Overview:

* This is a list/file matching website in python
* the matching engine will be Senzing, we will mount a docker into the same LAN in GCP in Kubernetes
* DB will be on VM for now, we will migrate to cloud managed DB in the future
* we will build a member area with signup page
* Uploaded data sources should be stored in a bucket
* I will build a wordpress site for marketing on a separate machine
* Member site
  + You can sign in (with 2FA) or sign up
  + User can create a project
  + upload files (csv, txt, xlsx, xls) to the project. They are called Data Sources.
  + Run projects which will send data sources to API for comparison. I believe we will upload data to a table
  + Payments
    - Membership will be based on 3 months or 12 months
    - Cost will be based on Max records uploaded to a project
      * Customers can have multiple projects but only to the Max # of records purchased
  + Customer will pay the project cost and then they can download their results
    - all data will be deleted after 60 days, we should connect a Google Bucket for data storage (uploaded data and results)
* Uploaded data
  + we will need to read the 1st row and ensure there is a column header and we will need to do a fuzzy mapping of the column header to the senzing mapping convention
  + We will use python pandas and perhaps other libraries to give a quick profile of the data.uploaded and run a quick check on integrity

Resources:

* <https://senzing.zendesk.com/hc/en-us/articles/231925448-Generic-Entity-Specification-Data-Mapping>
* sample web app: <https://github.com/Senzing/entity-search-web-app>
* sample stats: <https://senzing.zendesk.com/hc/en-us/articles/360052040553-Exploratory-Data-Analysis-Overview-EDA-Tools->
* Docker setup: <https://senzing.zendesk.com/hc/en-us/articles/12938524464403-Quickstart-For-Docker>
* API docs: https://docs.senzing.com/

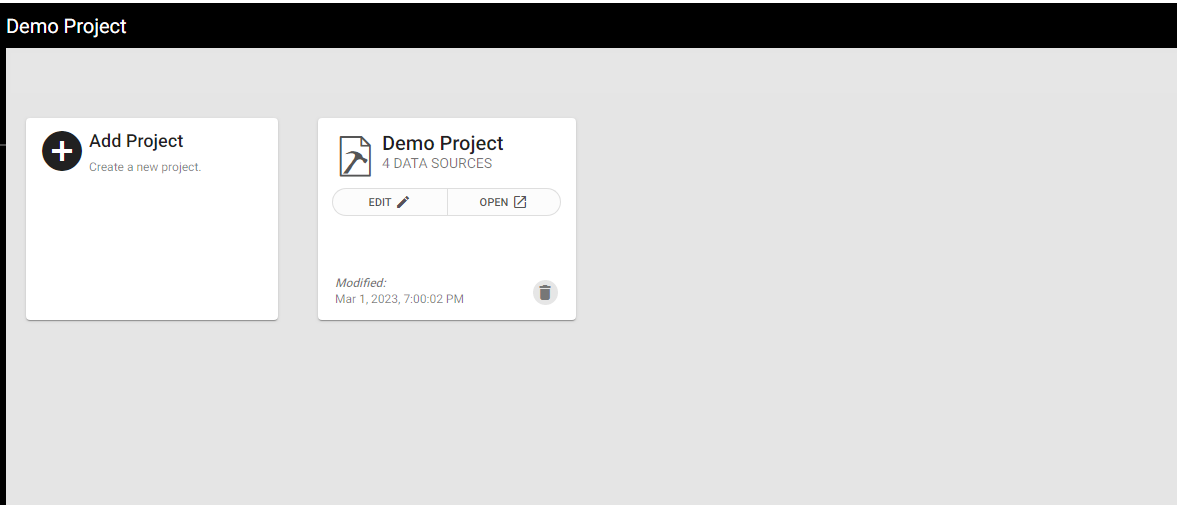
Member site pages needed:

* Database
  + PostgreSQL 14
  + IP:10.59.96.2
  + User: member-master
  + pass: abs<Za\_%FIXg435i
* Dashboard showing the projects
  + open existing project
  + create new project
* Project page
  + show data sources
  + review data profile
  + review/download previous results
* Data Profile page
  + profile is attached to data source, not project
  + I can show you a demo of this via screen share
* Matching screen
  + Map sources
  + Match Configuration
* login/register
  + 1 field for email}
    - if email exists ask for password
    - if not, open register form
  + 2FA always to email
  + reset password
* Billing page
  + stored payment methods
  + billing history
  + activity history
  + make a payment
  + auto top-up
* Settings
  + Notifications
  + account profile
  + invoicing options

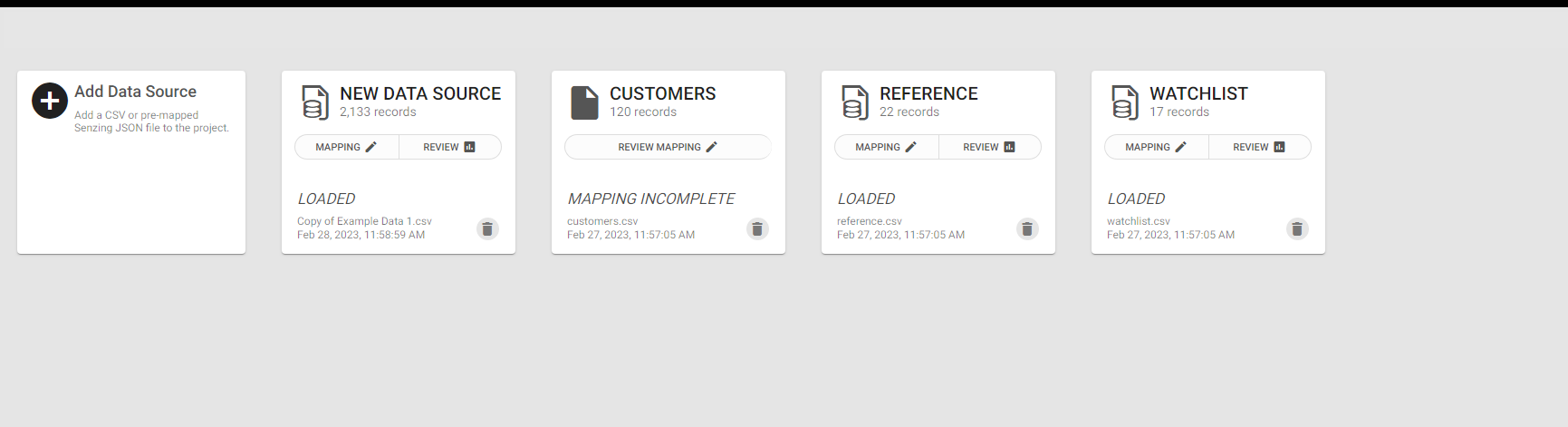
Notes:

* User friendliness is #1 priority
* <https://github.com/pytoolz/cytoolz/>

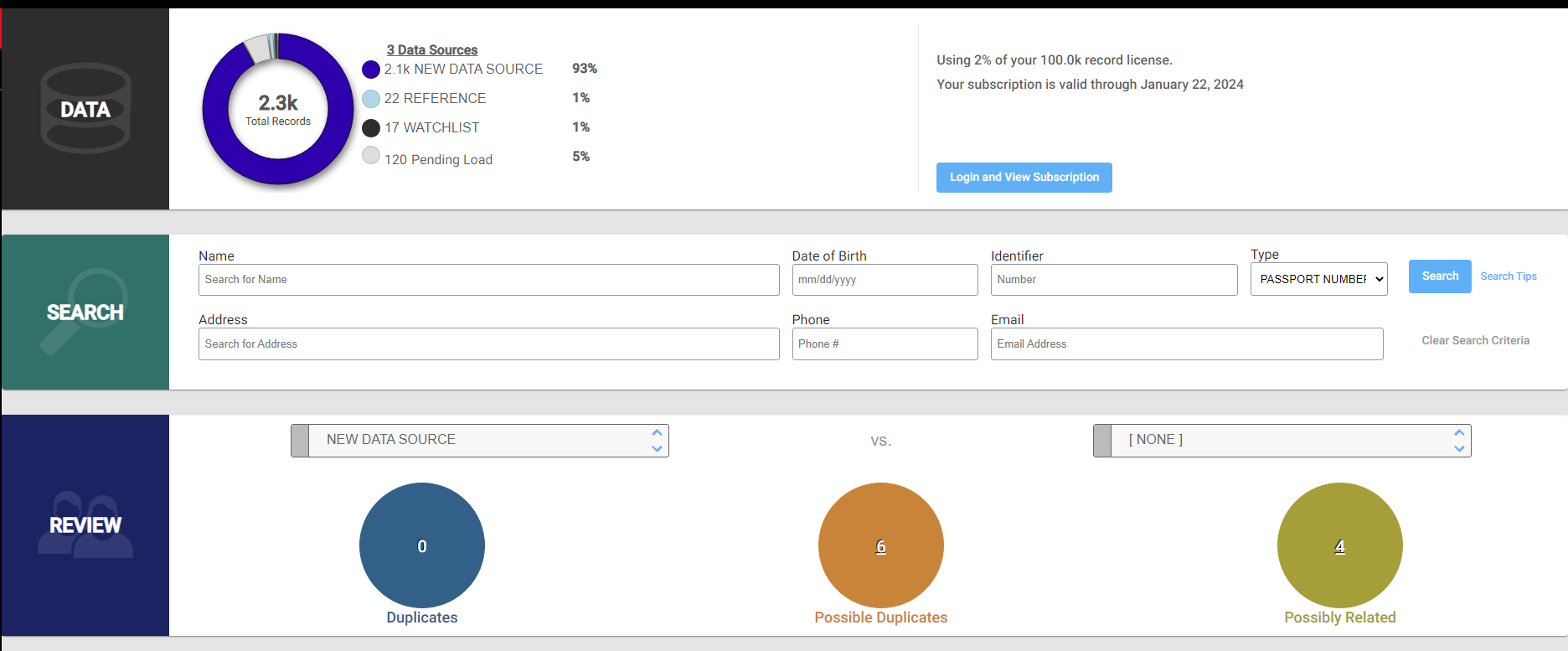
Sample dashboard



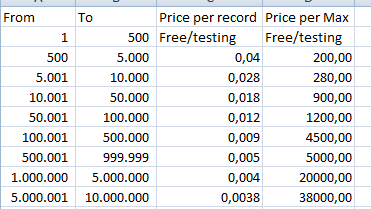
Sample existing project page



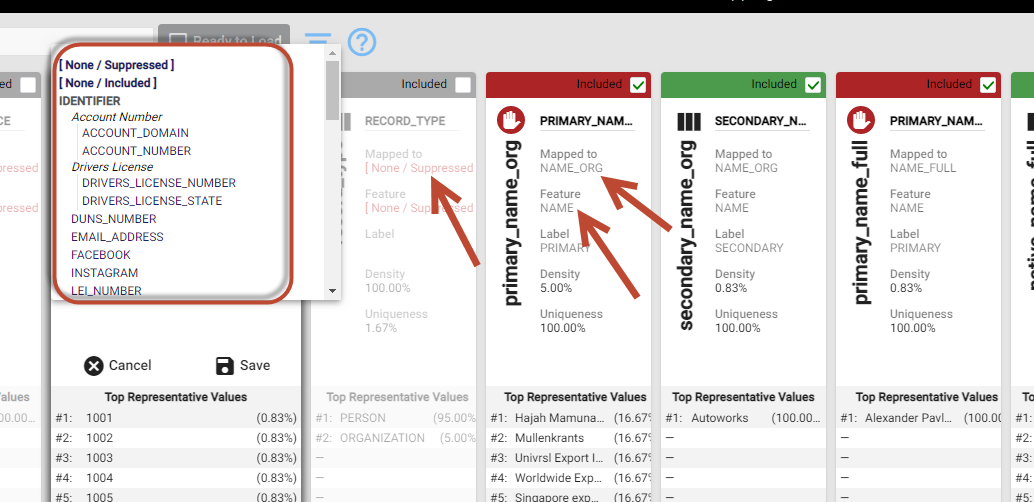
Example results page



Sample pricing



Sample mapping of a data source



**Features/Changelist/Menu Tree**

* Theme
  + Default should be the blue color
  + ”Dark” theme does not conserve the logo
  + Integrate thinking GIF
* Registration/login
  + Make the initial registration fields only **First Name, Last Name, Phone, email, password and password confirmation.** Once the user signs up, we will prompt to finish the user profile.
    - The idea is to require only critical info so user can start to test with the tool.
  + Phone should have a country code box before that is a dropdown, searchable by typing and localized to preselect the country based on IP.
  + The password confirmation should be after the password
  + OTP is not working. I uninstalled sendmail since we do not use it. You should be sending to localhost and the from email should be <MDP Authentication> “[2fa-mdp@matchdatapro.com](mailto:2fa-mdp@matchdatapro.com)”
  + OTP should expire if not used after 15 min
  + Allow user to define/restrict account access from countries and/or static IP address (default disabled)
* Menu 1 General Menu - this will show user general options available to the user
  + User Management
    - We will manage Organizations
    - Users can be a member of an organization
    - There will be one or more Admin users
    - Users will have permissions set,
      * Admin can see all organization projects, users can only see own projects
      * Read vs Write (Read just for viewers)
      * Admin can migrate projects from 1 user to another user
    - Admin can allocate paid resources to users
  + Projects
    - Fields for Project Creation
      * Project Title
      * Project Description
      * Option to Delete data after project completion and data exportation (Default Enabled)
        + If disabled we delete data when LastModifyDate<Today-n
      * Project Creation data (not entered by user)
      * Project Last modify date by user (ongoing)
      * Project last run date (ongoing based on runs)
      * Next Run Date/time (based on Automated projects only)
      * Project type is Manual, Automated or API
        + Manual will not do anything automatically, user an work with the project manually only
        + Automated is a cron job, we need to allow the user to select From Date, To Date (unlimited by default) and hour of day for runtime
        + API - Hide this for now. This will create a project and allow the user to make API requests and reference the project in order to execute the projects rules
      * Ability to share or not share between users, read and/or write
      * The button called Add, I would like to change this to Open, so when the user opens the project, he will see the data sources that have been uploaded and can add additional data sources.
      * Add a Delete button that will delete the data and all info regarding the project. We need double authorization, so add a step like Enter the word DELETE here in all caps to confirm you want to delete the project.
* Menu 2 Project Menu - This menu will show the purchased modules for use in projects
  + Data Import
    - Display uploaded data sources
      * refresh, delete, change data sources name, update custom query (for DBs)
    - Add new Data Source
      * Allow to hide column data from user view, but data exists in DB
  + Data Profiling
    - Purpose: Scan data sources and compile statistics based on the data
    - How we will price/charge for this service
      * Users will purchase a tier level and based on that tier level we will allow or reject the profiling
        + Example, if a user purchase tier 100K records, he cannot profile a data source with 1 million records, but he can profile a data source with 50K records…..up to 100K
    - Create New Profile
      * Profile Name
      * Add imported data source
      * add email addresses (multiple) for finished profile email (will email the report)
      * Description
    - There will be 5 tabs based on the dimensions of Data Quality. We will calculate a score that will be displayed on the Tab (Example: Accuracy - 82%), also highlighted in yellow and red when scores are low. In each tab, highlight sections where data quality issues exist. Also, an overall Data Quality score should be calculated
    - Tab 1 Accuracy
      * Total Rows
      * Pattern detection
        + This will apply a list of RegEx to the columns and report the qty of data that is valid (pattern detected) or not valid (not detected)
      * MAX LEN for a column
      * Character detection
        + Contains Numbers, numbers only, Contains Letters, Letters only, Contains letters and numbers
        + Punctuation

I would like to report record counts for each punctuation character where the count is bigger than 0

* + - * + Leading spaces
        + trailing spaces
        + non-printable characters
    - Tab 2 Completeness
      * Total Rows
      * Blank/null Fields
      * Outlier detection
        + Example: a column contains 2 distinct values over 1000 records, however, 1 additional value exists for a few records. These are outliers
    - Tab 3 Uniqueness
      * Distinct (how many distinct values are in a column)
      * Geo Map (this is low priority), if we can map ZIP/Postal codes to USA and Canada to show visualization of map with data points
    - Tab 4 Conformity
      * Data Type (String, integer, data, double, etc)
      * Date Format
    - Tab 5 Precision
      * Numerical statistics if a column is numeric (Min, Max, Mean, Median, Mode and Extreme)
    - Data preview is important, when a customer selects a metric, the data should filter and show only records affected by case
      * For Distinct records option, we should show pivot table with the distinct value, count of repeat and % of total data source this count occupies
      * user should be able to export data using the native data source connectors
    - Histogram, allow user to click to filter based on histogram bars
    - Need a nice report in PDF or Excel
  + Data Cleansing add rules to a list of cleansing options to include (recipe ingredients). Each rule will have an input columns, cleaning options and output column or replace existing values
    - Removal
      * Punctuation, spaces, numbers, letters, leading spacs, trailing spaces, non-printable chars, free field for user input
    - Replace
      * string with another string
        + Case sensitive yes/no, default yes
    - Copy Column
    - Case options
      * Upper, Lower, Capitalize
    - Count occurrences of char in COLUMN1
    - Fill Blanks with static test
    - Statement Generator this will allow the user to create 1 or more new columns based with data based on the following statement options .
      * Conditional If <COLUMN1> <OPERATOR (mathematical and string> COLUMN 2 (AND/OR/NOT…..) then output (Output can be static or column(s) value>
      * SUM(COLUMN1,COLUMN2,COLUMN(n))
        + Average, Difference, Divided, Multiply
      * Range(FromDateONumber,ToDateOrNumber,Output If In Range, Output if our of Range)
        + Output can be static or column(s) value
      * Merge(delimiter,1st not empty (boolean default=0),COLUMN1,COLUMN1,COLUMN(n)
      * Split(COLUMN1, delimiter)
      * Left/Right(COLUMN1, n digits)
      * Search(COLUMN1,char) return new column with True/False
      * Dynamic string (“Text”, (COLUMN1,(COLUMN2, (COLUMNn)
        + Text will include {0},{1},{n}
    - Patterns- apply a regular expression and allow user to extract detected patterns into 1 or more columns
      * We will supply a global or private library of regular expressions for the user to apply
      * user can copy and paste his own RegEx
    - Dictionaries - allow users to upload dictionaries in order to normalize/standardize values
      * Allow to replace existing value, extract to a new column (define column header), delete original value or create a flag column based on a value
      * tokenize the selected column and allow user to define delimiters for tokenizing
      * Allow users to upload created dictionaries for use
  + Data Matching
  + Group Overwriting/Enrichment
  + Project Auditing
    - allow customer to add new columns to track changes
    - allow customer to create project audit reports
  + Master Record Assignment/Survivorship
  + Export Options
    - All Records, All uniques, dedupe (unique + master), matches, matches between only, only duplicates
    - Export to via native connector (all except for files)
    - or public URL for files (n days before expire and delete)
    - or ftp
    - ZIP and email (size limitations
    - Export to drive, onedrive or dropbox

Image for Profiler ideas: <https://drive.google.com/file/d/13glyLurL7Vuj7ZKkmKoIcCJmJYwAf5n5/view?usp=sharing>