UNIVERSITÀ DEGLI STUDI
DI GENOVA

EasyBeer

Andrea Canepa (S 4185248)
Riccardo Bianchini (S 4231932)

Advanced Data Management a.y. 2018/2019
University of Study of Genoa

Structure of the project

1. **Introduction**
2. **Dataset**
3. **Cassandra**
   - *3.1. Aggregates*
   - *3.2. Workload*
4. **Spark**
   - *4.1. Machine Learning*
   - *4.2. Clustering*

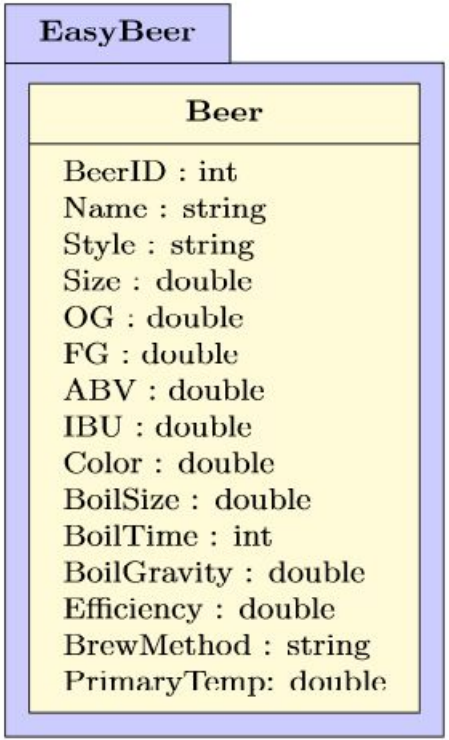
Introduction

Application Domain

We are interested in exploring the world of ***beer.***

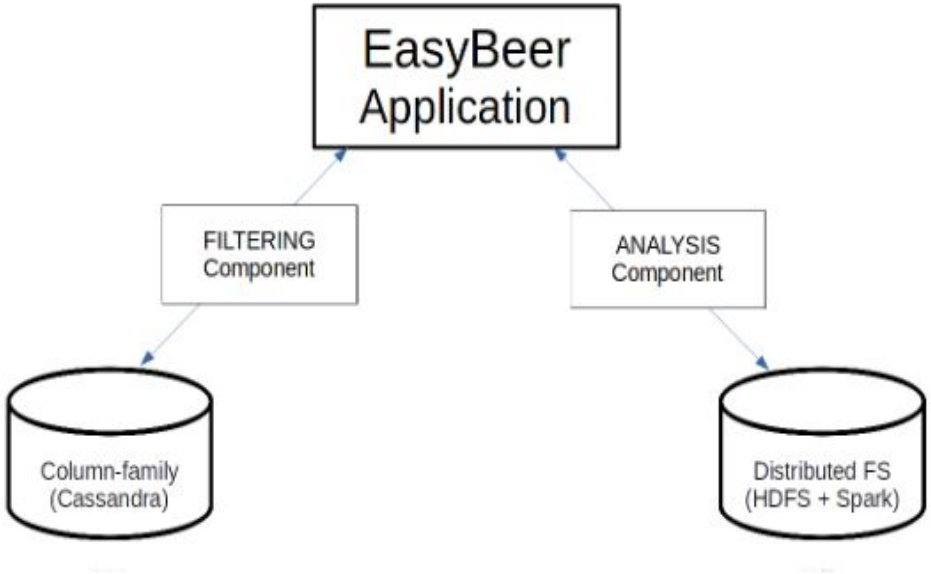
The image show the *Conceptual Schema* on which we have structured our workload.

The features present in the dataset are very specific and technical.



Easy Beer

Our application is divided in 2 components, a filter component, implemented with **Cassandra** and an analytical component, implemented with **Hdfs+Spark**, in order to exploit the specific peculiarities of the different technologies.



Dataset

Dataset

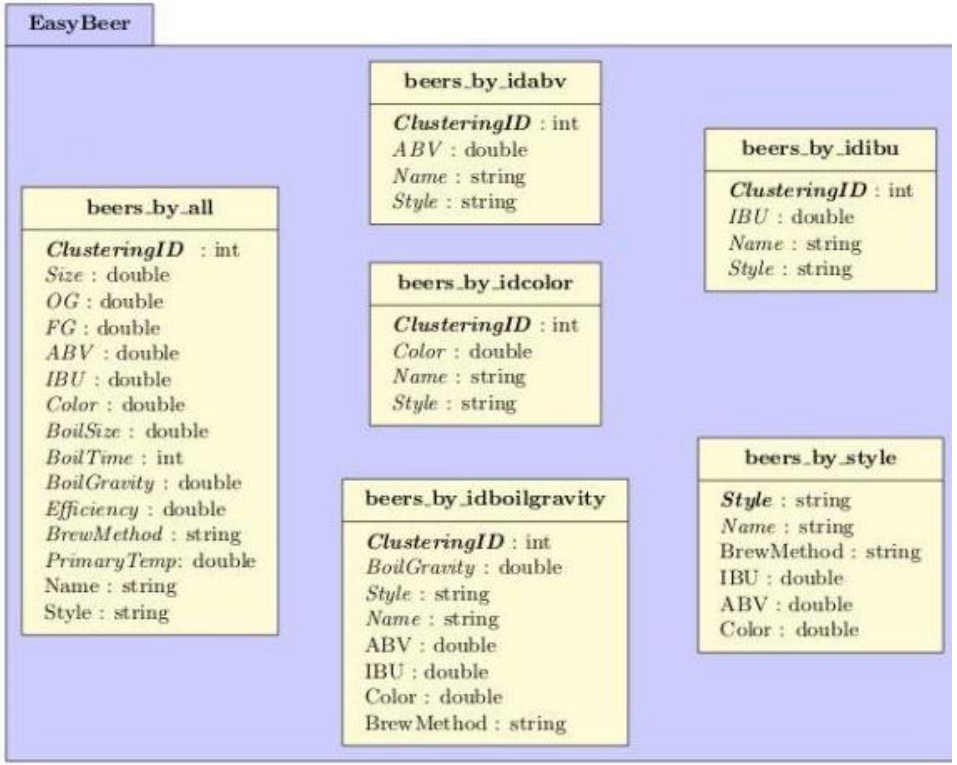
Dataset contains a lot of noise.

preprocessing.sh

- Feature selection
- Removal of non-printable characters
- Conversion from Specific Gravity to Plato Degrees
- Removal of rows that contain “N/A” values

Cassandra

Aggregates


EasyBeer
beers_by_idabv
ClusteringID : int
ABV : double
Name : string
Style : string
beers_by_idibu
ClusteringID : int
IBU : double
Name : string
Style : string
beers_by_all
ClusteringID : int
Size : double
OG : double
FG : double
ABV : double
IBU : double
Color : double
BoilSize : double
BoilTime : int
BoilGravity : double
Efficiency : double
BrewMethod : string
PrimaryTemp: double
Name : string
Style : string
beers_by_idcolor
ClusteringID : int
Color : double
Name : string
Style : string
beers_by_style
Style : string
Name : string
BrewMethod : string
IBU : double
ABV : double
Color : double
beers_by_idboilgravity
ClusteringID : int
BoilGravity : double
Style : string
Name : string
ABV : double
IBU : double
Color : double
BrewMethod : string

Workload

This is the workload as explained in the relation

- Find a beer by name or style
- Find a beer with abv or color or ibu in a given interval
- Find a beer with a specific color
- Order beers by style
- Find beers with a specific classification
- Remove from dataset beers with wrong values
- Obtain average,maximum and minimum color, ibu, abv

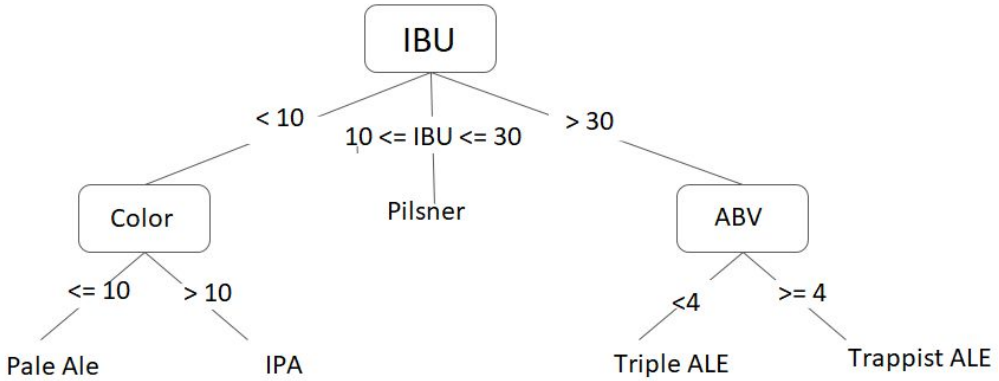
Spark

Spark

- Comparison between elements or subsets
- Aggregates functions user-defined (e.g. STD deviation)
- Self-Join
- Heavy iterative tasks
  - Machine Learning
  - Data Mining

Machine Learning

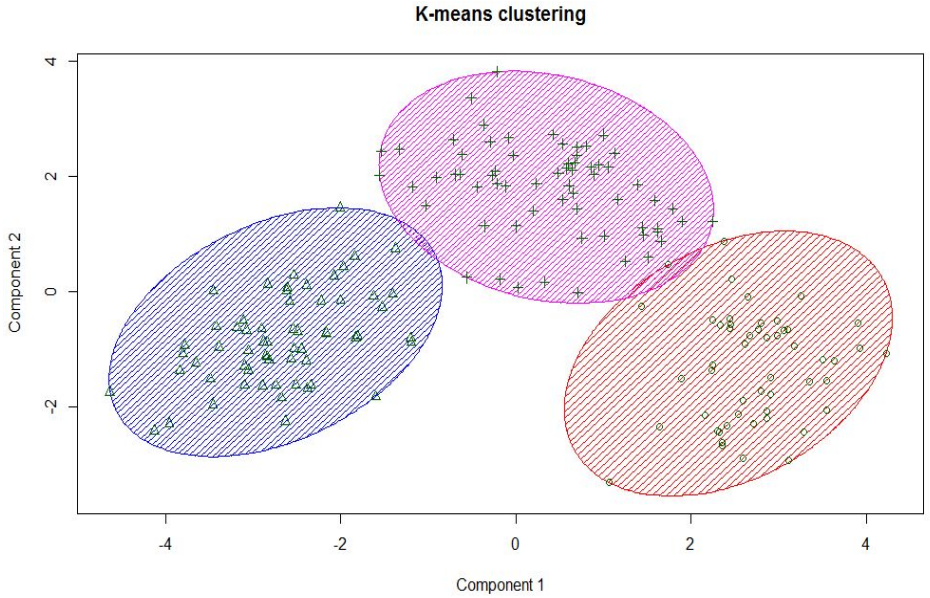
Decision tree model to predict the style of a beer using the other features



Clustering

Algorithms:

- *K-Means*
- *Bisecting K-Means*



Thanks for Your attention !