**Lab 7: Pig Install**

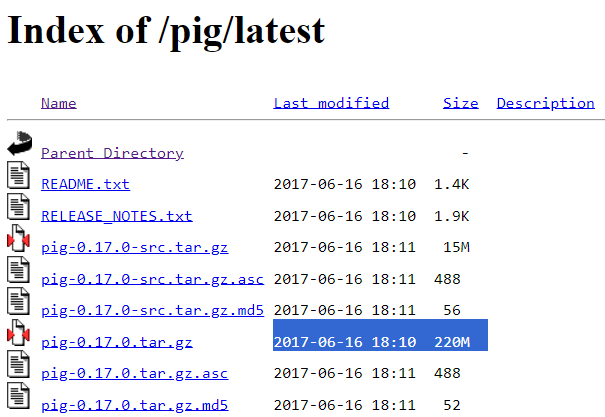
**Họ và Tên**: Trần Nguyễn Gia Long

**MSSV**: 1050080059

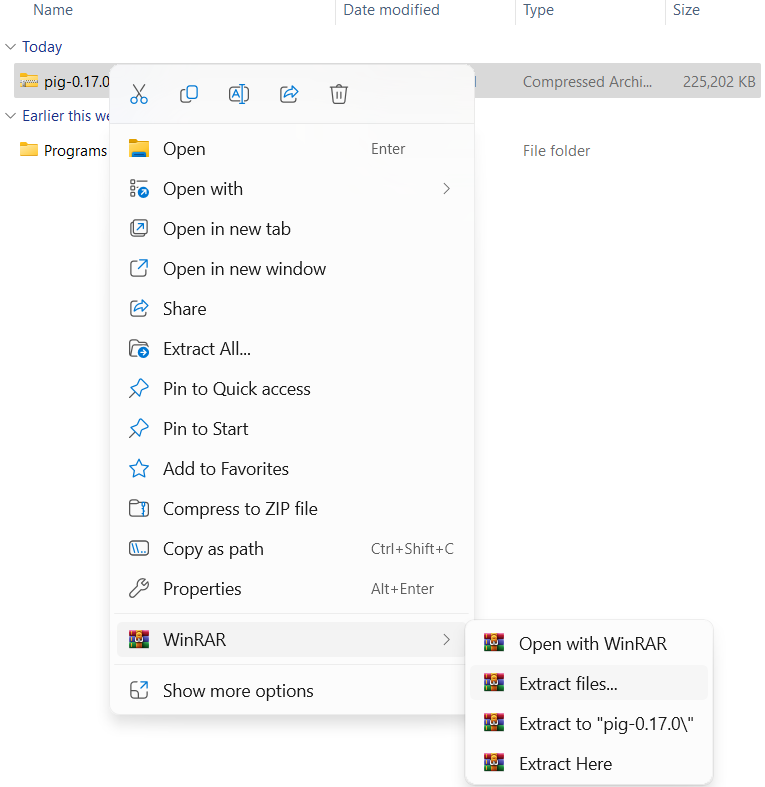
# 1. Cài đặt pig

*Ở đây ta sử dụng Pig trên Windows 11.*

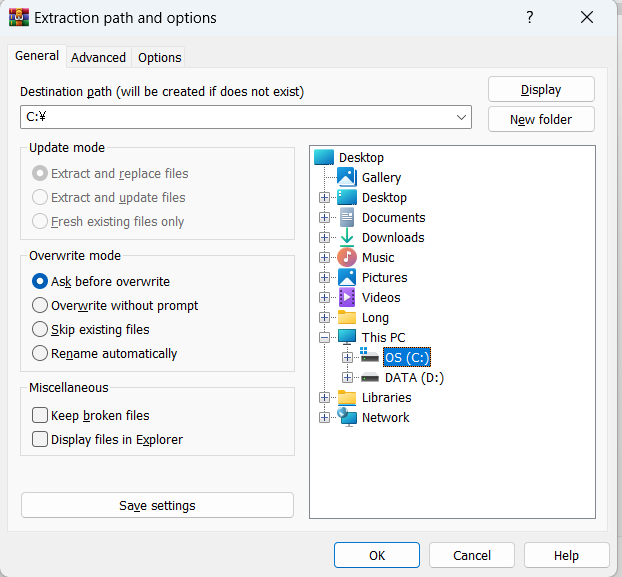
- Đầu tiên tải Pig về từ trang Apache Pig và tải bản mới nhất là 0.17.0.tat.gz.



- Giải nén Pig bằng cách chọn Extract file… và chọn vào ổ C, như thế Pig sẽ được giải nén với quyền quản trị.

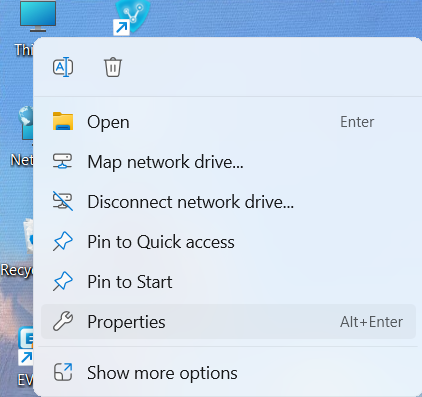


- Chọn ổ C.

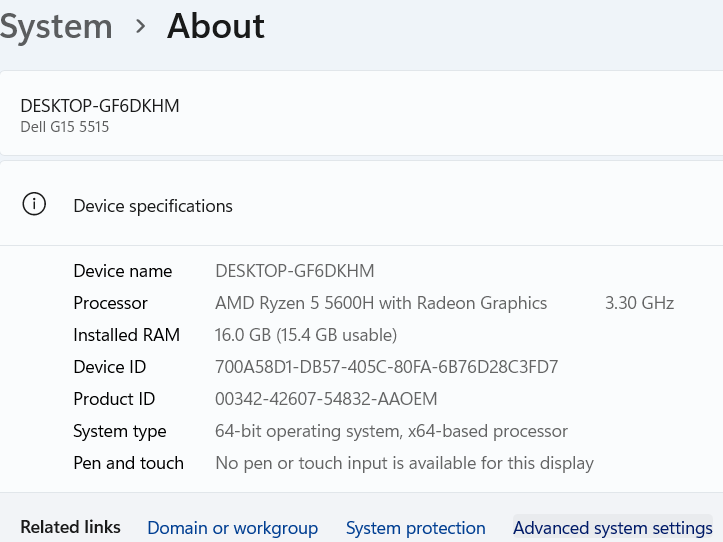


# 2. Cài đặt môi trường cho Pig.

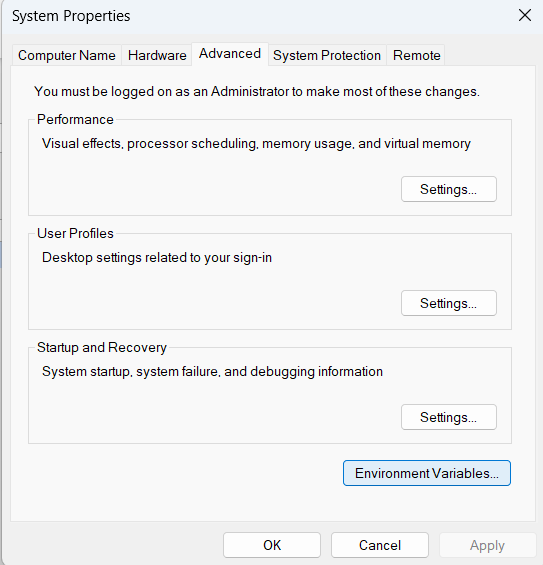
- Mở môi trường bằng cách phải chuột This PC, chọn **thuộc tính**.



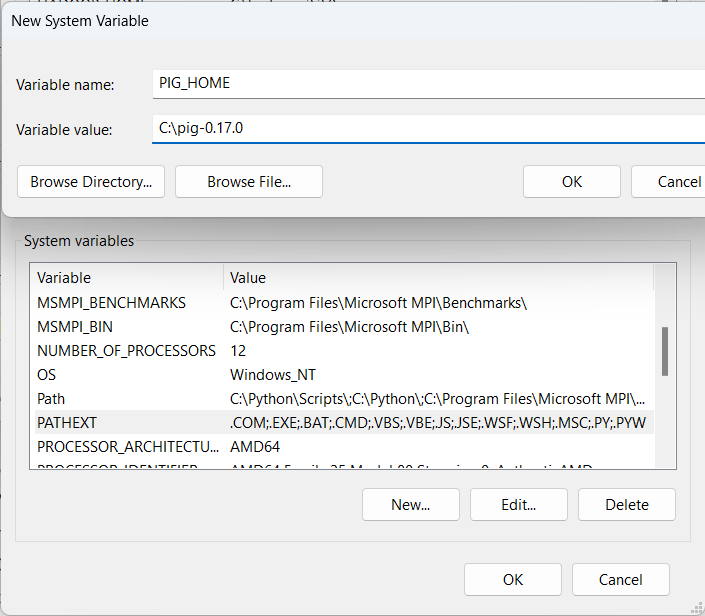
- Chọn **Advanced system settings**



- Chọn tiếp Environment Variables…

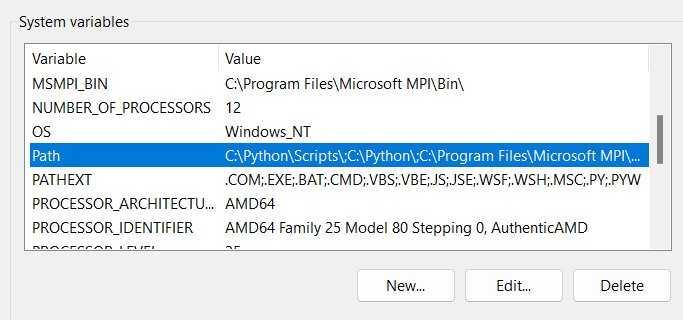


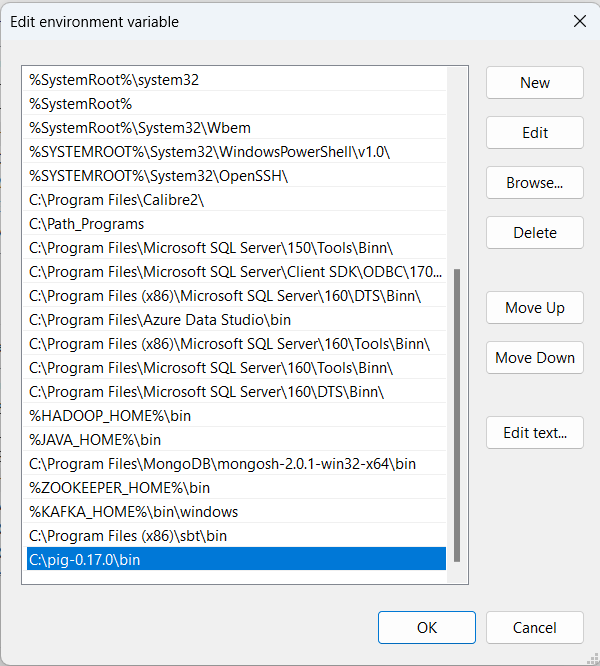
- Trong bảng *System Variables*, chọn New để tạo đường dẫn đến thư mục Pig đã giải nén.



- Thêm tên *PIG\_HOME* và đường dẫn chứa thư mục *C:\pig-0.17.0.*

- Tiếp theo cài đặt Variable **Path** ở phần *System variables* bằng cách chọn Edit, thêm đường dẫn *C:\pig-0.17.0\bin.*

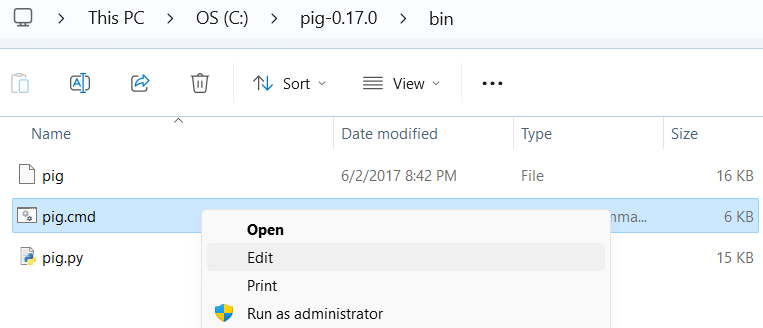




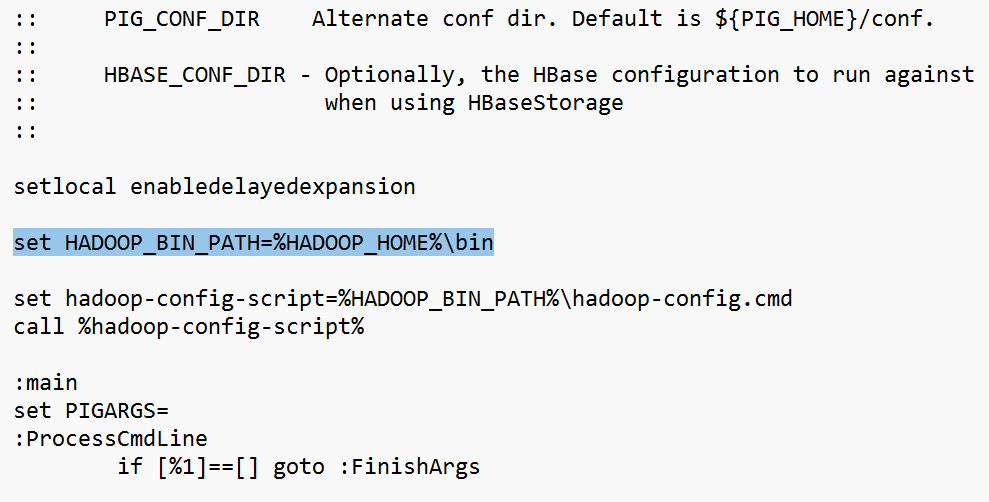
- Sau đó xác nhận bằng cách chọn OK tất cả.

# 3. Config tập tin trong Pig

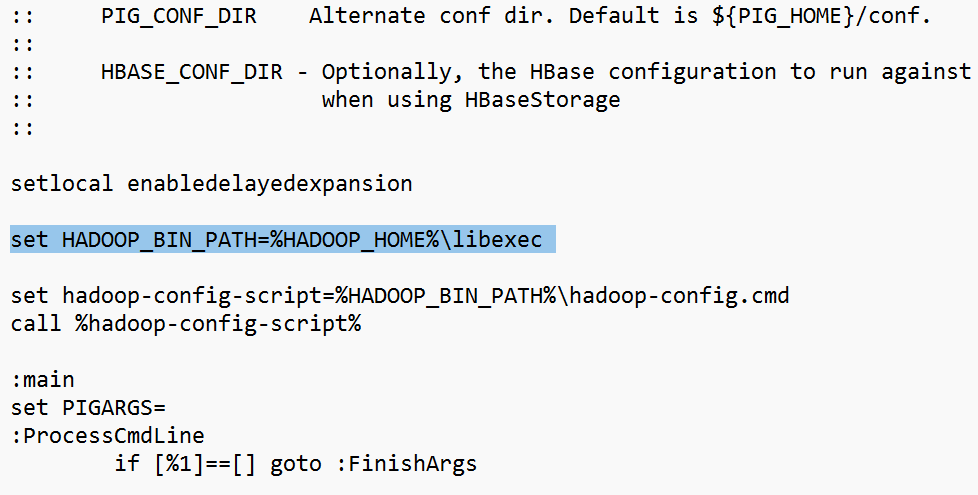
- Ta cấu hình Pig trong file pig.cmd ở đường dẫn *C:\pig-0.17.0\bin*. Phải chuột và nó chọn Edit.



- Tìm dòng *set HADOOP\_BIN\_PATH=%HADOOP\_HOME%\bin*.

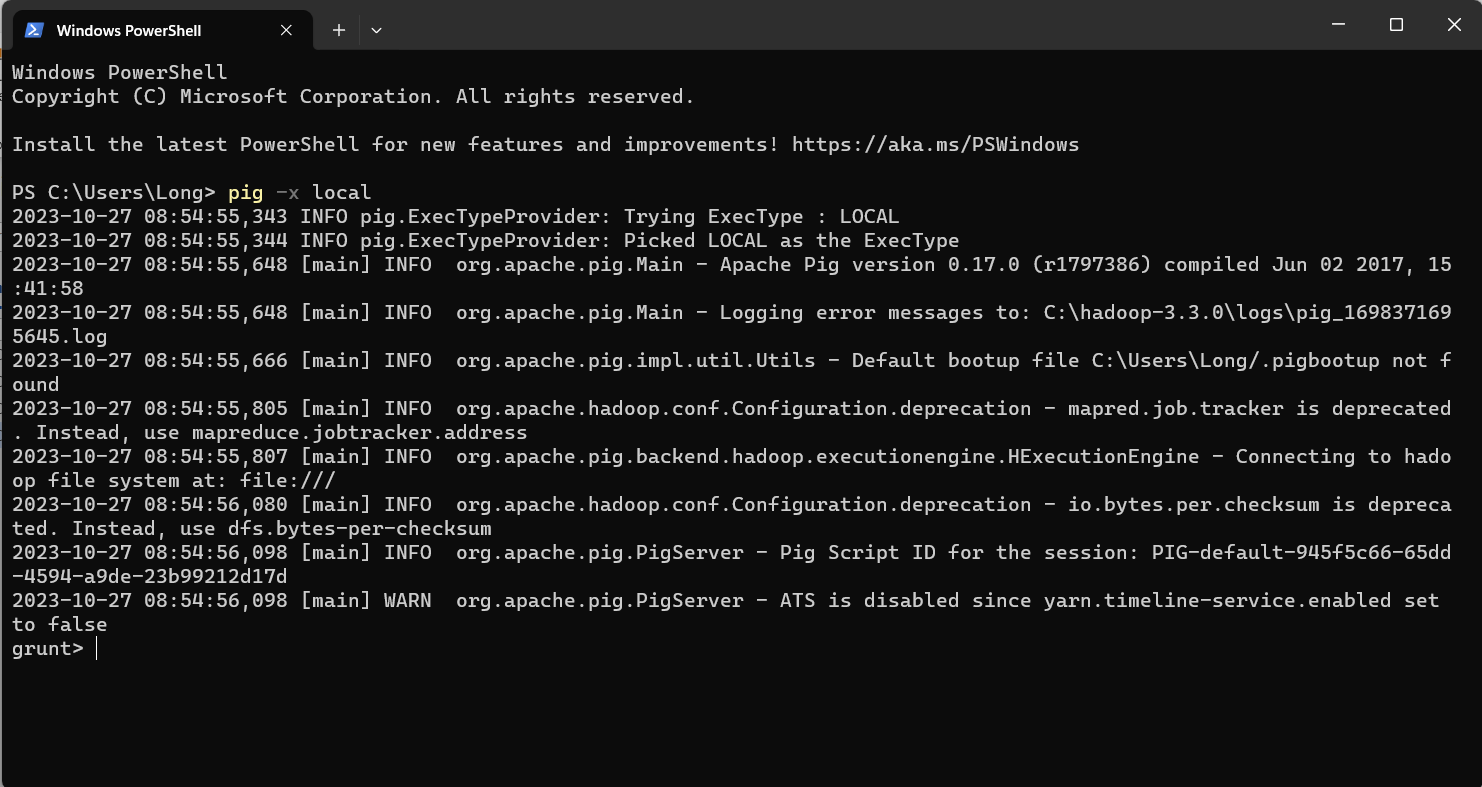


- Và thay đổi thành *set HADOOP\_BIN\_PATH=%HADOOP\_HOME%\libexec.*



# 4. Chạy Pig bằng grunt shell.

- Bắt đầu bằng việc mở **cmd**, nhập lệnh **pig -x local**.



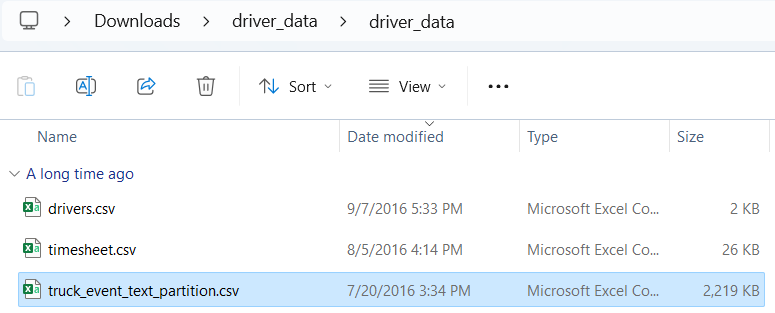
* **Vậy là đã chạy Pig thành công.**

# 5. Thực hành với Pig

- Chạy Pig bằng chế độ Local, **pig -x local**.

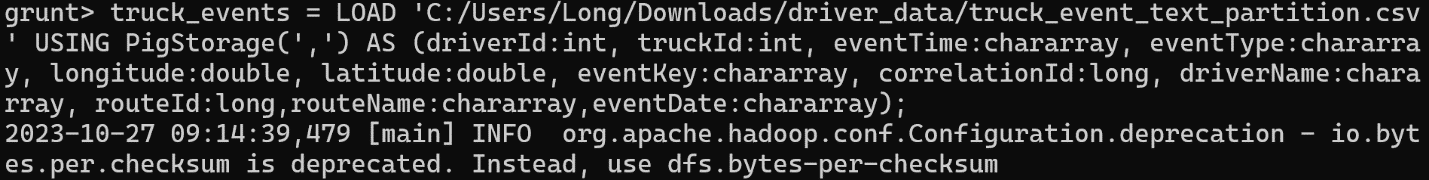
- Tải dữ liệu thực hành từ trang Github, <https://github.com/hortonworks/data-tutorials/blob/master/tutorials/hdp/beginners-guide-to-apache-pig/assets/driver_data.zip>.

- Giải nén và sử dụng file truck\_event\_text\_partition.csv.

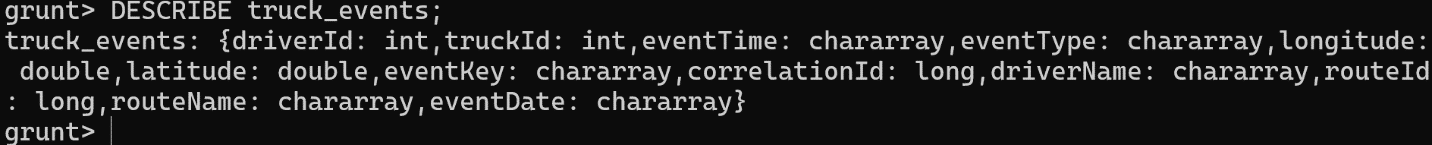


- Load file *truck\_event\_text\_partition* bằng với tên là **truck\_events** bằng lệnh LOAD**.**

truck\_events = LOAD 'C:/Users/Long/Downloads/driver\_data/truck\_event\_text\_partition.csv' USING PigStorage(',') AS (driverId:int, truckId:int, eventTime:chararray, eventType:chararray, longitude:double, latitude:double, eventKey:chararray, correlationId:long, driverName:chararray, routeId:long,routeName:chararray,eventDate:chararray);

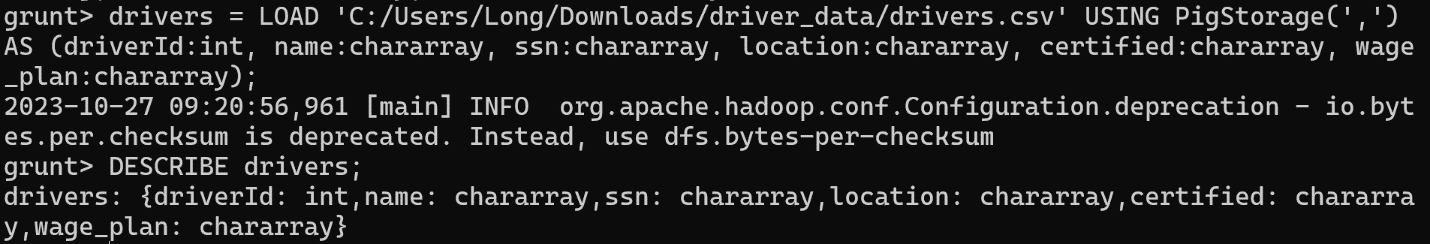


- Sau đó dùng lệnh DESCRIBE để kiểm tra file đã được load hay chưa.



- Tiếp tục load file *drivers*.

drivers = LOAD 'C:/Users/Long/Downloads/driver\_data/drivers.csv' USING PigStorage(',') AS (driverId:int, name:chararray, ssn:chararray, location:chararray, certified:chararray, wage\_plan:chararray);



- Dùng lệnh join *driverId* của file **drivers** vào file **truck\_events**.

join\_data = JOIN truck\_events BY (driverId), drivers BY (driverId);

- Sắp xếp thứ tự tăng dần theo tên driver là driverName.

ordered\_data = ORDER join\_data BY truck\_events::driverName asc;

- Lọc các dữ liệu có eventType là Normal.

filtered = FILTER ordered\_data BY NOT (eventType MATCHES 'Normal');

- Nhóm các dữ liệu.

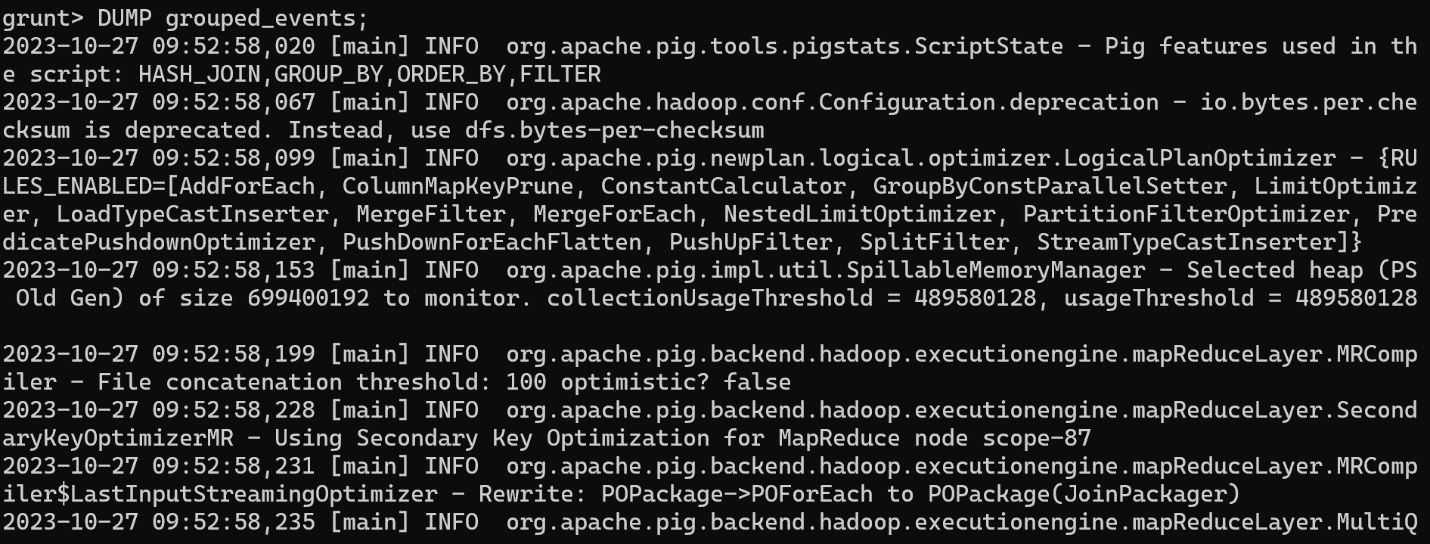
grouped\_events = GROUP filtered BY truck\_events::driverId;

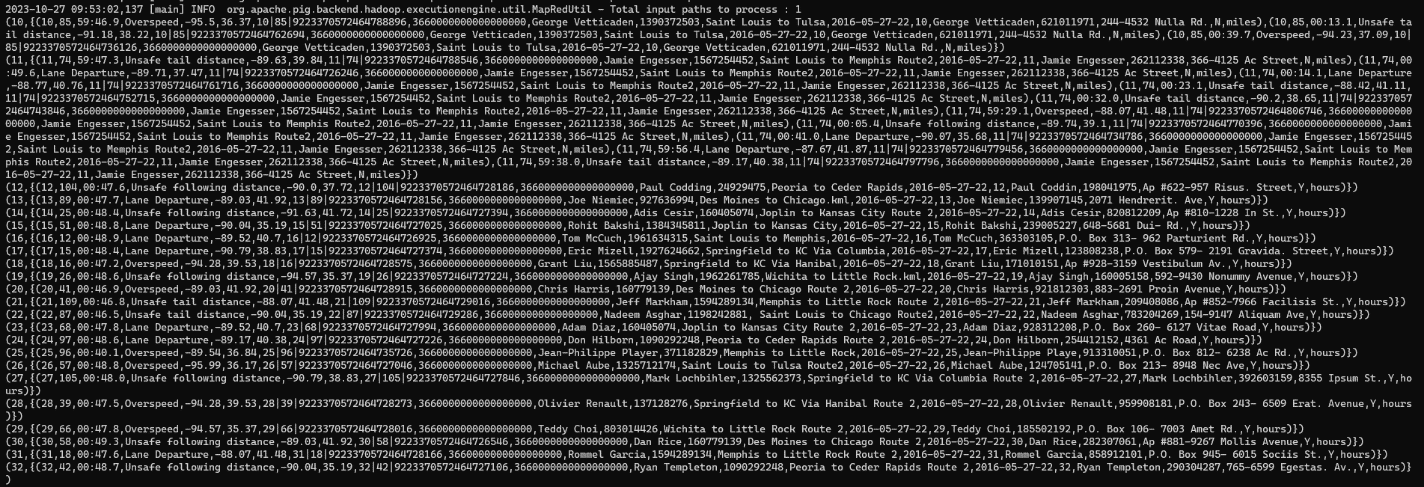
- Nhóm dữ liệu giới hạn 5.

grouped\_events\_subset = LIMIT grouped\_events 5;

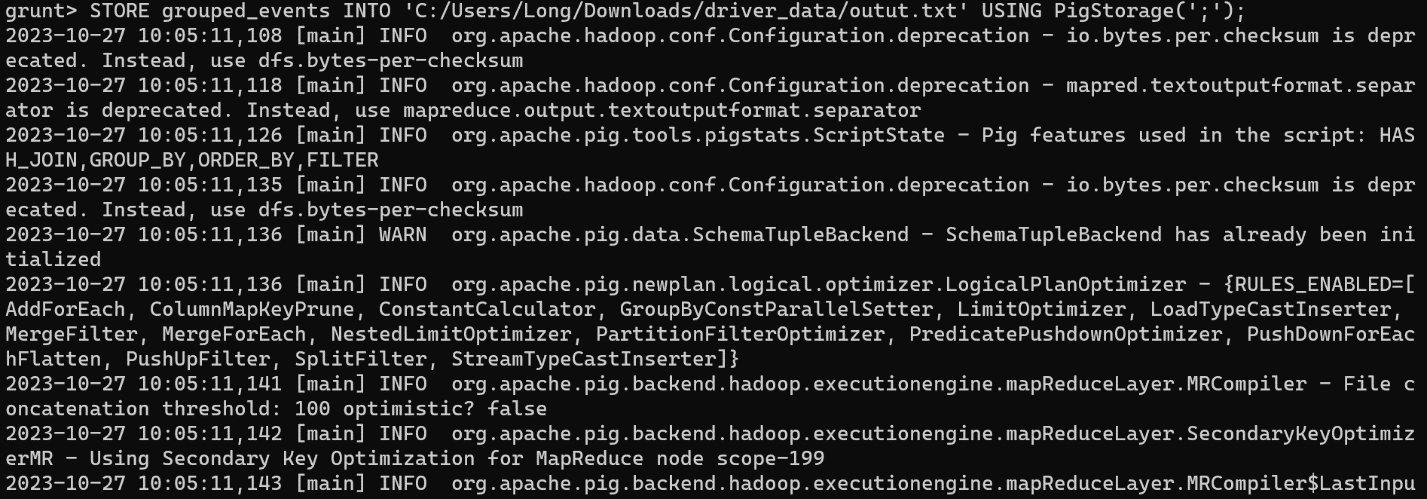
- Hiển thị dữ liệu đã thực hiện các câu lệnh lúc nãy.

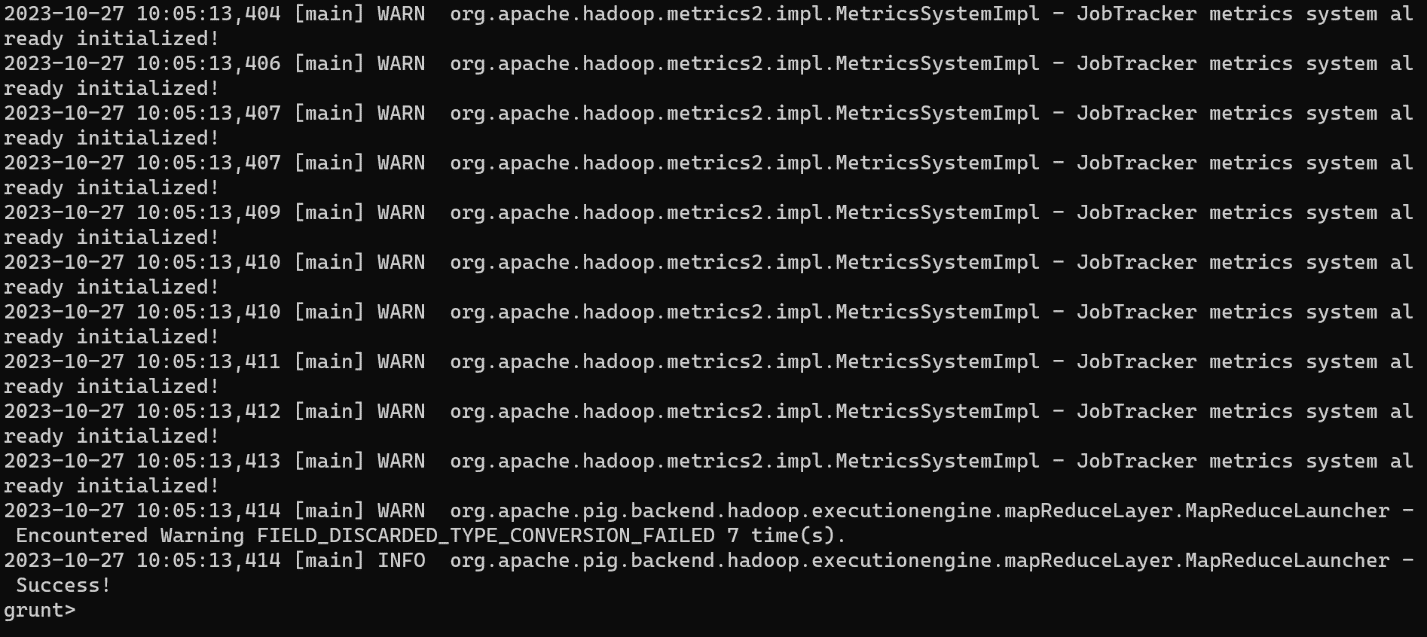
DUMP grouped\_events;

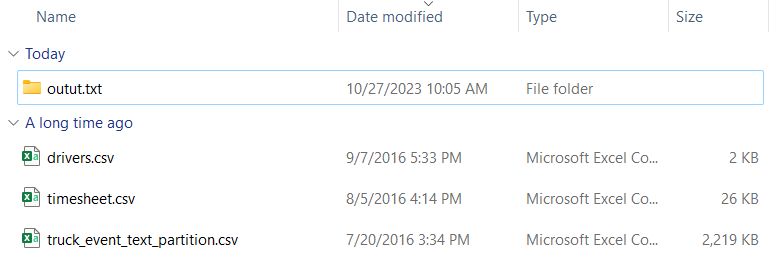


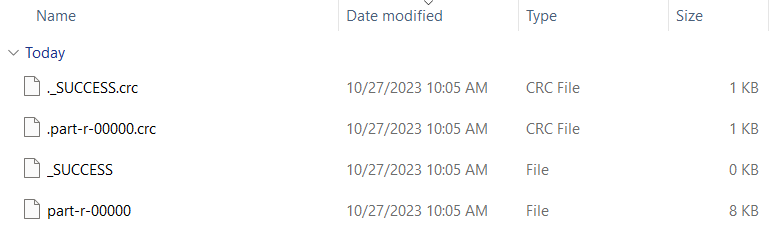


- Xuất file tên output.txt.









- Dữ liệu đã xuất.

