**Cars Dataset:**

The columns in cars.csv are mpg, cylinders, cubicinches, hp, weightlbs, time-to-60, year, brand

1.Remove the headers

2.Rename to '.txt'

3.Copy the local file in hdfs

#hadoop dfs -put 'file' /

**LOAD Dataset:**

cars = LOAD '/cars.txt' using PigStorage (',') as

(mpg:int,cylinders:int,cubicinches:int,hp:int,weight:int,time:int,year:int,brand:chararray);

****

**View Dataset:**

dump cars;



**Analysing Dataset:**

**1)Average MPG By Brand:**

GroupByBrand = GROUP cars BY brand;

dump GroupByBrand;

avg\_mpg = foreach GroupByBrand Generate group,AVG(cars.mpg);

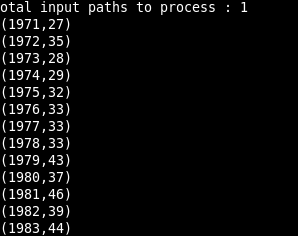
 dump avg\_mpg;

**2)Maximum HP By Year:**

GroupByYear = GROUP cars BY year;

dump GroupByYear;

max\_hp = foreach GroupByYear Generate group,MAX(cars.mpg);

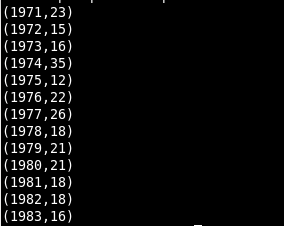
 dump max\_hp;

**3)Release Count by year:**

GroupByYear = GROUP cars BY year;

dump GroupByYear;

car\_release = foreach GroupByYear Generate group,COUNT(cars.mpg);

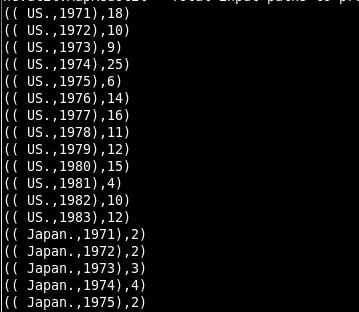
 dump car\_release;

**4) Release Count by year and Brand:**

GroupByBrandYear = GROUP cars BY (brand,year);

dump GroupByBrandYear;

car\_release = foreach GroupByBrandYear Generate group,COUNT(cars.mpg);

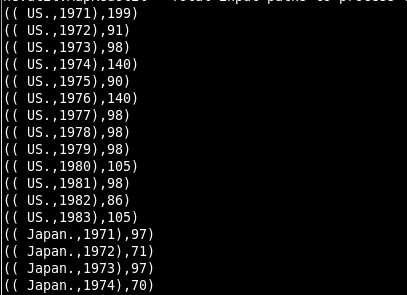
 dump car\_release;

**5) Minimum cubicinches by year and Brand:**

GroupByBrandYear = GROUP cars BY (brand,year);

dump GroupByBrandYear;

car\_release = foreach GroupByBrandYear Generate group,MIN(cars.cubicinches);

 dump car\_release;