

# 1 Inleiding

## 1.1 URL

`http://bondarenko.ga`

Voorbeeld accounts staan aan het einde van het rapport.  
**De server waar we op hosten heeft slechts 1Gb aan RAM en kan dus errors geven in geval van memory intensieve operaties!**

## 1.2 Geïmplementeerde (basis)vereisten

- user-service
  - Gebruikers kunnen een profiel aanmaken en de gegevens van hun profiel wijzigen.
  - Gegevens per gebruiker: voornaam (optioneel), achternaam (optioneel), gebruikersnaam (verplicht en uniek), email (verplicht en uniek), admin-rechten (boolean), (non-)actief (boolean)
  - Gebruikers met admin-rechten kunnen alle gebruikers bekijken en andere users: verwijderen, admin-rechten geven en op (non-)actief zetten.
- data-import service
  - Gebruikers kunnen projecten aanmaken (voorzien van een naam en beschrijving) die één of meerdere datasets kunnen omvatten.
  - Gebruikers kunnen in projecten data uploaden (voorzien van een naam en beschrijving) in volgende formaten: CSV, ZIP (met CSV's), SQL database dump
  - De types van de kolommen worden tijdens het uploaden bepaald.
  - Datasets kunnen gejoined worden tijdens het uploaden (als de bestanden meerdere tabellen bevatten) of achteraf.
  - De gebruiker heeft de mogelijkheid om projecten te delen met andere gebruikers. De gebruikers waarmee een project gedeeld wordt, kunnen de metadata hiervan niet wijzigen en dit enkel voor zichzelf verwijderen (= zichzelf access ontenemen).
  - Als de gebruiker, die oorspronkelijk het project heeft aangemaakt, het project verwijdert, wordt dit ook voor alle gebruikers waarmee het gedeeld is verwijderd.
- data-transform service
  - De gebruiker kan rijen verwijderen op basis van zelf samengestelde predicaten.
  - De gebruiker kan per kolom ontbrekende waarden invullen met een zelfgekozen waarde, het gemiddelde of de mediaan.
  - Elke transformatie die wordt uitgevoerd, wordt genoteerd in de geschiedenis van de dataset.
  - De dataset kan terug naar zijn originele toestand worden omgezet.

- Voor alle kolomtypes worden volgende operaties ondersteund:
  - \* De kolom verwijderen.
  - \* Find-and-replace.
  - \* Type veranderen
- Voor kolommen met een categorisch/text type worden volgende operaties ondersteund:
  - \* One-hot-encoding.
  - \* Find-and-replace a.h.v. reguliere expressies.
- Voor kolommen met een numeriek type worden volgende operaties ondersteund:
  - \* Normalisatie.
  - \* Discretisatie (equi-distant, equi-frequent en handmatige ranges).
  - \* Outliers op NULL zetten.
- Voor kolommen met een datum type worden volgende operaties ondersteund:
  - \* Extraheren van delen van de datum
- view-service
  - De ruwe gegevens van de dataset kunnen bekeken worden in een tabel-view.
  - Per kolom kunnen statistieken worden bekeken, samen met een visualisatie van de data vervat in de kolom.
  - De huidige toestand van de data in de dataset kan in CSV vorm gedownload worden.

### 1.3 Gebruikte frameworks/libraries

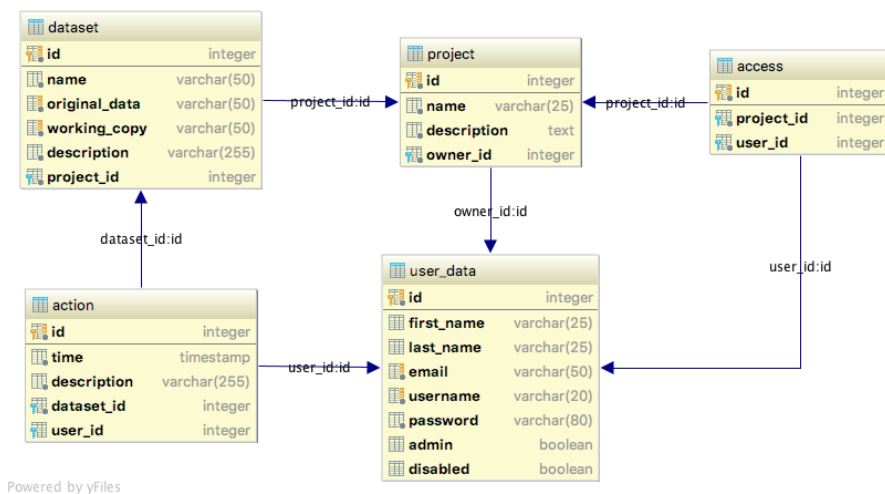
- |                    |                                 |
|--------------------|---------------------------------|
| • Flask            | • SQLAlchemy-Datatables         |
| – Flask-Bootstrap4 |                                 |
| – Flask-SQLAlchemy | • Datatables                    |
| – Flask-WTF        |                                 |
| – Flask-Login      | • Chart.js                      |
| – Flask-Testing    |                                 |
| • Psycpg2          | • Font Awesome                  |
| • Psycpg2-binary   |                                 |
| • WTForms          | • Python Unit-Testing Framework |
| • Pandas           | • Apache HTTP Server            |

## 1.4 Specialisaties

- Andrei: Front- & Backend
- Igor: Backend & Testing
- Stijn: Backend & Hosting

## 2 Design

### 2.1 Database design



**Verandering** Sinds het laatste rapport is de view entiteit weggefallen wegens het feit dat deze niet essentieel was om de basisvereisten te implementeren. Omdat we met een tekort aan tijd te kampen hadden, hebben we dus besloten om deze voorlopig te elimineren en indien nodig voor extra functionaliteiten terug te introduceren.

### Entiteiten

- *user\_data*: Houdt de gegevens van alle gebruikers bij.
- *project*: Houdt de naam en beschrijving van projecten bij.
- *access*: Helper tabel om toegang van gebruikers tot project bij te houden.
- *dataset*: Houdt de gegevens bij voor elke dataset, alsook de naam van de tabel waar de ruwe data wordt opgeslagen.
- *action*: Houdt alle acties (transformaties) bij, alsook de id van de view, waarop de transformatie is uitgevoerd, en de id van de user, die deze heeft uitgevoerd.

## Relaties

- *user\_data (many) to project (many) via access*: Een user kan toegang hebben tot meerdere projecten en een project kan toegankelijk zijn voor meerdere users.
- *project (one) to dataset (many)*: Een dataset kan enkel tot één project behoren.
- *dataset (one) to action (many)*: Er kunnen meerdere transformaties op één dataset uitgevoerd worden.
- *user\_data (one) to action(many)*: Gebruikers kunnen meerdere transformaties uitvoeren.

## Constraints

- *user\_data*:
  - *email*: Een gebruiker **moet** een **uniek** emailadres hebben (unique not null)
  - *username*: Een gebruiker **moet** een **unieke** username hebben (unique not null)
  - *password*: Een gebruiker **moet** een wachtwoord hebben (not null)
- *project*:
  - *name*: Een project **moet** een naam hebben (not null)
  - *description*: Een project **moet** een beschrijving hebben (not null)
  - *owner\_id*: Een project **moet** een eigenaar hebben, die bestaat in de user\_data tabel (not null foreign key referencing user\_data.id)
- *dataset*:
  - *name*: Een dataset **moet** een naam hebben (not null)
  - *description*: Een dataset **moet** een beschrijving hebben (not null)
  - *original\_data*: Een dataset **moet** een **unieke** tabelnaam bevatten waar de originele ruwe data in wordt opgeslagen (unique not null)
  - *working\_copy*: Een dataset **moet** een **unieke** tabelnaam bevatten waar de huidige toestand van de data in wordt opgeslagen (unique not null)
  - *project\_id*: Een dataset **moet** gekoppeld zijn aan een project, dat bestaat in de project tabel (not null foreign key referencing project.id)

- *action*:
  - *time*: Een actie **moet** een tijdstip hebben waarop deze is uitgevoerd (not null)
  - *description*: Een actie **moet** een beschrijving hebben (not null)
  - *dataset\_id*: Een actie **moet** gekoppeld zijn aan een dataset, die bestaat in de dataset tabel (not null foreign key referencing project.id)
  - *user\_id*: Een actie **moet** gekoppeld zijn aan een gebruiker, die bestaat in de user\_data tabel (not null foreign key referencing user\_data.id)

## 2.2 Programma design

**Packages** De backend van onze website is opgedeeld in vijf hoofdpackages:

- Admin
- Data
- Main
- Project
- User

De Data-package wordt verder nog onderverdeeld in volgende subpackages:

- Import
- Transform
- View

Elk van deze (sub)packages kan één of meerdere van de volgende modules bevatten:

- *models*: Bevat klassen die de database tabellen , die nodig zijn voor de werking van de packages, voorstellen.
- *controllers*: Bevat de verschillende routes waarmee onderdelen van de package kunnen worden aangesproken.
- *operations*: Bevat alle functies die berekeningen/omzettingen/database operaties/etc. uitvoeren.
- *forms*: Bevat klassen die vaste forms, die gebruikt worden door de package, voorstellen.

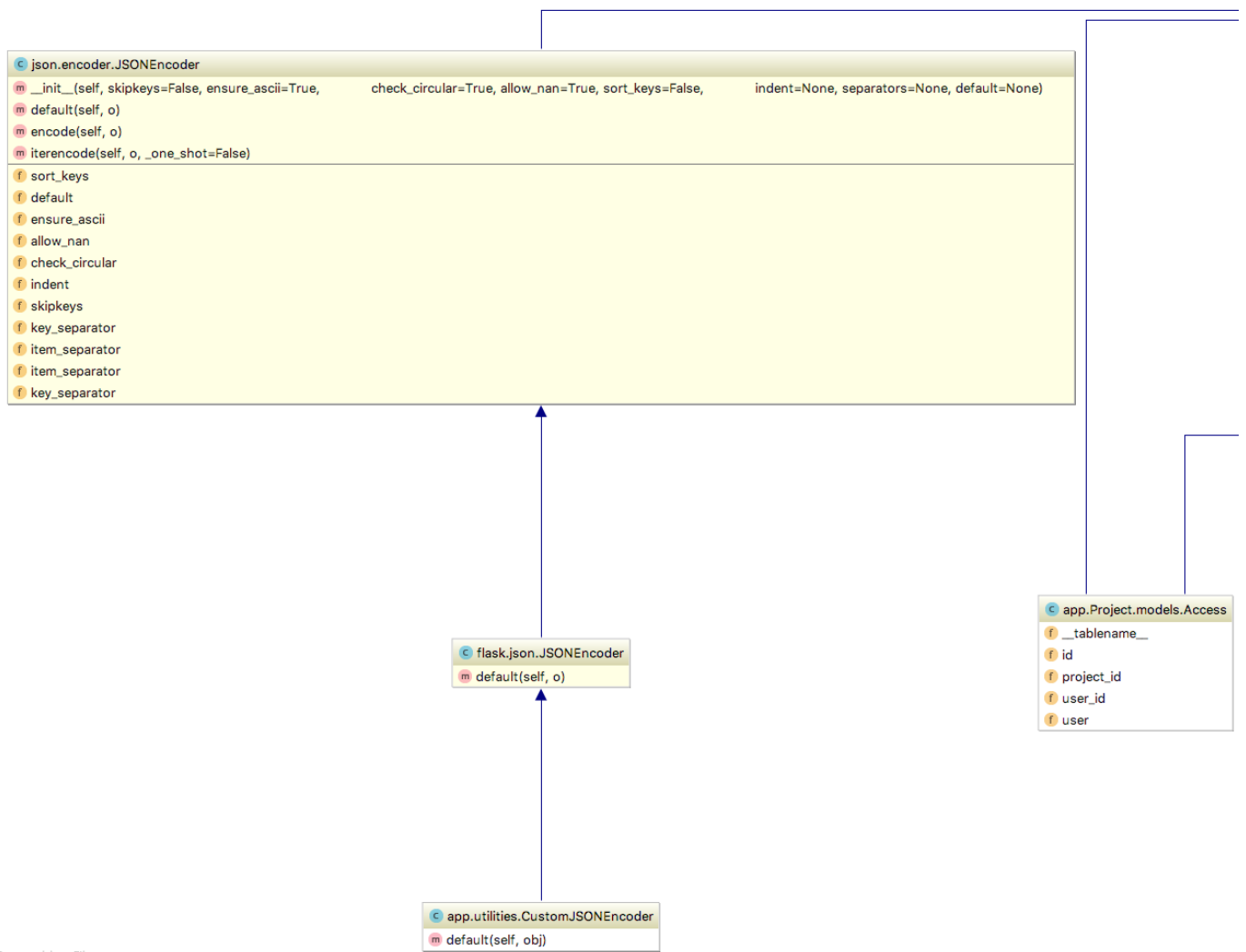


Figure 1: Class diagram left

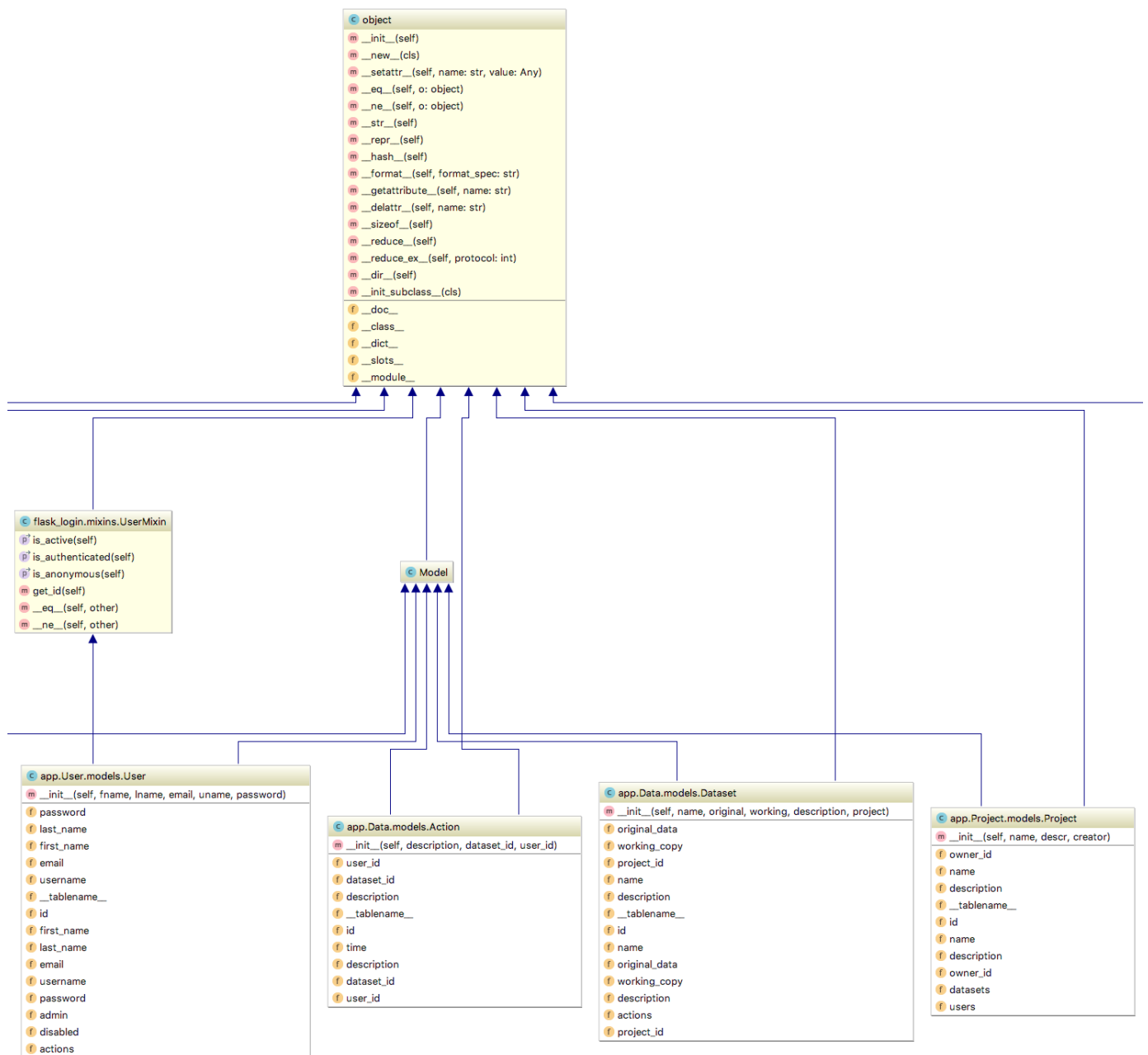


Figure 2: Class diagram mid

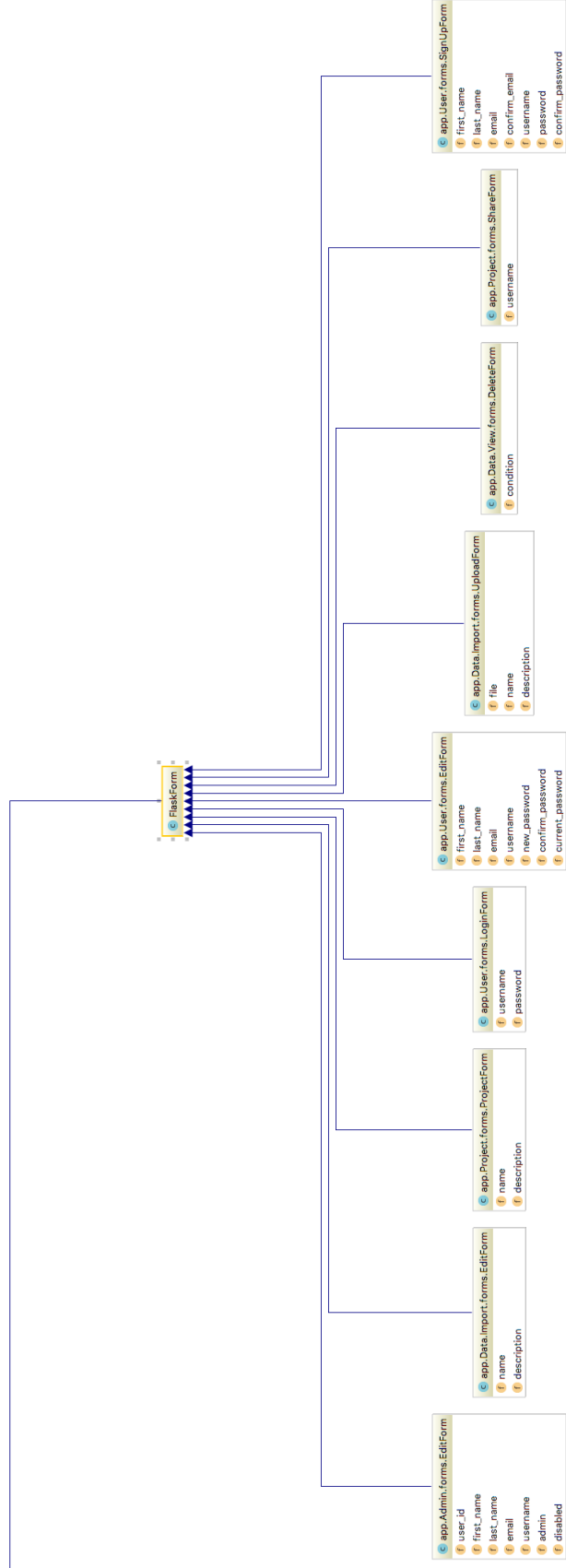
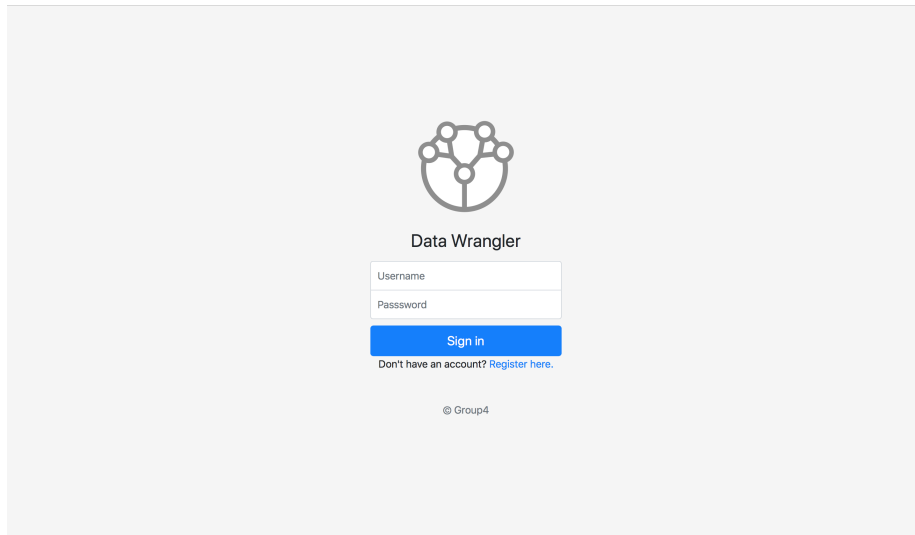


Figure 3: Class diagram right



## 3 Manual

### 3.1 Login



Wanneer een gebruiker voor de eerste keer onze website bezoekt, zal hij/zij terecht komen op de login pagina, waar hij/zij het volgende kan doen:

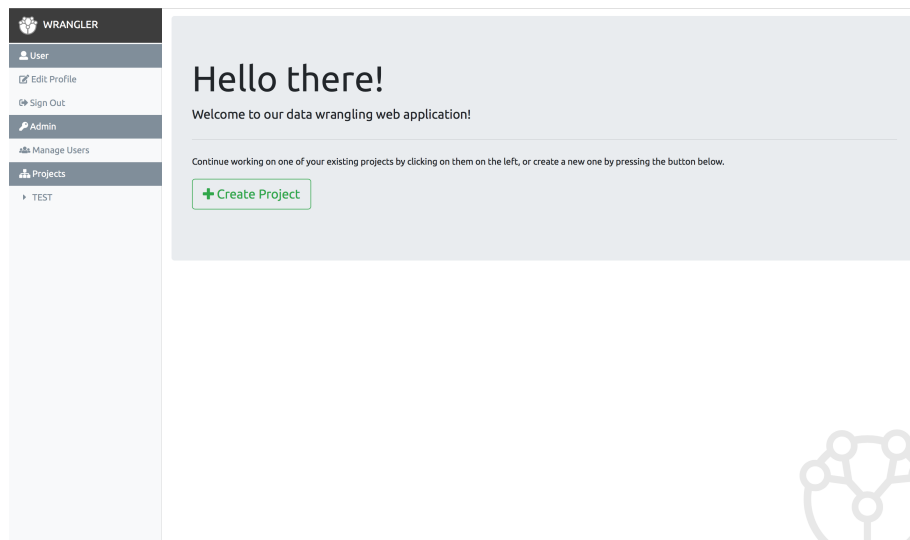
- D.m.v. op de “Register here”-link te drukken kan de gebruiker zich registreren.
- Mits een succesvolle registratie, kan de gebruiker zich aanmelden met zijn gegevens, door deze in hun respectievelijke velden in te vullen en op de “Sign In”-knop te duwen.

### 3.2 Dashboard

Eens de gebruiker zich succesvol heeft aangemeld zal hij/zij begroet worden door het dashboard. De structuur die men op deze pagina ziet zal behouden worden op de rest van de website.

**Operations sidebar** Aan de linkerkant, in de operations sidebar, zullen steeds alle acties die men kan uitvoeren te vinden zijn.

**Content view** Aan de rechterkant, in de content view, zal steeds de informatie specifiek aan de pagina weergegeven worden.



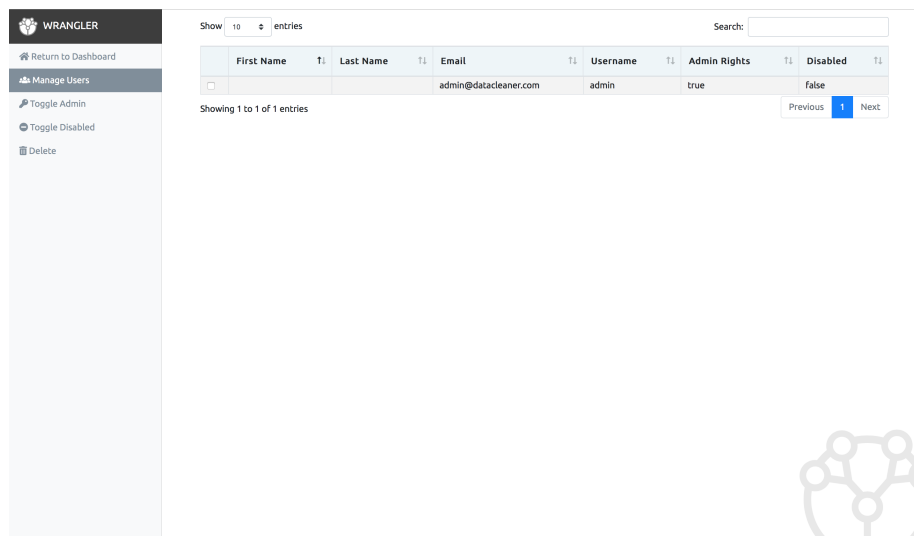
Vanuit de dashboard kan de gebruiker:

- Nieuwe projecten aanmaken door op de “Create Project”-knop de drukken.
- Bestaande project openen, bewerken, delen en verwijderen in de “Projects”-sectie in de operations sidebar.
- Zijn/haar gegevens wijzigen.
- Zich afmelden.

Indien de gebruiker een admin is, heeft hij/zij ook de mogelijkheid:

- Zich te begeven naar de administratiepagina .

### 3.3 Admin



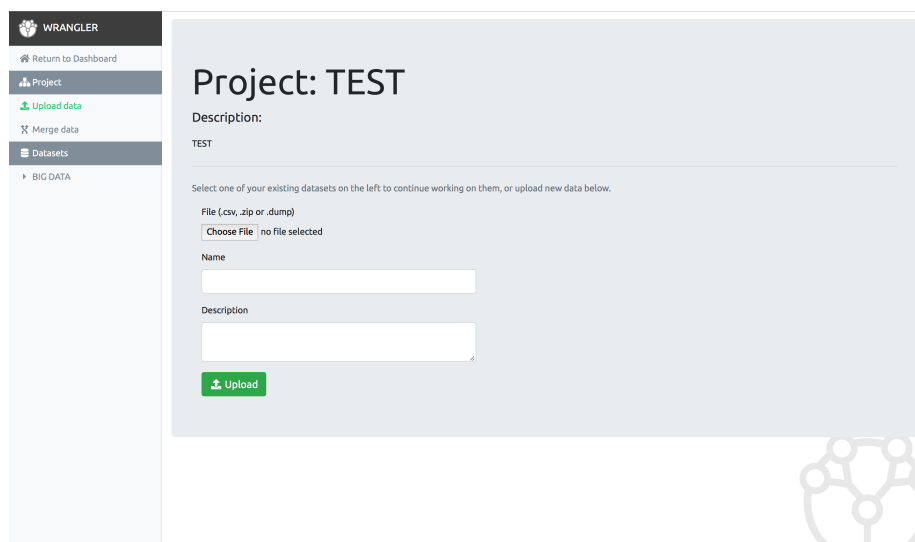
De administratiepagina biedt volgende mogelijkheden:

- Alle gebruikers bekijken in de content view.
- Gebruikers selecteren d.m.v. checkboxen en te bewerken met de operaties, die terug te vinden zijn in de operations sidebar.
- Terugkeren naar het dashboard.

### 3.4 Project

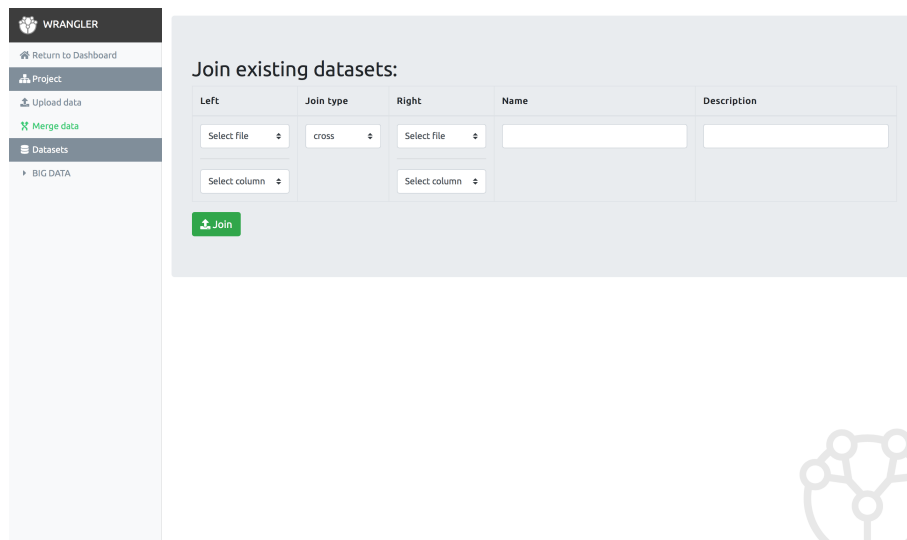
Voor de projectpagina's zal de operations sidebar steeds dezelfde mogelijkheden bieden:

- Men kan wisselen tussen de data-upload en data-merge pagina's.
- Men kan de geplote datasets openen, bewerken en verwijderen.
- Men kan terugkeren naar het dashboard



The screenshot shows the 'WRANGLER' interface. On the left, a sidebar contains links: 'Return to Dashboard', 'Project' (active), 'Upload data', 'Merge data', and 'Datasets'. The main area is titled 'Project: TEST'. Below the title, there's a 'Description:' field with the text 'TEST'. A message says 'Select one of your existing datasets on the left to continue working on them, or upload new data below.' Below this, there's a file upload section with the text 'File (.csv, .zip or .dump)' and a 'Choose File' button. Below the file section, there are input fields for 'Name' and 'Description', followed by an 'Upload' button. A faint logo is visible in the bottom right corner.

De upload pagina in een project biedt de mogelijkheid om bestanden te uploaden.

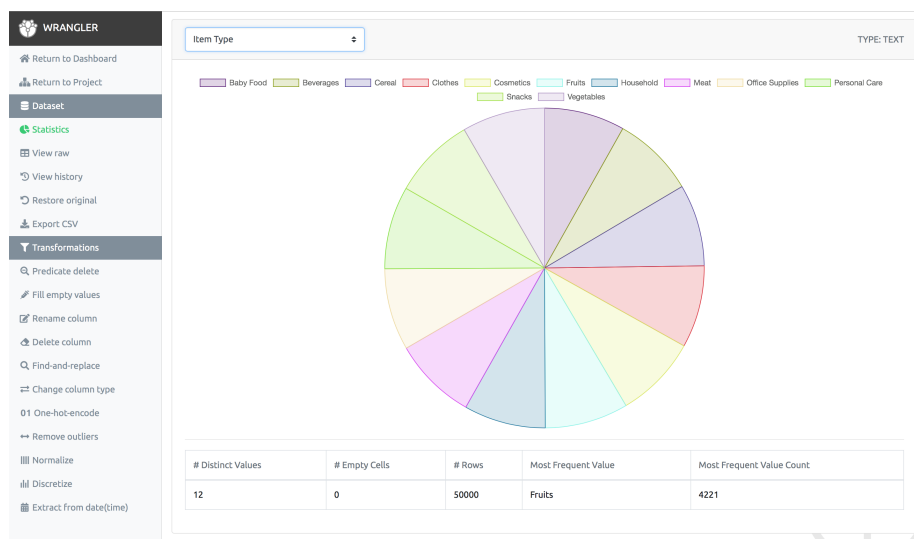


De merge pagina in een project biedt de mogelijkheid om reeds bestaande datasets te joinen.

### 3.5 Dataset

Voor de datasetpagina's zal de operations sidebar steeds dezelfde mogelijkheden bieden:

- Men kan wisselen tussen de statistics-, raw- en history-view pagina's.
- Men kan de geplote datasets openen, bewerken en verwijderen.
- Men kan terugkeren naar het dashboard



In de statistics-view kan men per kolom statistieken en een grafische representatie bekijken.

WRANGLER

Return to Dashboard

Return to Project

Dataset

Statistics

View raw

View history

Restore original

Export CSV

Transformations

Predicate delete

Fill empty values

Rename column

Delete column

Find-and-replace

Change column type

One-hot-encode

Remove outliers

Normalize

Discretize

Extract from date(time)

Show 25 entries

Search:

	Region T1	Country T1	Item Type T1	Sales Channel T1	Order Priority T1	Order Date T1	Order ID T1	Ship Date T1	Units Sold
<input type="checkbox"/>	Asia	Brunei	Cosmetics	Online	M	2013-03-04 00:00:00	485189139	2013-03-06 00:00:00	7103
<input type="checkbox"/>	Asia	Cambodia	Clothes	Online	H	2015-12-31 00:00:00	288541162	2016-01-25 00:00:00	4597
<input type="checkbox"/>	Asia	Indonesia	Meat	Offline	C	2011-03-06 00:00:00	885269582	2011-03-18 00:00:00	9797
<input type="checkbox"/>	Asia	Taiwan	Beverages	Offline	C	2010-03-14 00:00:00	323083293	2010-04-16 00:00:00	6381
<input type="checkbox"/>	Asia	Kazakhstan	Cosmetics	Offline	M	2013-04-20 00:00:00	647978627	2013-05-05 00:00:00	130
<input type="checkbox"/>	Asia	Malaysia	Beverages	Offline	L	2016-06-15 00:00:00	328939092	2016-07-19 00:00:00	9242
<input type="checkbox"/>	Asia	Sri Lanka	Household	Offline	H	2015-07-02 00:00:00	829262912	2015-08-21 00:00:00	6179
<input type="checkbox"/>	Asia	Taiwan	Snacks	Online	L	2011-02-03 00:00:00	274351203	2011-03-21 00:00:00	5472
<input type="checkbox"/>	Asia	Kyrgyzstan	Personal Care	Offline	H	2010-06-18 00:00:00	311560952	2010-08-04 00:00:00	6664
<input type="checkbox"/>	Asia	Tajikistan	Office Supplies	Online	M	2012-05-28 00:00:00	351776921	2012-06-14 00:00:00	7463
<input type="checkbox"/>	Asia	Sri Lanka	Office Supplies	Online	L	2016-10-28 00:00:00	363430154	2016-10-28 00:00:00	5914
<input type="checkbox"/>	Asia	India	Vegetables	Online	C	2010-03-21 00:00:00	897317636	2010-04-05 00:00:00	5084
<input type="checkbox"/>	Asia	Kazakhstan	Baby Food	Online	M	2016-01-30 00:00:00	446566617	2016-02-16 00:00:00	7841
<input type="checkbox"/>	Asia	Malaysia	Snacks	Offline	C	2013-11-18 00:00:00	163463103	2013-12-05 00:00:00	1220
<input type="checkbox"/>	Asia	Singapore	Meat	Offline	C	2010-06-13 00:00:00	133662259	2010-06-19 00:00:00	2488
<input type="checkbox"/>	Asia	Taiwan	Baby Food	Offline	H	2013-01-16 00:00:00	355735862	2013-01-19 00:00:00	9002
<input type="checkbox"/>	Asia	North Korea	Vegetables	Offline	M	2013-04-15 00:00:00	509674060	2013-05-22 00:00:00	2735
<input type="checkbox"/>	Asia	Kazakhstan	Household	Offline	H	2014-07-29 00:00:00	526880967	2014-08-28 00:00:00	4641
<input type="checkbox"/>	Asia	Mongolia	Office Supplies	Online	M	2015-03-27 00:00:00	919973504	2015-04-05 00:00:00	9459
<input type="checkbox"/>	Asia	Malaysia	Clothes	Offline	M	2014-11-24 00:00:00	269806039	2015-01-10 00:00:00	457
<input type="checkbox"/>	Asia	Cambodia	Beverages	Online	C	2012-06-10 00:00:00	274321591	2012-07-21 00:00:00	6710
<input type="checkbox"/>	Asia	Brunei	Household	Online	M	2016-11-03 00:00:00	694549930	2016-11-17 00:00:00	1613
<input type="checkbox"/>	Asia	Indonesia	Meat	Online	H	2016-08-26 00:00:00	473074476	2016-08-27 00:00:00	7643

Showing 1 to 25 of 50,000 entries

Previous

1

2

3

4

5

...

2000

Next

In de raw-view kan men de ruwe data in tabelvorm bekijken/sorteren/doorzoeken.

WRANGLER

Return to Dashboard

Return to Project

Dataset

Statistics

View raw

View history

Restore original

Export CSV

Transformations

Predicate delete

Fill empty values

Rename column

Delete column

Find-and-replace

Change column type

One-hot-encode

Remove outliers

Normalize

Discretize

Extract from date(time)

## History

Time	Description	User
2018-04-18 14:10:25.691527	rows deleted with condition "'Region'='1' or '1' = '1'"	admin
2018-04-20 08:49:17.590088	restored dataset to original state	admin
2018-04-20 08:51:42.133157	restored dataset to original state	admin
2018-04-20 14:20:26.799703	Deleted column time from Ship Date	admin
2018-04-20 14:27:54.148195	replaced 1 with Monday in column isodow from date from Order Date	admin
2018-04-20 16:58:48.139414	Deleted column Region	DELETED USER
2018-04-21 06:55:00.593993	Deleted column Country	admin
2018-04-21 09:25:28.903002	restored dataset to original state	admin

In de history-view kan men de historiek van uitgevoerde operaties bekijken.

## 4 Status

Onderdeel	Uitvoerder	Status
<b>user-service:</b>		
Registratie en login	Andrei	klaar
Gegevens wijzigen	Igor	klaar
Admin-operaties	Igor	klaar
<b>data-import-service:</b>		
CSV upload	Stijn	klaar
ZIP upload	Andrei	klaar
SQL-dump upload	Stijn	WIP
Join bij upload en achteraf	Andrei	klaar voor zip, WIP voor SQL-dump
Projecten aanmaken	Stijn	klaar
Projecten delen	Andrei	klaar
<b>data-transform-service:</b>		
Verwijderen op basis van predicaat	Igor	klaar
Ontbrekende data vullen	Andrei	Klaar
Historiek bijhouden	Igor	klaar
Tabeloperaties	Andrei + Igor	klaar
<b>view-service:</b>		
Tabel weergave ruwe data	Andrei + Igor	klaar
Grafiek per kolom	Andrei	klaar
Statistieken per kolom	Andrei	klaar
Download naar CSV	Andrei	klaar
<b>varia:</b>		
Hosting	Stijn	klaar
Zelfopkuisend systeem	Andrei	klaar
SQL-injectie preventie	Igor	klaar
Testen	Igor	klaar
Frontend	Andrei	klaar
Rapport	Andrei	klaar
Mockups	Thomas	klaar

## 5 Geplande extra functionaliteit

Om effectief te beslissen welke extra functionaliteit we gaan implementeren, gaan we wachten tot we feedback hebben gekregen op de presentatie, maar enkele ideeën die wel al hadden waren:

- Uitbreiding visualizaties.
- Omzetting van tekst naar features d.m.v. bag-of-words.
- Data-deduplicatie.
- Vullen van ontbrekende data d.m.v. het k-nearest neighbors algoritme.
- Een compare-view om statistieken tussen kolommen te vergelijken.

## 6 Demo Accounts

Username	Password	Admin	Disabled
admin	admin	True	False
normal	normal	False	False
disabled	disabled	False	True