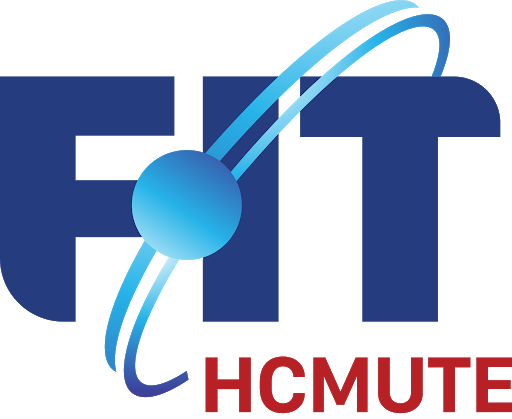
.

**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**

**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: **Nguyễn Phạm Duy Khiêm** MSSV 1: **19133027**

Họ và tên Sinh viên 2: **Nguyễn Thanh Tùng** MSSV 2: **19133065**

Họ và tên Sinh viên 2: **Lê Tuấn Hiệp** MSSV 2: **19133023**

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: **Nguyễn Phạm Duy Khiêm** MSSV 1: **19133027**

Họ và tên Sinh viên 2: **Nguyễn Thanh Tùng** MSSV 2: **19133065**

Họ và tên Sinh viên 2: **Lê Tuấn Hiệp** MSSV 2: **19133023**

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 2022*

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***

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

***NGUYỄN THANH TÙNG - 19133065***

***LÊ TUẤN HIỆP - 19133023***

**MỤC LỤC**

[LỜI CẢM ƠN 4](#_Toc90051286)

[MỤC LỤC 5](#_Toc90051287)

[DANH MỤC CÁC HÌNH 7](#_Toc90051288)

[PHẦN MỞ ĐẦU 8](#_Toc90051289)

[**1. Tính cấp thiết của đề tài** 8](#_Toc90051290)

[**2. Đối tượng nghiên cứu** 8](#_Toc90051291)

[**3. Phạm vi nghiên cứu** 8](#_Toc90051292)

[**4. Kết quả dự kiến đạt được** 8](#_Toc90051293)

[PHẦN NỘI DUNG 9](#_Toc90051294)

[CHƯƠNG 1: CƠ SỞ LÝ THUYẾT VỀ APACHE HIVE 9](#_Toc90051295)

[**1.1** **Tổng quan về Hive** 9](#_Toc90051296)

[**1.2** **Kiến trúc của Hive** 10](#_Toc90051297)

[**1.3** **Hoạt động của Hive** 11](#_Toc90051298)

[**1.4** **Mô hình dữ liệu trong Hive** 12](#_Toc90051299)

[*1.4.1. Tổ chức dữ liệu* 12](#_Toc90051300)

[*1.4.2. Kiểu dữ liệu* 13](#_Toc90051301)

[**1.5** **HiveSQL (HQL)** 14](#_Toc90051302)

[CHƯƠNG 2: CÀI ĐẶT, THIẾT KẾ VÀ XÂY DỰNG DATA WAREHOUSE 15](#_Toc90051303)

[**2.1 Cài đặt** 15](#_Toc90051304)

[*2.1.1 Cài đặt môi trường Hadoop* 15](#_Toc90051305)

[*2.1.2 Cài đặt Apache Hive* 22](#_Toc90051306)

[*2.1.3 Lỗi phát sinh trong quá trình cài đặt* 25](#_Toc90051307)

[**2.2 Triển khai Hive Web Interface** 27](#_Toc90051308)

[**2.3 Thiết kế Data warehouse** 32](#_Toc90051309)

[**2.4 Xây dựng Data warehouse** 33](#_Toc90051310)

[*2.4.1 Tạo các file dữ liệu* 33](#_Toc90051311)

[*2.4.2 Tạo Database trong Hive* 34](#_Toc90051312)

[*2.4.3 Tạo Table và load dữ liệu* 34](#_Toc90051313)

[**2.5 Truy vấn và báo cáo** 41](#_Toc90051314)

[PHẦN KẾT LUẬN 45](#_Toc90051315)

[**1.** **Kết quả đạt được** 45](#_Toc90051316)

[*1.1. Kiến thức tìm hiểu được* 45](#_Toc90051317)

[*1.2. Chương trình đã làm được* 45](#_Toc90051318)

[**2.** **Ưu điểm** 45](#_Toc90051319)

[**3.** **Nhược điểm** 45](#_Toc90051320)

[**4.** **Hướng phát triển** 46](#_Toc90051321)

[TÀI LIỆU THAM KHẢO 47](#_Toc90051322)

# **DANH MỤC CÁC HÌNH**

[**Hình 1: Kiến trúc của Hive** 10](#_Toc90049634)

[**Hình 2: Sơ đồ luồng hoạt động của Hive** 10](#_Toc90049635)

[**Hình 3: Tổ chức dữ liệu trong Hive** 11](#_Toc90049636)

# **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**

## **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.

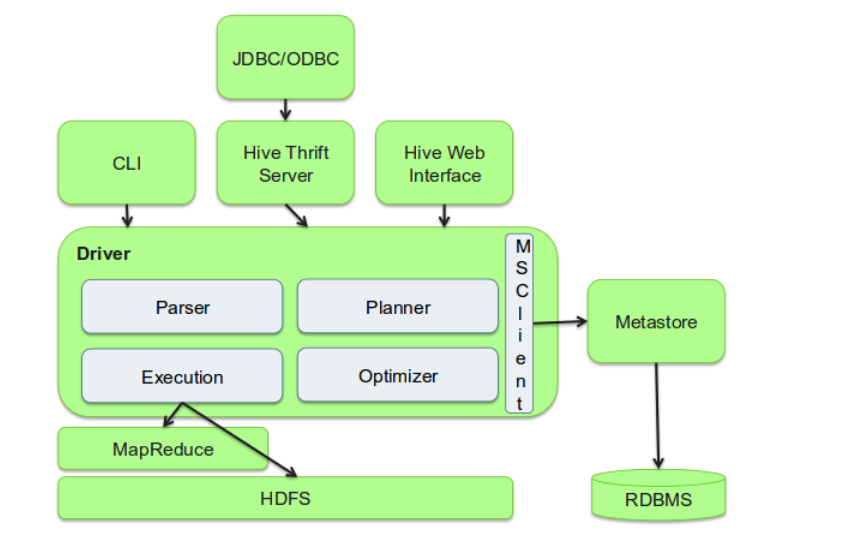
Hive không được thiết kế để cho các giao dịch online (OLTP workloads) và không nên dùng cho các real-time queries và các cập nhật trên từng dòng trong 1 table (row-level). Hive hoạt động tốt nhất cho các batch jobs trên các tập dữ liệu lớn, mà ở đó dữ liệu được thêm vào liên tục (append-only data) ví 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.

## **Kiến trúc của Hive**

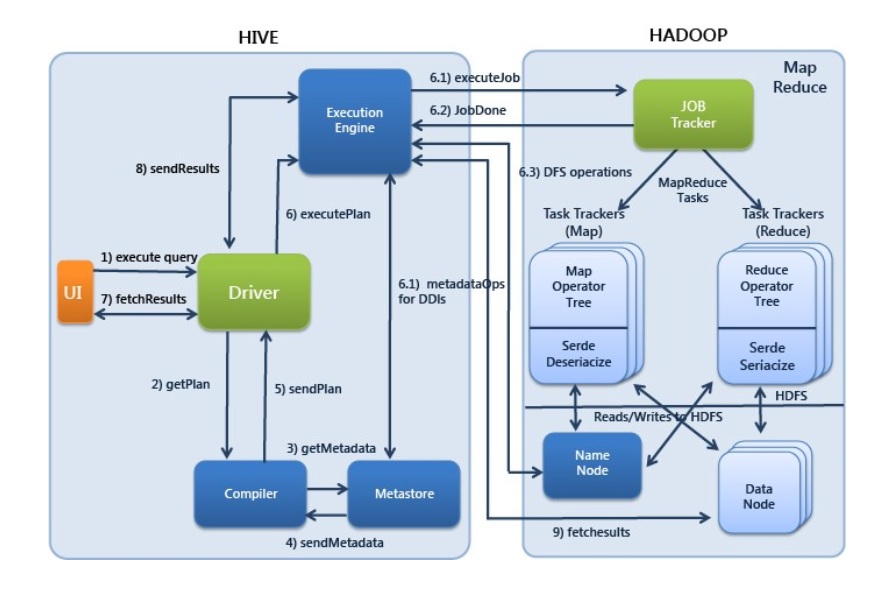
Hive có các thành phần chính là:

* 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**

## **Hoạt động của Hive**



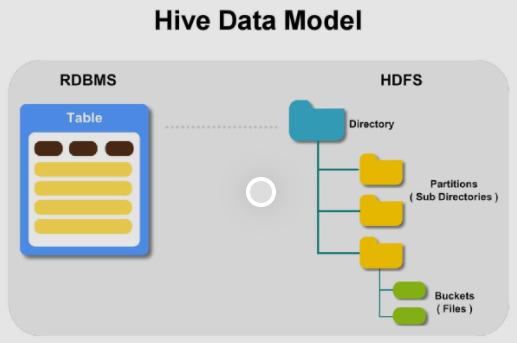
**Hình 2: Sơ đồ luồng hoạt động của Hive**

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)

## **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.

## **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

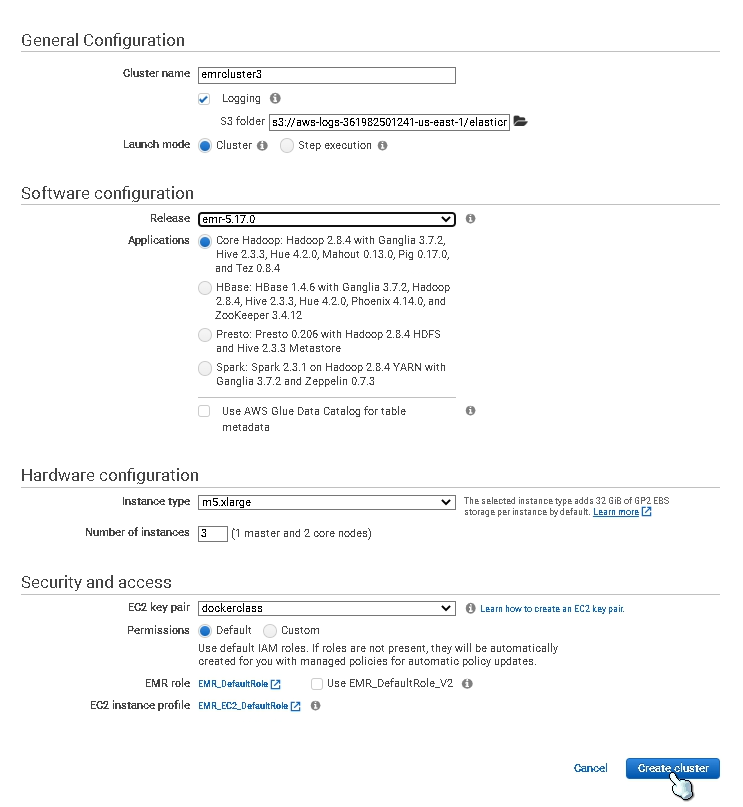
GROUP BY category;

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

## **2.1 Cài đặt môi trường Hadoop và Hive:**

## **M dán cái bài hadoop với hive của m zo đây**

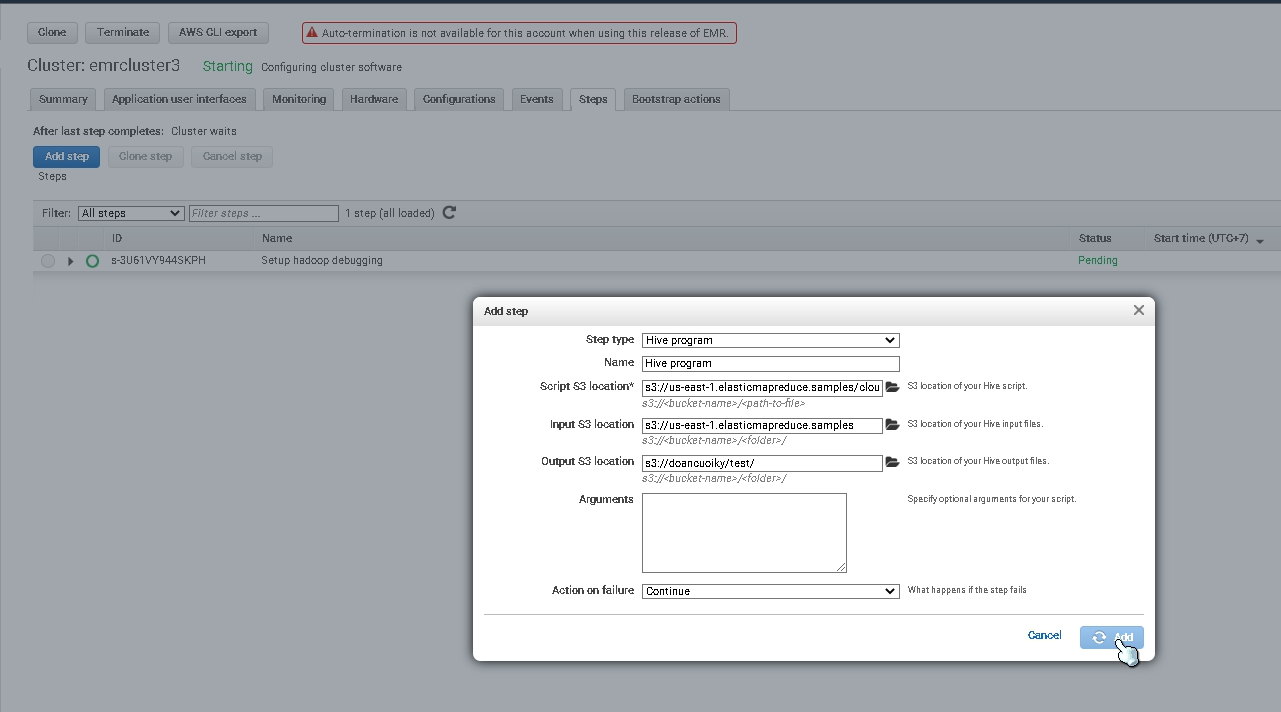
## **2.2 Cài đặt AWS EMR:**



S3://us-east-1.elascticmapreduce.samples/cloudfront/code/Hive\_CloudFront.q

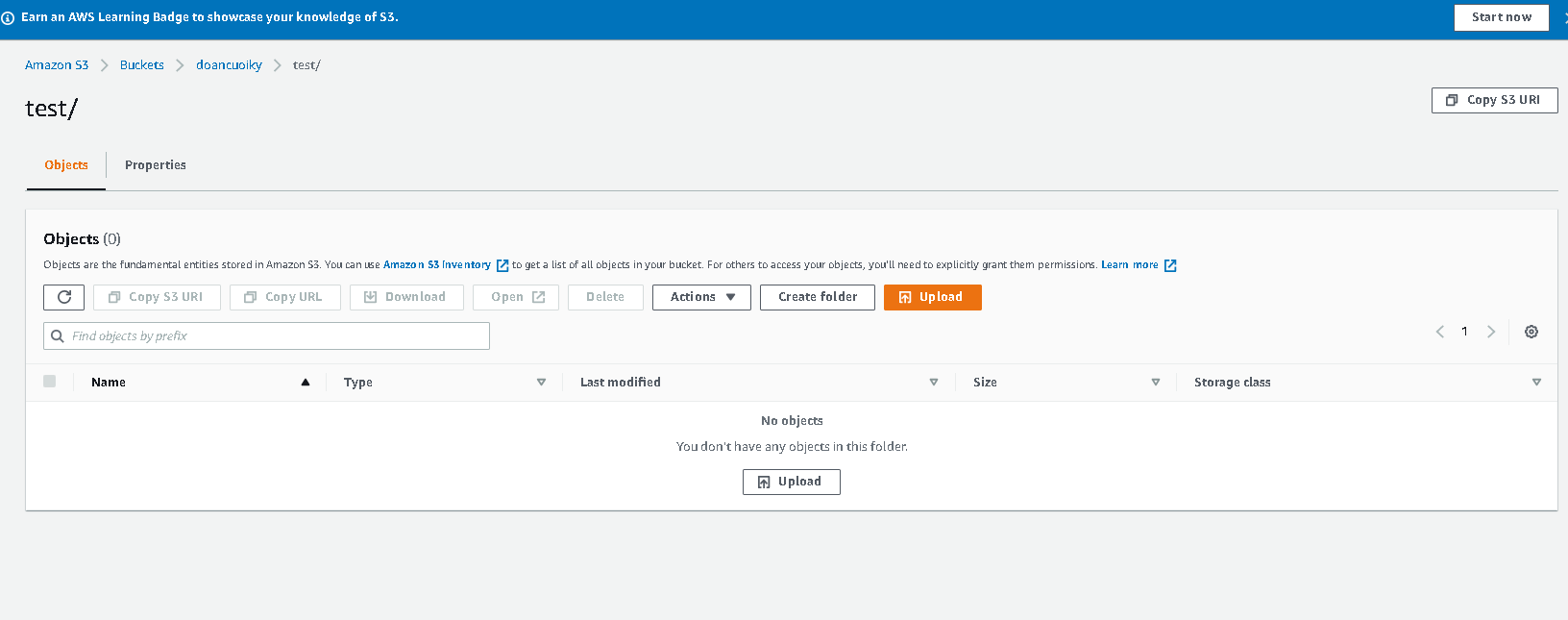
S3://us-east-1.elascticmapreduce.samples

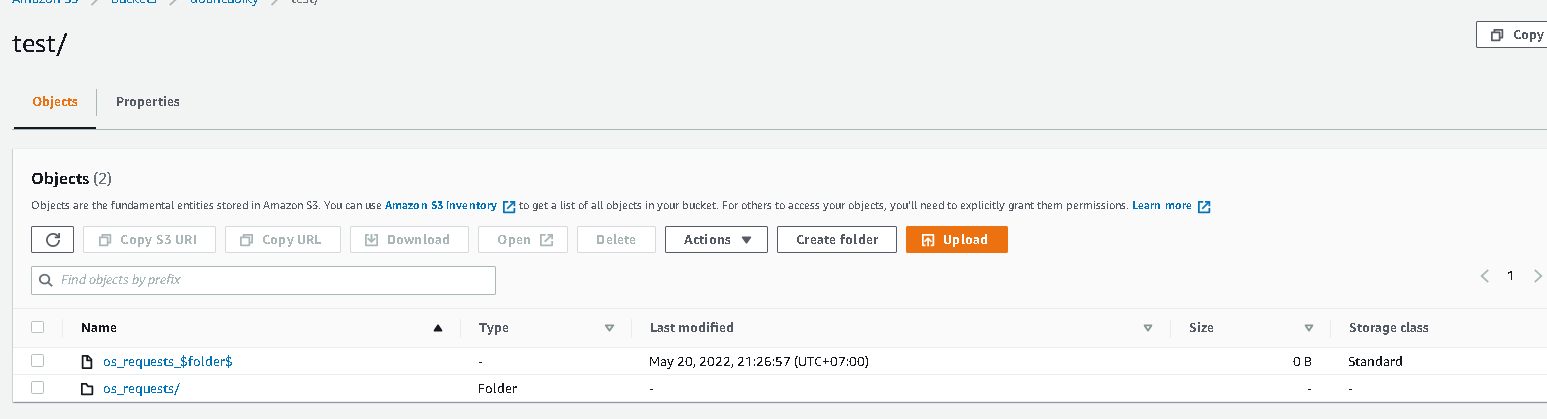
S3://doancuoiky/test



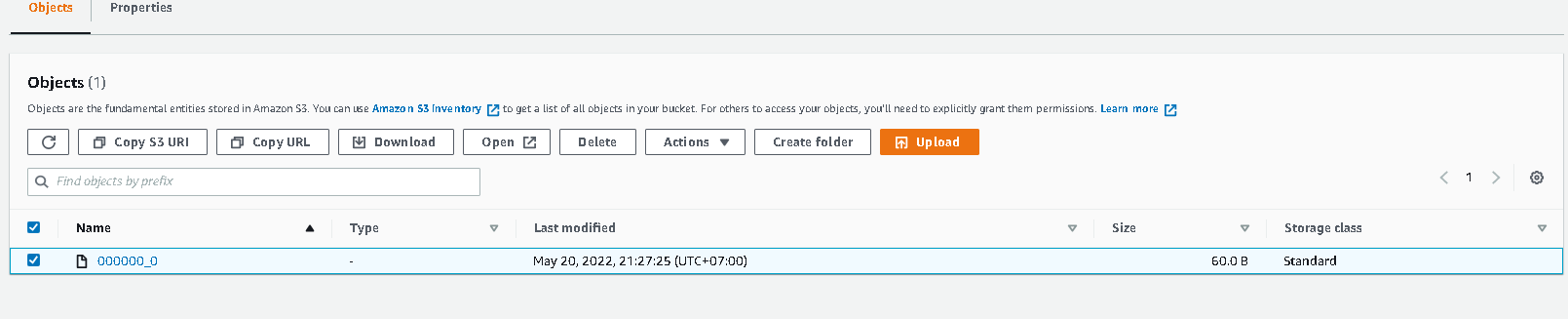
Trong phần Output S3 location, emd sử dụng 1 bucket tên doancuoiky đã tạo sẵn:

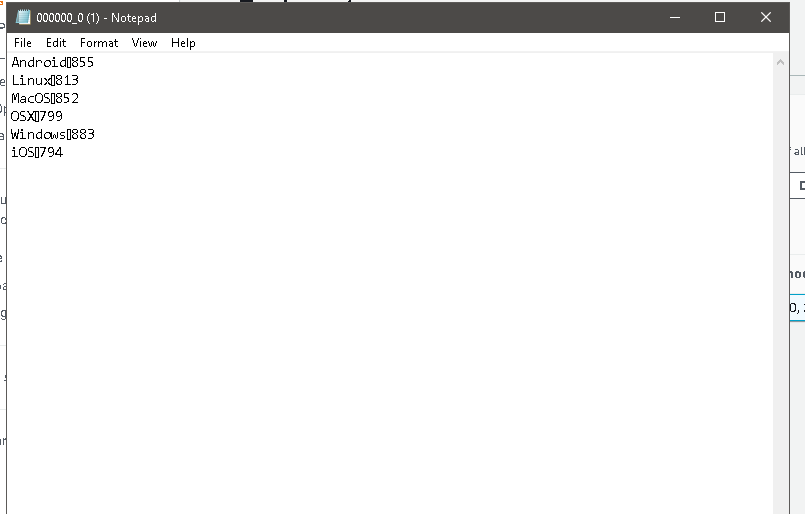
Sau khi tạo xong step ta phải chờ 4 5 phút để nó hoàn thành, sau đó ta vào trong thư mục test/ trong bucket doancuoiky/ se co file os\_request





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

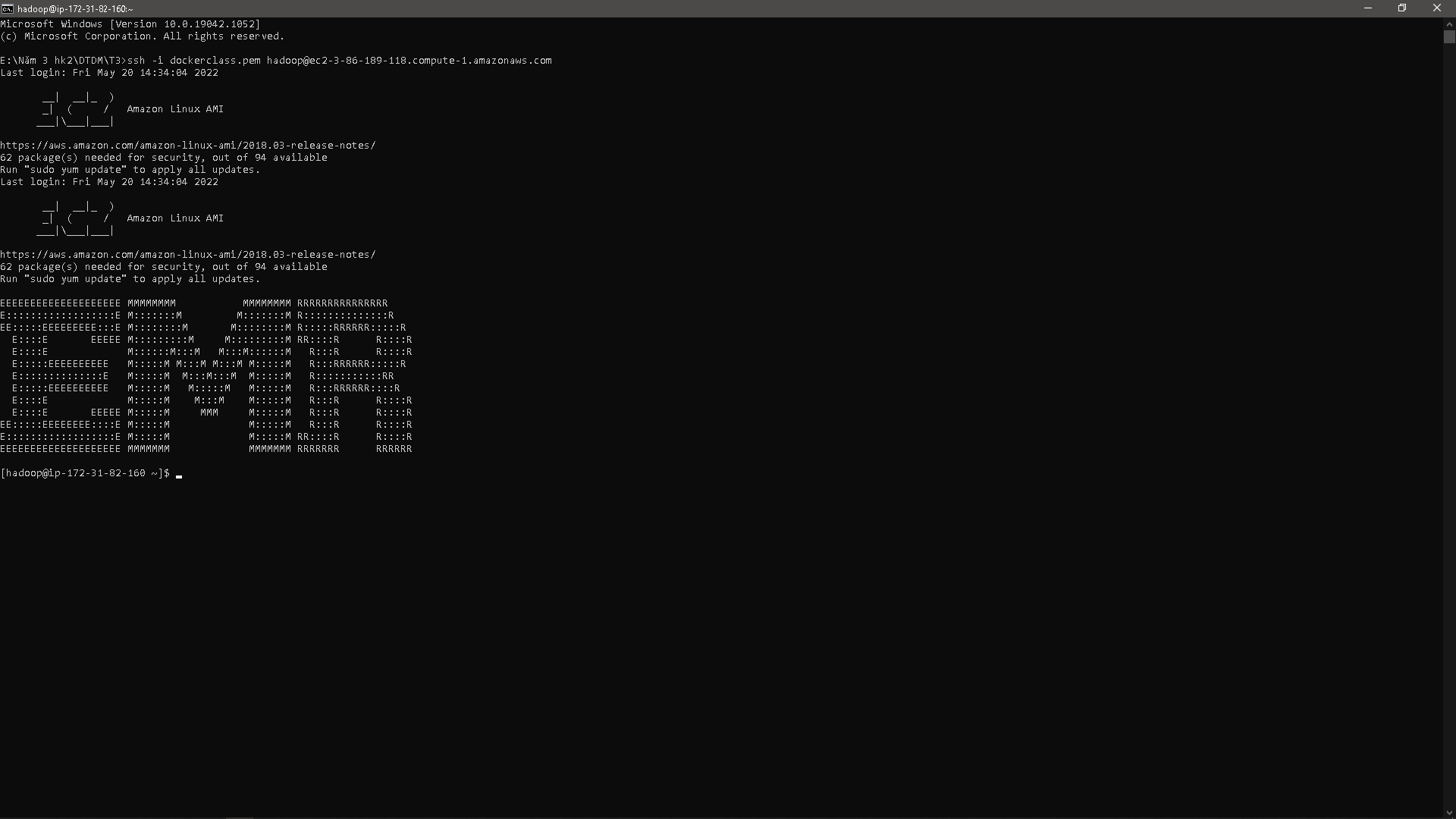




Giờ ta sẽ chạy kết nối tới Node Master sử dụng SSH

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

## **2.2 Cài đặt database:**



# **PHẦN KẾT LUẬN**

## **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*

## **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.

## **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.

## **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**