.golden@msn.com  12 Ridgeview Ave.       Fullerton       CA      92831
33      Deloris Burke   NULL    deloris.burke@hotmail.com       895 Edgemont Drive      Palos Verdes Peninsula  CA      90274
34      Brittney        Woodward        NULL    brittney.woodward@aol.com       960 River St.   East Northport  NY      11731
35      Guillermina     Noble   NULL    guillermina.noble@msn.com       6 Del Monte Lane        Baldwinsville   NY      13027
36      Bernita Mcdaniel        NULL    bernita.mcdaniel@hotmail.com    2 Peg Shop Ave.         Liverpool       NY      13090
37      Melia   Brady   NULL    melia.brady@gmail.com   907 Shirley Rd.         Maspeth NY      11378
38      Zelma   Browning        NULL    zelma.browning@aol.com  296 Second Street       Astoria NY      11102
39      Janetta Aguirre 717 670-2634    janetta.aguirre@aol.com 214 Second Court        Lancaster       NY      14086
40      Ronna   Butler  NULL    ronna.butler@gmail.com  9438 Plymouth Court     Encino  CA      91316
41      Kathie  Freeman NULL    kathie.freeman@msn.com  667 Temple Dr.  Queensbury       NY      12804
42      Tangela Quinn   NULL    tangela.quinn@aol.com   4 S. Purple Finch Road  Richmond Hill   NY      11418
43      Mozelle Carter  281 489-9656    mozelle.carter@aol.com  895 Chestnut Ave.       Houston TX      77016
44      Onita   Johns   NULL    onita.johns@msn.com     32 Glen Creek Lane      Elmont  NY      11003
45      Bennett Armstrong       NULL    bennett.armstrong@aol.com       688 Walnut Street       Bethpage        NY      11714
46      Monika  Berg    NULL    monika.berg@gmail.com   369 Vernon Dr.  Encino  CA      91316
```

- Bảng stores

```
hive> select * from stores;
OK
1       Santa Cruz Bikes        (831) 476-4321  santacruz@bikes.shop    3700 Portola Drive      Santa Cruz      CA      95060
2       Baldwin Bikes   (516) 379-8888  baldwin@bikes.shop      4200 Chestnut Lane      Baldwin NY      11432
3       Rowlett Bikes   (972) 530-5555  rowlett@bikes.shop      8000 Fairway Avenue     Rowlett TX      75088
Time taken: 1.558 seconds, Fetched: 3 row(s)
```

- Bảng staffs

```
hive> select * from staffs;
OK
1       Viet Anh        Tran    vanh@gmail.com  012345  0       1       1
2       Duc     Bui     duc@gmail.com   023456  1       1       1
3       Kien    Pham    kien@gmail.com  034567  1       1       1
4       Bang    Huynh   bang@gmail.com  045678  0       2       1
5       Toan    Vo      toan@gmail.com  056789  0       2       1
6       An      Tran    an@gmail.com    067891  1       2       1
7       Khuong  Tran    khuong@gmail.com        078912  1       2       1
8       Tung    Diep    Tung@gmail.com  089123  1       3       1
9       Truc    Phan    truc@gmail.com  091234  1       3       1
Time taken: 0.174 seconds, Fetched: 9 row(s)
```

- Bảng orders

```
hive> select * from orders;
OK
1       259     4       20160101        20160103        20160103        1       2
2       1212    4       20160101        20160104        20160103        2       6
3       523     4       20160102        20160105        20160103        2       7
4       175     4       20160103        20160104        20160105        1       3
5       1324    4       20160103        20160106        20160106        2       6
6       94      4       20160104        20160107        20160105        2       6
7       324     4       20160104        20160107        20160105        2       6
8       1204    4       20160104        20160105        20160105        2       7
9       60      4       20160105        20160108        20160108        1       2
10      442     4       20160105        20160106        20160106        2       6
11      1326    4       20160105        20160108        20160107        2       7
12      91      4       20160106        20160108        20160109        1       2
13      873     4       20160108        20160111        20160111        2       6
14      258     4       20160109        20160111        20160112        1       3
15      450     4       20160109        20160110        20160112        2       7
16      552     4       20160112        20160115        20160115        1       3
17      1175    4       20160112        20160114        20160114        1       3
18      541     4       20160114        20160117        20160115        1       3
19      696     4       20160114        20160117        20160116        1       2
20      923     4       20160114        20160116        20160117        1       2
21      1250    4       20160115        20160116        20160118        2       6
22      1035    4       20160116        20160118        20160117        1       2
23      1149    4       20160116        20160119        20160119        1       2
24      636     4       20160118        20160120        20160119        2       7
25      657     4       20160118        20160121        20160121        2       6
26      1280    4       20160118        20160121        20160119        2       7
27      57      4       20160119        20160121        20160120        2       7
28      252     4       20160119        20160120        20160121        2       6
29      437     4       20160120        20160122        20160121        2       6
30      1348    4       20160120        20160121        20160121        2       6
31      1238    4       20160120        20160122        20160122        3       8
32      1259    4       20160121        20160124        20160122        1       3
33      236     4       20160121        20160122        20160122        2       6
34      80      4       20160122        20160125        20160123        2       6
35      813     4       20160122        20160125        20160124        2       7
36      1321    4       20160123        20160124        20160124        2       6
37      164     4       20160125        20160128        20160126        2       6
38      583     4       20160125        20160127        20160126        2       7
39      1296    4       20160125        20160126        20160126        2       7
40      348     4       20160127        20160128        20160129        1       3
41      979     4       20160127        20160130        20160129        2       6
42      1095    4       20160127        20160128        20160130        2       7
43      1434    4       20160127        20160128        20160130        2       7
44      861     4       20160128        20160131        20160130        2       7
45      1220    4       20160128        20160131        20160131        2       7
46      746     4       20160129        20160131        20160131        2       7
47      1234    4       20160129        20160130        20160131        2       7
```

- Bảng order_items

```
hive> select * from order_items;
OK
1       1       20      1       599.99  0.2
1       2       8       2       1799.99 0.07
1       3       10      2       1549.0  0.05
1       4       16      2       599.99  0.05
1       5       4       1       2899.99 0.2
2       1       20      1       599.99  0.07
2       2       16      2       599.99  0.05
3       1       3       1       999.99  0.05
3       2       20      1       599.99  0.05
4       1       2       2       749.99  0.1
5       1       10      2       1549.0  0.05
5       2       17      1       429.0   0.07
5       3       26      1       599.99  0.07
6       1       18      1       449.0   0.07
6       2       12      2       549.99  0.05
6       3       20      1       599.99  0.1
6       4       3       2       999.99  0.07
6       5       9       2       2999.99 0.07
7       1       15      1       529.99  0.07
7       2       3       1       999.99  0.1
7       3       17      2       429.0   0.1
8       1       22      1       269.99  0.05
8       2       20      2       599.99  0.07
9       1       7       2       3999.99 0.1
10      1       14      1       269.99  0.1
11      1       8       1       1799.99 0.05
11      2       22      2       269.99  0.1
11      3       16      2       599.99  0.2
12      1       4       2       2899.99 0.1
12      2       11      1       1680.99 0.05
13      1       13      1       269.99  0.1
13      2       17      2       429.0   0.05
13      3       20      2       599.99  0.1
13      4       16      2       599.99  0.05
14      1       6       1       469.99  0.07
15      1       12      2       549.99  0.07
15      2       8       1       1799.99 0.07
15      3       18      2       449.0   0.05
15      4       23      2       299.99  0.2
16      1       8       1       1799.99 0.2
16      2       21      1       269.99  0.05
16      3       13      2       269.99  0.07
16      4       14      1       269.99  0.07
17      1       8       1       1799.99 0.07
17      2       23      1       299.99  0.1
17      3       5       1       1320.99 0.1
```

- Bảng stocks

```
hive> select * from stocks;
OK
1       1       27
1       2       5
1       3       6
1       4       23
1       5       22
1       6       0
1       7       8
1       8       0
1       9       11
1       10      15
1       11      8
1       12      16
1       13      13
1       14      8
1       15      3
1       16      4
1       17      2
1       18      16
1       19      4
1       20      26
1       21      24
1       22      29
1       23      9
1       24      10
1       25      10
1       26      16
1       27      21
1       28      20
1       29      13
1       30      30
1       31      2
1       32      0
1       33      10
1       34      2
1       35      18
1       36      26
1       37      12
1       38      13
1       39      2
1       40      24
1       41      10
1       42      0
1       43      2
1       44      1
1       45      15
1       46      19
```

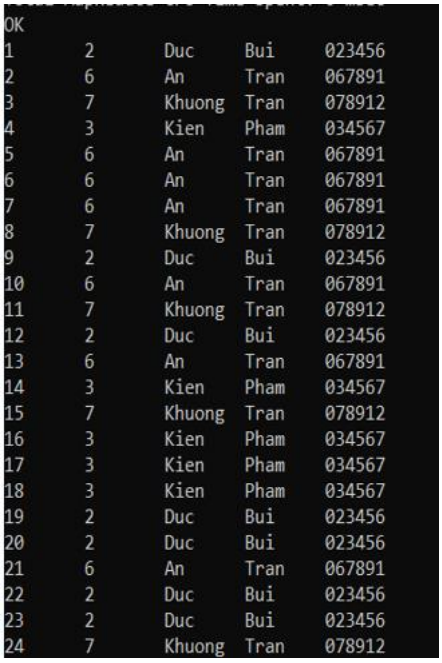
2.5 Truy vấn và báo cáo

- Tìm kiếm thông tin nhân viên và ID đơn hàng mà nhân viên đó thực hiện

Truy vấn HQL:

```
hive> select a.order_id, a.staff_id, b.first_name, b.last_name, b.phone from orders_mart a join staffs_mart b on (a.staff_id=b.staff_id);
```

Sau khi thực hiện truy vấn, kết quả như sau:



```
OK
1       2       Duc     Bui     023456
2       6       An      Tran    067891
3       7       Khuong  Tran    078912
4       3       Kien    Pham    034567
5       6       An      Tran    067891
6       6       An      Tran    067891
7       6       An      Tran    067891
8       7       Khuong  Tran    078912
9       2       Duc     Bui     023456
10      6       An      Tran    067891
11      7       Khuong  Tran    078912
12      2       Duc     Bui     023456
13      6       An      Tran    067891
14      3       Kien    Pham    034567
15      7       Khuong  Tran    078912
16      3       Kien    Pham    034567
17      3       Kien    Pham    034567
18      3       Kien    Pham    034567
19      2       Duc     Bui     023456
20      2       Duc     Bui     023456
21      6       An      Tran    067891
22      2       Duc     Bui     023456
23      2       Duc     Bui     023456
24      7       Khuong  Tran    078912
```

Cột đầu tiên là mã hóa đơn, cột tiếp theo là mã nhân viên, 2 cột tiếp theo lần lượt là tên, họ của nhân viên, cuối cùng là số điện thoại của nhân viên ấy

Nhìn vào kết quả trên, ta có thể biết được từng đơn hàng được nhân viên nào quản lý

- Xem sản phẩm bán chạy nhất

Truy vấn HQL:

```
hive> select a.product_id, a.product_name, count(b.quantity)
    > from products a join order_items b on (a.product_id=b.product_id)
    > group by a.product_id, a.product_name;
```

Kết quả sau truy vấn:

```
2       Ritchey Timberwolf Frameset - 2016      77
3       Surly Wednesday Frameset - 2016 86
4       Trek Fuel EX 8 29 - 2016        97
5       Heller Shagamaw Frame - 2016    91
6       Surly Ice Cream Truck Frameset - 2016   110
7       Trek Slash 8 27.5 - 2016        101
8       Trek Remedy 29 Carbon Frameset - 2016   85
9       Trek Conduit+ - 2016    101
10      Surly Straggler - 2016  97
11      Surly Straggler 650b - 2016     97
12      Electra Townie Original 21D - 2016      104
13      Electra Cruiser 1 (24-Inch) - 2016      103
14      Electra Girls Hawaii 1 (16-inch) - 2015/2016    86
15      Electra Moto 1 - 2016   91
16      Electra Townie Original 7D EQ - 2016    95
17      Pure Cycles Vine 8-Speed - 2016 91
18      Pure Cycles Western 3-Speed - Womens - 2015/2016        89
19      Pure Cycles William 3-Speed - 2016      78
20      Electra Townie Original 7D EQ - Womens - 2016   84
21      Electra Cruiser 1 (24-Inch) - 2016      90
22      Electra Girls Hawaii 1 (16-inch) - 2015/2016    94
23      Electra Girls Hawaii 1 (20-inch) - 2015/2016    100
24      Electra Townie Original 21D - 2016      89
25      Electra Townie Original 7D - 2015/2016  98
26      Electra Townie Original 7D EQ - 2016    90
27      Surly Big Dummy Frameset - 2017 21
28      Surly Karate Monkey 27.5+ Frameset - 2017       26
29      Trek X-Caliber 8 - 2017 24
30      Surly Ice Cream Truck Frameset - 2017   22
31      Surly Wednesday - 2017  13
32      Trek Farley Alloy Frameset - 2017       22
33      Surly Wednesday Frameset - 2017 22
34      Trek Session DH 27.5 Carbon Frameset - 2017     11
35      Sun Bicycles Spider 3i - 2017   17
36      Surly Troll Frameset - 2017     29
37      Haro Flightline One ST - 2017   20
38      Haro Flightline Two 26 Plus - 2017      18
39      Trek Stache 5 - 2017    16
40      Trek Fuel EX 9.8 29 - 2017      21
41      Haro Shift R3 - 2017    24
42      Trek Fuel EX 5 27.5 Plus - 2017 24
43      Trek Fuel EX 9.8 27.5 Plus - 2017       25
44      Haro SR 1.1 - 2017      16
45      Haro SR 1.2 - 2017      32
46      Haro SR 1.3 - 2017      19
47      Trek Remedy 9.8 - 2017  14
48      Trek Emonda S 4 - 2017  28
49      Trek Domane SL 6 - 2017 22
```

Cột đầu tiên là mã sản phẩm, cột thứ hai là tên sản phẩm, cột tiếp theo là số lượng sản phẩm được bán ra được group theo mã sản phẩm và tên sản phẩm

Nhìn vào kết quả truy vấn, ta có thể biết được sản phẩm bán chạy nhất và sản phẩm bán được ít nhất, từ đó có thể điều chỉnh số lượng hàng nhập

- Đếm số lượng đơn hàng và tổng tiền bán ra của từng nhân viên

Truy vấn HQL:

```
hive> select a.staff_id, b.first_name, b.last_name, count(a.order_id), sum(c.list_price)
    > from orders a join staffs b on (a.staff_id=b.staff_id) join order_items c on (a.order_id=c.order_id)
    > group by a.staff_id, b.first_name, b.last_name;
```

Kết quả sau khi truy vấn:

```
2       Duc     Bui     462     565457.7399999973
3       Kien    Pham    544     625915.9299999966
6       An      Tran    1615    1955964.1399999878
7       Khuong  Tran    1580    1938990.5799999882
8       Tung    Diep    269     337904.5099999988
9       Truc    Phan    252     302173.66999999934
```

Đầu tiên là mã nhân viên, hai cột tiếp theo lần lượt là tên, họ của nhân viên, cột thứ tư là số lượng sản phẩm bán ra, cuối cùng là tổng tiền bán ra của từng nhân viên

Dựa vào kết quả truy vấn trên, ta có thể biết được nhân viên nào có doanh thu cao nhất, dựa vào đó ta có thể thưởng hoặc thăng chức cho nhân viên đó

PHẦN KẾT LUẬN

1. Kết quả đạt được

Sau một thời gian nghiên cứu và thực hiện đề tài "***Tìm hiểu Apache Hive và viết ứng dụng demo***", nhóm chúng em đã đạt được những kết quả như sau:

1.1. Kiến thức tìm hiểu được

Nắm bắt được các kiến thức cũng như những vấn đề liên quan trọng về Apache Hive, truy vấn với HQL và áp dụng kiến thức để thiết kế và xây dựng một Data warehouse. Biết được cách Hive hoạt đông trên Hadoop, luồng dữ liệu của Hive, đặc trưng, kiến trúc, cách tổ chức dữ liệu trong Hive.

Nắm bắt được quy trình xử lý Big Data, thiết kế và xây dựng các bảng fact, các dimension, thực hiện tích hợp dữ liệu bằng ETL, HQL để tạo truy vấn phân tích dữ liệu từ những câu truy vấn.

1.2. Chương trình đã làm được

Xây dựng hoàn chỉnh một data warehouse bằng Hive với các chức năng cơ bản như:

- Quản lý các database, quản lý các bảng trong database.
- Thực hiện các câu lệnh truy vấn và tiến hành phân tích từ các câu lệnh truy vấn đó.

2. Ưu điểm

- Hoạt động của Hive diễn ra một cách trơn tru và chính xác, không xảy ra tình trạng lỗi trong hệ thống.
- Lưu trữ được lượng dữ liệu lớn.
- Xử lý thông tin, truy vấn dữ liệu chính xác và nhanh chóng.

3. Nhược điểm

- Chưa cấu hình được Hive Web Interface (HWI) do phiên bản nhóm cài đặt hiện tại quá cao, không hỗ trợ HWI.
- Để có thể sử dụng được HWI, cần phải sử dụng phiên bản Hive 2.2.0 trở xuống. Nhưng các phiên bản Hive 2.2.x không hỗ trợ Derby và Ant nên phải cài thêm. Trong quá trình cài đặt HWI, do bị giới hạn về mặt thời gian và con người nên nhóm vẫn chưa cài đặt được.

4. Hướng phát triển

- Tiếp tục hoàn thiện các chức năng còn thiếu.
- Xây dựng và quản lý data warehouse lớn hơn.
- Tìm hiểu và cài đặt giao diện HWI ở các phiên bản cũ hơn hoặc liên kết với các giao diện hỗ trợ ngôn ngữ khác.

TÀI LIỆU THAM KHẢO

[1]. Admin, *Ngôn Ngữ Lập Trình Hive Là Gì? Cách Thức Làm Việc Của Hive*, Blog.itnavi, Ngày đăng: 28/07/2020.

Link: https://blog.itnavi.com.vn/ngon-ngu-lap-trinh-hive-la-gi

[2]. Apache Hive TM

Link: https://hive.apache.org

[3]. Apache Hive - Apache Software Foundation

Link: https://cwiki.apache.org/confluence/display/Hive

[4]. *How to install Hadoop*, Phoenixnap.com

Link: https://phoenixnap.com/kb/install-hadoop-ubuntu#ftoc-heading-7

[5]. *How to install Apache Hive*, Phoenixnap.com

Link: https://phoenixnap.com/kb/install-hive-on-ubuntu

[6]. *How to create a table in Hive*, Phoenixnap.com

Link: https://phoenixnap.com/kb/hive-create-table

[7]. Pham Thi Hong Anh, *Một số câu lệnh cmd HDFS trong ngôn ngữ Hive*, Viblo.asia, Ngày đăng: 11/06/2020.

Link:https://viblo.asia/p/mot-so-cau-lenh-cmd-hdfs-trong-ngon-ngu-hive-eW65G1XJZDO