Data Discovery Classifier: Vision and Scope

Lucky 7
Computer Science Department
California Polytechnic State University
San Luis Obispo, CA USA

October 8, 2018

Contents

Credits

Name	Date	Role	Version
Adam Beymer	October 8, 2018	Lead Author of Business Requirements	1.0
Max Loumena	October 8, 2018	Lead Author of User Description	1.0
Kyle Maxwell	October 8, 2018	Coauthor of Vision of Solution	1.0
Samantha Koski	October 8, 2018	Coauthor of Vision of Solution	1.0
Jacob Territo	October 8, 2018	Lead Author of Scope and Limitations	1.0
Poojitha Karumanchi	October 8, 2018	Lead Author of Business Context	1.0
Katie Mei	October 8, 2018	Lead Author of Competitive Analysis	1.0

Revision History

Name	Date	Reason for Changes	Version

1 Business Requirements

1.1 Background

Our customer, MarkLogic, is a consulting company that helps its clients integrate their databases. MarkLogic's customers usually have databases across different physical sites and in many various formats. This information is known to be siloed. MarkLogic will take the siloed data and integrate it into a single noSQL database that has strong organization and querying for max performance.

1.2 Business Opportunity

MarkLogic sees us as a great business opportunity because we will create 5 different POC's to the proposed problem and will act as a free think-tank for them in expanding their business.

1.3 Business Objectives and Success Criteria

MarkLogic wants a tool for the easy visualization of data. Often times when MarkLogic is integrating databases, the different datasets come with different column headers that may actually contain the same data (Example 'Gender' versus 'Sex'). MarkLogic wants to be able to easily combine these different categories into one. Our project would be a success if we can create a system to accurately combine separate data categories. Another category of success for ourselves should be how much we can learn from this opportunity to work with a real company.

1.4 Customer or Market Needs

The main customers of this data visualization tool are data scientists. Data scientists need to look at data that comes from many separate silos for analyzing.

1.5 Business Risks

This is a relatively low risk investment for both MarkLogic and ourselves. MarkLogic's goal is to use us as a brainstorming tool. MarkLogic isn't in-

vesting any money into this project and is giving us a relatively small amount of time. Our goal is to seek a learning experience in requirement elicitation and get practice working with a company. Over the next year we will be investing a lot of time into this project, however it should be treated as a class and the measure of success will be in how much we learned and not how much we earned.

2 User Description

2.1 User/Market Demographics

Our main demographics are companies and researchers, as those groups are the most likely to have large amounts of unorganized data. Therefore our target demographic is educated, though their education is not necessarily software-focused.

2.2 User Personas

Josh is a researcher who needs to conduct a meta-analysis on different engineering techniques. He has collected thousands of examples of meta-data on projects using different engineering techniques. These examples do not have a standard format for how they store and collect data. Josh would like to look through his data and quickly and easily compare the projects he has collected.

2.3 User Environment

Our users will not be independent, they will either have a company or university backing them. They will be somewhat familiar to very familiar with data analysis techniques, but will not necessarily have any experience dealing with large amounts of data or unorganized data. They will be using this software in conjunction with other software, and will not necessarily have the time to get fully acquainted with a complicated conceptual model. The data that is put into the system will come from the user, and the data that exits the system will be used by the same user.

2.4 Key User Needs

Our users mainly need a user-intuitive way of traversing through large, unorganized quantities of data. They need the system to automatically find trends in the data in order to classify them. They will also need a user interface to correct any mistakes that our system generates.

3 Vision of the Solution

3.1 Vision Statement

To allow data scientists to extract insights from massive data sets by enabling interactive classification of data categories within the data.

3.2 Major Features

FE-1	The Data Classifier will use unsupervised machine learning techniques to identify classes of information contained in selected data sets fed as input to the Data Classifier.
FE-2	The Data Classifier will allow users to browse, edit, and add to the classifications that were performed automatically. A user may remove an erroneous classification and add a new classification for a data element that was not automatically classified.
FE-3	Allow users to see predefined and learned categories of data elements that will be recognized by the Data Classifier.

3.3 Assumptions and Dependencies

AS-1	The user is a data scientist with domain knowledge who understands not only the data they are working with but the context surrounding it.
AS-2	The user has clearance to view all data fed to the Data Classifier.
DE-1	The Data Classifier has been provided database(s).

4 Scope and Limitations

4.1 Scope of Initial and Subsequent Releases

Release one targets the end of 405. Release two will occur in 406.

Feature	Release One	Release Two
FE-1	Fully Implemented	
FE-2	Not Implemented	Fully Implemented
FE-2	Not Implemented	Fully Implemented

4.2 Limitations and Exclusions

LI-1	TBD
EX-1	TBD

5 Business Context

5.1 Stakeholder Profiles

Stakeholder	Value	Attitudes	Constraints
Lucky 7 Developer	learning opportunities, maintainability of project	excited	none
MarkLogic Developer	portability, data categorization	excited	none
Data Scientists	ease of use, data categorization	excited	none

5.2 Project Priorities

Dimension	Driver	Constraint	Degree of Freedom
Schedule	We must release our first iteration by the end of quarter two and have the final release by the end of quarter three.		
Features		Creating a tool with an easy visualization of data.	
Quality			TBA
Staff		Lucky Seven is composed of 7 software engineers.	
Cost		Each student must spend 8-12 hour a week on this project.	

5.3 Operating Environment

OE-1	This will be a web-based application

6 Competitive Analysis

6.1 Overview

As a primarily customer focused database driven company, MarkLogic has a number of existing competitors that centralize around managing customer data efficiently. Further, as a company that develops and enterprises a NoSQL database, there exists several competing companies that create commercialized NoSQL databases as well. MarkLogic founded in 2001 while not necessarily a new company, is less mature than companies such as Oracle that have a longstanding reputation in customer database management. However, MarkLogic with over a thousand customers has established itself as a leading business in NoSQL database management. Other competitors

such as DataBricks founded in 2013, are still in the stages of raising funding and lack the customer reputation in comparison to MarkLogic.

6.2 Competitor 1

Oracle is one of the leading competitors in database management and enterprise software systems. A number of different products advertised within Database management include Oracle Autonomous Database, Oracle Autonomous Dta Warehouse, Oracle Autonomous NoSQL Database, and Oracle Database 18c. Marketing their NoSQL database as one that can "instantly scale to meet dynamic application workloads for demanding cloud applications", Oracle emphasizes their solution strengths of security, performance, and autonomy.

6.3 Competitor 2

DataStax Inc., a clear competitor of MarkLogic centralizes itself around customer data management and enterprise applications to provide businesses full data autonomy. Utilizing the open source NoSQL database Apache Cassandra, DataStax created a proprietary version called DataStax Enterprise (DSE). Steming off of DSE, they released DataStax Enterprise Graph which included a graph database, management, graph visualization and language support drivers. With a number of key partnerships with companies such as Hewlett-Packard, Microsoft and Oracle, DataStax identifies itself as a platform to drive the "Right-Now" economy.

6.4 Competitor 3

Databricks, a company founded by the creators of Apache Spark emphasizes their data processing as "unifying data science, engineering, and business". Currently, Databricks offers a Databricks platform that provides business a zero-management cloud platform built around Spark. Another product they offer is a Community Edition for students and universities who want to experience using Spark. The Databricks Unified Analytics Platform focuses on handling analytic processes such as model training and deployment and also help automate complex data pipelines for simplified job scheduling, monitoring, and debugging.