`

Big Data
PES University
Bangalore

Hands-On Session 1
MapReduce

24th August 2020

OVERVIEW

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster. A MapReduce program is composed of a map procedure, which performs **filtering** and **sorting**, and a reduce method, which performs an **aggregate** operation.

In this session, we aim to solve a real world problem using mapreduce.

What will you learn?

1. Viewing a problem in a MapReduce perspective.
2. Implementation and working of MapReduce.
3. Intricacies of the Hadoop Distributed File System.

PROBLEM STATEMENT

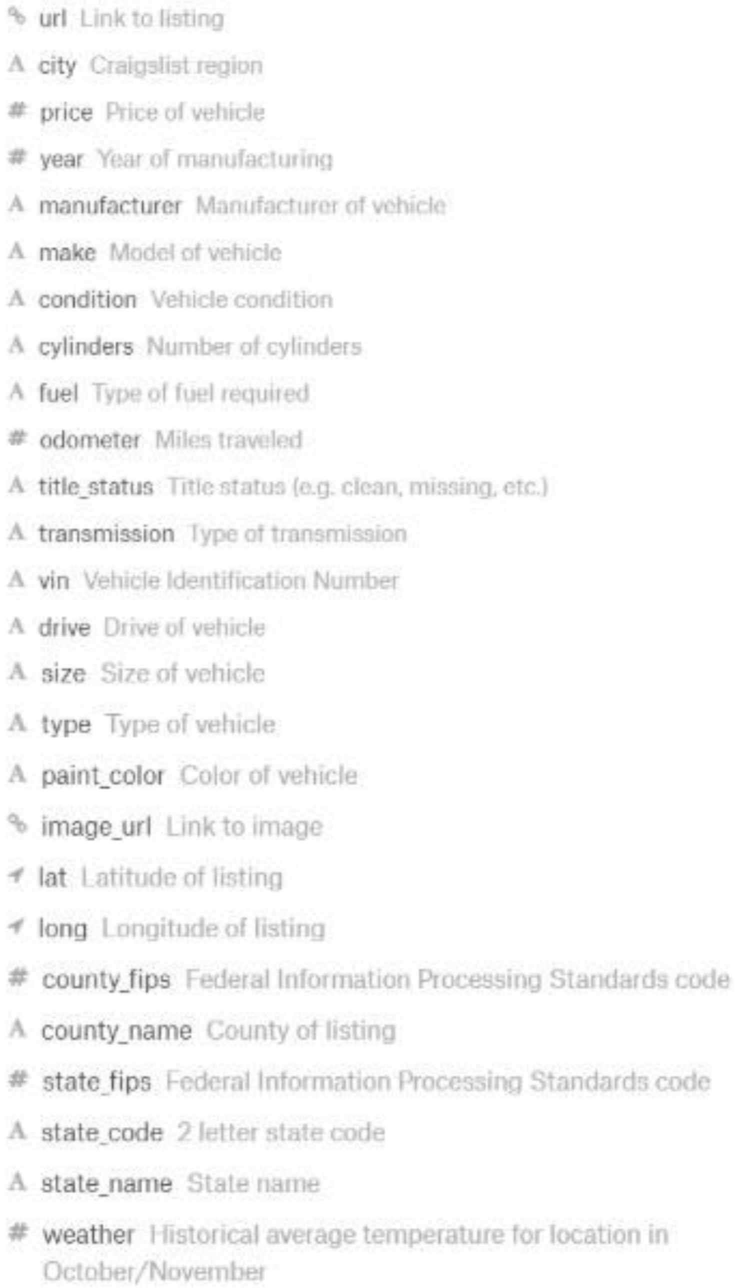
Find the number of cars in every city which use gas as a mode of fuel using MapReduce.

SPECIFICATIONS

1. Ubuntu 16.04+
2. Hadoop: 3.2
3. Python: 2.x/3.x
4. Java: 1.8
5. Dataset: You will be using the modified “Craigslist Used Cars Dataset” for this session. Please download the dataset from the below Google Drive link:
https://drive.google.com/open?id=1GxEaY_aAlkMHfJN2Z1Cvt1O1yNtCp1gN

Please do come with the above requirements installed on your local machines, and use the given dataset only. The dataset is voluminous and an average machine will have a memory shortage in viewing the file.

Columns of the Dataset



🔗 url Link to listing

A city Craigslist region

# price Price of vehicle

# year Year of manufacturing

A manufacturer Manufacturer of vehicle

A make Model of vehicle

A condition Vehicle condition

A cylinders Number of cylinders

A fuel Type of fuel required

# odometer Miles traveled

A title_status Title status (e.g. clean, missing, etc.)

A transmission Type of transmission

A vin Vehicle Identification Number

A drive Drive of vehicle

A size Size of vehicle

A type Type of vehicle

A paint_color Color of vehicle

🔗 image_url Link to image

➤ lat Latitude of listing

➤ long Longitude of listing

# county_fips Federal Information Processing Standards code

A county_name County of listing

# state_fips Federal Information Processing Standards code

A state_code 2 letter state code

A state_name State name

# weather Historical average temperature for location in October/November

Disclaimer: The columns are indexed from [0-25]
(Ex. Transmission is the 11th index)

OUTPUT EXAMPLE

City	Number of Cars that use Gas
Bangalore	10
Chennai	12

The table is solely for representational purposes. We expect the actual output to be in a text file with each line of the answer having the pair <cityname> <number>.