ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA

Scuola di Ingegneria e Architettura
Laurea Magistrale in Ingegneria e Scienze Informatiche

# USED CARS ANALYSIS

*Elaborato in*
BIG DATA

*Presentata da*
GABRIELE GUERRINI

Anno Accademico 2019 – 2020

# Table of contents

# Chapter 1

# Dataset and query

## 1.1 Dataset

The dataset can be downloaded at:
**https://www.kaggle.com/austinreese/craigslist-carstrucks-data**.

## 1.2 Query

The query to be executed is the following one:

"For each region, it must be found the OPI of the most widespread brand in such region, considering cars that use gas fuel only.

OPI is an acronym for Odometer-Price Index and represents the average ratio odometer/price. It is calculated upon all cars of a given brand in the country, regardless of other car features (fuel type, number of cylinders...)."

# Chapter 2

# MapReduce

## 2.1 Workflow

Figure 2.1 shows the adopted workflow[1].

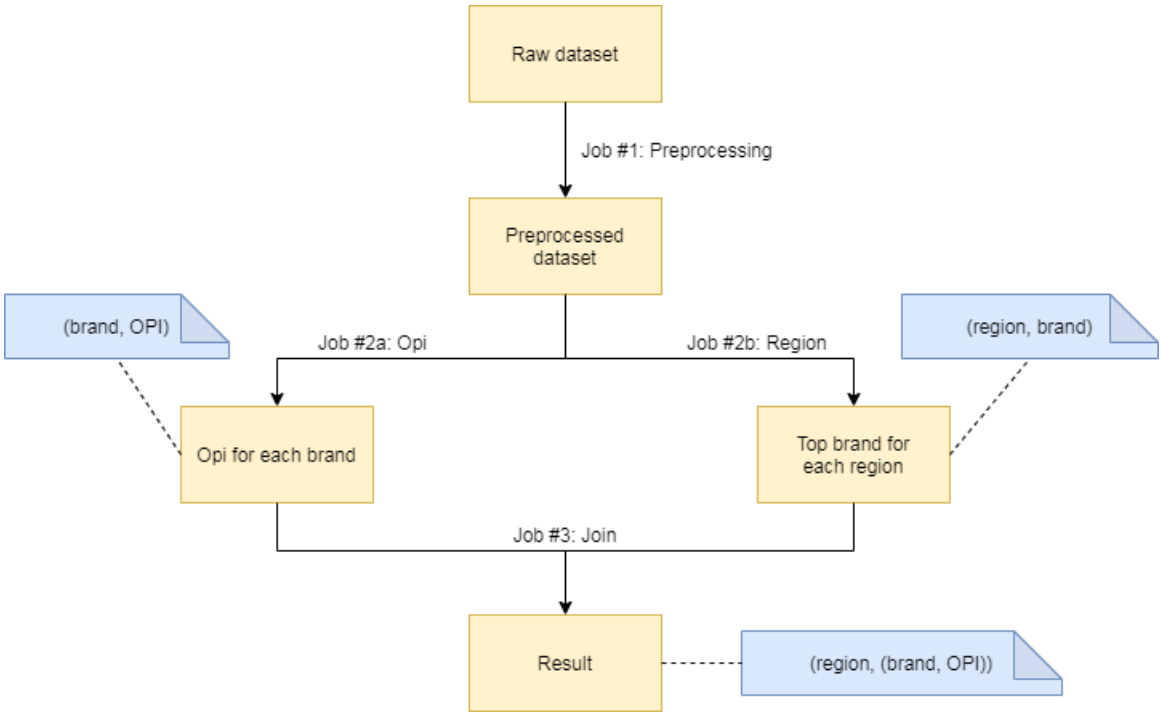


Figure 2.1: Adopted workflow for MR jobs.

[1]It is not the optimal one since the same query could be executed in less than four job. The idea is to structure the workflow this way so that few different MR algorithms can be applied (e.g. filtering, projection, summarization, join).

## 2.2 Jobs

### 2.2.1 Job 1: Preprocessing

The job executes all preprocessing operation needed to correctly use the dataset in next jobs. In particular, it fulfills the following aspects:

- **Cut out header**: the raw dataset is a csv file. The header must be eliminated.

- **Drop useless columns**.

- **Drop incomplete records**: record that have missing values on mandatory fields are simply dropped.

  Job execution and interfaces are described in figure 2.2.

  Raw records are read and parsed into "Car", i.e. custom "Writable" objects, during map stage. Each car stores data about:

  - Region
  - Price
  - Brand
  - Fuel
  - Odometer

  The map output is a pair where the key is the default key used by Hadoop when reading text files and the value is the car itself.

  The reduce stage just replaces the default key with a "NullableWritable" so that the whole job output is a set of records yet.

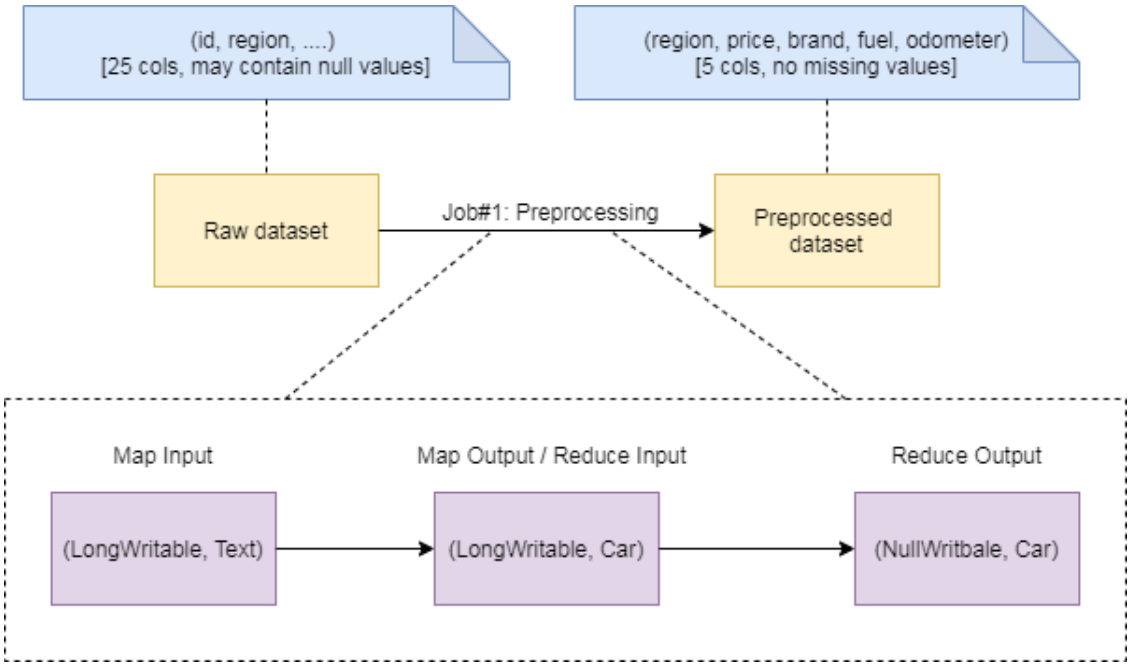


Figure 2.2: Description of "job 1: preprocessing" workflow and interfaces.

### 2.2.2 Job 2a: Opi

### 2.2.3 Job 2b: Region

### 2.2.4 Job 3: Join

## 2.3 Performance evaluation

# Chapter 3

# Spark

## 3.1 Workflow

## 3.2 Performance evaluation

# Chapter 4

# Conclusions