VIT UNIVERSITY
(Estd. u/s 3 of UGC Act 1956)

Software Requirements Specification

for

Data clustering and statistical modelling of CSV datasets

Version 1.0

Prepared by

Group Name: **siDBa**

Nishant Rohan Rodrigues	16BCE0098	rohan.rodrigues2016@vitstudent.ac.in
Annam Sai Kaushik	16BCE0527	annamsai.kaushik2016@vitstudent.ac.in
Shubham Vishwakarma	15BCE0334	shubhamvishwakarma.2015@vit.ac.in

Instructor: Prof. Karthikeyan T

Course: Software engineering CSE 3001

Lab Section: L3 + L4

Date: 21 January 2018

Contents

1 Introduction

This software siDBa, aims create a mongoDB from a csv file and also provide statistical insights to the user. siDBa is open source and modular to allow experienced user to utilize it for various database related purposes. A csv file is a text version of an excel sheet where each column is separated by a “ , ”, hence the name csv (comma separated values). The general problem faced with this file type is that all the values are converted to string and then stored in the text based file with the .csv extension.
siDBa aims to allow the user to import a csv dataset into a mongoDB without having to check and convert all the data-types of each and every entry manually. The mongoDB can be local or hosted depending on the preference of the user. This also means that a good statistical model can be prepared for the user to have a better understanding of the dataset as now he/she does not have to manually run through the dataset.
The following sections contain each of the modules and working of the software in greater detail, hence explaining the wireframe models and the data flow through the project.

1.1 Document Purpose

siDBa Version 1.0 is the first version of the software and implements all the core concepts of the software. This includes the core module to convert csv to mongoDB. It also provides a report on the data to help the user have a better understanding of the imported dataset. This document will provide particulars about the software functionalities and the requirements to create and utilize the software.

1.2 Product Scope

siDBa 1.0 is aimed at all csv dataset users. This applies to large fraction of developers, data scientists and students as csv is the most common file type to share datasets over the internet. It is due to the simple text file type of the dataset making it completely harmless and easy to convert back to an excel format. The simple interface is designed to allow even inexperienced user to have a great experience in converting their dataset.
One of the largest benefits about siDBa is that the new db can be local or hosted on various services such as mlab etc. This means that once a dataset is converted it can be shared easily with the entire team without the extra hassle of each member creating their own database and manually importing it over and over again.

1.3 Intended Audience and Document Overview

This document is directed mainly toward developers who are looking to get a better understanding about behind the scenes working of the software for various purposes such as to make this an embedded part of some other software of modify the file types during conversion for user specific tasks. As the interface is simple and aimed to make the experience hassle free there is absolutely no need for general clients to have a look at the documentation.
For the developers the best practice is to skip forward to the module definitions and the dataflow diagram only after going through the section 1.4 to have a better understanding of the content.

1.4 Definitions, Acronyms and Abbreviations

Abbreviations

csv – comma separated values
DB/db – database

Definitions

csv - In computing, a comma-separated values (CSV) file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas.
MongoDB - MongoDB (from humongous) is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas.

1.5 Document Conventions

This document follows the IEEE standard with the following formatting requirements

Font

Arial font size 12 for all the content
Bold heading and sub heading

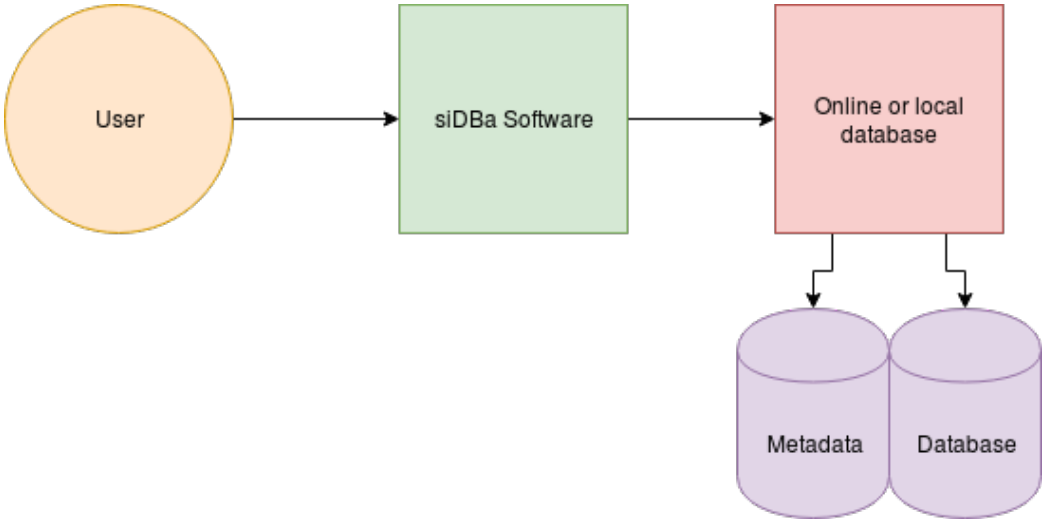
Others

Added hyperlinks for digital references and extra information about various topics.

2 Overall Description

2.1 Product Perspective

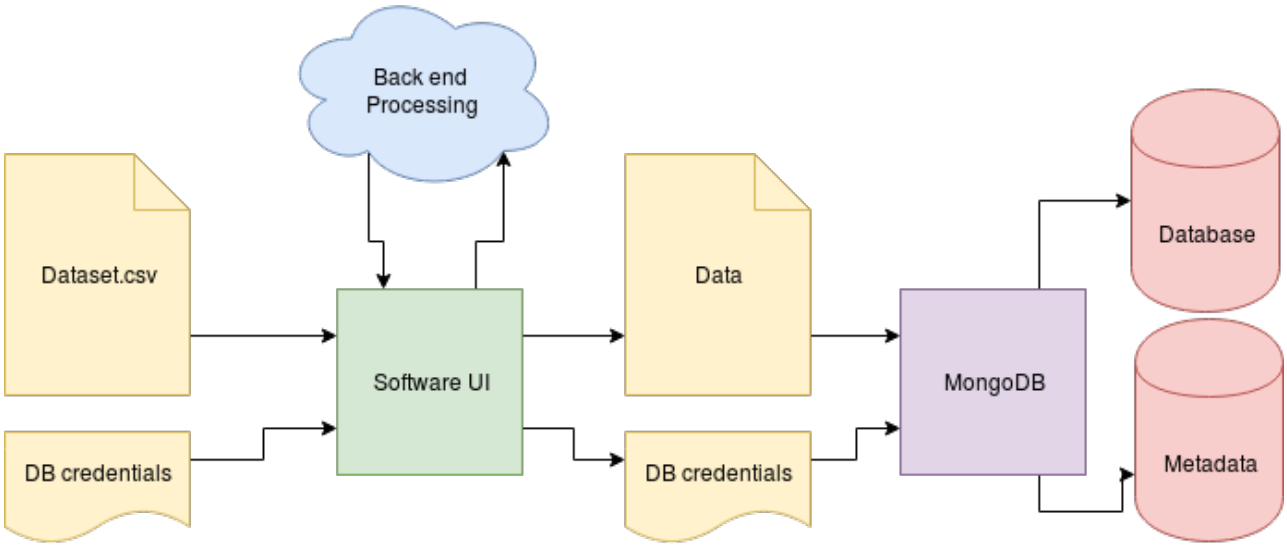
The software is a first of its kind and the idea is to make it easier for any user to deal with the hassle of data wrangling and database creation. This document defines the entire system with all its modules for version 1.0 for siDBa. The software works with a mongo instance and can be integrated with any other form of databases also. It can work with both local mongoDB's as well as any of the online mongoDB service providers. The integration with mongoDB can be better understood with the help of the following illustration.



2.2 Product Functionality

Functionality can be broken down into the following modules

- User input
  Includes all the user input data such as the csv file, mongoDB instance link, databese name and the collection details.
- Import module
  It reads the data from the csv and also checks for data consistency and readability issues.
- Data wrangling
  This makes sure that the imported data in text format is correctly converted to its appropriate format that may include integer, float, date etc.
- Database connection
  This create the connection to the database which may be local or hosted online. Also it handles any connection errors or other exceptions.
- Statistical analysis
  Provides a statistical report on the database which allows the user to have a better understanding of the data.



2.3 Users and Characteristics

siDBa is a very simple to use software and this allows all users to have a very fast and hassle free experience. Due to this it is easy to classify the users into mainly two categories.
These categories include

- Casual users
  All the users that integrate this software into the database creation and statistical report point of view and use the software as a standalone.
- Developers
  As the software is open source, it is free for anyone to use it in their own way. This may include integration with other third party software's, automation for web and online services and improving and adding new features to the existing software.

The casual users make up the largest part of all users and is also the most important section as they directly experience the software as it is. This means that the software must be easy to use and simple to understand.

2.4 Operating Environment

The software is very simple and does not require and special software or hardware requirements. A stable internet connection is a must when dealing with online data bases and poor RAM and processing power can lead to longer execution times when dealing with large datasets. In-case of dealing with local mongoDB's it is a must to have the local instance updated and running at all times.

Recommended hardware specifications

- Processor - 4 (Quad) Core Intel i3 or better
- Memory – 2GB RAM or better

2.5 Design and Implementation Constraints

Constraints include

- Basic internet facilities
- Simple hardware constraints
- Proper input csv data set
- Correct online database credentials
- Security constraints

2.6 User Documentation

The software is aimed at all users and to allow this, it has a very simple and easy to understand user interface. Even though the software allows a hassle free experience, delaying with database setup and credential can be a problem. To help with this, users will be directed to the mongoDB official website documentation through the software's "Help" section. It will also contain the basic instructions required to test the software and a sample video and data set to test the software is provided on the website.

2.7 Assumptions and Dependencies

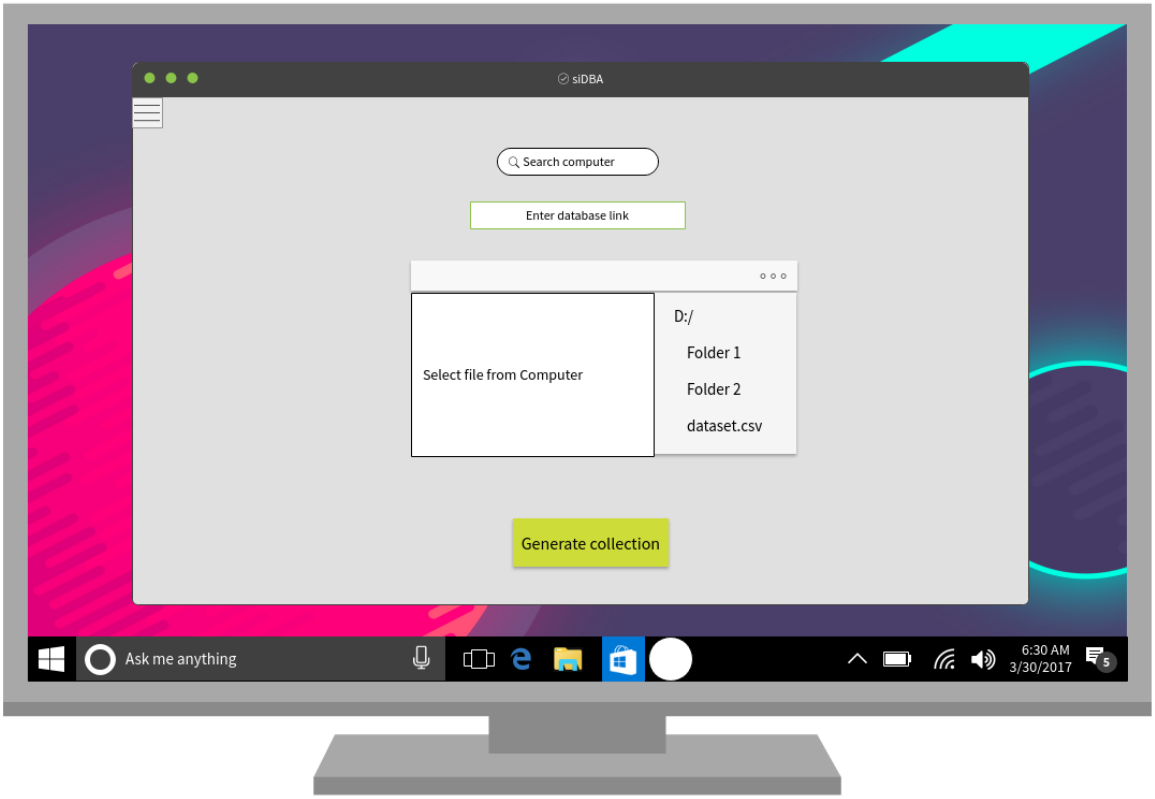
- User has basic knowledge of computers.
- The system is fast enough to deal with large data sets.
- Internet connection is available to all the users who use this subsystem.
- The user is assumed to give system correct information about his details such as the database credentials.
- The system will have simple and easy to use interfaces.
- All the necessary information is present in the input csv file.
- Provides accurate data.

3 Specific Requirements

3.1 External Interface Requirements

3.1.1 User Interfaces

siDBa contains only two user interfaces. These include the interface to generate the collection in the database and one to have a look at the statistical model of the generated data base.
The database generation interface is very simple to use. The user only has to select the file from the computer or enter a file URL. Next he/she has to mention the mongoDB url with the proper credentials and hit the “Generate collection” button. This will work its magic on the csv file and convert it to a mongoDB collection in the user specified database. The mock user interface can be seen in the illustration below.



3.1.2 Hardware Interfaces

Libraries include

- pymongo
  Creates a web socket to link the server with the local or online hosted database.
- Pandas
  For advanced data modelling functionalities like creating data frames etc.
- Numpy

Numerical python that utilizes the power of C for a quicker and more optimized mathematical solution.

3.1.3 Software Interfaces

siDBa can run on both Linux as well as Windows systems. It only requires python and the pymongo library as mentioned in the above specifications. In-case it is hosted as web service, this is not required as the file is uploaded to the server the handles all the back end processing. But this method is not feasible in-case the file has very large size making it very difficult to upload it to the server.

3.1.4 Communications Interfaces

The communication interface is handled by the pymongo library used in Python. It take care of securing all the connections to the database in case it a local database or hosted online by a service provider.
In-case the software is hosted online as a web page, it will also have a secure https connection with each and every user as the data flowing through the connection need to be secure and its integrity must be preserved. The library also takes care of setting up the web socket that allows for an uninterrupted connection to the database.

3.2 Functional Requirements

Functionality can be broken down into the following modules

- User input
  Includes all the user input data such as the csv file, mongoDB instance link, database name and the collection details.
- Import module
  It reads the data from the csv and also checks for data consistency and readability issues.
- Data wrangling
  This makes sure that the imported data in text format is correctly converted to its appropriate format that may include integer, float, date etc.
- Database connection
  This create the connection to the database which may be local or hosted online. Also it handles any connection errors or other exceptions.
- Statistical analysis
  Provides a statistical report on the database which allows the user to have a better understanding of the data.

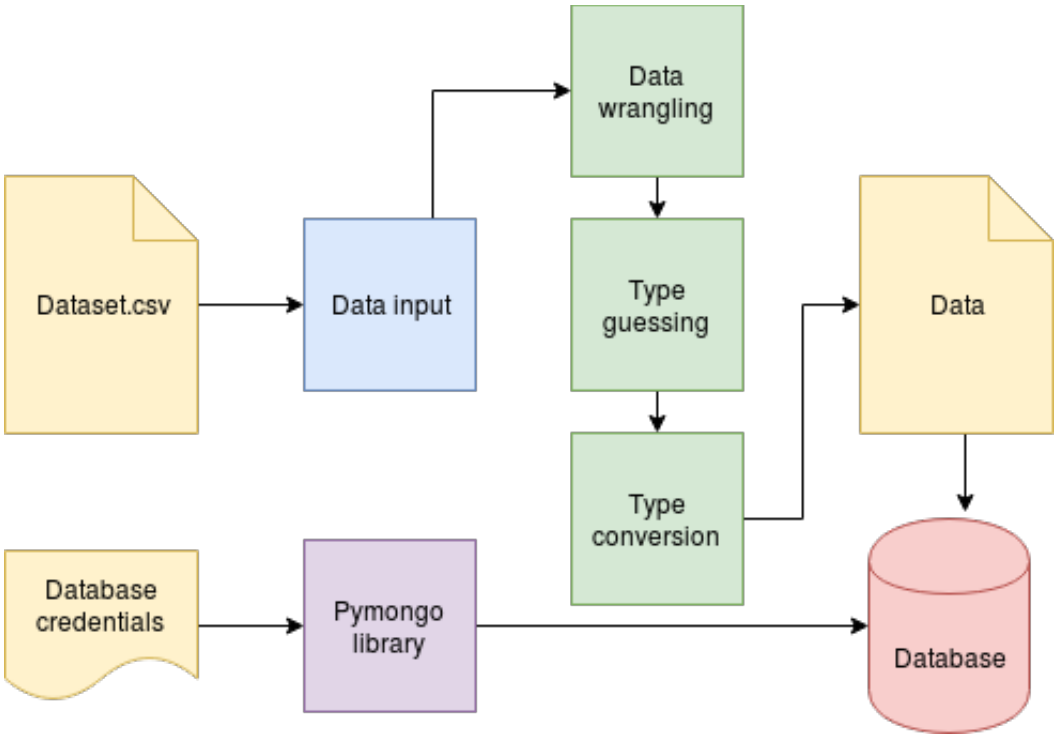
3.3 Behaviour Requirements

3.3.1 Use Case View

Actors include

- User – Person that interacts with the software
- Server – Does the data crunching and modelling. It may be hosted online or a local program depending upon the user preference

- Database – Local database or an online instance depending upon the user preference



4 Other Non-functional Requirements

4.1 Performance Requirements

Performance requirements

- Upload file size must not be greater than 50MB
- File uploaded must be a of csv format
- Proper database credentials
- Time taken to complete the task depends upon the size of the input file
- File data must be correct

4.2 Safety and Security Requirements

Requirements include

- The database credentials are only used once and not stored for any other purpose. Hence each time the user must mention the database link.
- Incase of large files, the user can run the software on a local system. This allows high performance and also prevents the database from corruption due to timeouts.
- Incase of a local instance, do not store the credentials for reuse as this may lead to issues during updates and can lead to bugs in the software.

4.3 Software Quality Attributes

Software quality is maintained by

- Creating a timeout hence maximizing the server performance. This allows a larger number of users to utilize the software at the same time.
- The software is completely open source and users can run it on a local server to compute large data sets and highly time consuming tasks.
- Regular maintenance and updates for bugs
- Also issues can be opened on the github repo. and this allows the community to interact and understand the software and its functioning.

APPENDIX A – Data Dictionary

Variable	Values
db_link	Database link with embedded credentials
File path	Path to local file (OR) URL to online file File must be csv format
Return	Return values 200 – for completed request 400 – for input error 500 – for internal server error

APPENDIX B - Group Log

TOTAL HOURS =7 HOURS	TIME UTILISED FOR
2 HOURS	DISCUSSING AND DEICIDING THE SOFTWARE MODEL
2 HOURS	FOR GATHERING INFORMATION ABOUT THE TOPIC
3 HOURS	FOR SORTING AND MAKING THE DOCUMENT