Nội dung thực hành

[I. Viết câu truy vấn Hive sử dụng Shell, Scripts và Hue 1](#_Toc9349996)

[Bài 1: Chạy câu truy vấn trong Hive shell 1](#_Toc9349997)

[Bài 2: Chạy truy vấn trực tiếp từ Terminal. 2](#_Toc9349998)

[Bài 3: Chạy truy vấn được viết trong file \*.hql 3](#_Toc9349999)

[Bài 4: Chạy truy vấn sử dụng Hue 3](#_Toc9350000)

[II. Quản trị dữ liệu với Hive 5](#_Toc9350001)

[Bài 1: Dùng Sqoop import để tạo bảng và dữ liệu trong Hive 5](#_Toc9350002)

[Bài 2: Tạo bảng external trong Hive 6](#_Toc9350003)

[Bài 3: Tạo vào đẩy dữ liệu vào bảng dạng hive-managed. 7](#_Toc9350004)

[Bài 4: Tạo, đẩy dữ liệu và truy vấn bảng có cột dạng complex 9](#_Toc9350005)

[III. Phân tích cảm xúc (sentiment analysis) dùng Hive 11](#_Toc9350006)

[Bài 1: Phân tích rating sản phẩm 11](#_Toc9350007)

[Bài 2: Phân tích bình luận 11](#_Toc9350008)

# **Viết câu truy vấn Hive sử dụng Shell, Scripts và Hue**

Quan trọng: để chuẩn bị dữ liệu cho bài tập dưới bạn chạy lệnh dưới. Khi màn hình hiện ra các lựa chọn gõ số **6.**

~/scripts/analyst/catchup.sh

## Bài 1: Chạy câu truy vấn trong Hive shell

Dualcore tạo cuộc thi cho khách hàng gửi những đoạn video về cách họ sử dụng sản phẩm tablet mua từ Dualcore. Một giải thưởng 5000 USD được trao cho khách hàng nhận được bình chọn cao nhất. Tuy nhiên, không may là dữ liệu đăng ký bị mất, Dualcore chỉ còn thông tin về video của người đoạt giải. Trong video này, khách hàng trúng thưởng ghi tên là “Bridget from Kansas City”. Bài tập này, bạn sẽ viết câu truy vấn vào bảng customers để tìm ra thông tin chính thức về người chiến thắng để trao giải cho cô ấy.

1. Mở Terminal window và chuyển đến thư mục chứa dữ liệu cho bài thực hành này

cd /home/training/training\_materials/analyst/exercises/analyzing\_sales

1. Bạn mở Hive shell để viết truy vấn bằng cách gõ từ khóa *hive* trong Terminal

hive

1. Để kết quả truy vấn dễ đọc, bạn có thể hiển thị tên cột trong kết quả bằng cách

set hive.cli.print.header=true;

1. Bạn hãy viết truy vấn vào bảng customers để tìm ra bản ghi của người có tên là Bridget và sống ở Kansas (để biết chính xác cần filter theo trường nào trong bảng customers, bạn có thể xem cấu trúc của bảng này trước khi viết truy vấn bằng lệnh *describe customers;*)

*Trường hợp bạn gặp khó khăn có thể tham khảo đáp án:*

select \* from customers where fname like '%Bridget%' and city like '%Kansas%';

1. Thoát khỏi hive shell bằng lệnh *quit*

quit;

## Bài 2: Chạy truy vấn trực tiếp từ Terminal.

Mở Terminal window và có thể chạy trực tiếp lệnh HiveQL với cú pháp *hive -e ‘câu\_lệnh’*

hive -e 'select price, brand, name from products order by price desc limit 3'

Kết quả như sau:

975149 Byteweasel Hadoop Cluster, Economy (4-node)

614559 Gigabux Server (2U rackmount, eight-core, 64GB, 12TB)

599319 Krustybitz Server (2U rackmount, eight-core, 64GB, 12TB)

## Bài 3: Chạy truy vấn được viết trong file \*.hql

File *verify\_tablet\_order.hql* viết truy vấn để tìm xem Bridget (có Id=1139477) có mua sản phẩm tablet (Id = 1274348) hay không. Chúng ta có thể chạy truy vấn được viết trong file này bằng cách

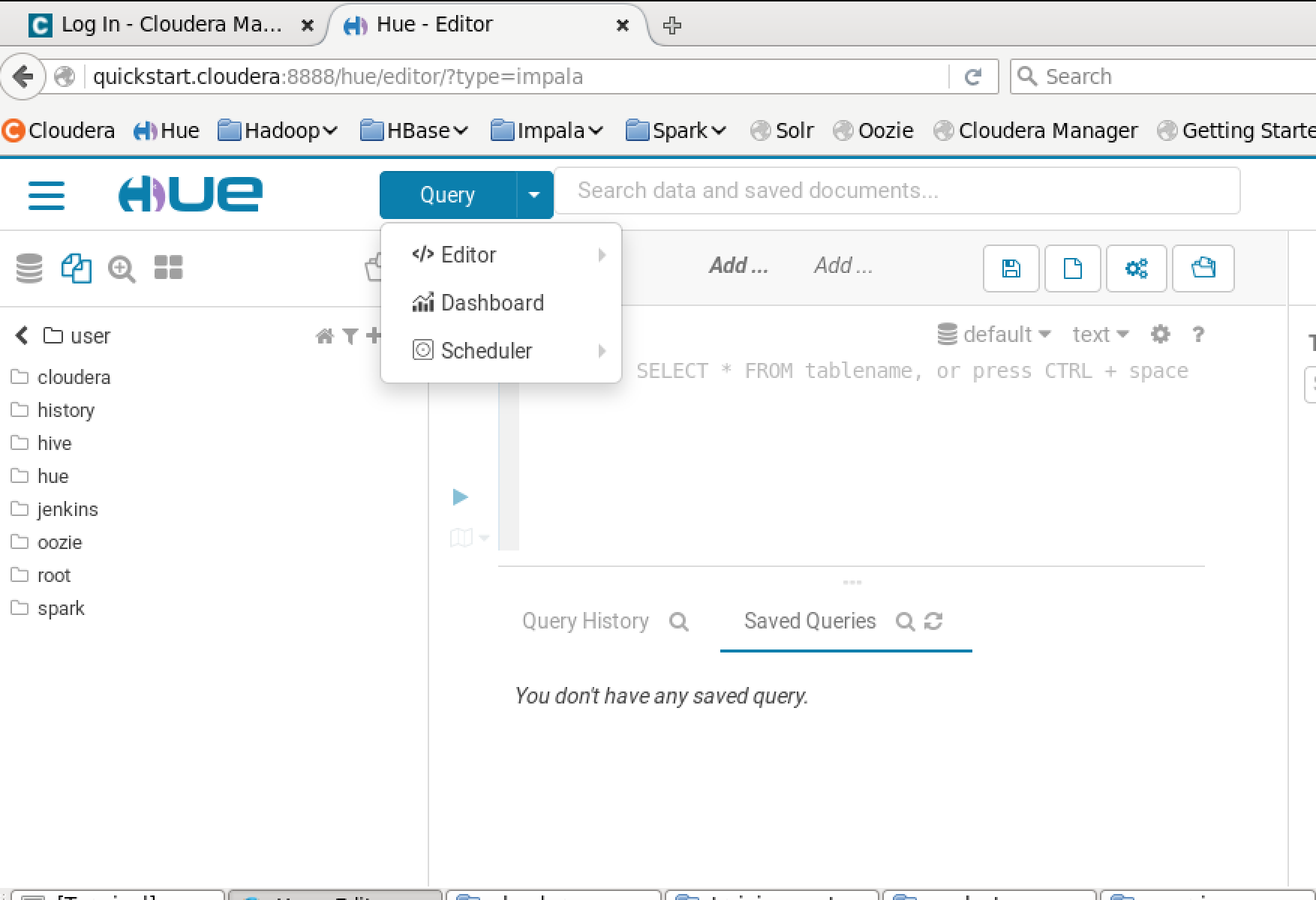
hive -f verify\_tablet\_order.hql

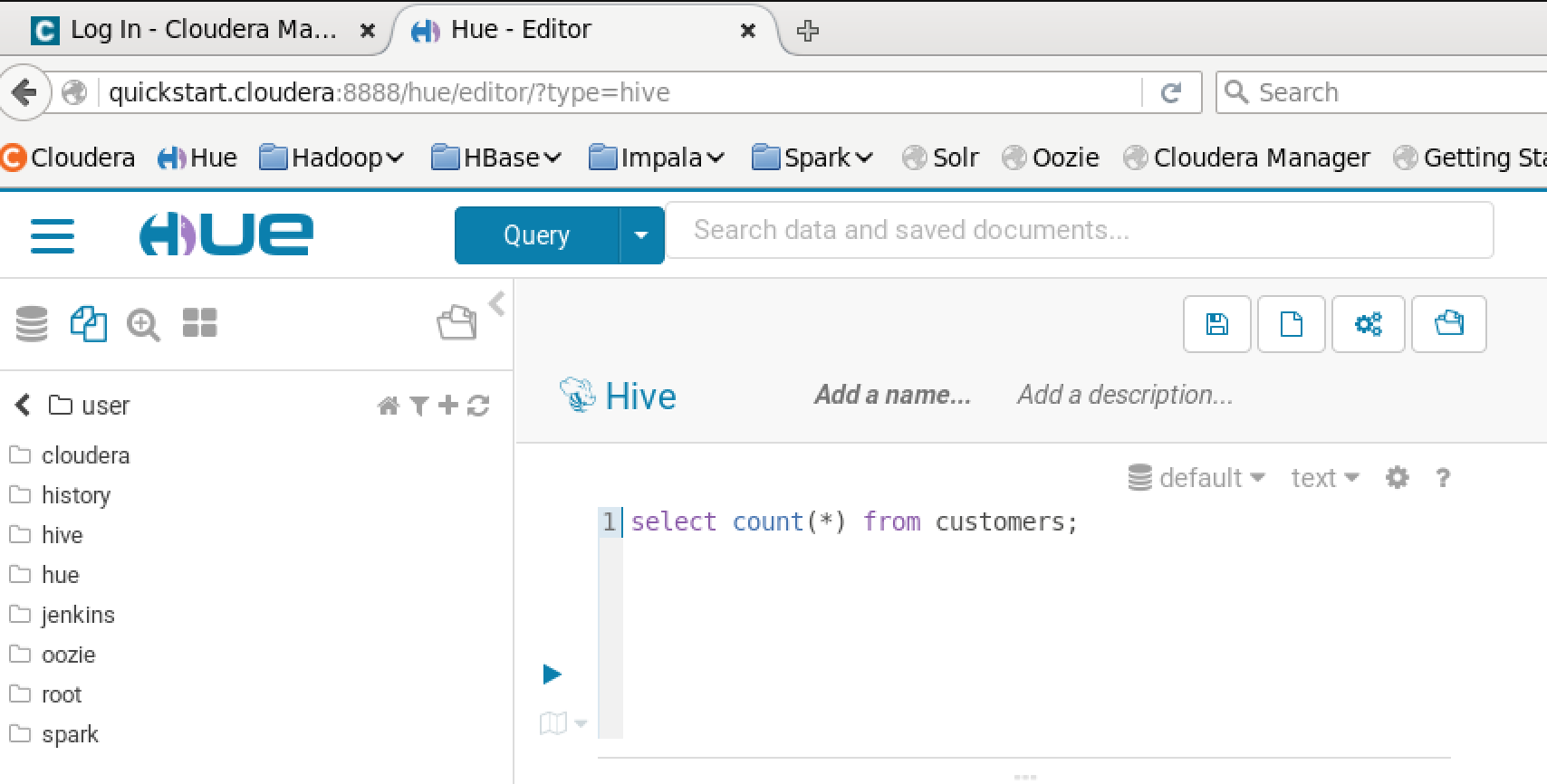
## Bài 4: Chạy truy vấn sử dụng Hue

Chúng ta có thể dùng Hue để viết và chạy truy vấn.

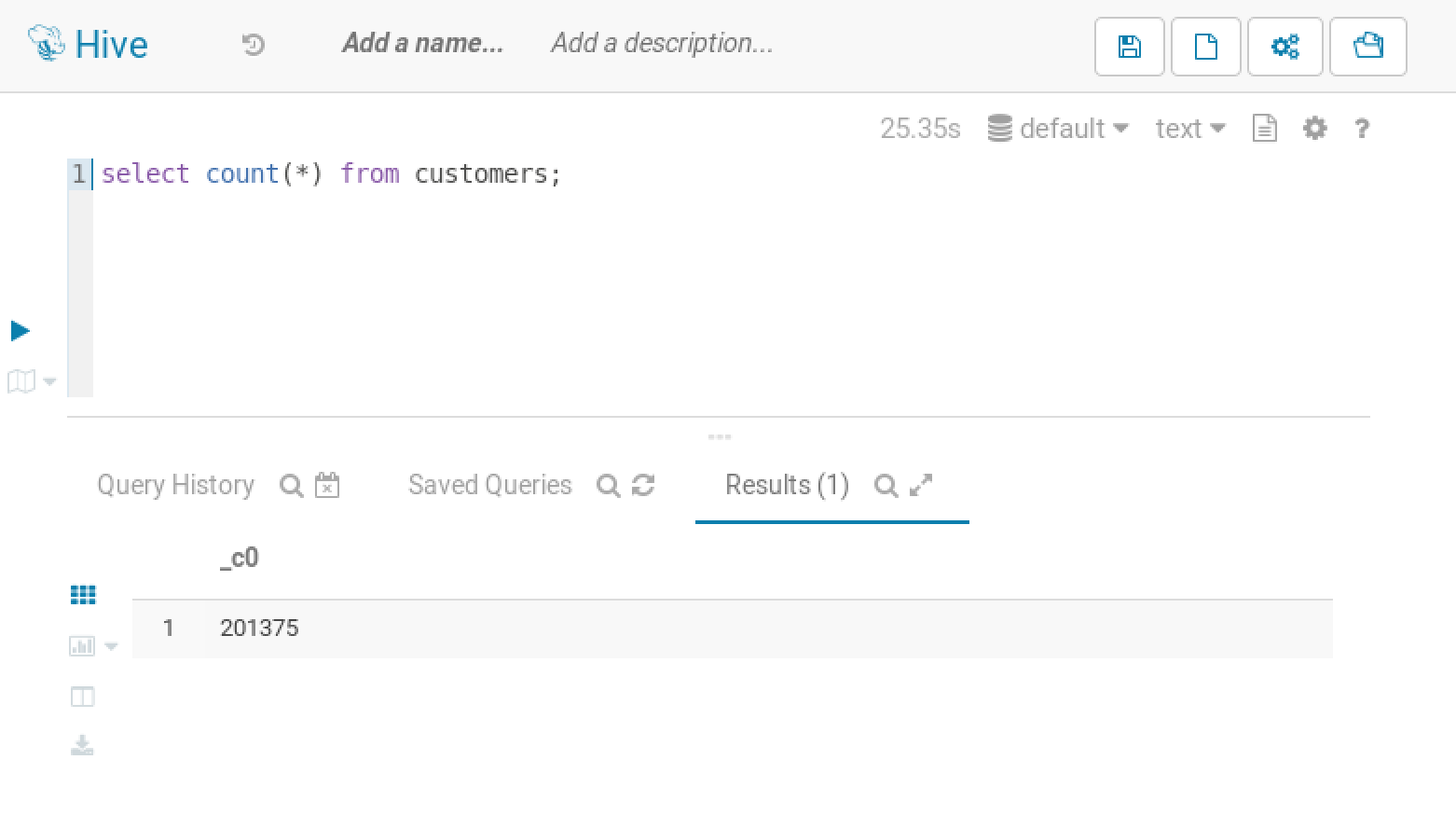
1. Bạn mở firefox (biểu tượng trái đất và chuột) trên menu trên cùng. Sau đó nhập đường dẫn phía dưới. Account để đăng nhập là *cloudera/cloudera*

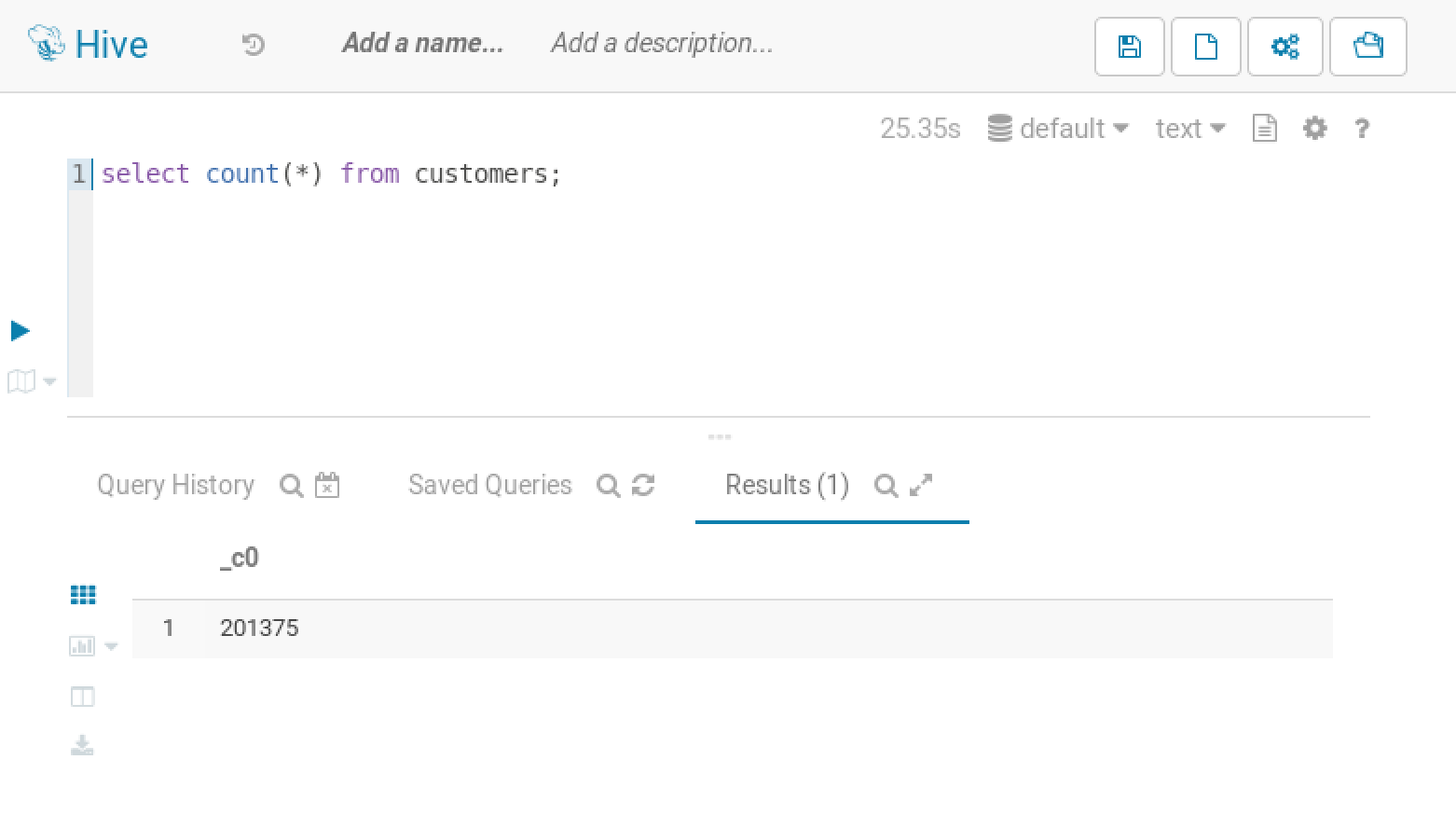
http://localhost:8888

1. Trong Hue bạn mở Hive query từ Query > Editor > Hive (như hình dưới)
2. Trong Hive editor bạn viết lệnh đếm số lượng customer: *select count(\*) from customers;* rồi bấm nút chạy (hình tam giác xanh bên trái editor) hoặc *Ctrl + Enter*



1. Kết quả hiện ra như sau (**201375**)





# **Quản trị dữ liệu với Hive**

Trong phần này bạn sẽ thực hành việc tạo và đẩy dữ liệu vào bảng trong Hive. Bạn cũng sẽ học cách viết truy vấn với các cột complex: array, map, struct

Quan trọng: để chuẩn bị dữ liệu cho bài tập dưới bạn chạy lệnh dưới. Khi màn hình hiện ra các lựa chọn gõ số **7.**

~/scripts/analyst/catchup.sh

## Bài 1: Dùng Sqoop import để tạo bảng và dữ liệu trong Hive

1. Mở Terminal và chuyển đến thư mục thực hành

cd /home/training/training\_materials/analyst/exercises/data\_mgmt

1. Chạy *sqoop* để đẩy dữ liệu từ bảng *suppliers* trong MySQL tới Hive. Lệnh này cũng sẽ tạo luôn bảng *suppliers* trên Hive

sqoop import --connect jdbc:mysql://localhost/dualcore

--username root --password cloudera

--fields-terminated-by '\t'

--table suppliers

--hive-import

1. Mở Hive shell

hive

1. Kiểm tra xem sqoop đã import và tạo bảng suppliers trên Hive chưa. Nếu mọi thứ chạy đúng, lệnh dưới sẽ trả ra 9 bản ghi.

SELECT COUNT(\*) FROM suppliers WHERE state=’TX’;

## Bài 2: Tạo bảng external trong Hive

Chúng ta sẽ thực hành việc tạo bảng external trong Hive. Trên HDFS đã có dữ liệu về employees trong thư mục /dualcore/employees, bài tập này chúng ta sẽ tạo bảng external có tên là employees và trỏ đến dữ liệu nằm ở /dualcore/employees.

1. Bạn hãy viết lệnh HiveQL để tạo bảng external có tên là *employees* trỏ đến thư mục */dualcore/employees* trên HDFS. Bảng employees có các cột là:

|  |  |
| --- | --- |
| **Cột** | **Kiểu dữ liệu** |
| emp\_id | STRING |
| fname | STRING |
| lname | STRING |
| address | STRING |
| city | STRING |
| state | STRING |
| zipcode | STRING |
| Job\_title | STRING |
| Email | STRING |
| Active | STRING |
| Salary | INT |

*Nếu bạn gặp khó khăn có thể tham khảo đáp án*

CREATE EXTERNAL TABLE employees

(emp\_id STRING,

fname STRING,

lname STRING,

address STRING,

city STRING,

state STRING,

zipcode STRING,

job\_title STRING,

email STRING,

active STRING,

salary INT)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t'

LOCATION '/dualcore/employees';

1. Chạy lệnh dưới để kiểm tra việc tạo bảng của bạn ok

SELECT job\_title, COUNT(\*) AS num

FROM employees

GROUP BY job\_title

ORDER BY num DESC LIMIT 3;

## Bài 3: Tạo vào đẩy dữ liệu vào bảng dạng hive-managed.

1. Tạo bảng ratings để lưu dữ liệu cách nhau bởi dấu tab. Cấu trúc bảng rating như sau:

|  |  |
| --- | --- |
| **Cột** | **Kiểu dữ liệu** |
| posted | TIMESTAMP |
| cust\_id | INT |
| prod\_id | INT |
| rating | TINYINT |
| message | STRING |

*Nếu bạn gặp khó khăn có thể tham khảo đáp án*

CREATE TABLE ratings

(posted TIMESTAMP,

cust\_id INT,

prod\_id INT,

rating TINYINT,

message STRING)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '\t';

1. Mở Terminal window và đẩy dữ liệu rating năm 2012 lên HDFS. (chú ý, dấu \ trong terminal cho phép bạn gõ câu lệnh trên 2 dòng, trường hợp bạn gõ lệnh trên cùng 1 dòng thì không cần dấu \ này). Khi bạn đẩy dữ liệu vào */user/hive/warehouse/ratings*, mặc định dữ liệu này sẽ phục vụ cho bảng ratings mà bạn tạo ở bước 1.

hadoop fs -put /home/training/training\_materials/analyst/data/ratings\_2012.txt \

/user/hive/warehouse/ratings

1. Bạn có thể kiểm tra lại dữ liệu trong bảng ratings. Kết quả trả ra phải là 464 bản ghi.

SELECT COUNT(\*) FROM ratings;

1. Ngoài cách đẩy thẳng dữ liệu vào thư mục khai báo bảng ratings trên HDFS. Chúng ta có thể dùng lệnh *LOAD DATA INPATH*. Trước khi thực hiện lệnh này, bạn hãy đẩy file *rating\_2013.txt* lên thư mục */dualcore* trên HDFS.

hadoop fs -put /home/training/training\_materials/analyst/data/ratings\_2013.txt /dualcore

1. Kiểm tra lại xem file đã được đẩy lên HDFS hay chưa

hadoop fs -ls /dualcore/ratings\_2013.txt

1. Sử dụng lệnh LOAD DATA INPATH để đẩy dữ liệu từ /dualcore/ratings\_2013.txt vào bảng ratings trong Hive. Bạn cần mở Hive shell hoặc chạy lệnh sau trong Hue

LOAD DATA INPATH ‘/dualcore/ratings\_2013.txt’ INTO TABLE ratings;

1. Khi dùng lệnh LOAD DATA INPATH thì file ratings\_2013.txt sẽ bị xóa trong thư mục /dualcore và chuyển tới thư mục của bảng rating. Bạn có thể kiểm tra bằng việc chạy lệnh sau trong Terminal

hadoop fs -ls /dualcore/ratings\_2013.txt

hadoop fs -ls /user/hive/warehouse/ratings

1. Kiểm tra trong bảng ratings có bao nhiêu bản ghi. Lúc này kết quả cần là: 21997 bản ghi.

SELECT COUNT(\*) FROM ratings;

## Bài 4: Tạo, đẩy dữ liệu và truy vấn bảng có cột dạng complex

Dualcore triển khai chương trình khách hàng thân thiết, rất nhiều khách hàng đăng ký tham gia chương trình. Với mỗi khách hàng, Dualcore thu thập được số điện thoại (cố định, di động…), danh sách đơn hàng đã mua, và thông tin về số tiền thấp nhất, lớn nhất, trung bình và tổng tiền của các đơn hàng trong quá khứ. Trong bài tập chúng ta sẽ tạo bảng có thể chứa được những thông tin trên, đồng thời đẩy dữ liệu và truy vấn trên bảng đó.

1. Mở Terminal và chuyển thư mục

cd /home/training/training\_materials/analyst/exercises/data\_mgmt

1. Bạn hãy tạo bảng loyalty\_program như dưới đây, chú ý số điện thoại chúng ta dùng kiểu MAP. Danh sách đơn hàng order\_ids dùng mảng (ARRAY), và giá trị đơn hàng chúng ta dùng STRUCT.

CREATE TABLE loyalty\_program

(cust\_id INT,

fname STRING,

lname STRING,

email STRING,

level STRING,

phone MAP<STRING, STRING>,

order\_ids ARRAY<INT>,

order\_value STRUCT<min:INT,

max:INT,

avg:INT,

total:INT>)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY '|'

COLLECTION ITEMS TERMINATED BY ','

MAP KEYS TERMINATED BY ':';

1. Đẩy dữ liệu từ file loyalty\_data.txt vào bảng trên. Chú ý bạn cần mở Hive shell để làm lệnh dưới.

LOAD DATA LOCAL INPATH ‘loyalty\_data.txt’ INTO TABLE loyalty\_program;

1. Viết truy vấn để lấy số điện thoại ở nhà của khách hàng có Id = 1200866. Bạn sẽ thấy kết quả là 408-555-4914

SELECT phone[“HOME”] FROM loyalty\_program WHERE cust\_id = 1200866;

1. Viết lệnh để lấy ra phần tử thứ 3 của cột order\_details cho khách hàng cust\_id=1200866. Kết quả cần trả ra là 5278505
2. Lấy ra thuộc tính total của cột order\_value cho khách hàng cust\_id = 1200866. Kết quả trả ra cần là 401874

# **Phân tích cảm xúc (sentiment analysis) dùng Hive**

Dualcore cho phép người dùng có thể đánh giá (rating) và bình luận (comment) về các sản phẩm của họ. Tuy nhiên những bình luận ở dạng tự do nên có thể phải xử lý khác nhau cho từng trường hợp. Trong bài này chúng ta sẽ sử dụng một số hàm dựng sẵn trong Hive để phân tích bình luận (comment) và đánh giá (rating) của khách hàng để từ đó tìm ra vấn đề và đưa ra giải pháp.

Quan trọng: để chuẩn bị dữ liệu cho bài tập dưới bạn chạy lệnh dưới. Khi màn hình hiện ra các lựa chọn gõ số **8.**

~/scripts/analyst/catchup.sh

## Bài 1: Phân tích rating sản phẩm

1. Mở Terminal và chuyển thư mục

cd /home/training/training\_materials/analyst/exercises/sentiment

1. Lấy ra sản phẩm được rating cao nhất, và có tối thiểu 50 người rating. Kết quả Productid = 1273880, rating trung bình là 4.12

SELECT prod\_id, FORMAT\_NUMBER(avg\_rating, 2) AS avg\_rating

FROM (SELECT prod\_id, AVG(rating) AS avg\_rating, COUNT(\*) AS num

FROM ratings

GROUP BY prod\_id) rated

WHERE num >= 50

ORDER BY avg\_rating DESC

LIMIT 1;

1. Lấy danh sách sản phẩm được rating thấp nhất, và có tối thiểu 50 người rating. Kết quả trả ra cần là ProductId = 1274673 và rating trung bình là 1.10

## Bài 2: Phân tích bình luận

Trong bài 1 chúng ta thấy khách hàng không hài lòng với sản phẩm 1274673, nhưng chúng ta không hiểu tại sao họ không thích sản phẩm đó. Để tìm hiểu lý do chúng ta sẽ cần dùng Hive để phân tích bình luận của người dùng.

1. Truy vấn dưới chúng ta sẽ chuẩn hóa tất cả bình luận về dạng chữ thường, sau đó tách chúng thành tập hợp các từ (word) sử dụng hàm SENTENCE, và truyền những từ này vào hàm NGRAMS để xem 5 tổ hợp **hai từ** nào hay xuất hiện nhiều nhất.

SELECT EXPLODE(NGRAMS(SENTENCES(LOWER(message)), 2, 5))

AS bigrams

FROM ratings

WHERE prod\_id = 1274673;

1. Bạn sẽ thấy với truy vấn trên, hầu hết đều có từ “expensive” (đắt). Từ này không mang nhiều ý nghĩa, nên chúng ta thử phân tích như trên nhưng lần này lấy ra 5 tổ hợp **ba từ** hay xuất hiện cùng nhau

SELECT EXPLODE(NGRAMS(SENTENCES(LOWER(message)), 3, 5))

AS trigrams

FROM ratings

WHERE prod\_id = 1274673;

1. Bạn sẽ thấy cụm từ có ý nghĩa là “ten times more” (đắt hơn 10 lần), điều đó có nghĩa là khách hàng nói sản phẩm này quá đắt. Giờ chúng ta sẽ đi sâu vào các bình luận để xem

SELECT message

FROM ratings

WHERE prod\_id = 1274673 AND message LIKE '%ten times more%'

LIMIT 3;

1. Bạn sẽ thấy có 3 bình luận là: “why does the red one cost ten times more than the others” (“Tại sao sản phẩm **màu đỏ** lại đắt hơn 10 lần so với các sản phẩm khác”). Bạn thấy rằng sản phẩm “màu đỏ” (red) có lẽ có vấn đề. Vì vậy, viết truy vấn lấy ra thông tin chi tiết về sản phẩm 1274673

SELECT \* FROM products WHERE prod\_id=1274673;

1. Bạn sẽ thấy sản phẩm trên là “16GB USB Flash Drive (Red)” của hãng “Orion”. Giờ chúng ta sẽ xem các sản phẩm cùng là “16GB USB Flash Drive” và của hãng Orion

SELECT \* FROM products

WHERE name LIKE '%16 GB USB Flash Drive%'

AND brand='Orion';

Bạn sẽ thấy rằng các sản phẩm khác cùng loại, cùng hãng chỉ khác màu và có giá tương đương như nhau, nhưng riêng sản phẩm màu đỏ này giá cao hơn 10 lần. Điều đó có thể là do nhân viên nhập liệu đã gõ sai giá sản phẩm. Như vậy, bằng việc phân tích bình luận và rating của người dùng, chúng ta có thể tìm ra được lỗi sai trong hệ thống.