Data Classifier: Vision and Scope

Team Mark
Computer Science Department
California Polytechnic State University
San Luis Obispo, CA USA

October 8, 2018

Contents

Credits

Name	Date	Role	Version
Spencer Schurk	October 6, 2018	Lead Author of Business Requirements	1.0
Geraldo Macias	October 6, 2018	Lead Author of Scope and Limitations	1.1
Matt Yarmolich	October 7, 2018	Competitive Analysis	1.1
Brad Foster	October 7, 2018	Vision of the Solution	1.1
Landon Gerrits	October 7, 2018	Lead Author of Business Context	1.0
Jake Veazey	Octover 7, 2018	User Description	1.1

Revision History

Name	Date	Reason for Changes	Version
Spencer Schurk	October 6, 2018	Initial version of Section 1	1.0
Geraldo Macias	October 6, 2018	Initial version of Section 4	1.1
Matt Yarmolich	October 7, 2018	Initial version of Section 6.0	6.1
Brad Foster	October 7, 2018	Initial version of Vision of the Solution	1.1
Landon Gerrits	October 7, 2018	Intial version of Section 5	1.0
Jake Veazey	Octover 7, 2018	Initial Version of User Description	1.1

1 Business Requirements

1.1 Background

Ever-growing data sets are causing an unnecessary burden on data scientists and data analysts. As data sets are expanding, and the number of sources where these sets come from is increasing, it is becoming harder for professionals working on this data to find the information they need. Much of this work currently is done manually, and makes it very difficult for someone new coming into a data position at an existing company with a large data sets to start working productively.

1.2 Business Opportunity

Developing a new data classifier tool will remove many of the manual hardships involved with analyzing large, and often unorganized data sets. Instead of spending hours searching for the proper data classification a data analyst might be looking for, the new data classifier tool will allow for automated classification. This tool would allow recently hired data analysts who don't have experience with a company's complex data sets to do analysis quicker. Developing this data classifier as a web-based interface allows data scientists and data analysts to spend less time looking for the data, and more time providing useful insights and metrics to the company.

1.3 Business Objectives and Success Criteria

BO-1	Develop a machine-learning powered data classifier.
BO-2	Develop a web-based GUI that interfaces with the data classifier and allows edits to classification.
BO-3	Data classifications can be exported into a Data Catalog.

SC-1	Increase productivity of data scientists and data analysts.
SC-2	Classifier scales over various data sets and successfully categorizes data.
SC-3	Open-source classifier sees adoption and adaptation by outside companies.

1.4 Customer or Market Needs

CN-1	Interface should be viewable from a modern web-browser.
CN-2	Machine-Learning techniques should be used to classify incoming data.
CN-3	Users should be able to edit classifications before they're stored and sent to Data Catalog.

1.5 Business Risks

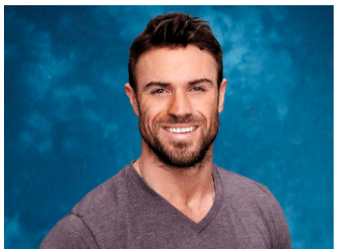
No known business risks at present.

2 User Description

2.1 User/Market Demographics

Market Demographics for the Data Classification tool are wide as we are not targeting a niche group of people, but instead creating an open-source project for the anyone to use. However, the typical market demographic that we will be targeting will be between the age of 20 and 50 with a college degree in data science or a similar technical related field. In addition, the Data Classifier will be used majorly by data analysts who plan on taking raw data and classifying it by wanted aspects.

Chad Chaderson



" I like the work that I do, I just wish i could do it faster!"

Age: 28
Work: Data Analyst
Family: Single
Location: Palo Alto, CA
Character: The Planner

Personality

Introvert	Extrovert
Thinking	Feeling
Sensing	Intuition
Judging	Perceiving

Charismatic | Fast-Paced | Work Oriented

Goals

- Take user data and relay information from it
- Create a business that streamlines data analytics from research

Frustrations

- Taking data from a user and then working on it can take a lot of time
- Vast amounts of user data that is not yet sorted

Bio

Chad is a software engineer who focused on data science and data analytics. He currently lives alone in the city and commutes to work every day at 9:00 AM. Chad works at XYZ Technologies where he specialists in taking customer data and then relaying pertinent information form them.

Motivation

Incentive
Fear
Growth
Power
Social

Preferred Channels

Traditional Ads
Online & Social Media
Referral
Guerrilla Efforts & PR

2.3 User Environment

The User Environment is more selective and users will be using the Data Classifier while working to classify raw data for work or personal use. This service will be used by data analysts at work whenever new raw data is presented.

2.4 Key User Needs

Users will use this Data Classification tool to take raw data, in multiple file formats, and classify each piece so that it can be used for other tasks and in other tools afterwards. Users will need to be able to upload files and then, after processing time, be able to download a classified file from our software.

3 Vision of the Solution

3.1 Vision Statement

The solution involves a web-based interface that uses a data classifier to groom data that is sourced from various different data "silos." These silos will have data that is in various different non-standard formats.

Data analysts will be able to interact with the web interface to select parameters that will alter the presentation of the data. Through machine learning techniques, the program will make intelligent automated classifications of the data based on a respective industry. The data will also be tagged to preserve its origin.

3.2 Major Features

FE-1	Data analysts will provide data sets from various sources.
FE-2	The Data Classifier will use machine learning to consolidate data categories.
FE-3	The Data Classifier will output in a machine-readable standardized format.
FE-4	Data analysts will be able to use the web interface to change the presentation of the data.

3.3 Assumptions and Dependencies

FE-1	Users of the Data Classifier will have an account with MarkLogic.

4 Scope and Limitations

4.1 Scope of Initial and Subsequent Releases

405 initial release targets.

1. Develop machine learning Python program which can detect different types of field types and classify different datasets.

2. Create basic front end which can execute the python program on local datasets.

406 release 1 targets.

1. The system will prompt the user to verify the contents and label of dataset columns. This edit will be applied and improve the machine learning on future datasets.

2. The system will catalog the each dataset according to a type defined by the machine learning algorithm.

3. A data scientist user may search for datasets which include specific fields, or search by types of datasets.

406 release 2 targets.

1. The system will allow account creations with different permission settings.

2. The system will redact sensitive information according to account permission settings. A system administrator will have the highest privilege to information but no modification privileges. A data scientist will have some data redacted but modification privileges. An employee will have minimal data access and no modification privileges.

4.2 Limitations and Exclusions

1. All datasets will be of a .csv file type.

2. Datasets must be stored within customers maintained database.

5 Business Context

5.1 Stakeholder Profiles

Stakeholder	Value	Attitudes	Interests	Constraints
Developers	A college degree and work experience	Concerned	Machine learning and web development	Marklogic's requirements
Data Scientists	Holistic view and insight of data input, categories, and classifications	Intrigued	Data classifications	None
Data Companies	Faster processing time for their data scientists	Excited	Business and rapid deployment	None

5.2 Project Priorities

Dimension	Driver	Constraint	Degree of Freedom
Schedule	We must release our first iteration by the end of quarter 2 and have the final release by the end of quarter 3		
Features		The industry and the type of data set our team chooses to use	We can choose our industry and data sets
Quality		Undefined at this time	Unknown
Staff		Project team is composed of 5 student developers and 3 Marklogic representatives	
Cost		Students will distribute the weight of this assignment equally, spending an equal amount of hours each week (approx. 5-10 hrs)	

5.3 Operating Environment

OE-1	The system will parse a .CSV data set and do basic data classification
OE-2	The system will "learn" data classifications based on data input
OE-3	The system will process multiple data sets with the same classification parameters
OE-4	The system will have a web GUI with the option select input data sets. The system will output classified data sets

6 Competitive Analysis

6.1 Overview

Right now in terms of data classifiers, there are not a ton of viable competitors besides some open source plugins and Impervia, which is a product that specialized in data classification. this product is mature and interfaces with multiple databases such as Oracle, Microsoft SQL, SAP Sybase, IBM DB2 and MySQL.

6.2 Impervia Data Classifier

Impervia appears to be the biggest at market competitor being the first Google search result and seem to have a fair amount of existing marketable customers. Regarding features they are a post database entry classifier meaning you log into your database after logging in and analyze what is currently stored in that database. This differentiates this solution from the solution for MarkLogic as the solution for MarkLogic is meant to serve as a buffer between the input data and the database. Regarding pricing, Impervia seems to offer the tool for free but then charges for additional use and training for the product.

6.3 Scikit-learn

Another competitor for our solution for MarkLogic is Scikit-learn and TensorFlow. These are open source plugin for Python which allows you to train a MachineLearning model for datasets. In terms of marketing/marketability, since these products are open source, it is available for both commercial and private use for free. This means these product are not a direct competitor to our solution for MarkLogic, and can probably be used in our implementation pending on the desired feature set.

6.4 Conclusion

In conclusion there are not a ton of commercially viable solutions for MarkLogic that we will be competing with. Most existing solutions require a team of Software Engineers to train the model and build it into a usable tool.