

# Rapport Groep 1 – 24/05/2018

## 1. Inleiding

Webapp url: <http://projectdb18.tk>

Test account 1: Email: [john.doe@hotmail.com](mailto:john.doe@hotmail.com) Password: dbinlog1  
Test account 2: Email: [janedoe@gmail.com](mailto:janedoe@gmail.com) Password: dbinlog2  
Test account 3: Email: [henkjansen@gmail.com](mailto:henkjansen@gmail.com) Password: dbinlog3

Het werk dat we sinds de vorige presentatie hebben verzet heeft vooral te maken met het implementeren van de extra functionaliteiten, het opschonen van de basis functionaliteiten en het hervormen van de front-end user interface.

Wat de extra functionaliteiten betreft zijn we erin geslaagd om alle opgegeven functionaliteiten te implementeren. Binnen de view-service zijn er nieuwe opties om data te downloaden. Downloaden in sql-dump formaat is nu mogelijk, en men kan nu een dataset download in zip formaat in plaats van enkel individuele tables te moeten downloaden. Binnen de transform-service heeft men nu meer mogelijkheden om tables te joinen. Het is ook mogelijk om een attribute conversion te forceren, indien er een value niet geconverteerd kan worden zal deze op NULL gezet worden, of kan de rij verwijderd worden. Het is nu ook mogelijk om user-generated queries op te geven.

Het opschonen van bestaande functionaliteiten houdt vooral in dat we de code opgeruimd hebben en dat we SQL-queries van UserManager, DatasetManager, DatasetInfo en DatasetPermissionsManager nu vervat hebben in de Model laag van de applicatie waardoor deze nu volledig gescheiden is van de Controller laag.

Het herstructureren van de frontend heeft vooral als gevolg dat de userinterface zoveel mogelijk gecentreerd is in plaats van verspreid te zijn over meerdere paginas. Zo hebben we nu gebruikt gemaakt van modals. Deze hebben we voornamelijk gebruikt op de table pagina om transformaties en statistieken weer te geven. Ook is de site nu gebruiksvriendelijker door het gebruik van loading screens.

## Specialisaties

Atisha:

- Upload en Download klasse
- Statistieken en Grafieken
- Data Deduplicatie
- Join Tables

Chris:

- Transformaties
  - Psycopg2
  - SQL Alchemy
  - Postgresql

Kasper:

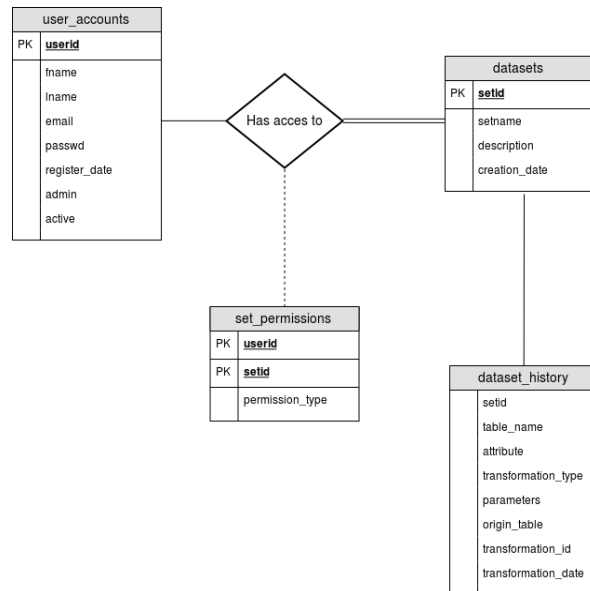
- User systeem
- Dataset systeem
- Communicatie tussen frontend en backend
  - Flask

Luuk:

- Clientside code
  - HTML/CSS
  - Javascript/JQuery
  - Ajax

## 2. Design

### Databank ontwerp



De databank is redelijk simpel, met maar 3 entiteiten: `user_accounts`, `datasets` en `dataset_history`. De relatie tussen `user_accounts` en `datasets` is een many-to-”at least one”-relatie, omdat een gebruiker aan meerdere datasets moet kunnen en een dataset moet door minstens 1 gebruiker ge-accessed worden. De relatie tussen deze twee wordt weergegeven door `set_permissions`. `Permission_type` kan een van drie waarden zijn: ‘admin’, ‘write’ of ‘read’. Elke `dataset_history` entry hangt vast aan een dataset door de `setid` attribute. De `table_name` attribute geeft weer voor welke table de transformatie gebeurd is.

De datasets van de user worden opgeslagen in aparte postgresql schema’s zodat die van elkaar gescheiden blijven. Hierdoor treden er geen problemen op in verband met tabelnamen of permissions. De naam van zo’n dataset schema is de `setid`, die gevonden kan worden in de `datasets` tabel.

Per schema zal er ook een extra original (back-up) schema zijn die de oorspronkelijke data bevat vóór eventuele transformaties. De naam van dit schema is redelijk simpel: `original_setid`, met `setid` de id van de dataset waaraan de backup gerelateerd is.

# Software Stack

## Development

- Git
- Bitbucket

## Server-side

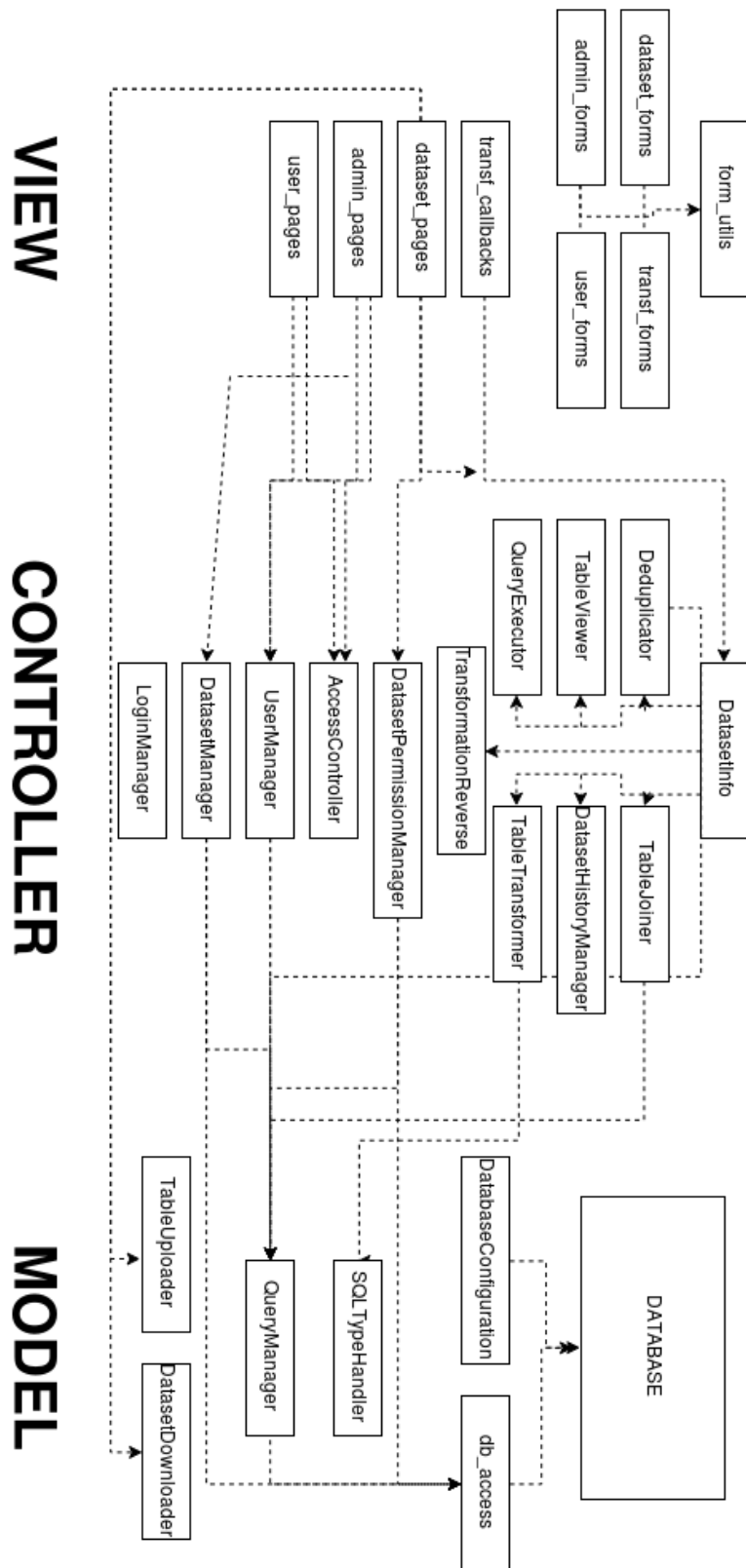
- Python 3
- Flask
- Postgresql
- Psycpg2
- WTForms
- Flask-WTF
- Pandas
- Passlib
- SQL Alchemy
- ConfigParser
- PSQLParse (Custom Queries)
- Recordlinkage (Data Deduplication)

## Client-side:

- HTML/CSS
- Javascript/JQuery
- Ajax
- Chart.JS
- DataTables

## Programma ontwerp

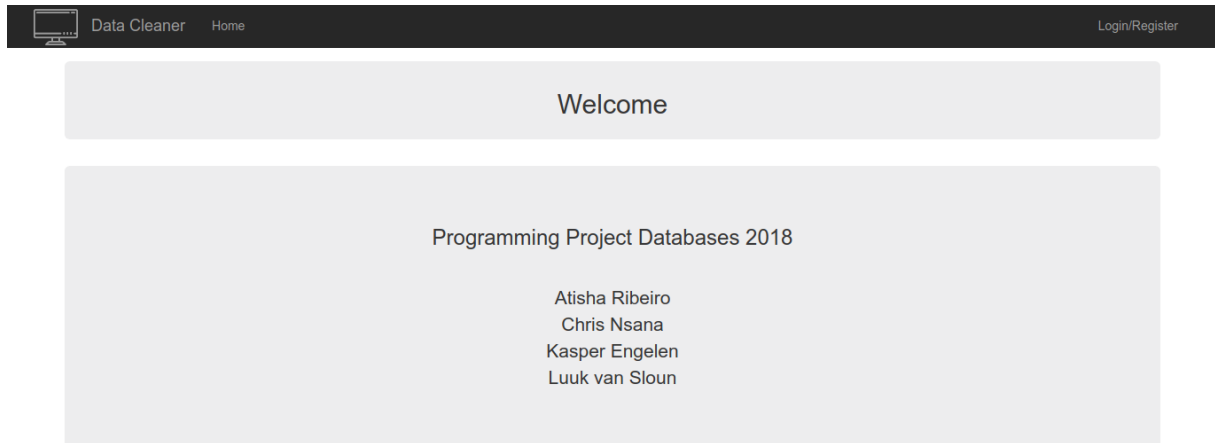
Voor het ontwerp van de code gebruiken we het MVC patroon, alle code is dus opgedeeld in directories die Model, View en Controller heten. De rootdirectory is er een file, client.py, die het programma start.



### 3. Manual

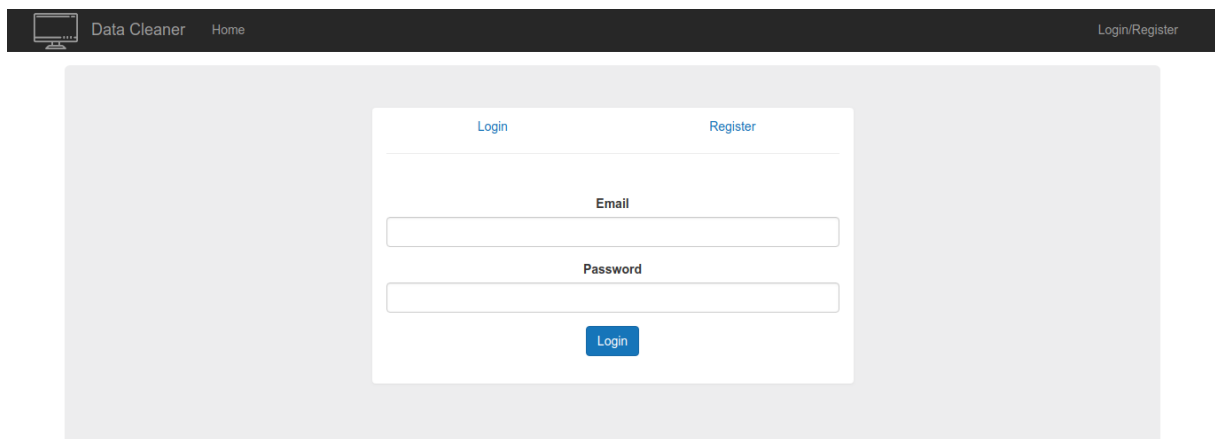
- Standaard home page

Op de standaard home page kan men enkel inloggen en registreren.



- Login Pagina

Hier kunnen gebruikers inloggen. Wanneer men een foutief wachtwoord invoert of een emailadres dat niet bestaat, zal een flash message ze hiervan op de hoogte brengen.



- Registratie Pagina

Nieuwe gebruikers kunnen zich hier registreren. Ieder invoerveld heeft de restrictie dat het verplicht ingevuld moet worden. Daarnaast wordt er gecontroleerd of het wachtwoord minimaal 6 en maximaal 50 tekens bevat en of de ingevoerde wachtwoorden overeenkomen.

The screenshot shows the 'Register' page of the 'Data Cleaner' application. The top navigation bar includes a logo, 'Data Cleaner', 'Home', and a 'Login/Register' link. The main content area has a central white box with a 'Register' form. The form includes fields for 'Firstname', 'Lastname', 'Email', 'Password', and 'Confirm Password', each with a blue 'Register' button below them. The background is a light gray.

- Home Pagina wanneer ingelogd

Wanneer een gebruiker ingelogd is, zal de home pagina veranderen ten op zichte van de standaard home pagina. De View Datasets knop zal de gebruiker verplaatsen naar zijn overzicht van beschikbare datasets.

The screenshot shows the 'Home' page of the 'Data Cleaner' application when a user is logged in. The top navigation bar includes a logo, 'Data Cleaner', 'Home', 'Datasets', 'Profile', and a 'Logout' link. The main content area has a light gray background with a 'Getting Started' section and an 'Existing Datasets' section. The 'Existing Datasets' section contains a blue 'View Datasets' button.

- Dataset List Pagina

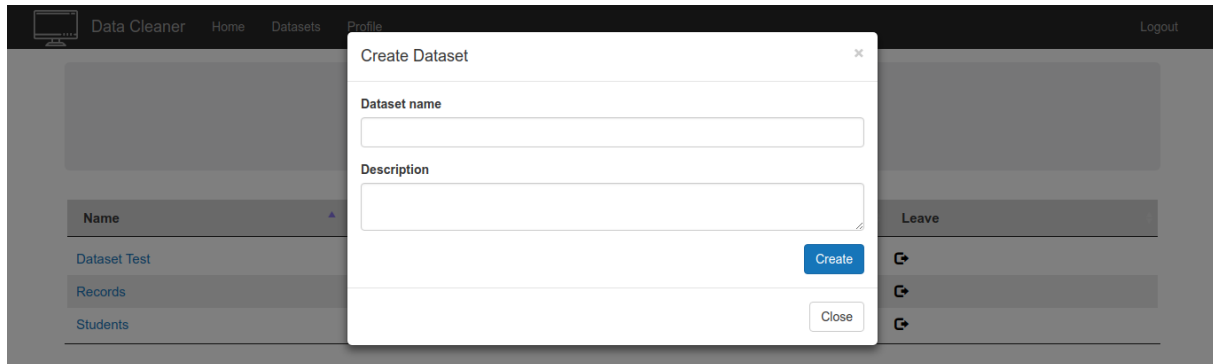
Op deze pagina kan de gebruiker zijn/haar reeds aangemaakte datasets bekijken. Ook staan hier de datasets tussen waar ze rechten toe hebben. De knop Create Dataset zal een popup openen waarin de gebruiker snel een nieuwe dataset zal kunnen aanmaken.

The screenshot shows the 'Datasets' page of the 'Data Cleaner' application. The top navigation bar includes a logo, 'Data Cleaner', 'Home', 'Datasets', 'Profile', and a 'Logout' link. The main content area has a light gray background with a 'Datasets' section containing a blue 'Create Dataset' button. Below this is a table listing datasets.

Name	Description	Permission	Leave
<a href="#">Dataset Test</a>	Description Test	admin	
<a href="#">Records</a>	Sales records	admin	
<a href="#">Students</a>	Student database of UA	admin	

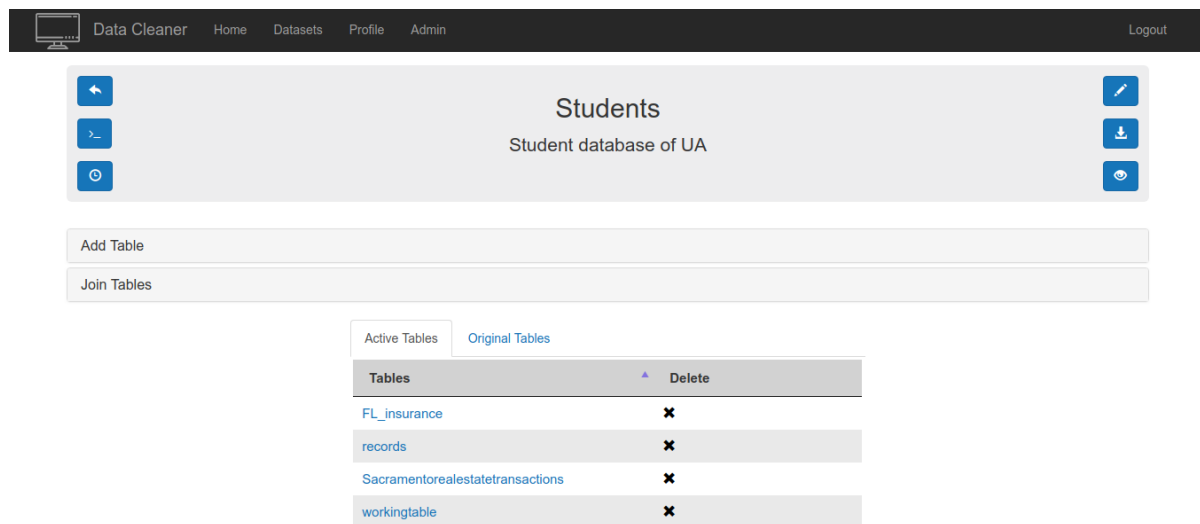
- Create Dataset Pagina

Hier kunnen gebruikers een nieuwe dataset aanmaken. Net als bij het registratieformulier dienen beide velden ingevuld te worden.



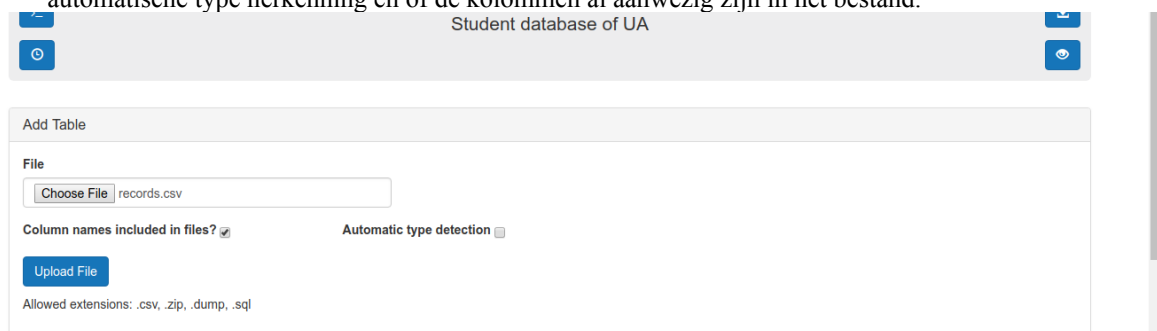
- Dataset Home Pagina

Op deze pagina staan de tables die in de dataset zitten. Hier kunnen tevens tables toegevoegd of gejoined worden. De tab 'Original Tables' zal alle ruwe data files weergeven. Deze kan de gebruiken bekijken en indien gewenst ook downloaden. In de rechtbovenhoek vinden we de knoppen voor het editen van de metadata van de dataset als ook de permissions. Tevens kan de dataset gedownload worden in de vorm van csv-bestanden of een dump-file. Linksonder in de titlebar vinden we de knop die de gebruiker naar de historiekpagina brengt en daarboven de knop voor het uitvoeren van custom queries.



- Add-Table Panel

In dit paneel kan met een bestand kiezen om te uploaden, zij het een csv-, sql-, dump- of zip-file van de drie types. Bij het uploaden van een csv-bestand zullen twee opties beschikbaar worden, namelijk automatische type herkenning en of de kolommen al aanwezig zijn in het bestand.



- Join-Tables Panel

In dit paneel kan men tables joinen. Hiervoor zijn 2 opties: Normal Join en Natural Join. Voor beide opties zijn 4 types beschikbaar, namelijk Inner-, Left Outer-, Right Outer- en Full Outer Join.

- Edit Dataset

In deze popup kan men de metadata van de dataset wijzigen of de gehele dataset verwijderen.

- Dataset Download

Hier kan men kiezen om de dataset als csv's te downloaden of als dump. Standaardwaarden voor de csv staan hier al ingevuld, maar de gebruiker kan daar van maken wat hij/zij wenst.



- Dataset History

Op deze pagina kan met de historiek bekijken van de volledige dataset en van alle tables los die in de dataset zitten. Men kan hier tussen wisselen, door een table of 'Entire Dataset' te selecteren in het menu rechtsboven. De undo-knop zal beschikbaar zijn indien een undo ook echt mogelijk is. Undo's kunnen enkel gedaan worden op de historiekpagina voor een specifieke table.

Date	Description
2018-04-24 12:06:35.529431	Performed One-hot-encoding using attribute "Region" of table "records".
2018-04-24 12:42:42.965337	Converted attribute "Order_ID" of table "records" to type 'FLOAT'.
2018-04-24 12:43:10.750128	Converted attribute "Order_Priority" of table "records" to type 'CHAR(1)'.
2018-04-24 22:51:40.181094	Deleted attribute "Asia" of table "records"
2018-04-24 22:52:44.002706	Converted attribute "Unit_Price" of table "records" to type 'FLOAT'.
2018-04-25 05:53:56.470147	Deleted rows from table "workingtable" using the following predicate: "last_name" = 'Nsana' And "age" = '20' Or "age" < '20'.
2018-04-25 06:11:17.805411	Deleted attribute "Country" of table "records"
2018-04-25 06:11:31.547667	Deleted attribute "Middle East and North Africa" of table "records"

- Dataset View/Transform Table Pagina

Hier kan men de daadwerkelijke tabel bekijken en bewerken (indien de gebruiker daar de rechten toe bezit). In de rechterbovenhoek vindt men wedermaal drie knoppen, net als op de Dataset Home page: één voor download functionaliteiten (deze zijn gelijkaardig aan die van de datasets), één voor het kopiëren van een table en één voor het weergeven van de grafieken en statistieken. De knoppen links zijn voor het teruggaan naar de de home page van de dataset, het uitvoeren van custom queries en het toeglen van het transformatie menu.

Region	Country	Item_Type	Sales_Channel	Order_Priority	Order_Date	Order_ID	Ship_Date
Asia	Singapore	Meat	Offline	C	6/13/2010	133662259	6/19/2010
Asia	Brunei	Household	Online	M	11/3/2016	694549930	11/17/2016
Asia	Mongolia	Office Supplies	Online	M	3/27/2015	919973504	4/5/2015
Asia	Malaysia	Clothes	Offline	M	11/24/2014	269806039	1/10/2015
Asia	Indonesia	Meat	Online	H	8/20/2010	472974574	8/27/2010
Asia	South Korea	Cereal	Offline	H	1/23/2015	733410203	2/24/2015

- View Popup

In deze kan de gebruiker van de statistieken van de kolommen in de tabel bekijken. Denk daarbij aan numerieke statistieken, zoals het gemiddelde. Daarnaast is het frequentie-diagram beschikbaar, om aan te geven hoe vaak een entry voor komt. Een numeriek histogram zal gegenereerd worden, indien de kolom ook van een numeriek type is.




- Transformations Menu


In dit menu kan men transformaties kiezen om uit te voeren op de table. Iedere transformatie zal een popup openen waarin de gebruiker snel de waardes kan invoeren. Tevens krijgt de gebruiker de mogelijkheid om de table te laten kopiëren en de transformatie op de kopie te laten uitvoeren.

Data Cleaner								
Attribute Type Conversion	Baby Food	Offline	M	12/8/2015	975060927	1051	255.28	159.42
Custom Query/Queries	Baby Food	Online	H	5/31/2015	176939809	9276	255.28	159.42
Data Deduplication	Baby Food	Online	M	3/5/2012	362985551	4346	255.28	159.42
Delete Rows using Predicate	Baby Food	Offline	M	6/10/2010	161330640	176	255.28	159.42
Delete Outliers	Baby Food	Online	C	7/21/2014	804983985	1505	255.28	159.42
Discretize +	Baby Food	Online	H	11/20/2010	599480426	8435	255.28	159.42
Extract Part of Date	Baby Food	Offline	H	1/16/2013	355735862	9002	255.28	159.42
Fill Nulls +	Baby Food	Offline	M	5/7/2017	581689441	3687	255.28	159.42
Find and Replace +	Baby Food	Online	H	5/13/2015	908131444	5130	255.28	159.42
Normalize By Z-Score	Baby Food	Online	M	1/30/2016	446566617	7841	255.28	159.42
One Hot Encoding	Baby Food	Online	C	8/23/2011	320065915	3472	255.28	159.42
	Baby Food	Online	L	7/18/2012	889963662	8879	255.28	159.42
	Baby Food	Online	L	11/8/2014	333288991	6474	255.28	159.42
	Baby Food	Online	H	10/16/2011	217504251	3196	255.28	159.42
	Baby Food	Online	L	8/20/2012	334428393	3648	255.28	159.42

- Dataset Permissions Pagina

Op deze pagina kan een admin van een dataset gebruikers toevoegen met bepaalde rechten tot de dataset. Aan de hand van het emailadres van een gebruiker krijgt deze rechten tot de dataset. De mogelijke rechten zijn 'admin', 'write' en 'read'. Een admin kan hier tevens gebruikers die toegang hebben tot de dataset hun rechten ontnemen door ze te verwijderen.


[Data Cleaner](#)
[Home](#)
[Datasets](#)
[Profile](#)
[Admin](#)
Logout



## Permissions

Email

Permission Type


Admin

Add User



ID	Email	Permission	Remove
1	luukvansloun@live.nl	admin	✕
11	henk@henk.com	write	✕

- Admin Manage Users

De admin van de website moet ten alle tijden de huidige, geregistreerde gebruikers kunnen wijzigen/verwijderen. Denk aan iemand die zijn wachtwoord vergeten is of een gebruiker dient verwijderd te worden. Daarnaast kan de admin een gebruiker ook activeren/deactiveren. Dit zal hun account blokkeren, maar niet verwijderen. Enkel de toegang wordt tijdelijk geweigerd.



[Data Cleaner](#)
[Home](#)
[Datasets](#)
[Profile](#)
[Admin](#)
Logout

## Manage Users









Name	Email	Admin	Status	Activate/Deactivate	Edit	Delete
henk henk	henk@henk.com	No	Active	■		✕
Luuk van Sloun	luukvansloun@live.nl	Yes	Active	■		✕

- Admin Manage Datasets

De admin van de website moet ten alle tijden alle datasets kunnen aanpassen. Net zoals bij het beheren van de gebruikers kunnen datasets verwijderd worden en kan hun metadata gewijzigd worden.



[Data Cleaner](#)
[Home](#)
[Datasets](#)
[Profile](#)
[Admin](#)
Logout


## Manage Datasets

ID	Name	Description	Edit	Delete
5	Students	Student database of UA		✕
6	Records	Sales records		✕
8	Records	Sales Records		✕
9	Chart Screenshot	cs		✕
11	Sacramento	Financial, Real Estate and Crime Rate		✕
12	United States of America	US Population Statistics		✕
15	Test Dataset	test		✕
22	Records	test records table		✕

- Admin Creation

Om een nieuwe admin aan te maken of een admin zijn rechten te ontnemen op de website, moet men de ‘verborgen’ pagina bereiken. Wanneer men ingelogd is en zich op de homepage bevindt, kan de Admin Promote/Demote pagina bereikt worden door driemaal te klikken op de tekst ‘Getting Started’. (Dit werkt niet op mobiele apparaten, aangezien dit de zoom-functie togglet.) Door middel van een pincode kan een gebruiker zijn rechten ontvangen/ontnomen worden.

 Data Cleaner   Home   Datasets   Profile   Admin   Logout

 Promote/Demote User

Email

Action

PROMOTE ▾

Pincode

Promote/Demote

## 4. Status

### Load/View functionaliteiten:

- **[Klaar]** Een downloadfunctie voor dump-files voor tables zonder het interne systeem bloot te geven.
- **[Klaar]** Een downloadfunctie voor van een dump-file voor een gehele dataset.
- **[Klaar]** Een downloadfunctie voor een zip van csv's van een gehele dataset.
- **[Klaar]** Het verschaffen van een beter overzicht van view-functionaliteiten (Chart.JS)

### Transform functionaliteiten:

- **[Klaar]** Meerdere keuzes voor joins
- **[Klaar]** Coerce een attribuut conversie (Slechte rows verwijderen/op null zetten)
- **[Klaar]** Dynamisch tables reverten aan de hand van Dataset History
- **[Klaar]** Data deduplicatie
- **[Klaar]** Mini SQL-parser voor user-generated queries.

### Frontend functionaliteiten:

- **[Klaar]** Een cleaner design aanbieden voor de gebruiker:
  - Compacter door middel van modals/popups en meer functionaliteiten binnen dezelfde pagina.
- **[Klaar]** Loading Screen voor tijdrovende operaties