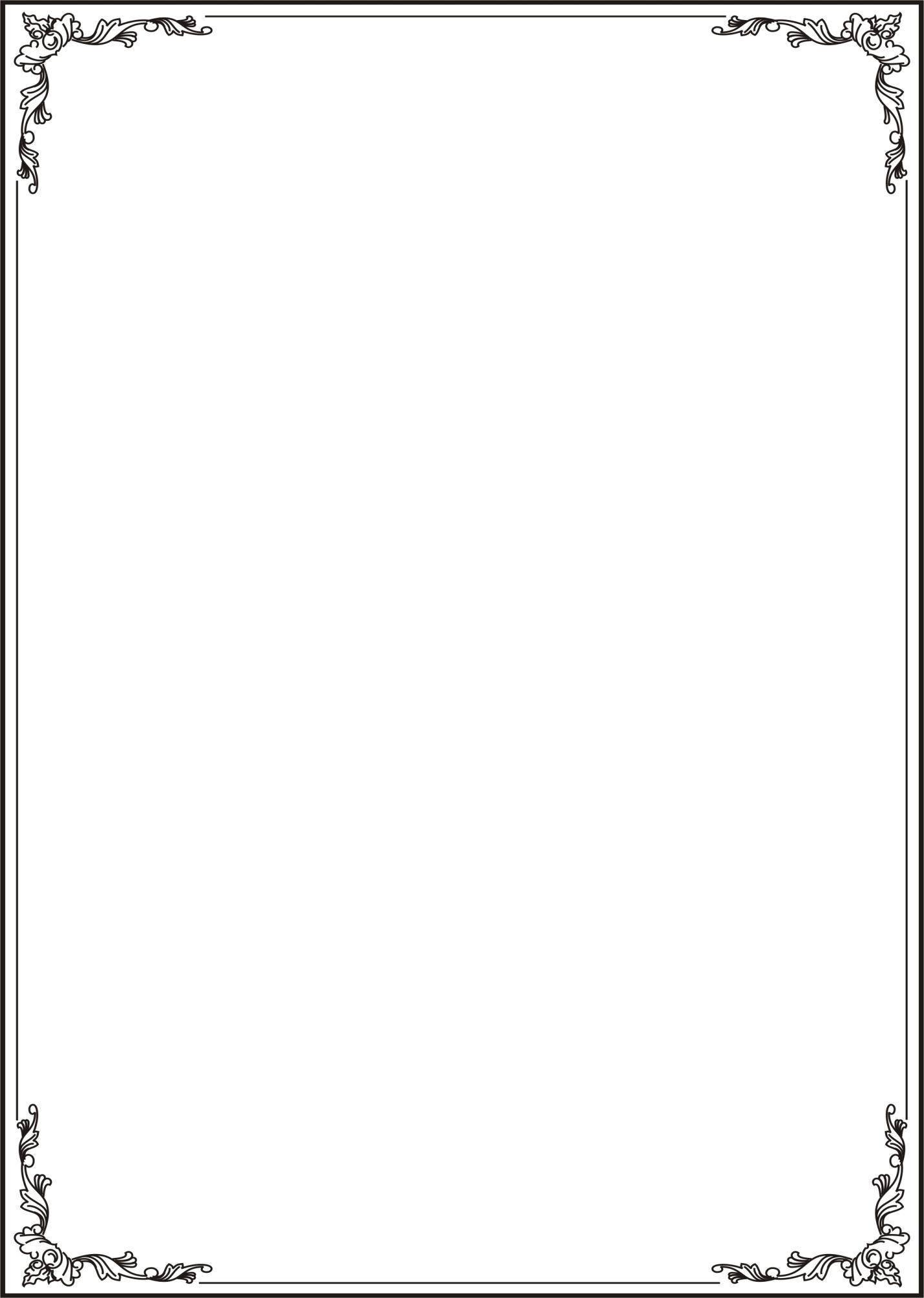
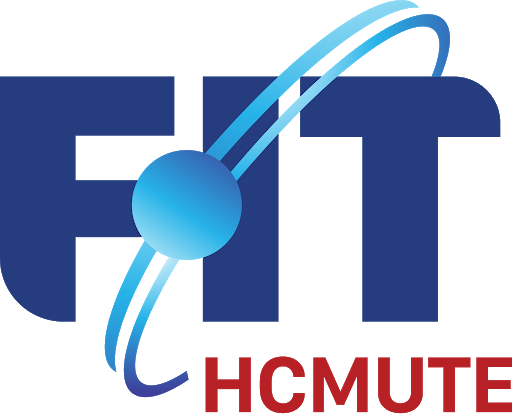
.

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

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN ĐIỆN TOÁN ĐÁM MÂY**

-----🙞🙜🕮🙞🙜-----

****

**NGUYỄN PHẠM DUY KHIÊM - 19133027**

**NGUYỄN THANH TÙNG - 19133065**

**LÊ TUẤN HIỆP - 19133023**

**Đề Tài:**

**TÌM HIỂU APACHE HIVE VÀ VIẾT ỨNG DỤNG DEMO**

**GIẢNG VIÊN HƯỚNG DẪN**

**TS. HUỲNH XUÂN PHỤNG**

**TP HCM, tháng 06 năm 2022**

**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**  \*\*\*\*\*\*\* |
| --- | --- |

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

Họ tên sv 1: **Nguyễn Phạm Duy Khiêm** Mã số sinh viên 1: **19133027**

Họ tên sv 2: **Nguyễn Thanh Tùng** Mã số sinh viên 2: **19133065**

Họ tên sv 3: **Lê Tuấn Hiệp** Mã số sinh viên 3: **19133023**

Ngành: **Kỹ thuật dữ liệu**

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

Giáo viên hướng dẫn: **TS. Huỳnh Xuân Phụng**

**LỜI NHẬN XÉT CỦA GIẢNG VIÊN**

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

2. Ưu điểm của đề tài:

3. Khuyết điểm cần sửa đổi trong đề tài:

5. Đánh giá xếp loại:

6. Điểm:

| *Tp. HCM, ngày tháng năm 2022* | | |
| --- | --- | --- |
|  | Giáo viên hướng dẫn  *(Ký & ghi rõ họ tên)* |

# **LỜI CẢM ƠN**

*Đầu tiên nhóm chúng em xin được phép gửi lời cảm ơn chân thành nhất 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 tốt nhất cho nhóm chúng em được học tập, phát triển nền tảng kiến thức sâu sắc từ đó có thể được đề tài kết thúc môn học này.*

*Bên cạnh đó nhóm chúng em xin gửi đến thầy Huỳnh Xuân Phụng lời cảm ơn sâu sắc nhất trong suốt thời gian dài học tập cũng như thực hiện đề tài trong thời gian qua.*

*Với sự chỉ bảo nhiệt tình, giảng dạy tận tình và đầy đủ kiến thức của thầy Huỳnh Xuân Phụng, nhóm 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.*

*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 đề tài này. Chính vì vậy việc xảy ra những thiếu sót là điều khó có thể tránh khỏi. Nhóm chúng em hi vọng nhận được sự góp ý, nhận xét và phê bình tận tình của quý thầy (cô) qua đó chúng em có thể rút ra được các bài học kinh nghiệm từ đó hoàn thiện và cải thiện nâng cấp lại đề tài của mình một cách tốt nhất có thể.*

*Nhóm chúng em xin chân thành cảm ơn quý thầy(cô)!*

***Sinh viên thực hiện đồ án***

***NGUYỄN PHẠM DUY KHIÊM - 19133027***

***NGUYỄN THANH TÙNG - 19133065***

***LÊ TUẤN HIỆP - 19133023***

**MỤC LỤC**

[**LỜI CẢM ƠN 3**](#_heading=h.1fob9te)

[**PHẦN MỞ ĐẦU 5**](#_heading=h.tyjcwt)

[**1.**](#_heading=h.3dy6vkm) **Tính cấp thiết của đề tài** 5

[**2.**](#_heading=h.1t3h5sf) **Đối tượng nghiên cứu** 5

[**3.**](#_heading=h.4d34og8) **Phạm vi nghiên cứu** 5

[**4.**](#_heading=h.2s8eyo1) **Kết quả dự kiến đạt được** 5

[**PHẦN NỘI DUNG 6**](#_heading=h.17dp8vu)

[**Chương 1 : CƠ SỞ LÝ THUYẾT VỀ APACHE HIVE 6**](#_heading=h.3rdcrjn)

[**1.**](#_heading=h.26in1rg) **Hive là gì ?** 6

[**2.**](#_heading=h.lnxbz9) **Ra đời giải quyết vấn đề gì ?** 6

[**3.**](#_heading=h.35nkun2) **Các tính năng của Hive** 6

[**4.**](#_heading=h.1ksv4uv) **Kiến trúc của Hive** 6

[**5.**](#_heading=h.44sinio) **Quy trình làm việc của Hive** 7

[**Chương 2 : CÀI ĐẶT, THIẾT KẾ VÀ XÂY DỰNG DATA 9**](#_heading=h.2jxsxqh)

[**1.**](#_heading=h.z337ya) **Cài đặt môi trường Hadoop** 9

[**2.**](#_heading=h.3j2qqm3) **Cài đặt Hive** 23

[**3.**](#_heading=h.1y810tw) **Cài đặt AWS EMR** 27

[**4.**](#_heading=h.4i7ojhp) **Sử dụng HUE để tạo cơ sở dữ liệu và thực hiện các truy vấn** 31

[**PHẦN KẾT LUẬN 42**](#_heading=h.2xcytpi)

[**1.**](#_heading=h.1ci93xb) **Kết quả đạt được** 42

[**2.**](#_heading=h.3whwml4) **Ưu điểm** 42

[**3.**](#_heading=h.2bn6wsx) **Nhược điểm** 42

[**4.**](#_heading=h.qsh70q) **Hướng phát triển** 42

[**TÀI LIỆU THAM KHẢO 43**](#_heading=h.3as4poj)

# **PHẦN MỞ ĐẦU**

1. **Tính cấp thiết của đồ án**

Hiện tại, dữ liệu lớn đã và đang dần phát triển rộng rãi trên thế giới, nhằm giải quyết khối lượng lớn dữ liệu này, Apache Software Foundation đã giới thiệu một khuôn khổ được gọi là Hadoop và bên trong nó là Apache Hive để giải quyết các thách thức về quản lý và xử lý Dữ liệu lớn.

1. **Đối tượng nghiên cứu**

Đối tượng nghiên cứu của đề tài này là Dữ liệu lớn. Đồng thời kèm theo đó là các công nghệ áp dụng để xây dựng một 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 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 cho Metastore.

1. **Phạm vi nghiên cứu**

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

1. **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 cho đồ án.

# **PHẦN NỘI DUNG**

# **Chương 1 : CƠ SỞ LÝ THUYẾT VỀ APACHE HIVE**

1. **Hive là gì ?**

Hive là một trong các công cụ cơ sở hạ tầng về data warehouse để xử lý dữ liệu có cấu trúc trong hệ thống Hadoop. Nó nằm trên Hadoop để tóm tắt big data từ đó giúp dễ dàng truy vấn và phân tích.



Hive không phải là :

* Một Cơ sở dữ liệu quan hệ
* Một thiết kế để xử lý các giao dịch trực tuyến (OnLine Transaction Processing - OLTP)
* Một ngôn ngữ cho các truy vấn thời gian thực và cập nhật cấp hàng

1. **Ra đời nhằm giải quyết những vấn đề gì ?**

Hive là một dự án phần mềm cung cấp các truy vấn và phân tích dữ liệu. Nó tạo điều kiện thuận lợi cho việc đọc, viết và xử lý các tập dữ liệu rộng được lưu trữ trong bộ lưu trữ phân tán và được truy vấn bằng các cú pháp ngôn ngữ truy vấn cấu trúc (SQL).

Hive không được xây dựng cho khối lượng công việc xử lý giao dịch trực tuyến (OLTP). Nó thường được sử dụng cho các tác vụ lưu trữ dữ liệu như đóng gói dữ liệu, truy vấn đặc biệt và phân tích các tập dữ liệu khổng lồ.

Nó được thiết kế để nâng cao khả năng mở rộng, khả năng mở rộng, hiệu suất, khả năng chịu lỗi và kết hợp lỏng lẻo với các định dạng đầu vào của nó.

1. **Các tính năng của Apache Hive**

• Hive lưu trữ lược đồ trong CSDL và xử lý các dữ liệu vào HDFS.

• Hive được thiết kế giành cho OLAP.

• Hive cung cấp các ngôn ngữ kiểu SQL dùng để truy vấn được gọi là HiveQL hoặc là HQL.

• Hive là quen thuộc, nhanh chóng và có khả năng mở rộng cao

1. **Kiến trúc của Apache Hive**

***Sơ đồ sau đây miêu tả kiến trúc của Hive :***

Diagram

Description automatically generated

Sơ đồ thành phần này miêu tả kiến trúc của Hive chứa các đơn vị khác nhau.

• User Interface: Hive là một trong những phần mềm cơ sở hạ tầng kho dữ liệu có thể tạo ra được sự tương tác giữa người dùng và HDFS. Các giao diện thông dùng được người dùng sử dụng mà Hive hỗ trợ hiện nay là Web UI, command line, HD Insight (máy chủ Window).

• Meta Store: Hive chọn các máy chủ cơ sở dữ liệu tương ứng để lưu trữ lược đồ hoặc metadata của các bảng, cơ sở dữ liệu, các cột ở trong một bảng dữ liệu, các loại dữ liệu của chúng và ánh xạ lên HDFS.

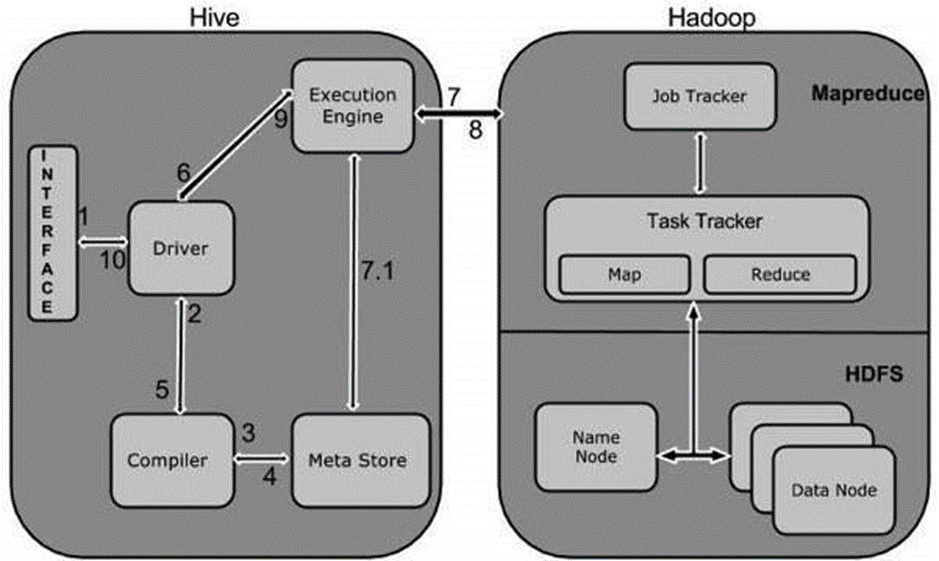
• HiveQL Process Engine: HiveQL tương tự như SQL để truy vấn thông tin lược đồ trên Metastore. Đây thay thế của chương trình MapReduce. Thay vì chúng ta phải viết chương trình MapReduce bằng Java, thay vào đó có thể viết một truy vấn cho MapReduce và xử lý nó một cách nhanh chóng.

• Execution Engine: Phần kết hợp của công cụ xử lý HiveQL và MapReduce là Công cụ thực thi Hive (Hive Execution Engine). Một công cụ thực thi xử lý truy vấn và tạo ra các kết quả giống như kết quả của chương trình MapReduce.

• HDFS hoặc HBASE: Hệ thống tệp phân tán Hadoop hoặc HBASE là các kỹ thuật lưu trữ dữ liệu để lưu trữ dữ liệu vào hệ thống tệp.

1. **Quy trình làm việc của Apache Hive**

***Sơ đồ sau mô tả quy trình làm việc giữa Hive và Hadoop.***

******

Cách Hive tương tác với framework Hadoop:

• Thực thi query: Giao diện Hive như command line hoặc giao diện người dùng web gửi truy vấn đến trình điều khiển để thực thi.

• Nhận kế hoạch: Trình điều khiển có sự trợ giúp của trình biên dịch truy vấn để phân tích cú pháp truy vấn để kiểm tra cú pháp và kế hoạch để thực hiện truy vấn hoặc các yêu cầu của truy vấn.

• Nhận metadata: Trình biên dịch gửi yêu cầu metadata đến Metastore (bất kỳ cơ sở dữ liệu nào).

• Gửi metadata: Metastore gửi metadata như một phản hồi cho trình biên dịch.

• Gửi kế hoạch: Trình biên dịch gửi kiểm tra yêu cầu và gửi lại các kế hoạch cho trình điều khiển thực hiện. Ở đây, việc phân tích các cú pháp và biên dịch một truy vấn đã hoàn thành.

• Kế hoạch thực hiện: Trình điều khiển gửi các kế hoạch thực hiện đến các công cụ thực thi.

• Thực thi công việc: Trong nội bộ, các quá trình thực thi công việc là một công việc MapReduce. Công cụ thực thi gửi các công việc đến JobTracker, trong các node Name và ở đây nó gán công việc này cho TaskTracker xử lý, trong node Data. Đến đây, truy vấn thực thi công việc MapReduce. Hoạt động metadata: Trong khi xử lý, công cụ thực thi có thể thực thi các hoạt động metadata đối với Metastore.

• Lấy kết quả: Công cụ thực thi ghi nhận kết quả từ các node Data.

• Gửi kết quả: Công cụ thực thi gửi các giá trị kết quả đó đến trình điều khiển.

• Gửi kết quả: Trình điều khiển gửi kết quả đến Giao diện người dùng Hive.

# **Chương 2 : CÀI ĐẶT, THIẾT KẾ VÀ XÂY DỰNG DATA**

1. **Cài đặt môi trường Hadoop**
   1. **Cài đặt cụm Hadoop một Node trên Ubuntu Server**Text

      Description automatically generated

-kiể m tra ip

Text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

- Thêm vào các nội dung sau trong ip address:

Text

Description automatically generated

* 1. **Cài đặt Oracle Java 16**

Text

Description automatically generated with medium confidence

* 1. **Cài đặt SSH**

Text

Description automatically generated

* 1. **Cấu hình host/host name**
     1. **Kiểm tra ip của các máy master, slave**

# ifconfig

kiểm tra coi nó giống với trên ip network không

* + 1. **Cấu hình host**

-Text

Description automatically generated

* + 1. **hostname cho master**

Graphical user interface

Description automatically generated with medium confidence

* + 1. **hostname cho slave**

A picture containing shape

Description automatically generated

* 1. **Tạo User Hadoop**

- Tạo user hadoopuser để quản lý các permission cho đơn giản

* 1. **Cài đặt Hadoop 3.3.1**

- Chuyển qua hadoopuser

- Chuyển qua thư mục /home/hadoopuser để download filevà giải nénText

Description automatically generated

* 1. **Cấu hình các thông số cho Hadoop**
     1. **File .bashrc**

Text

Description automatically generated

* + 1. **File hadoop-env.sh**

- Tìm đoạn export JAVA\_HOME=... rồi sửa lại thành



* + 1. **File core-site.xml**

# vim ~/hadoop/etc/hadoop/core-site.xml

rồi chỉnh lại cấu hình cho phù hợp với môi trường hadoop

* + 1. **File maped-ste(chỉ cấu hình ở master)**

# cd ~/hadoop/etc/hadoop/

# vim mapred-site.xml

rồi chỉnh lại cấu hình cho phù hợp với môi trường hadoop

* + 1. **File hdfs-site.xml**

# vim ~/hadoop/etc/hadoop/hdfs-site.xml

rồi chỉnh lại cấu hình cho phù hợp với môi trường hadoop

* + 1. **File yarn-site.xml**

~/hadoop/hadoop-yarn-project/hadoop-yarn/conf

# cd

# vim ~/hadoop/etc/hadoop/yarn-site.xml

rồi chỉnh lại cấu hình cho phù hợp với môi trường hadoop

* 1. **Tạo máy khiem-slave1**

Text

Description automatically generated

* 1. **Cài đặt ssh key giữa các node**

Text

Description automatically generated

Text

Description automatically generated

* 1. **Format namenode**

- Chỉ làm 1 lần

* 1. **Kiểm tra mọi thứ đã ổn**

- Trên master chúng ta chạy lệnh start-all.sh bắt đầu kiểm tra

# jps

Text

Description automatically generated

# hdfs dfsadmin –report

Text

Description automatically generated

1. **Cài đặt Hive**
   1. **Download và install Hive**

Tải Hive phiên bản 3.2.1

Text

Description automatically generated

* 1. **Cài đặt biến môi trường cho Hive**

A screenshot of a computer

Description automatically generated with medium confidence

* 1. **Cấu hình Hive**

Text

Description automatically generated

* 1. **Download và install Apache Derby**

Hệ thống sử dụng apache derby 10.15.2.0

Text

Description automatically generated

Text

Description automatically generated

* 1. **Tạo thư mục để lưu Metastore**

Text

Description automatically generated

Text

Description automatically generated

Kiểm tra Hive

Text

Description automatically generated

Show tables

Text

Description automatically generated

Show databases

Graphical user interface, text

Description automatically generated

1. **Cài đặt AWS EMR**

Ta vào services rồi tìm EMR

Graphical user interface, text, application

Description automatically generated

Ta bắt đầu tạo cluster

Graphical user interface, text

Description automatically generated

General Configuration

Cluster name là tên của cluster

S3 folder tức là folder cluster

Software configuration

Release thì chọn phiên bản mà ta muốn sử dụng, ở đây em chọn emr-5.17.0

Hardware Configuration

Instance type thì em chọn m5.xlagre và chọn 3 instance trong đó có 1 master và core nodes

Security and Access

Chonj key pair là cái file dockerclass.pam đã tạo từ đầu học kỳ.

Rồi sau đó nhấn chọn

Graphical user interface, application, table

Description automatically generated with medium confidence

s3://us-east-1.elasticmapreduce.samples/cloudfront/code/Hive\_CloudFront.q

s3://us-east-1.elasticmapreduce.samples

s3://doancuoiky/test

Graphical user interface, text, application

Description automatically generated Trong phần Output S3 location, em sử dụng 1 bucket tên doancuoiky đã tạo sẵn:

Graphical user interface, text, application, email

Description automatically generated

Sau khi tạo xong step ta phải chờ 4 5 phút để nó hoàn thành, sau đó ta vào trong thư mục test/ trong bucket doancuoiky/ se co file os\_request

Graphical user interface, text, application, email

Description automatically generated

Nếu ra được như hình thì chứng tỏ là step đã chạy thành công

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, application

Description automatically generated

Giờ ta sẽ chạy kết nối tới Node Master sử dụng SSH

ssh -i dockerclass.pem hadoop@ec2-3-86-189-118.compute-1.amazonaws.com

#note: mỗi lần tạo lại cluster thì câu lệnh trên sẽ khác ip.

#note: nhớ bỏ ~/ trong câu lệnh từ trang aws

Text

Description automatically generated

1. **Sử dụng HUE để tạo cơ sở dữ liệu và thực hiện các truy vấn**
   1. **Tìm hiểu và sử dụng HUE**

Giờ ta mở HUE theo đường dẫn từ trang Application user interfaces, nhờ là trong phần đặt port 8888 trong inbound rules trước

Text, table

Description automatically generated

Sau khi vào được Hue thì nó sẽ ra giống như hình

Graphical user interface, application

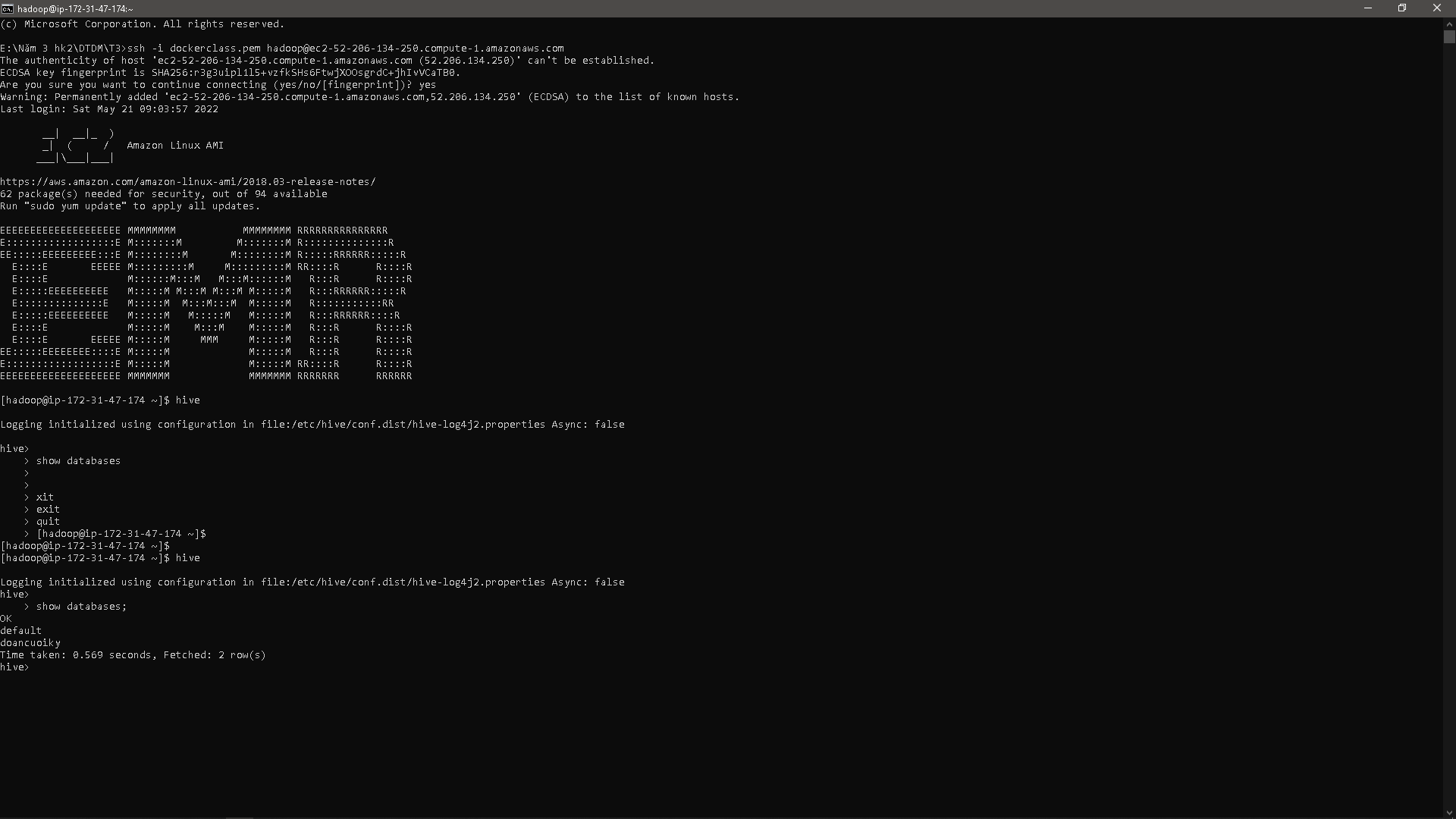
Description automatically generated

Tiếp đến ta tạo database có tên là doancuoiky

Graphical user interface, text, application

Description automatically generated

Rồi ta xuống cmd test thử, nếu ra kết quả giống bên dưới là ta đã tạo thành công



Tiếp đến ta import một table từ file stores.txt

Table

Description automatically generated with low confidence

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Rồi ta xuống cmd kiểm tra, nếu hiện ra bên dưới tức là table đã tạo thành công

Text

Description automatically generated

Giờ ta sẽ truy vấn mẫu, vd: truy vấn tất cả có trong table stores

Text

Description automatically generated

Text

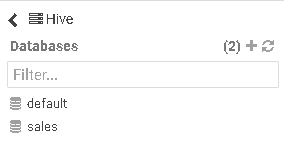
Description automatically generated

Hình dưới đây cho ta thấy 3 cluster(1 master- 2 core) đang chạy.

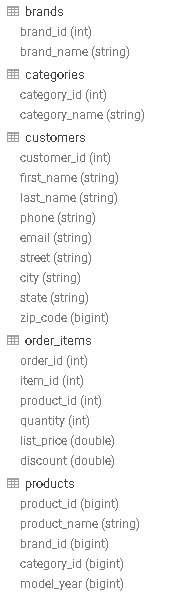
Graphical user interface, application

Description automatically generated

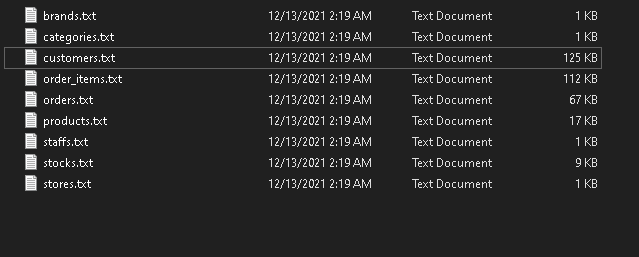
* 1. **Tạo cơ sở dữ liệu để truy vấn**

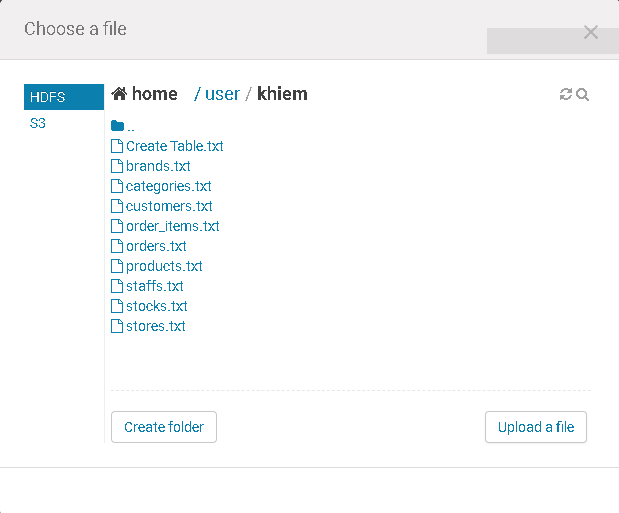
**Đầu tiên ta sẽ tạo 1 data base tên là Sales**

**Trong data base đó ta tạo các tables:**

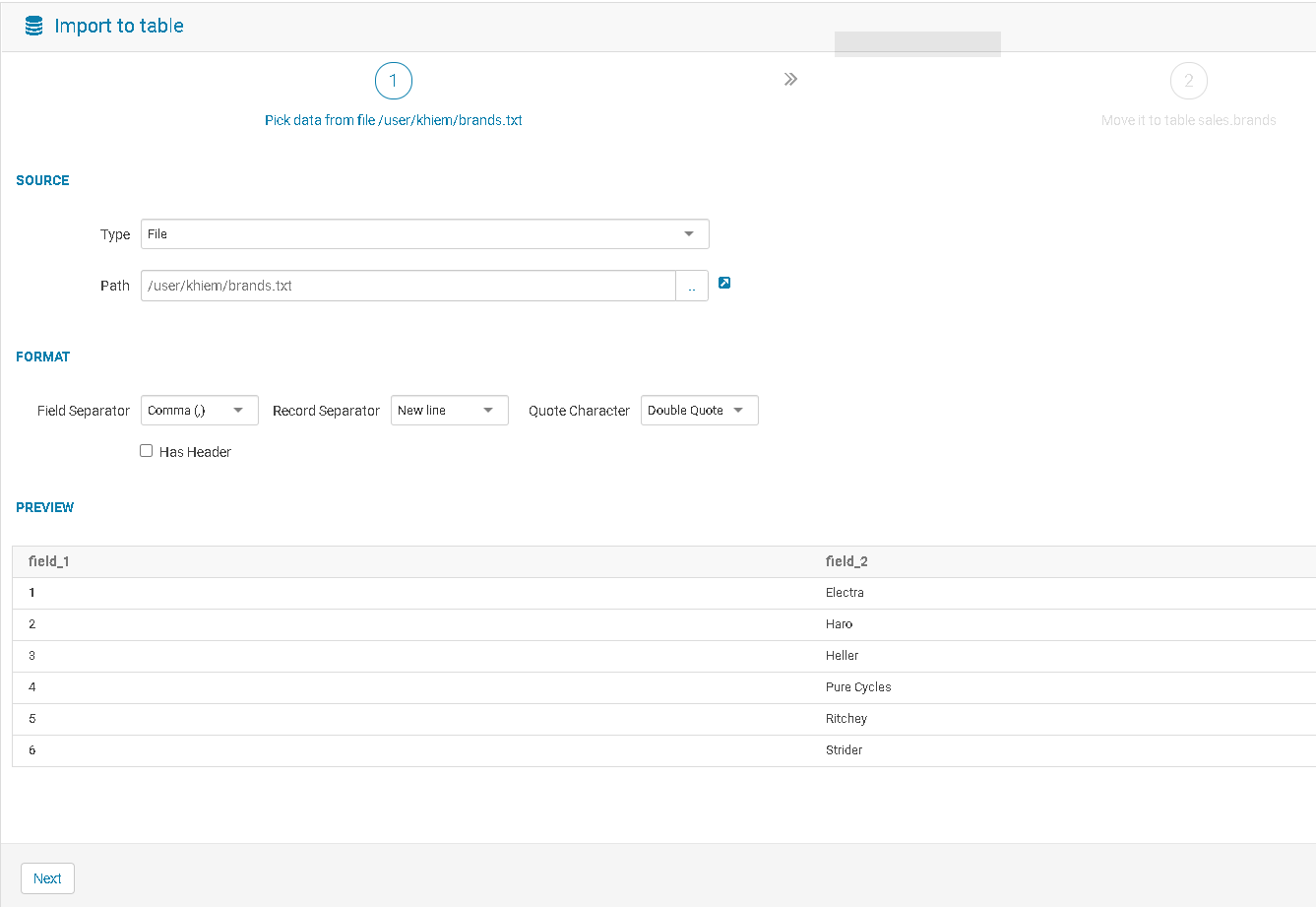
****

**Tụi em sẽ tạo các table này bằng cách tải file txt chuẩn bị sẵn từ máy lên**

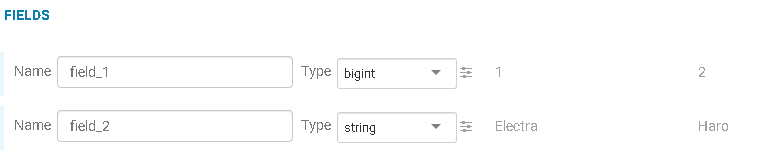
****

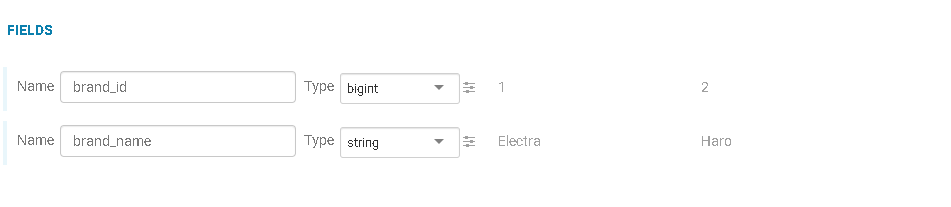
****

**Sau đó ta chọn 1 file để bắt đầu tạo table, ví dự như file brand.txt:**

****

**Sau khi chọn file txt sẽ ra giao diện như hình để ta có thể nhìn tổng quát file, tiếp đên ta sẽ chỉnh lại tên các thuộc tính do file mà ta upload là chỉ có data nhưng ko có định dạng thuộc tính nên nó sẽ mặc định là field\_1,field\_2,… .**

****

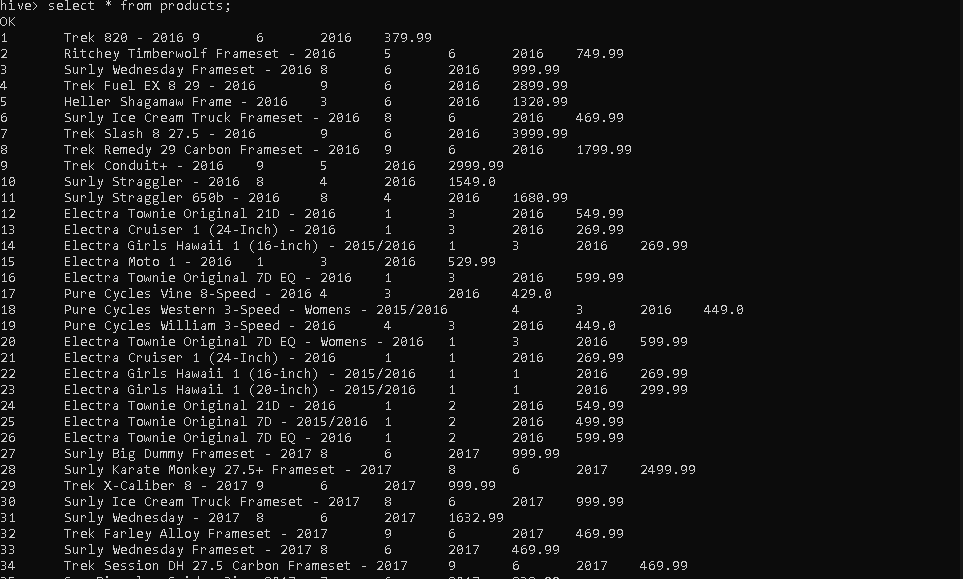
****

**Sau khi tạo xong các tables, ta xuống cmd để kiêm tra**

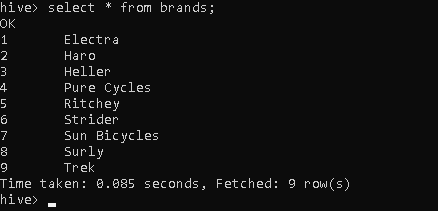
****

**Thực hiện các truy vấn đơn giản**

**select \* from products;**



select \* from brands;



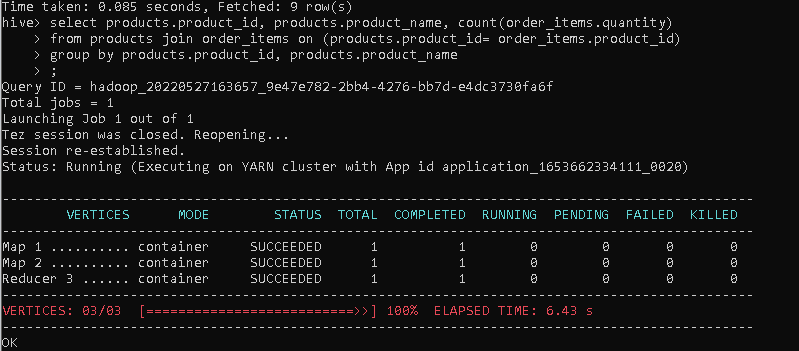
Xem sản phẩm bán chạy nhất

Truy vấn HQL:

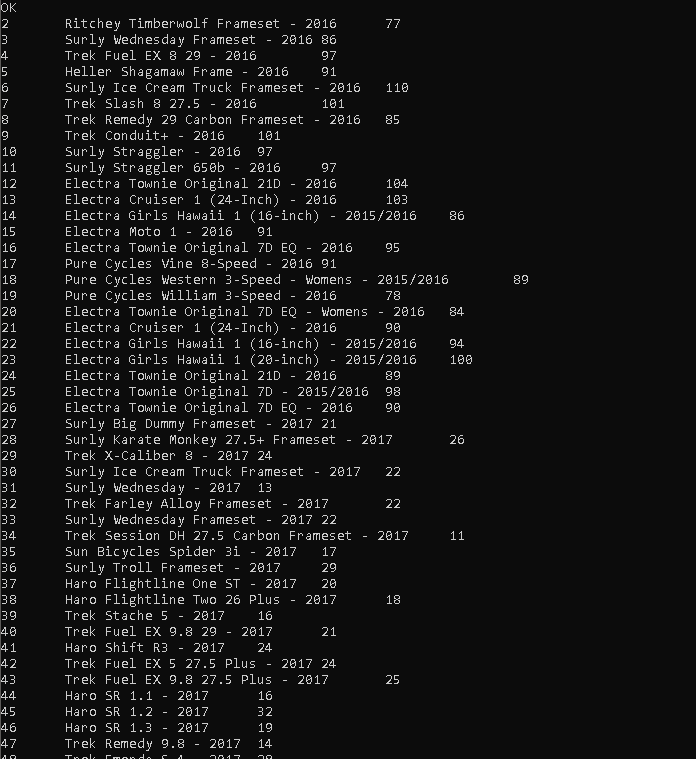
select products.product\_id, products.product\_name, count(order\_items.quantity)

from products join order\_items on (products.product\_id= order\_items.product\_id)

group by products.product\_id, products.product\_name;



Kết quả:



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 biêt được sản phẩm bán chạy nhất và sản phẩm bán được ít nhất, từ đó có thể điều chỉnh số lượng hàng nhập

# **PHẦN KẾT LUẬN**

1. **Kết quả thực hiện và thực nghiệm được**

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

Nắm bắt được các kiến thức quan trọng về Apache Hive, truy vấn với HQL và áp dụng kiến thức để thiết kế và xây dựng một Data warehouse đơn giản. Biết được cách Apache Hive hoạt động trên Hadoop, luồng dữ liệu của Hive, kiến trúc, đặc trưng, 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 dim, thực hiện tích hợp dữ liệu bằng ETL, HQL để tạo truy vấn phân tích dữ liệu từ những câu truy vấn.

1. **Ưu điểm**

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

1. **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 đã làm hết sức có thể để thực hiện đồ án.

1. **Hướng phát triển**

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

# **TÀI LIỆU THAM KHẢO**

1. Apache Hive Tool: <https://www.jetbrains.com/datagrip/features>
2. Apache Hive: <https://hive.apache.org>
3. <https://www.youtube.com/watch?v=dU40df0Suoo>
4. <https://docs.aws.amazon.com/emr/latest/ReleaseGuide/emr-hue.html>
5. <https://viblo.asia/p/hadoop-cai-dat-hadoop-tren-ubuntu-phan-1-4dbZNo3qlYM>
6. <https://phambinh.net/bai-viet/huong-dan-cai-dat-hadoop-cluster-tren-ubuntu-18-04/>
7. <https://phoenixnap.com/kb/install-hive-on-ubuntu>