DataMaster:
Vision and Scope

Data and Datasets (DnD)
Computer Science Department
California Polytechnic State University
San Luis Obispo, CA USA

December 10, 2018

Contents

Credits

Name	**Date**	**Role**	**Version**
Griffin Aswegan	October 3, 2018	Author	1.0
Steven Bradley	October 3, 2018	Author	1.0
Christina Daley	October 3, 2018	Author	2.0
Larry Hu	October 3, 2018	Author	1.0
Shane Villalpando	October 3, 2018	Author	1.0
Dustyn Zierman-Felix	October 3, 2018	Author	1.0

Revision History

Name	**Date**	**Reason for Changes**	**Version**
Christina Daley	December 2, 2018	Final review for prototype relese	2.0
Team DnD	October 8, 2018	Initial baseline approved by Customer	1.0

1 Business Requirements

1.1 Background

Data has become one of the biggest markets in the last decade or so. Companies like Facebook, Amazon, and Google use data to an extreme extent. However, while data has become a massive market, classifying groups of data sets and organizing data sets based on the data within has been almost nonexistent. Being able to tag, organize, and search through different types of data sets would be extremely beneficial to many companies around the world in multiple different potential ways.

1.2 Business Opportunity

There are very few systems present in modern technology that allow you to take multiple data sets and organize them based on the information they contain. There's a unique opportunity to allow data sets to be organized to allow for further analysis, as well as making connections that may have been previously unknown to companies.

1.3 Business Objectives and Success Criteria

This product will be considered successful if the following occur. Users are able to add data sets and classify the data so that it is organized into categories. Users can edit and add to these categories. The user is able to search through the data sets to find specific categories.

1.4 Customer or Market Needs

This product is simple to use for data analysts and database managers.
This product supplies a way to allow multiple data sets to be integrated into one. This allows easier interaction with customizing the data to get desired results.

1.5 Business Risks

Business risks involve potential losses of income. These include the following. Competing companies could take away business and/or supply a better solution. Insufficient advertising or misuse of this product could also be a problem.

2 User Description

2.1 User/Market Demographics

Some of the key users we expect to use this program heavily are data analyst and database managers. Others users include people who are required to search through and/or interact with large data sets often. These could include firms for organizing their clients, producers sorting through a customer databases base and in general data bases that involve people as one item.

2.2 User Personas

There is a Data analyst who is asked to draw conclusions and categorize a local firm on its clients for property management. The clients have categories based on worth, location, list of property, priority, along with personal information. So the data Analyst would add the information into the database catalog and from there use the systems machine learning to connect and sort through the data.

2.3 User Environment

Our user Environment or user interface would consist of a web app. This should be multi-platformed able to be accessed through chrome, Mozilla Firefox, IE 8 or up. Our web app should be connecting the back-end with SQL and front end with JavaScript react. But for the user it should just be a website link for them to access from computer or mobile.

2.4 Key User Needs

Users should be able to classify and organize different data sets based on the data contained within them.

Users should be able to search for data sets based on the classifications given to those data sets.
Users should be able to let the program do the classification and discovery of new types of data using machine learning.
Users should be able to intervene if the machine learning results in an improper classification occurs.
Users should be able to interact with the program using a web-based GUI.

3 Vision of the Solution

3.1 Vision Statement

DataMaster is a software suite designed to provide users with a simplistic and intuitive system that can automatically identify, classify, and organize data sets together.
DataMaster allows for searching through collections of data sets to find sets that have similar features as well as sets that utilize similar information.
Users of DataMaster will be able to interact with the information gathered about the data sets via a graphical data catalog.

3.2 Solution Overview

DataMaster will consist be a data classifier web application.

3.2.1 Data Classifier

The data classifier will create classification models about the information found in multiple data sets. It will do so by using existing machine learning techniques. The data classifier will be built using Python 3 and Django.

3.2.2 Web Application

The web application will be where users will interact with DataMaster. The web application will provide the users with a graphical interface where they will be able to view and interact with the information gathered by the data classifier. The web application will be built using JavaScript and React.

3.3 Major Features

[**FE-1**] DataMaster will be able to take in data sets and classify those data sets based on the type of data within them.

[**FE-2**] DataMaster will be able to let users search through multiple data sets and return those that have the specified data type(s).

[**FE-3**] DataMaster will use machine learning to learn and classify data sets that it had not seen previously.

3.4 Assumptions and Dependencies

Data sets are assumed to have valid names and correct data.
Data sets are assumed to be relevant to the users.
Data sets are given in a commonly used format(CSV, JSON, XML).

4 Scope and Limitations

4.1 Scope of Initial and Subsequent Releases

Feature	**Release 1**	**Release 2**
FE-1	Implement	Implement
FE-2		Implement
FE-3		Implement

4.2 Limitations

Time will be a big limiting factor in completing DataMaster. DataMaster will require machine learning concepts that the DnD team and other specific knowledge that will require lots of research on top of development.

5 Business Context

5.1 Stakeholder Profiles

DataMaster's has several different stakeholders.

5.1.1 MarkLogic

DataMaster can be used by MarkLogic both internally and externally. MarkLogic will be able to use DataMaster to serve the different data organization needs of their clients.

5.1.2 MarkLogic Clients

Depending on the competence of client, MarkLogic's client may be physically interacting with DataMaster.

5.1.3 MarkLogic Competitors

DataMaster will impact the field of data analysis businesses. Companies like MongoDB and Amazon Web Service's Glue.

5.2 Project Priorities

5.2.1 Release 1

Release 1 will focus on creating a functional web application that classifies different inputted data with categories and can output a simple overview display.

5.2.2 Release 2

Release 2 will focus on improving the complexity of the data classification and overall user interface. This release will spend time on creating a specific search for user's to result in different displays of the specific data they want.

5.3 Operating Environment

DataMaster will operate as a web application. It will be available online and be open source.

6 Competitive Analysis

6.1 Overview

Data analysis is a very competitive business and a very large field. As a result, the competitive environment of data analysis is very fierce. The most prominent and current competitors to this project are Amazon Web Service's Glue, Informatica, and Alooma

6.2 AWS's Glue

AWS Glue is a fully managed extract, transform, and load (ETL) service that makes it easy for customers to prepare and load their data for analytics. AWS Glue lets you collect your data sets together, manage transformations on those data sets, and then do "batch" processing on those data sets.

6.3 Informatica

Informatica is a cloud data management system. Working with companies such as JLL, Nissan, and Kelly Services, Informatica strives to provide analytics, foresight, and insight into the large, ever-expanding world that is Data management.

6.4 Alooma

Alooma is a data management company that focuses on providing data visibility, security, transparency, and versatility. Alooma makes it a proiroty to ensure that customers are aware of what they're data is doing and where it is at every step of the management process.