Big Data Lab 2 | Hochschule Luzern | Master of Science in Applied Information and Data Science

FINISHED

Projekttitel
Datenvorbereitung & -Analyse im Bereich US-Hausmieten

Autor
Ramon Schildknecht
Ebenaustrasse 10
6048 Horw

Dozent
Bruno Grossniklaus | Dozent | Big Data Lab 2

Abgabe der Projektarbeit
14. Februar 2020

Quellen: Die Mehrheit der in diesem Dokument vorliegenden Inhalte sind Teil des Kurses Big Data Lab 2 (https://bdl03-2-hs19.it-grossniklaus.ch/doc/index.html). Die restlichen Inhalte sind gekennzeichnet.

Took 0 sec. Last updated by admin at February 14 2020, 7:54:03 PM.

Inhaltsverzeichnis

FINISHED

- Ausblick

Took 0 sec. Last updated by admin at February 14 2020, 7:54:04 PM.

Arbeitsschritte (Plan)

FINISHED

1. Arbeitsschritte definieren
2. Datenset evaluieren & auswählen
3. Fragestellungen definieren
4. Systemarchitektur & Datenfluss erarbeiten
5. Datenset von Kaggle herunterladen & unzippen
6. Datenset samplen um auf Zielgrösse zu kommen (zw. 50 und 200 MB)
7. Daten auf den Gateway und HDFS laden
8. Daten explorieren (NA-Werte, ...)
9. Fragen beantworten
10. Dokumentation finalisieren
11. Abgabe gemäss Vorgaben (4 Zip-Files) auf Ilias

Took 0 sec. Last updated by admin at February 14 2020, 7:54:04 PM.

Projektablauf (Realität)

FINISHED

1. Arbeitsschritte definieren: schwerfälliger als gedacht infolge Ecosystem-Komplexität
2. Datenset evaluieren & auswählen: Dauerte 3x länger als gedacht!
3. Fragestellungen definieren
4. Systemarchitektur & Datenfluss erarbeiten: mehrere Iterationen benötigt bis zum finalen Ergebnis
5. Datenset von Kaggle herunterladen & unzippen
6. Datenset samplen um auf Zielgrösse zu kommen (zw. 50 und 200 MB): Gleichmässiges Samplen im Verhältnis der Häufigkeit der US-Bundesstaaten als Herausforderung
7. Daten auf den Gateway und HDFS laden: hat nicht über Zeppelin-%sh-Befehl geklappt. Daher bash als Alternative genutzt
8. Daten explorieren (NA-Werte, ...): aufgrund Zeitknappheit nur NULL Werte entfernt
9. Fragen beantworten
10. Dokumentation finalisieren: Management Summary als Herausforderung
11. Abgabe gemäss Vorgaben (4 Zip-Files) auf Ilias

Took 0 sec. Last updated by admin at February 14 2020, 7:54:06 PM.

FINISHED

Management Summary

Zielgruppe: Top Management eines KMUs | Der Vorteil der nachfolgend genutzten Technologien (z. B. Hadoop Distributed File System (HDFS) oder Spark) ist eine eine Zeiteinsparung bei der Berechnung der Abfragen durch paralelle Verarbeitung sowie eine schnellere Time2Market. Für das bessere Verständnis für die parallele Verarbeitung verwende ich folgendes Beispiel:

```
Sie müssen die Buchstaben a und e der pro Harry Potter Buch lesen. Alleine haben Sie
Sie erhalten CHF 10'000 wenn sie es innerhalb eines Tages schaffen. Sie mobilisieren
Die Anzahl e und a tragen die einzelnen Personen in einer zentralen Excel-Liste ein u
Das war ein guter Stundenlohn ;-)
```

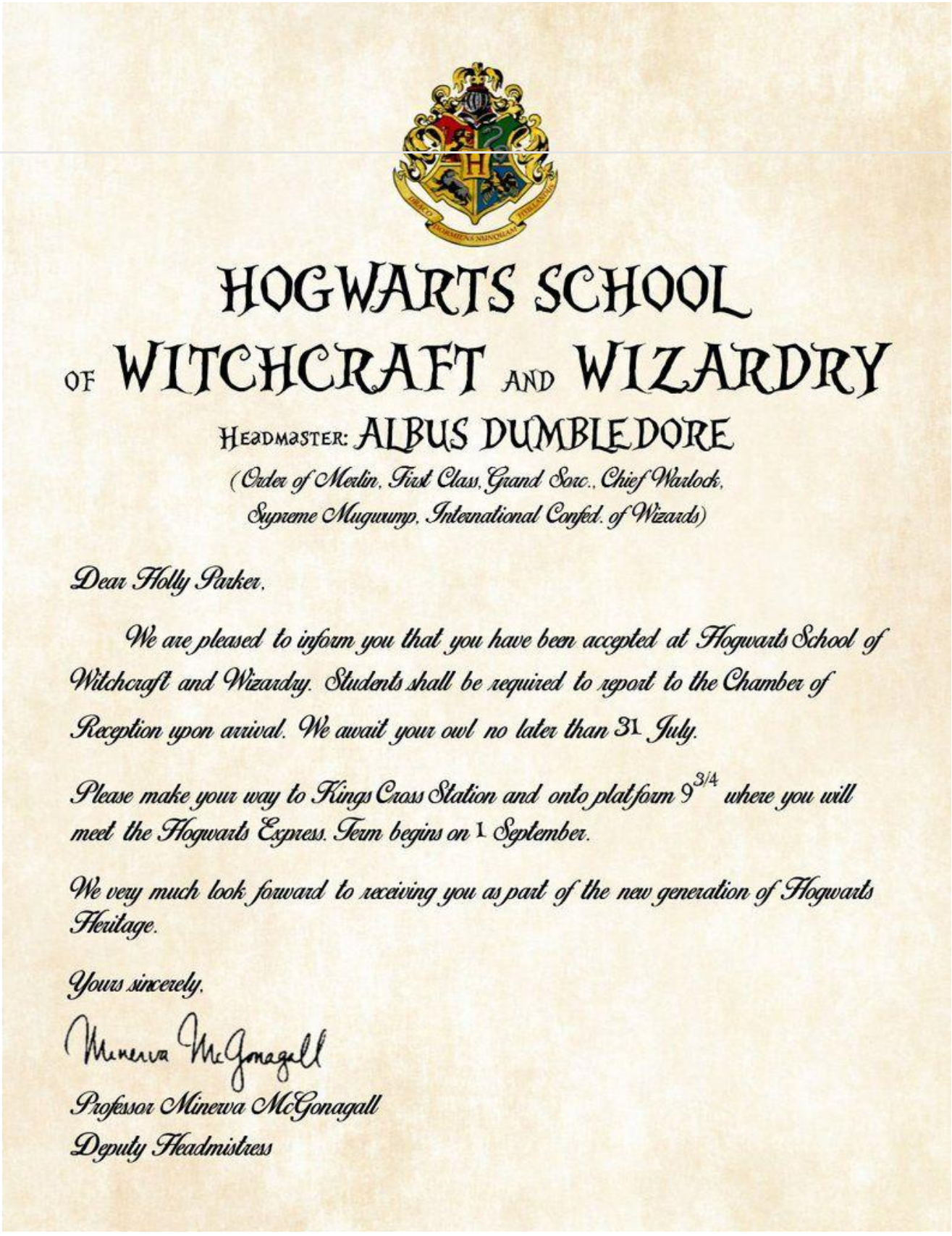
Ähnlich wie in diesem Beispiel funktionert das verteilte Rechnen auf Hadoop. Sie erkennen, dass Sie einerseits viel schneller im Erreichen der Aufgaben und dadurch schneller in der Entscheidungsfindung sind. Weitere Vorteile von Hadoop (https://www.mindsmapped.com/hadoop-advantages-and-disadvantages/) sind: Skaliberbarkeit, Flexibilität, Fehlerresistenz.

Die nachfolgende Arbeit zeigt eine End-zu-End-Lösung für Hadoop ausgehend von Fragestellungen im Hausmietpreisbereich in den USA auf. Die Basis dafür legt Systemarchitektur und die eingesetzten Technologien. Die Daten werden entsprechend ins HDFS gelesen und in eine Datenbank abgelegt. Die Analyse erfolgte direkt über die HDFS-Datei und/oder über die angelegte Datenbank. Dank dem verteilten Rechnen sind die Ergebnisse schnell erzielt.
Die Bewältigung der vielen Technologien und deren Einsatzzweck im jeweiligen Anwendungsfall stellt eine Herausforderung dar. Nach einer herausfordenden Einarbeitungszeit können Data Engineers und Data Scientists schnell die investierte Zeit durch Effizienzsteigerung einsparen. Lessons Learned: Anwender sollen sich auf die wesentlichen Punkte konzentrieren und sich schnell darin einarbeiten. Es ist leicht möglich sich im Hadoop-Technologien-Dschungel zu verirren.

Als Aufpassfeld sind sicherlich die Gesamtkosten zu berücksichtigen. Gerade KMU können

Took 0 sec. Last updated by admin at February 14 2020, 7:54:07 PM.

FINISHED



HOGWARTS SCHOOL
OF WITCHCRAFT AND WIZARDRY
HEADMASTER: ALBUS DUMBLEDORE
(Order of Merlin, First Class, Grand Sorc., Chief Warlock,
Supreme Mugwump, International Confed. of Wizards)

Dear Holly Parker,

We are pleased to inform you that you have been accepted at Hogwarts School of Witchcraft and Wizardry. Students shall be required to report to the Chamber of Reception upon arrival. We await your owl no later than 31 July.

Please make your way to Kings Cross Station and onto platform $9^{3/4}$ where you will meet the Hogwarts Express. Term begins on 1 September.

We very much look forward to receiving you as part of the new generation of Hogwarts Heritage.

Yours sincerely,

Minerva McGonagall

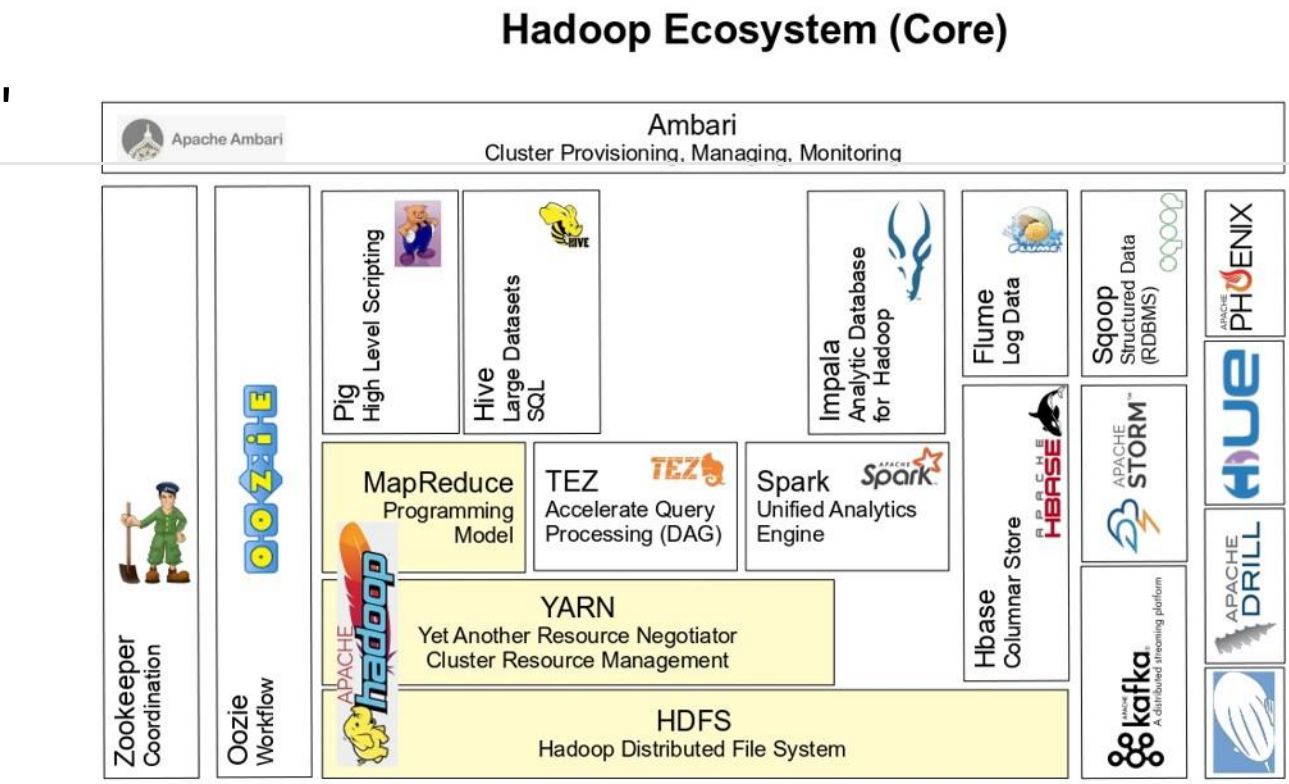
Professor Minerva McGonagall
Deputy Headmistress

Took 0 sec. Last updated by admin at February 14 2020, 7:54:07 PM.

Systemarchitektur

FINISHED

Wir gehen von folgender Architektur aus, von welcher gezielt Bausteine verwendet werden:



Took 0 sec. Last updated by admin at February 14 2020, 7:54:08 PM.

Technologien

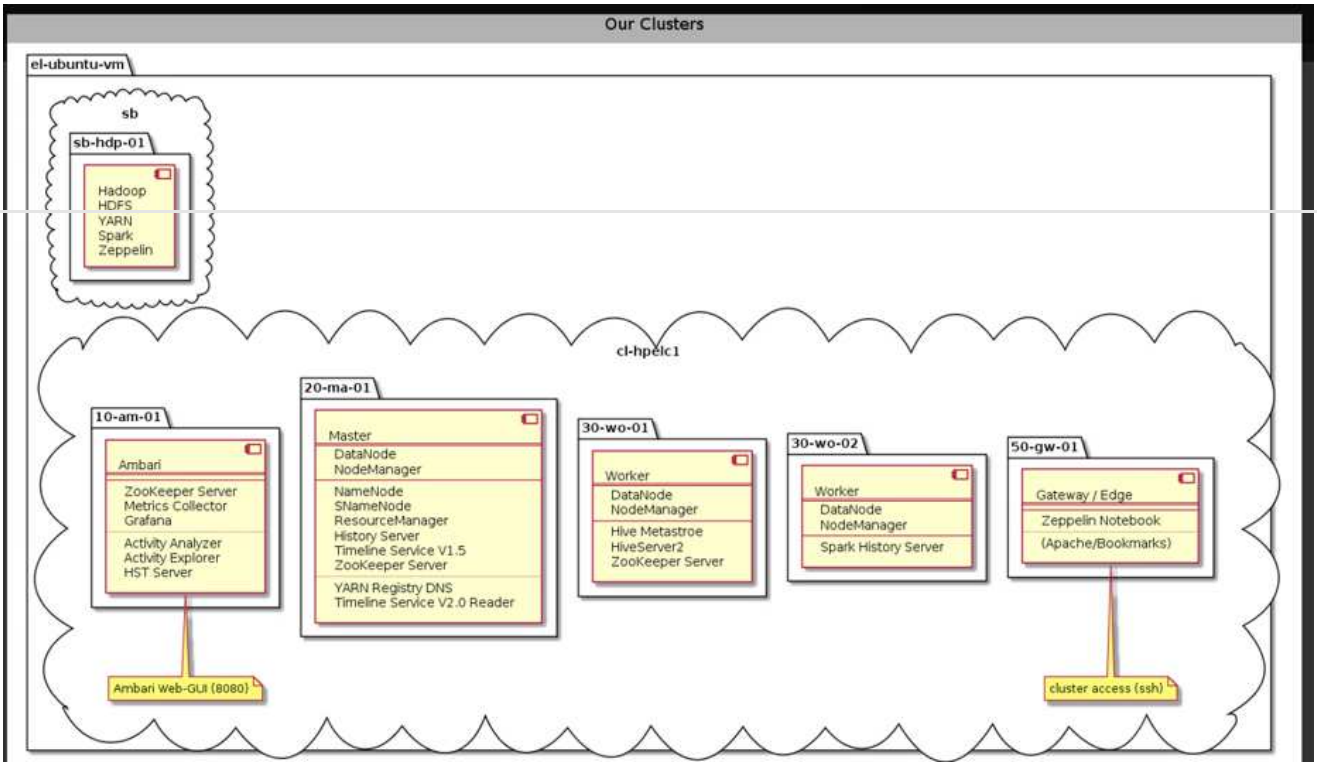
FINISHED

In dieser Arbeit werden folgende Komponenten aus dem links gezeigten System verwendet:

- R (Technologie 1)
- HDFS (Technologie 2)
- pySpark / MapReduce (Technologie 3)
- Spark SQL / MapReduce (Technologie 4)
- Zeppelin (Technologie 5
- Hive (Technologie 6)

Zudem werden YARN und der Zookeeper verwendet.

Die genannten Technologien sind in den folgenden fünf Linux Container auf Linux Ubuntu 18.04 aufgesetzt, welche als Alleinstellungsmerkmal verteiltes Rechnen ermöglichen. Die erwähnten Container und Technologien werden nachfolgend aufgeführt.



Took 0 sec. Last updated by admin at February 14 2020, 7:54:08 PM.

Daten

FINISHED

Took 0 sec. Last updated by admin at February 14 2020, 7:54:08 PM.

Datenfluss

FINISHED

1. Daten von Kaggle herunterladen (https://www.kaggle.com/austinreese/usa-housing-listings)
2. Datei entzippen
3. Mit R Sample erstellen (Reduktion Datengrösse um 70%) um den Datenvorgaben >= 50 & <= 200 MB zu entsprechen
4. Sampledatei von Ubuntu auf Gateway kopieren
5. Sampledatei von Gateway auf HDFS kopieren
6. Anhand Sampledatei eine Datenbank sowie eine Tabelle anlegen
7. Eine Tabelle in Hive anlegen

Die Daten können am Schluss über Zeppelin als ein möglicher Datenzugang eingelesen und für die Beantwortung der Fragestellungen verwendet werden.

Took 0 sec. Last updated by admin at February 14 2020, 7:54:09 PM.

Datenquellen

FINISHED

Der Grund für die Datenwahl ist primär das persönliche Interesse an einem Haus- und/oder Grundstückkauf. Zuerst habe ich während rund 15 Minuten Daten für Grundstückpreise gesucht aufgrund dem äusserst spannenden Artikel https://www.hausinfo.ch/de/home/gebaeude/bauplanung/grundstueckwahl/grundstueckpreise.html (https://www.hausinfo.ch/de/home/gebaeude/bauplanung/grundstueckwahl/grundstueckpreise.html). Danach Strategie geändert und bei Kaggle nach ähnlichem Datenset mit Einschränkung Datei zwischen 50 und 200 MB gesucht. Nach weiteren 30 Minuten und nach dem Sortieren nach «Usablity» (wie gut Datenset beschrieben ist) kamen diese drei Datensets in die engere Auswahl:

1. https://www.kaggle.com/austinreese/usa-housing-listings (https://www.kaggle.com/austinreese/usa-housing-listings)
2. https://www.kaggle.com/tianhwu/brooklynhomes2003to2017 (https://www.kaggle.com/tianhwu/brooklynhomes2003to2017)
3. https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critics-datasets (https://www.kaggle.com/stefanoleone992/rotten-tomatoes-movies-and-critics-datasets)

Das erste Datenset wurde trotz der Grösse von knapp 532 MB ausgewählt. Die entscheidenden Kriterien sind die Kategorie Immobilien und die Usablity. Bei letzterem ist der Maximalwert von zehn vorhanden. Betreffend Immobilien sind zwar nur die Mietpreise vorhanden. Trotzdem sind diese zusammen mit den restlichen Merkmalen wie Anzahl Schlafzimmer und Nutzfläche wertvoll für einen möglichen Hauskauf.

Took 0 sec. Last updated by admin at February 14 2020, 7:54:09 PM.

Sampling

FINISHED

Das Datenset wurde mittels folgenden R-Befehlen *(Technologie 1)* gesampelt. R mit dem Paket-Universum eignet sich wegen der Einfachheit und der Geschwindigkeit bestens für diese Aufgabe. Dadurch konnte die Vorgabe der Datengrösse >= 50 und <= 200 MB sichergestellt werden. Das File ist nach der Transformation noch 162.3 MB gross.

```
library(tidyverse)
library(skimr)
df <- read_csv("data/housing.csv")

# explore
glimpse(df)
skim(df)

df$state <- factor(df$state)

summary(df$state)

set.seed(22)
df_sample_thirty_percent <- sample_frac(df, size = 0.3)
write.csv(df_sample_thirty_percent, file = "data/housing_sample_thirty_percent.csv")

# show if data is equally picked --> success
summary(df_sample_thirty_percent$state)/summary(df$state)
```

Das Skript ist im *ProjectSourceFiles.zip* unter dem Namen *data_sampling.R* zu finden.

Took 0 sec. Last updated by admin at February 14 2020, 7:54:09 PM.

Datenverständnis

FINISHED

Das Codebook (Merkmalbeschreibung) ist wie folgt. Die wichtigen Merkmale für die oben stehenden Fragestellungen sind **fett** hervorgehoben.

- id: listing id
- url: listing URL
- region: craigslist region
- region_url: region URL
- **price: rent per month**
- **type: housing: type**
- **sqfeet: total square footage**
- **beds: number of beds**
- **baths: number of bathrooms**
- cats_allowed: cats allowed boolean (1 = yes, 0 = no)
- dogs_allowed: dogs allowed boolean
- smoking_allowed: smoking allowed boolean
- wheelchair_access: has wheelchair access boolean
- electric_vehicle_charge: has electric vehicle charger boolean
- comes_furnished: comes with furniture boolean

- laundry_options: laundry options available
- parking_options: parking options available
- image_url: image URL
- description: description by poster
- lat: latitude
- long: longitude
- **state: state of listing**

Took 0 sec. Last updated by admin at February 14 2020, 7:54:10 PM.

Fragestellungen

FINISHED

Die aus dem Datenset zentral abgeleitete Fragestellung lautet:

- Wie sieht die Preisverteilung (Monatsmiete in $) der Häuser aus?
- Was ist der Haus-Durchschnittspreis (monatliche Miete in $) pro amerikanischer Bundesstaat?
- Was bekommt ein Mieter bzw. eine Mieterin für das erste, zweite und dritte Quartil der monatlichen Wohnungsmiete (price) für Häuser?
- Welche Unterkünfte gibt es neben Wohnungen und Häusern? Wie sieht die Häufigkeit dieser Unterkünfte aus?

Weitergehende mögliche Fragestellungen sind wie folgt, werden aber im Rahmen dieser Arbeit nicht behandelt:

- Welches sind die häufigsten Wörter der Hausbeschreibungen?
- Welcher Bundesstaat bietet die günstigsten Häuser an?
- Welcher Bundesstaat bietet die teuersten Häuser an?

Took 0 sec. Last updated by admin at February 14 2020, 7:54:11 PM.

Daten laden

FINISHED

Diese Befehle absetzen in Konsole um die Daten vom Ubuntu Server auf den Linux Container mit dem Gateway zu kopieren:

```
stud-10@bdl03-10:~/Project/data$ scp housing_sample_thirty_percent.csv bd01@cl-hpelc1
```

Weiter erstelle ich einen Ordner Projekt im HDFS, gebe dem Zeppelin-User Adminrechte, kopiere das .csv-File vom Gateway auf HDFS und entferne schliesslich den Header der .csv-Datei und schreibe eine neue Datei mit dem Suffix _no_header.csv.

```
bd01@cl-hpelc1-50-gw-01-lx-ub18.lxd:~/repos$ hdfs dfs -mkdir /user/bd01/examples/proj
bd01@cl-hpelc1-50-gw-01-lx-ub18.lxd:~/repos$ hdfs dfs -chmod 777 /user/bd01/examples/
bd01@cl-hpelc1-50-gw-01-lx-ub18.lxd:~$ hdfs dfs -put housing_sample_thirty_percent.cs
bd01@cl-hpelc1-50-gw-01-lx-ub18.lxd:~$ hdfs dfs -cat  /user/bd01/examples/projekt/hou
```

Nachfolgend sehe ich, dass es geklappt hat:

Took 0 sec. Last updated by admin at February 14 2020, 7:54:11 PM.

FINISHED

```
-rw-r--r--   1 bd01 hdfs  170227402 2020-02-13 15:42 /user/bd01/examples/projekt/housing_sample_thi
rty_percent.csv
-rw-r--r--   1 bd01 hdfs  170227132 2020-02-13 17:21 /user/bd01/examples/projekt/housing_sample_thi
rty_percent_no_header.csv
```

Took 3 sec. Last updated by admin at February 14 2020, 7:54:14 PM.

Die Daten sind nun auf dem HDFS (Technologie 2) und können mittels pySpark (Technologie 3) gelesen werden.

FINISHED

Took 0 sec. Last updated by admin at February 14 2020, 7:54:15 PM.

SPARK JOBS FINISHED

```
+---+----------+--------------------+--------------------+--------------------+----+---------+--
--+---+---+----+----+----+----+----+----+----------+-----------------+-------------------+---
-----------------+-------+--------+----+
|_c0|       _c1|                 _c2|                 _c3|                 _c4| _c5|      _c6| _
c7|_c8|_c9|_c10|_c11|_c12|_c13|_c14|_c15|      _c16|              _c17|               _c18|
_c19|   _c20|    _c21|_c22|
+---+----------+--------------------+--------------------+--------------------+----+---------+--
--+---+---+----+----+----+----+----+----+----------+-----------------+-------------------+---
-----------------+-------+--------+----+
|  1|7011325029|https://lafayette...|           lafayette|https://lafayette...| 850|    house|15
00|  2|  1|   1|   1|   1|   0|   0|   0|        NA|   attached garage|https://images.cr...|2 B
EDROOMS 1 BATH...|30.2813| -92.016|  la|
|  2|7035895496|https://dallas.cr...| dallas / fort worth|https://dallas.cr...|1158|apartment| 7
12|  1|  1|   1|   1|   1|   0|   0|   0|w/d in unit|   detached garage|https://images.cr...|Con
tact info:Matt...|32.9227|-96.6248|  tx|
|  3|7046851456|https://spokane.c...|spokane / coeur d...|https://spokane.c...| 950|apartment| 7
25|  2|  1|   1|   1|   0|   0|   0|   0|w/d in unit|    street parking|https://images.cr...|ava
ilable jan 1st   |47.7032|-117.368|  id|
```

settings

_c0	_c1	_c2	_c3	_c4
1	7011325029	https://lafayette.craigslist.org/apa/d/lafayette-new-executive-	lafayette	https://la list.org

		home/7011325029.html		
2	7035895496	https://dallas.craigslist.org/ndf/apa/d/garland-6-weeks-free-brand-new-luxury/7035895496.html	dallas / fort worth	https://d.org
3	7046851456	https://spokane.craigslist.org/apa/d/spokane-hillyard-2-bdrm-	spokane / coeur d'alene	https://slist.org

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 0 sec. Last updated by admin at February 14 2020, 7:54:16 PM. (outdated)

Prototyp

FINISHED

Der Prototyp dient als minimal funktionsfähiges Produkt. Konkret soll ein Data Frame zusammen mit dem Schema (Spaltennamen) für das Housing-Datenset erstellt werden. Dieses soll dann in den Ergebnissen verwendet werden, um die ersten Fragen zu beantworten.

Took 0 sec. Last updated by admin at February 14 2020, 7:54:17 PM.

FINISHED

```
['_c0', 'id', 'url', 'region', 'region_url', 'price', 'type', 'sqfeet', 'beds', 'baths', 'cats_allowed', 'dogs_allowed', 'smoking_allowed', 'wheelchair_access', 'electric_vehicle_charge', 'comes_furnished', 'laundry_options', 'parking_options', 'image_url', 'description', 'lat', 'long', 'state']
```

Took 0 sec. Last updated by admin at February 14 2020, 7:55:50 PM.

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=261) FINISHED

```
(115495, 23)
```

Took 5 sec. Last updated by admin at February 14 2020, 7:55:55 PM.

FINISHED

```
root
 |-- _c0: integer (nullable = true)
 |-- id: float (nullable = true)
 |-- url: string (nullable = true)
 |-- region: string (nullable = true)
 |-- region_url: string (nullable = true)
 |-- price: integer (nullable = true)
 |-- type: string (nullable = true)
 |-- sqfeet: integer (nullable = true)
```

```
|-- beds: integer (nullable = true)
|-- baths: integer (nullable = true)
|-- cats_allowed: integer (nullable = true)
|-- dogs_allowed: integer (nullable = true)
|-- smoking_allowed: integer (nullable = true)
|-- wheelchair_access: integer (nullable = true)
|-- electric_vehicle_charge: integer (nullable = true)
|-- comes_furnished: integer (nullable = true)
|-- laundry_options: string (nullable = true)
|-- parking_options: string (nullable = true)
|-- image_url: string (nullable = true)
|-- description: string (nullable = true)
|-- lat: float (nullable = true)
|-- long: float (nullable = true)
|-- state: string (nullable = true)
```

Took 1 sec. Last updated by admin at February 14 2020, 7:55:56 PM.

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=262) FINISHED

settings

_c0	id	url	region	regi
		-6-weeks-free-brand-new-luxury/7035895496.html		
3	7046851600	https://spokane.craigslist.org/apa/d/spokane-hillyard-2-bdrm-freshly/7046851456.html	spokane / coeur d'alene	https://slist.org
4	7048502800	https://greenville.craigslist.org/apa/d/simpsonville-swimming-pool-close-to/7048502923.html	greenville / upstate	https://gslist.org
5	7043846100	https://spacecoast.craigslist.org/apa/d/melbourne-close-to-park/7043846137.html	space coast	https://sigslist.o
7	7019951100	https://gulfport.craigslist.org/apa/d/gulfport-500-off-of-your-rent-for/7019951211.html	gulfport / biloxi	https://gst.org
8	7040095700	https://spacecoast.craigslist.org/apa/d/melbourne-all-want-from-santa-is-this/7040095550.html	space coast	https://sigslist.o

9	7048571900	https://desmoines.craigslist.org/apa/d/urban	des moines	https://dgslist.or

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 0 sec. Last updated by admin at February 14 2020, 7:55:56 PM.

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=263) FINISHED

```
(111292, 23)
```

Took 5 sec. Last updated by admin at February 14 2020, 7:56:02 PM.

Gewisse Spalten konnten nicht im passenden Typ gespeichert werden wie z. B. die laundry_options. Grund: Fehlende Werte. Diese werden im Rahmen dieser Arbeit nicht weiter angeschaut und könnten Teil der weiteren Aufgaben sein. Sie werden im folgenden Codeblock entfernt. Bedingung: mindestens ein NA-Wert pro Reihe. FINISHED

Took 0 sec. Last updated by admin at February 14 2020, 7:56:02 PM.

SPARK JOBS FINISHED

```
vorher:  (111292, 23)
nach dem Entfernen von none Werten:  (98492, 23)
```

Took 10 sec. Last updated by admin at February 14 2020, 7:56:12 PM.

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=266) FINISHED

settings

_c0	id	url	region	regio
1	7011324900	https://lafayette.craigslist.org/apa/d/lafayette-new-executive-home/7011325029.html	lafayette	https://list.org
2	7035895300	https://dallas.craigslist.org/ndf/apa/d/garland-6-weeks-free-brand-new-luxury/7035895496.html	dallas / fort worth	https://d.org
3	7046851600	https://spokane.craigslist.org/apa/d/spokane-hillyard-2-bdrm-freshly/7046851456.ht	spokane / coeur d'alene	https://slist.org

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Took 0 sec. Last updated by admin at February 14 2020, 7:56:13 PM. (outdated)

Ergebnisse

FINISHED

Ich zeige mir die für die Fragestellungen relevanten Spalten an:

Took 0 sec. Last updated by admin at February 14 2020, 7:56:14 PM.

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=267) FINISHED

```
+-----------+--------------------+-----+---------+------+----+-----+---------------+-----------
--------+-----+
|         id|              region|price|     type|sqfeet|beds|baths|comes_furnished|         des
cription|state|
+-----------+--------------------+-----+---------+------+----+-----+---------------+-----------
--------+-----+
|7.0113249E9|           lafayette|  850|    house|  1500|   2|    1|              0|2 BEDROOMS 1
BATH...|   la|
|7.0358953E9| dallas / fort worth| 1158|apartment|   712|   1|    1|              0|Contact inf
o:Matt...|   tx|
|7.0468516E9|spokane / coeur d...|  950|apartment|   725|   2|    1|              0|available ja
n 1st...|   id|
|7.0485028E9|greenville / upstate|  949|apartment|   737|   1|    1|              0|Special 1 Ca
ll to...|   sc|
|7.0438461E9|         space coast| 1205|    condo|   857|   2|    2|              0|Welcome to U
niver...|   fl|
|7.0199511E9|   gulfport / biloxi|  948|apartment|  1227|   2|    2|              0|Raise your e
xpect...|   ms|
|7.0400957E9|         space coast| 1915|apartment|  1451|   2|    2|              0|Highlands VI
era W...|   fl|
|7.0485719E9|          des moines|  765|apartment|   832|   2|    1|              0|To schedule
a tou...|   ia|
|7.0426916E9|             jackson|  820|apartment|  1050|   2|    2|              0|Our leasing
offic...|   ms|
|7.0483098E9|              boston| 2274|apartment|  1231|   1|    1|              0|Call now
```

Took 0 sec. Last updated by admin at February 14 2020, 7:56:14 PM. (outdated)

FINISHED

SPARK JOBS FINISHED

```
Hauspreisverteilung:
+-------+-----------------+
|summary|            price|
+-------+-----------------+
|  count|             7433|
|   mean|1747.337683304184|
| stddev|18661.604052599214|
|    min|                0|
|    max|          1089000|
```

```
+-------+------------------+

Wohnungspreisverteilung:
+-------+------------------+
|summary|             price|
+-------+------------------+
|  count|             88196|
|   mean|1895.8178035285048|
```

Took 10 sec. Last updated by admin at February 14 2020, 7:56:25 PM.

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=270) FINISHED

```
+--------------+
|Anzahl_Haeuser|
+--------------+
|          7433|
+--------------+
```

Took 6 sec. Last updated by admin at February 14 2020, 7:56:31 PM.

FINISHED

Es soll nun die zweite Frage beantwortet werden:
Was ist der Haus-Durchschnittspreis (monatliche Miete in $) pro amerikanischer Bundesstaat?
Antwort: New York City (nc), New Mexico (nm) und South Carolina (sc) sind die Bundesstaaten mit den teuersten Haus-Durchschnittspreisen. New York City ist dabei doppelt so teuer wie der Rang 2 Bundesstaat New Mexico. Die restlichen Bundestaaten sind unten einsehbar.

Took 0 sec. Last updated by admin at February 14 2020, 7:56:31 PM.

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=271) FINISHED

```
+-----+------------------+
|state|mietpreise_haeuser|
+-----+------------------+
|   nc|10206.229166666666|
|   nm| 4922.803571428572|
|   sc| 3632.852760736196|
|   hi| 2591.216494845361|
|   ca|2569.1148825065275|
|   ma|1987.8494623655913|
|   co| 1895.608040201005|
|   me|1839.2105263157894|
|   ri|1836.0869565217392|
|   wa|1795.5689655172414|
|   ct|1739.4545454545455|
|   dc| 1718.357142857143|
|   or|1681.3229166666667|
|   nv|1614.4102564102564|
|   ak|         1607.8125|
```

Took 6 sec. Last updated by admin at February 14 2020, 7:56:38 PM.

FINISHED

Abkürzungen und Namen der US-Bundesstaaten:
Alabama - AL
Alaska - AK
Arizona - AZ
Arkansas - AR
California - CA
Colorado - CO
Connecticut - CT
Delaware - DE
Florida - FL
Georgia - GA
Hawaii - HI
Idaho - ID
Illinois - IL
Indiana - IN
Iowa - IA
Kansas - KS
Kentucky - KY
Louisiana - LA
Maine - ME
Maryland - MD
Massachusetts - MA
Michigan - MI
Minnesota - MN
Mississippi - MS
Missouri - MO
Montana - MT

Took 0 sec. Last updated by admin at February 14 2020, 7:56:39 PM.

FINISHED

Nebraska - NE
Nevada - NV
New Hampshire - NH
New Jersey - NJ
New Mexico - NM
New York - NY
North Carolina - NC
North Dakota - ND
Ohio - OH
Oklahoma - OK
Oregon - OR
Pennsylvania - PA

Rhode Island - RI
South Carolina - SC
South Dakota - SD
Tennessee - TN
Texas - TX
Utah - UT
Vermont - VT
Virginia - VA
Washington - WA
West Virginia - WV
Wisconsin - WI
Wyoming - WY

Took 0 sec. Last updated by admin at February 14 2020, 7:56:40 PM.

FINISHED

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=272) FINISHED

```
housings
+---+-----------+--------------------+------------------+--------------------+-----+---------+-
-----+----+-----+------------+------------+--------------+-----------------+------------------
----+--------------+--------------+--------------+--------------------+--------------------+-
------+--------+-----+
|_c0|         id|                 url|            region|          region_url|price|     type|s
qfeet|beds|baths|cats_allowed|dogs_allowed|smoking_allowed|wheelchair_access|electric_vehicle_ch
arge|comes_furnished|laundry_options|parking_options|           image_url|         description|
lat|    long|state|
+---+-----------+--------------------+------------------+--------------------+-----+---------+-
-----+----+-----+------------+------------+--------------+-----------------+------------------
----+--------------+--------------+--------------+--------------------+--------------------+-
------+--------+-----+
|  1|7.0113249E9|https://lafayette...|         lafayette|https://lafayette...|  850|    house|
1500|   2|    1|           1|           1|             1|                0|
0|             0|            NA|attached garage|https://images.cr...|2 BEDROOMS 1 BATH...|30.2
813| -92.016|   la|
|  2|7.0358953E9|https://dallas.cr...|dallas / fort worth|https://dallas.cr...| 1158|apartment|
```

Took 0 sec. Last updated by admin at February 14 2020, 7:56:40 PM.

FINISHED

```
%spark2.pyspark
# Create table houses

spark.sql("DROP TABLE IF EXISTS houses")
spark.sql("""
CREATE TABLE IF NOT EXISTS houses
  (
      _c0                       INT
    , id                        FLOAT
    , url                       STRING
    , region                    STRING
    , region_url                STRING
    , price                     INT
    , type                      STRING
    , sqfeet                    INT
    , beds                      INT
    , baths                     INT
    , cats_allowed              INT
```

```
    , dogs_allowed              INT
    , smoking_allowed           INT
    , wheelchair_access         INT
    , electric_vehicle_charge   INT
    , comes_furnished           INT
    , laundry_options           STRING
    , parking_options           STRING
    , image_url                 STRING
    , description               STRING
    , lat                       FLOAT
    , long                      FLOAT
    , state                     STRING

  )
  COMMENT 'Employees created via pyspark'
  STORED AS TEXTFILE
""")
```

```
DataFrame[]
```

Took 0 sec. Last updated by admin at February 14 2020, 7:56:41 PM. (outdated)

```
%spark2.pyspark
# Insert Data from tmpHousings into Table houses
spark.sql("INSERT INTO TABLE houses SELECT * FROM tmpHousings")
```

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=273) FINISHED

```
DataFrame[]
```

Took 9 sec. Last updated by admin at February 14 2020, 7:56:50 PM. (outdated)

Die Tabelle wird zusätzlich noch für die Nutzung in Hive (Technologie 6) vorbereitet: FINISHED

Took 0 sec. Last updated by admin at February 14 2020, 7:56:51 PM.

```
%spark2.pyspark
df_houses_sql.write.mode("append").saveAsTable("houses_append", mode="append", format="hive")
df_houses_sql.write.mode("append").format("hive").saveAsTable("houses_append" )

# create hive table
df_houses_sql.write.mode("overwrite").format("hive").saveAsTable("houses_hive")
df_houses_sql.write.mode("append").format("hive").saveAsTable("houses_hive")
```

SPARK JOBS FINISHED

Took 28 sec. Last updated by admin at February 14 2020, 7:57:20 PM. (outdated)

In weiteren Schritten könnte Hue eingerichtet werden, so dass darüber via HiveQL Abfragen, z. B. von einem Manager, gemacht werden können. FINISHED

Took 0 sec. Last updated by admin at February 14 2020, 7:57:20 PM. (outdated)

```
%spark2.sql
-- Test ob alles geklappt hat
select *
from housings.houses
limit 2
-- Erfolg!
```

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=278) FINISHED

settings

_c0	id	url	region	regio
1	7011324900	https://lafayette.craigs	lafayette	https://la

		list.org/apa/d/lafayette-new-executive-home/7011325029.html		list.org
2	7035895300	https://dallas.craigslist.org/ndf/apa/d/garland-6-weeks-free-brand-new-luxury/7035895496.html	dallas / fort worth	https://d.org

Took 1 sec. Last updated by admin at February 14 2020, 7:57:21 PM. (outdated)

FINISHED

Nun beantworte ich die dritte Frage: Was bekommt ein Mieter bzw. eine Mieterin für das erste, zweite und dritte Quartil der monatlichen Wohnungsmiete (price) für Häuser?

Diese Frage scheint mit SQL nicht wirklich oder schwer beantwortbar zu sein. Dazu habe ich folgende Anweisungen in diesem Artikel gefunden: https://www.quora.com/How-do-I-calculate-first-quartile-and-3rd-quartile-using-SQL (https://www.quora.com/How-do-I-calculate-first-quartile-and-3rd-quartile-using-SQL)

Deshalb wird diese Frage via pyspark in Kombination mit Spark SQL untenstehend im Detail beantwortet.

Took 0 sec. Last updated by admin at February 14 2020, 7:57:21 PM.

SPARK JOBS FINISHED

```
%spark2.pyspark
# filter on 'house' type and generate quartiles
print("1. Quartil: ", df_houses.where(df_houses.type.contains('house')).stat.approxQuantile("price
print("2. Quartil: ", df_houses.where(df_houses.type.contains('house')).stat.approxQuantile("price
print("3. Quartil: ", df_houses.where(df_houses.type.contains('house')).stat.approxQuantile("price
```

```
1. Quartil:  [775.0]
2. Quartil:  [1100.0]
3. Quartil:  [1600.0]
```

Took 16 sec. Last updated by admin at February 14 2020, 7:57:37 PM. (outdated)

FINISHED

1. Quartil

Für $775 erhält ein Käufer bzw. eine Käuferin durchschnittlich rund 854 Quadratfüsse bzw. 79 Quadratmeter Wohnfläche, 1.64 Schlafzimmer und 1.3 Badzimmer. Für die Werte $675 bis $875 (+/- $100) gibt es die durchschnittlichen Ausstattungen gemäss der nachfolgenden SQL-Ausgabe:

Took 1 sec. Last updated by admin at February 14 2020, 7:57:38 PM.

```
%spark2.sql
SELECT price, mean(sqfeet), mean(beds), mean(baths)
FROM housings.houses
WHERE price BETWEEN 675 AND 875
GROUP BY price
ORDER BY price DESC
```

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=282) FINISHED

settings

price	avg(sqfeet)	avg(beds)
875	921.095087163233	1.8399366085578448
874	930.7936507936508	1.8412698412698412
873	733.6818181818181	1.1363636363636365
872	736.2666666666667	1.2666666666666666
871	868.3684210526316	1.736842105263158
870	823.3816425120773	1.4492753623188406
869	822.2974683544304	1.3101265822784811
868	781.7692307692307	1.3461538461538463

Took 25 sec. Last updated by admin at February 14 2020, 7:58:03 PM. (outdated)

2. Quartil

FINISHED

Für $1100 erhält ein Käufer bzw. eine Käuferin durchschnittlich rund 1002 Quadratfüsse bzw. 93 Quadratmeter Wohnfläche, rund 2 Schlafzimmer und 1.4 Badzimmer. Für die Werte $1000 bis $1200 (+/- $100) gibt es die durchschnittlichen Ausstattungen gemäss der nachfolgenden SQL-Ausgabe:

Took 0 sec. Last updated by admin at February 14 2020, 7:58:04 PM.

```
%spark2.sql
SELECT price, mean(sqfeet), mean(beds), mean(baths)
FROM housings.houses
WHERE price BETWEEN 1000 AND 1200
GROUP BY price
ORDER BY price DESC
```

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=283) FINISHED

settings

price	avg(sqfeet)	avg(beds)
1200	1137.8768736616703	2.215203426124197
1199	948.7885304659499	1.8172043010752688
1198	863.8181818181819	1.3636363636363635
1197	1010.7	1.8
1196	855.75	1.5

1195	992.5494949494949	1.97979797979798
1194	969.5588235294117	1.8529411764705883
1193	1135.9130434782608	2.1739130434782608

Took 21 sec. Last updated by admin at February 14 2020, 7:58:26 PM. (outdated)

3. Quartil

FINISHED

Für $1600 erhält ein Käufer bzw. eine Käuferin durchschnittlich rund 1280 Quadratfüsse bzw. 119 Quadratmeter Wohnfläche, rund 2.2 Schlafzimmer und 1.6 Badzimmer. Für die Werte $1500 bis $1700 (+/- $100) gibt es die durchschnittlichen Ausstattungen gemäss der nachfolgenden SQL-Ausgabe:

Took 0 sec. Last updated by admin at February 14 2020, 7:58:27 PM.

```
%spark2.sql
SELECT price, mean(sqfeet), mean(beds), mean(baths)
FROM housings.houses
WHERE price BETWEEN 1500 AND 1700
GROUP BY price
ORDER BY price DESC
```

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=284) FINISHED

settings

price	avg(sqfeet)	avg(beds)
1700	1146.8925373134327	2.080597014925373
1699	1108.5142857142857	2.1285714285714286
1698	976.7777777777778	1.6666666666666667
1697	1017.25	1.75
1696	1146.9285714285713	2
1695	1128.3266331658292	2.150753768844221
1694	1086.5	2
1693	928	1.6

Took 16 sec. Last updated by admin at February 14 2020, 7:58:43 PM. (outdated)

Ich komme bereits zur vierten und letzten Frage:
Welche Unterkünfte gibt es neben Wohnungen und Häusern? Wie sieht die Häufigkeit dieser Unterkünfte aus?

FINISHED

Diese Frage beantworte ich mit Spark SQL:

Took 0 sec. Last updated by admin at February 14 2020, 7:58:43 PM.

```
%spark2.sql
-- show unique Wohntypen
SELECT DISTINCT type
FROM housings.houses
```

SPARK JOBS FINISHED

settings

type
in-law
loft
duplex
townhouse
flat
condo
manufactured
land

Took 9 sec. Last updated by admin at February 14 2020, 7:58:52 PM. (outdated)

FINISHED

Die Übersetzungen der erhaltenen Wohntypen von https://www.deepl.com/ (https://www.deepl.com/) sind wie folgt:

- *in-law* → Wohnung als Hausteil
- *Loft*
- *duplex*
- Stadthaus
- *Wohnung*
- *Eigentumswohnung*
- hergestellt
- *Land*
- Haus
- *Wohnung*
- Häuschen/Kabine

Ich gehe davon aus, dass die *kursiv markierten Wohntypen* keine Häuser sind.

Took 0 sec. Last updated by admin at February 14 2020, 7:58:53 PM.

```
%spark2.sql
-- count by Wohntyp descending
SELECT type, count(type) as
    anzahl_pro_wohnungstyp
FROM housings.houses
WHERE type IN ("in-law", "loft", "duplex", "flat"
    , "condo", "land", "apartment")
GROUP BY type
ORDER BY anzahl_pro_wohnungstyp DESC
```

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=290) FINISHED

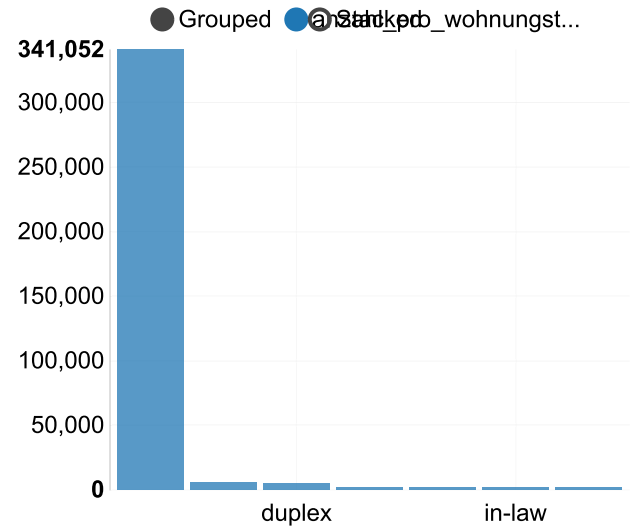
settings

type	anzahl_pro_wohnungstyp
apartment	341052
condo	5632
duplex	4912
loft	736
flat	452
in-law	172
land	12

Took 5 sec. Last updated by admin at February 14 2020, 7:58:59 PM. (outdated)

SPARK JOB (http://cl-hpelc1-50-gw-01-lx-ub18.lxd:4041/jobs/job?id=291) FINISHED

settings



Took 7 sec. Last updated by admin at February 14 2020, 7:59:06 PM. (outdated)

FINISHED

Zusammenfassung der Erkenntnisse anhand der oben stehenden Ausgaben: Der Wohnungstyp “aparment” kommt vorwiegend (Häufigkeit über 90%) vor. Es gibt weitere häufige Wohnungstypen wie “duplex” und “condo” (Eigentumswohnung). Die restlichen Typen wie “loft” oder “flat” sind vernachlässibar.

Took 0 sec. Last updated by admin at February 14 2020, 7:59:06 PM.

Erfahrungen

FINISHED

Über alles gesehen habe ich sehr viel gelernt und gute Erfahrungen gemacht. Es hat sich allerdings einmal mehr gezeigt, dass beim Einarbeiten in solche neue komplexen Technologien viel Geduld und Durchhaltewillen gefordert ist. Weitere Erkenntnisse:

- Fehlende Pakete für Visualisierungen (matplotlib, pandas, numpy), welche Limits mit sich bringen (z. B. ein einfaches Histogramm erstellen).
- Domänenwissen ist extrem wichtig. Im Rahmen dieser Arbeit habe ich mich lediglich auf den Haustyp “house” konzentriert. Es gibt aber weitere Typen wie “manufactured” oder “townhouse”.
- Bestimmte Technologien funktionieren gut bis sehr gut. Bestimmte Funktionalitäten sind noch nicht ausgereift, wie z. B. ein automatisiertes Inhaltsverzeichnis in Zeppelin einfügen.
- Kombination der verschiedenen Technologien an einem zentralen Ort wie Zeppelin ist äusserst wertvoll für Data Engineering und Data Science.
- Die Kenntnis der Zusammenhänge in der Hadoop-Welt sind unabdingbar für ein effektives und effientes Erarbeiten von Big Data-Problemlösungen.
- Beim automatischen Ausführen aller Zeppelin-Paragraphen werden die Codes teils nicht mehr angezeigt. Personen müssen das Notebook im Nachgang nochmals durchschauen, was unpraktisch ist.

Took 0 sec. Last updated by admin at February 14 2020, 7:59:06 PM.

Ausblick

FINISHED

Verschiedene Themen konnten nicht erledigt oder nur teilweise abgeschlossen werden. Die wichtigsten sind nachfolgend aufgeführt:

- Automatisiertes Inhaltsverzeichnis einfügen. Dieser Ansatz schlug fehl: https://zeppelin.apache.org/contribution/documentation.html (https://zeppelin.apache.org/contribution/documentation.html)
- Histogramm und Boxplot für die Verteilung als Ergänzung zur Antwort auf Frage 1 erstellen, z. B. mittels PySpark Dist Explore (https://github.com/Bergvca/pyspark_dist_explore) .

- Spalte für Quadratmeter anhand Quadratfüsse berechnen mit Spark SQL für Frage 3
- Gesamte Data Pipeline weiter automatisieren mit Cronjobs.
- Daten über Hive(QL) abfragbar machen.
- Die weiteren definierten Fragen beantworten, welche im Rahmen dieser Arbeit explizit nicht behandelt wurden:
    - Welches sind die häufigsten Wörter der Hausbeschreibungen?
    - Welcher Bundesstaat bietet die günstigsten Häuser an?
    - Welcher Bundesstaat bietet die teuersten Häuser an?

Took 0 sec. Last updated by admin at February 14 2020, 7:59:07 PM.