Hadoop and Hive

Igor Baranov
York University School of Continuing Studies
ivb@outlook.com

ABSTRACT

A dataset was scraped from several websites in Czech Republic and Germany over a period of more than a year. The dataset contains over 3.5 million records and has lots of missing data. The goal of this project was to apply such big data tools as Hadoop and Hive to load and query the data and prepare it to the analysis. During the analysis the questions like what is the most advertised vs sol car, car maker and car model vere answered.

1. INTRODUCTION

The first goal of the project was to use apache HIVE as an analytic tool to analyze realistic data. Second goal was to acquare experience working with opened end problems, that are similar to real problems that are faced by data professionals. At the end of this project we should gain sufficient confidence in using Hadoop and Apache Hive, get experience in working with large datasets and be aware of the potential and benefits of analyzing large datasets.

2. ENVIRONMENT PREPARATION

2.1 Hadoop

2.2 HIVE

3. LOADING CLASSIFIED ADS FOR CARS DATA TO HADOOP

3.1 Data understanding

The dataset sit[3] 16 attributes and 3.5 million instances. The data was scraped from several websites in Czech Republic and Germany over a period of more than a year. The scrapers were tuned slowly over the course of the year and some of the sources were completely unstructured, so as a result the data is dirty, there are missing values and some values are very obviously wrong (e.g. phone numbers scraped as mileage etc.) There are roughly 3,5 Million rows and the following columns:

- maker - normalized all lowercase
- model - normalized all lowercase
- mileage - in KM
- manufacture_year
- engine_displacement - in ccm
- engine_power - in kW
- body_type - almost never present, but I scraped only personal cars, no motorcycles or utility vehicles
- color_slug - also almost never present
- stk_year - year of the last emission control
- transmission - automatic or manual
- door_count
- seat_count
- fuel_type - gasoline, diesel, cng, lpg, electric
- date_created - when the ad was scraped
- date_last_seen - when the ad was last seen. Our policy was to remove all ads older than 60 days
- price_eur - list price converted to EUR

Lorem ipsum dolor sit[2] amet, consectetur adipiscing[1] elit.

```
ssh maria_dev@127.0.0.1 -p 2222
cd used-cars/
exit
```

Donec massa justo, ultricies quis facilisis sed, tristique nec metus. Vestibulum id condimentum diam. Integer semper augue id porttitor ultrices. Cras vulputate felis eu diam porttitor, ac pulvinar nisi imperdiet. Donec eros felis, imperdiet vel malesuada at, varius et quam. Phasellus facilisis non risus eu placerat. Sed ac mollis lorem.

4. EVALUATION

Nullam semper imperdiet orci, at lacinia est aliquet et. Sed justo nibh, aliquet et velit at, pharetra consequat velit. Nullam nec ligula sagittis, adipiscing nisl sed, varius massa. Mauris quam ante, aliquet a nunc et, faucibus imperdiet libero. Suspendisse odio tortor, bibendum vel semper sit amet, euismod ac ante. Nunc nec dignissim turpis, ac blandit massa. Donec auctor massa ac vestibulum aliquam. Fusce auctor dictum lobortis. Vivamus tortor augue, convallis quis augue sit amet, laoreet tristique quam. Donec id volutpat orci. Suspendisse at mi vel elit accumsan porta ac ut diam. Nulla ut dapibus quam.

```
CREATE EXTERNAL TABLE IF NOT EXISTS
used_cars.events (
    maker STRING,
```

```
    model STRING,
    mileage INT,
    manufacture_year INT,
    engine_displacement INT,
    engine_power INT,
    body_type STRING,
    color_slug STRING,
    stk_year STRING,
    transmission STRING,
    door_count INT,
    seat_count INT,
    fuel_type STRING,
    date_created TIMESTAMP,
    date_last_seen TIMESTAMP,
    price_eur DECIMAL(13,2)
)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY ','
STORED AS TEXTFILE
LOCATION '/user/maria_dev/cars/classified';

ALTER TABLE used_cars.events SET SERDEPROPERTIES
("timestamp.formats"="yyyy-MM-dd HH:mm:ss.SSSSSSZ");
```

Sed est odio, ornare in rutrum et, dapibus in urna. Suspendisse varius massa in ipsum placerat, quis tristique magna consequat. Suspendisse non convallis augue. Quisque fermentum justo et lorem volutpat euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Morbi sagittis interdum justo, eu consequat nisi convallis in. Sed tincidunt risus id lacinia ultrices. Phasellus ac ligula sed mi mattis lacinia ac non felis. Etiam at dui tellus.

5. ANALYZING DATA

To perform the analysis, certain R libraries were used. The code below was used to load and initialize the library, then loads the data. To pretty-print the tables in this report we used xtable cit[???] library.

```
nrow(cars.sample)
```

```
## [1] 29958
```

5.1 Check for missing values

The dataset has no missing values. Code below calculate number of rows with missing values and checks if there is at list one.

```
any(is.na(cars.sample))
```

```
## [1] TRUE
```

Creating additional columns for analysis

```
cars.sample$ListedTS <-
  strptime(cars.sample$Listed, '%Y-%m-%d %H:%M:%OS')
```

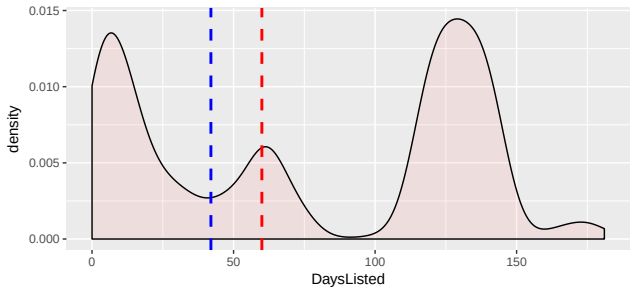


Figure 1: Days Cars Listed

```
cars.sample$RemovedTS <-
  strptime(cars.sample$Removed, '%Y-%m-%d %H:%M:%OS')

cars.sample$Age <- as.integer(ceiling(
  difftime(cars.sample$ListedTS,
           strptime(cars.sample$Year,'%Y'),
           units = "days")/365))

cars.sample$DaysListed <- as.integer(ceiling(
  difftime(cars.sample$RemovedTS,
           cars.sample$ListedTS, units = "days")))
```

How long the cars are usually listed?

```
ggplot(cars.sample, aes(x=DaysListed)) +
  geom_density(fill="#FF6666", alpha=.1) +
  geom_vline(aes(xintercept=42), color="blue",
             linetype="dashed", size=1) +
  geom_vline(aes(xintercept=60), color="red",
             linetype="dashed", size=1)
```

Let's consider cars listed less than 42 days (6 weeks) to be sold

```
cars.sample$Sold <- cars.sample$DaysListed <= 42
```

What is the distribition of advertized cars age?

```
ggplot(cars.sample, aes(x=Age)) +
  geom_density(fill="#FF6666", alpha=.1) +
  scale_x_continuous(limits = c(0, 30))+
    geom_vline(aes(xintercept=mean(Age, na.rm=T)),
               color="green", linetype="dashed", size=1)
```

What is the distribition of mileage of the sold cars?

```
ggplot(cars.sample, aes(x=Mileage)) +
  geom_density(fill="#FF6666", alpha=.1) +
  scale_x_continuous(limits = c(0, 250000))+
    geom_vline(aes(xintercept=mean(Mileage, na.rm=T)),
               color="green", linetype="dashed", size=1)
```

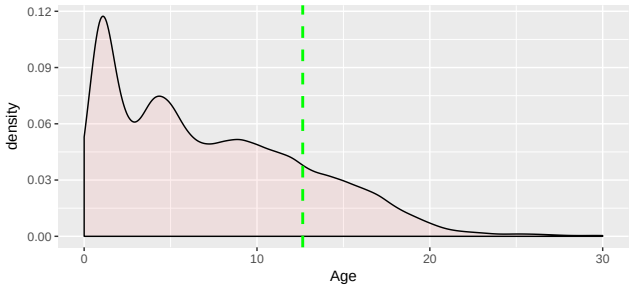


Figure 2: Number of Ads by Maker

Figure 3: Mileage distribution

	Maker	Model	Year	Mileage	Age	DaysListed	Sold
1	skoda	citigo	2014	10349.00	2	75	FALSE
2	fiat	marea	2000	300017.00	16	75	FALSE
3	skoda	octavia	2003	145665.00	13	75	FALSE
4	skoda	citigo	2015	9800.00	1	75	FALSE
5	kia	sportage	2001	1.00	15	75	FALSE
6	skoda	superb	2002	234000.00	14	75	FALSE

Table 1: Sample Car Ads Dataset - first rows

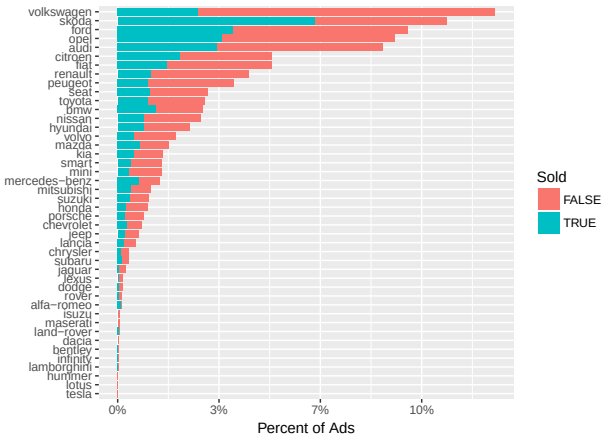


Figure 4: Number of Ads by Maker

```
# summary(cars.sample)
```

What is the most advertized vs sold car maker?

```
require(forcats)
```

```
## Loading required package: forcats
```

```
total <- nrow(cars.sample)
ggplot(cars.sample, aes(fct_rev(fct_infreq(Maker)), fill=Sold))
  geom_bar() +
  labs(x="", y="Percent of Ads") +
  scale_y_continuous(
    labels = function(x) sprintf("%.0f%%",x/total*100)) +
  coord_flip()
```

What is the 20 best advertived vs sold car models?

```
require(forcats)
total <- nrow(cars.sample)
cars.sample$Car <- paste(cars.sample$Maker, cars.sample$Model)
betsCarsList <- fct_infreq(cars.sample$Car)
cars.sample.bestCars <- cars.sample[cars.sample$Car %in%  level
ggplot(cars.sample.bestCars,
       aes(fct_rev(fct_infreq(Car)),fill=Sold))+
  geom_bar() +
  labs(x="", y="Percent of Ads in the Sample Set") +
  scale_y_continuous(labels =
    function(x) sprintf("%.0f%%",x/total*100)) +
  coord_flip()
```

What is the best 20 advertised vs sold cars?

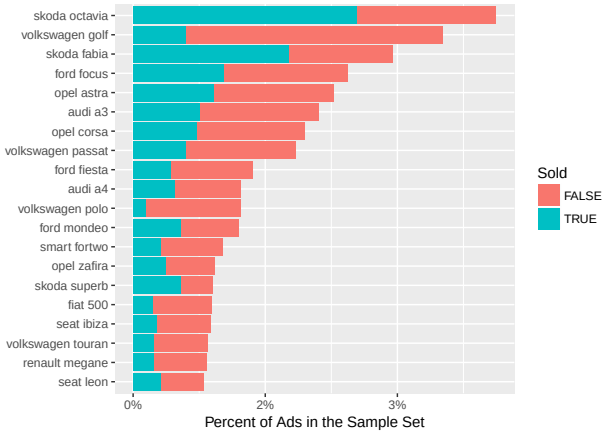


Figure 5: 40 Best Car Models

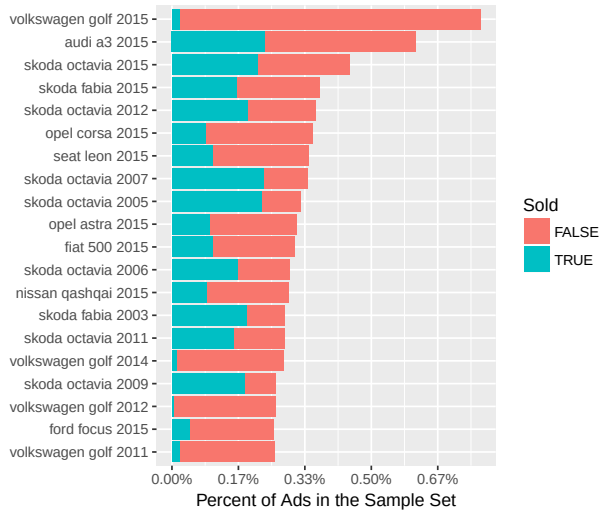
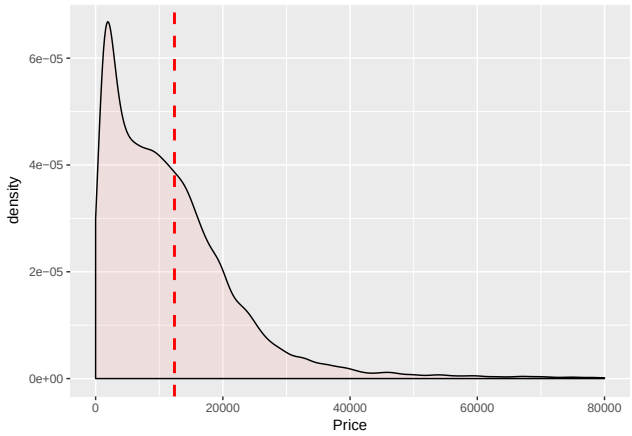


Figure 6: 20 Best Cars

```
require(forcats)
total <- nrow(cars.sample)
cars.sample$Car1 <- paste(cars.sample$Maker, cars.sample$Model, cars.sample$Year)
betsCarsList <- fct_infreq(cars.sample$Car1)
cars.sample.bestCars <- cars.sample[cars.sample$Car1 %in% levels(betsCarsList)[1:20],]
ggplot(cars.sample.bestCars, aes(fct_rev(fct_infreq(Car1)), fill=Sold)) +
       geom_bar() +
       labs(x="", y="Percent of Ads in the Sample Set") +
       scale_y_continuous(labels = function(x) sprintf("%.2f%%", x/total*100)) + coord_flip()
```

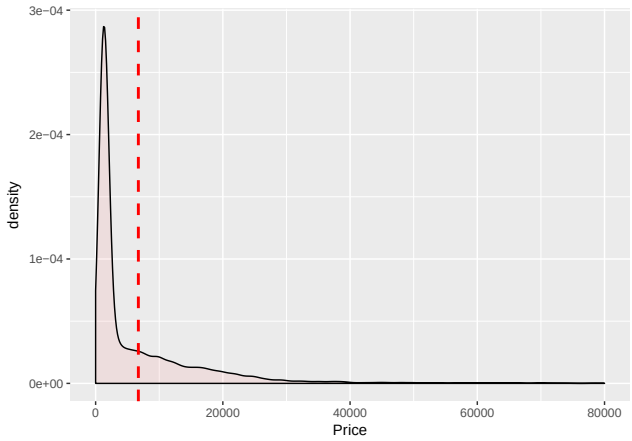
What is the distribution of car prices in the ads for the cars that were not sold?

```
ggplot(cars.sample[!(cars.sample$Sold),], aes(x=Price)) +
  geom_density(fill="#FF6666", alpha=.1) +
  scale_x_continuous(limits = c(0, 80000)) +
    geom_vline(aes(xintercept=mean(Price, na.rm=T)), color="red", linetype="dashed", size=1)
```



What is the distribution of car prices of the cars that were sold?

```
ggplot(cars.sample[cars.sample$Sold,], aes(x=Price)) +
  geom_density(fill="#FF6666", alpha=.1) +
  scale_x_continuous(limits = c(0, 80000)) +
    geom_vline(aes(xintercept=mean(Price, na.rm=T)), color="red
```



6. CONCLUSION

Duis nec purus sed neque porttitor tincidunt vitae quis augue. Donec porttitor aliquam ante, nec convallis nisl ornare eu. Morbi ut purus et justo commodo dignissim et nec nisl. Donec imperdiet tellus dolor, vel dignissim risus venenatis eu. Aliquam tempor imperdiet massa, nec fermentum tellus sollicitudin vulputate. Integer posuere porttitor pharetra. Praesent vehicula elementum diam a suscipit. Morbi viverra velit eget placerat pellentesque. Nunc congue augue non nisi ultrices tempor.

References

[1] Fenner, M. 2012. One-click science marketing. *Nature Materials*. 11, 4 (Mar. 2012), 261–263.

[2] Meier, R. 2012. *Professinal Android 4 Application Development*. John Wiley & Sons, Inc.

[3] Classified ads for cars kaggle.