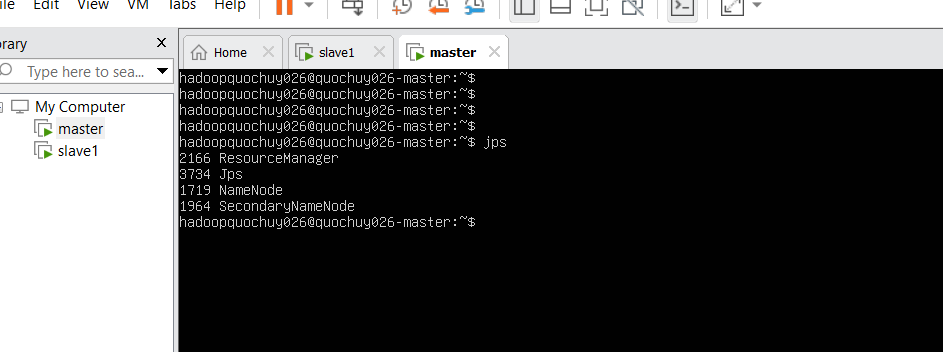
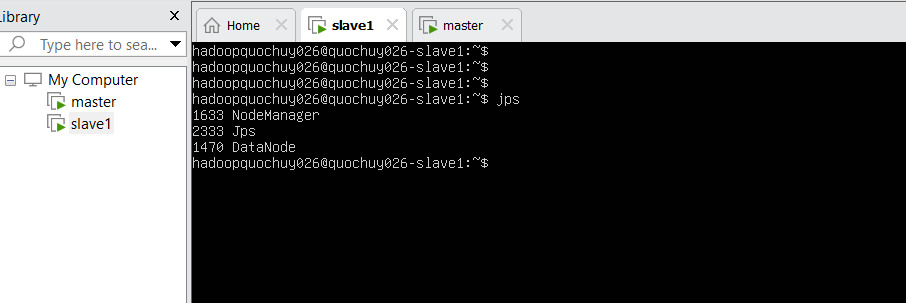
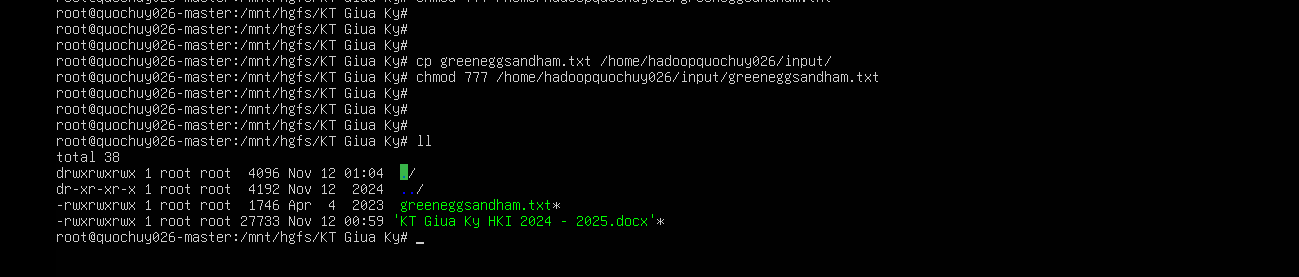
1c, 2d, 3d, 4c, 5b, 6b, 7b, 8c, 9b, 10d, 11d, 12a, 13d, 14a, 15d, 16b, 17b, 18a, 19d, 20c, 21a, 22d, 23b, 24a, 25d, 26d, 27a, 28a, 29b, 30d, 31b, 32c, 33b, 34b, 35c, 36b, 37a, 38b, 39b, 40b

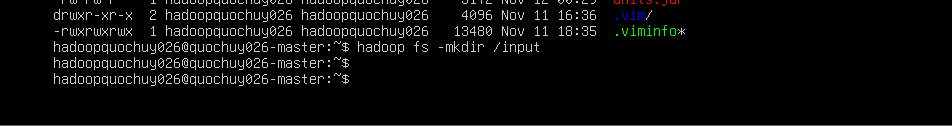




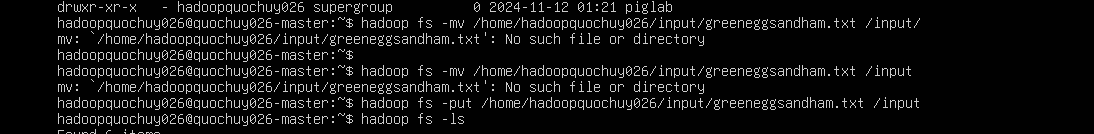


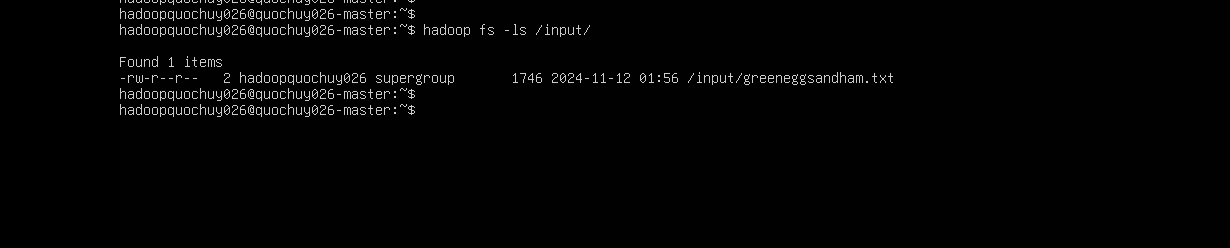


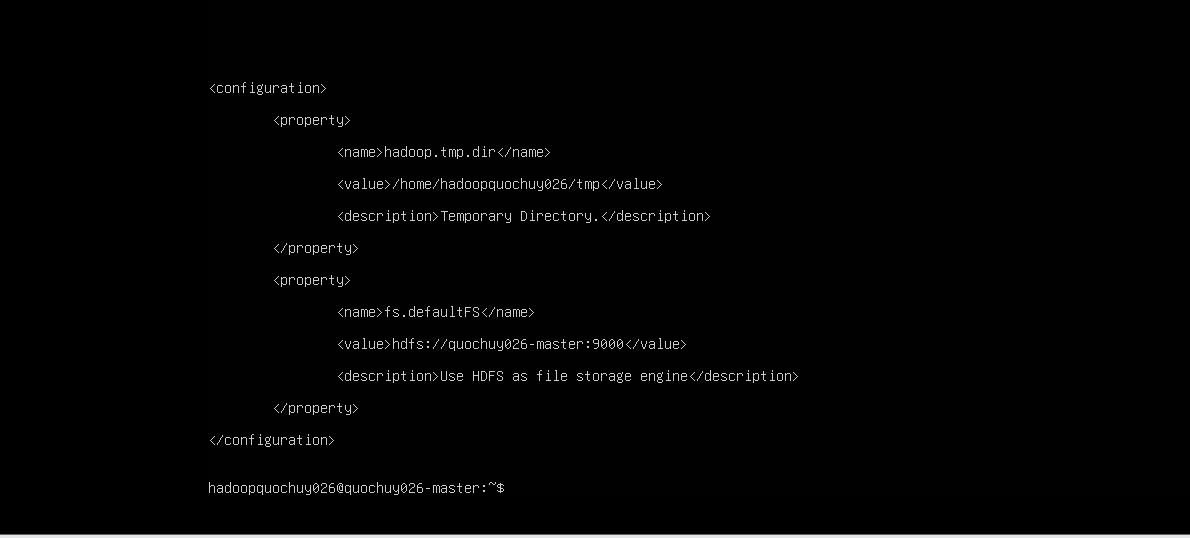
$ hadoop fs -mkdir input



$ hadoop fs -put /home/hadoopquochuy026/input/greeneggsandham.txt input







$ pig

1. NguyenQuocHuyInput = LOAD 'hdfs://quochuy026-master:9000/input/greeneggsandham.txt' USING TextLoader() AS (myword: chararray);



Giải thích:

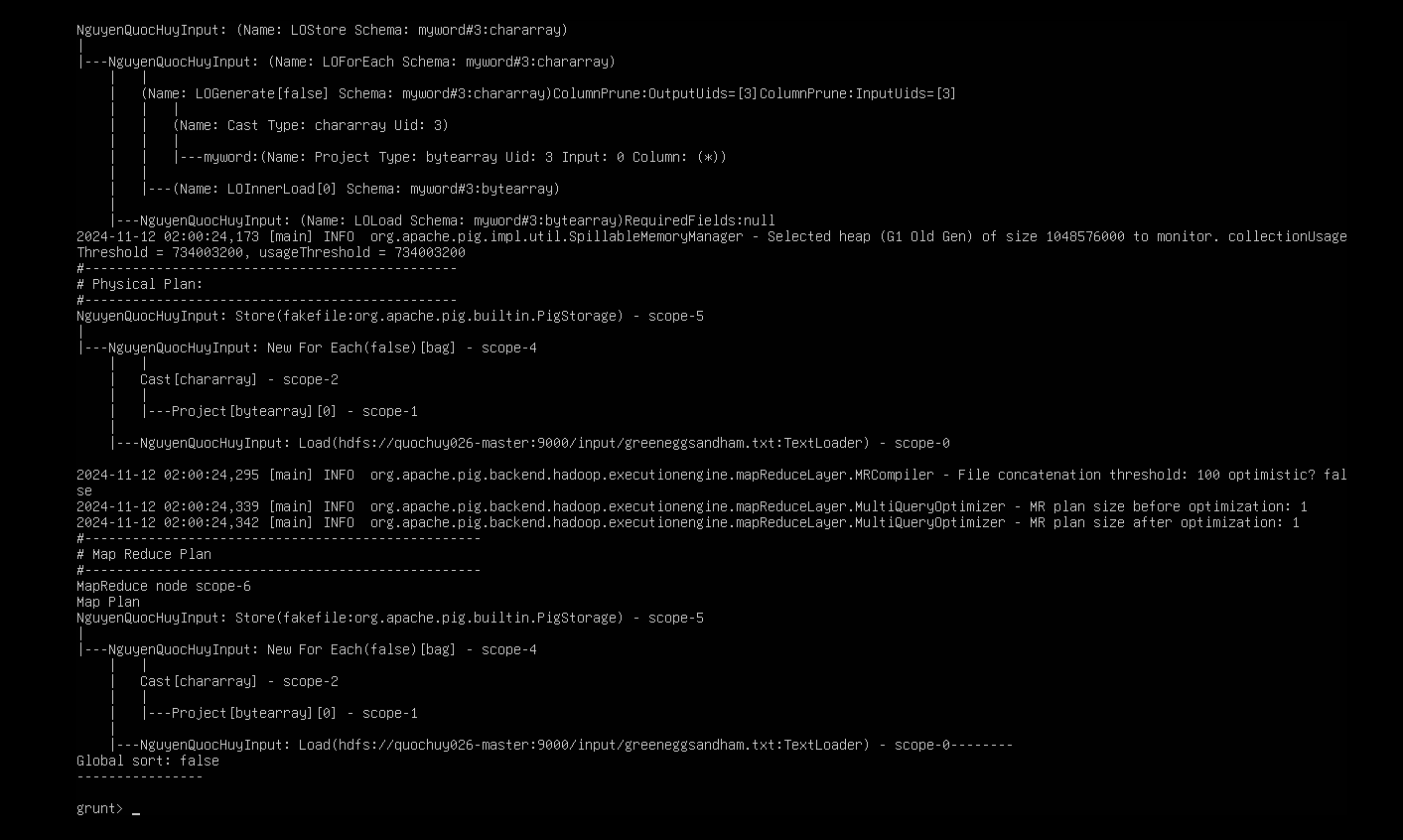
* lệnh trên là lệnh để tải dữ liệu từ hệ thống hdfs vào pig.
* 'hdfs://quochuy026-master:9000/input/greeneggsandham.txt' là đường dẫn đầy đủ tới tệp dữ liệu trong HDFS.
* USING TextLoader() là để dữ liệu được tải bằng cách sử dụng TextLoader
* AS (myword: chararray) định nghĩa tên và kiểu dữ liệu cho từng dòng trong mỗi dòng của file

2.1 describe NguyenQuocHuyInput;



Giải thích: hiển thị cấu trúc của đối tượng NguyenQuocHuyInput. Cho biết NguyenQuocHuyInputcó một trường duy nhất myword với kiểu dữ liệu là chararray

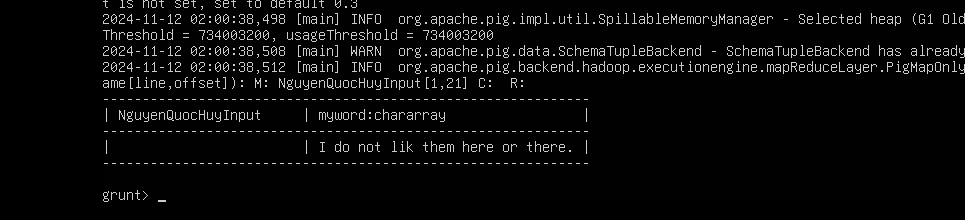
2.2 explain NguyenQuocHuyInput;



Giải thích:

* Cho biết chi tiết quá trình thực thi của lệnh LOAD. Xem được các bước xử lý của Pig và cách dữ liệu được tải từ hdfs.

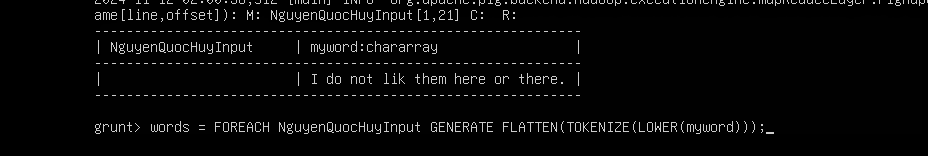
2.3 illustrate NguyenQuocHuyInput;



Giải thích:

* Cho biết cách thức hoạt động của các phép toán trên dữ liệu. Lệnh trên giúp mình xem Pig sẽ thực hiện gì đối với dữ liệu trong NguyenQuocHuyInput.

1. words = FOREACH NguyenQuocHuyInput GENERATE FLATTEN(TOKENIZE(LOWER(myword)));



Giải thích:

* Lệnh lấy mỗi từ trong quan hệ NguyenQuocHuyInput, chuyển đổi tất cả thành chữ thường (LOWER(\*))
* TOKENIZE sẽ tách mỗi từ thành phần tử riêng biệt. FLATTEN giúp mở rộng các từ thành các bản ghi (rows) riêng biệt trong words.

4. grouped = GROUP words BY $0;



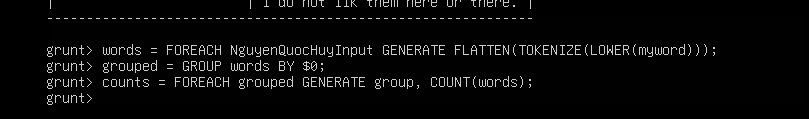
Giải thích:

- Nhóm các từ trong words theo cột đầu tiên ($0), tức là từ (word) trong words. Group sau khi nhóm chứa danh sách từ và số lượng bản ghi cho mỗi từ.

5. counts = FOREACH grouped GENERATE group, COUNT (words);

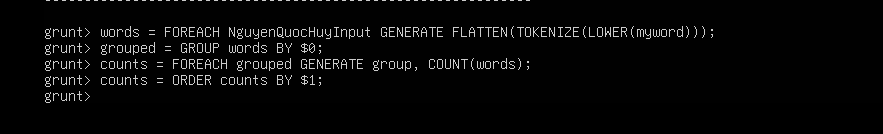
Giải thích:

* Đếm số lần xuất hiện của mỗi từ trong nhóm grouped. group giữ giá trị của từ, COUNT(words) tính số lần từ đó xuất hiện.



6. counts = ORDER counts BY $1;

- Sắp xếp counts theo số lượng từ xuất hiện ($1 là cột đếm), mặc định theo thứ tự tăng dần.



-- Giải thích câu lệnh

7. dump counts;

Giải thích:

- hiển thị nội dung của counts trên màn hình, cho phép xem kết quả trực tiếp mà không lưu vào file.

8. Lưu trữ kết quả counts ra file counts.txt và hiển thị kết quả.

****

Giải thích:

* Lưu dữ liệu của counts vào file counts.txt trên hệ thống lưu trữ HDFS. PigStorage() là định dạng lưu mặc định, lưu kết quả dưới dạng văn bản phân tách bằng dấu phẩy.