Software Design 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:	Professor. Karthikeyan T
Course:	Software engineering CSE 3001
Lab Section:	L3 + L4
Date:	18 February 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 datatypes 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 design requirements and the data flow through each module of the project.

1.1 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 design to create the individual modules and the User Interface (UI/UX) for the software.

1.2 System Overview

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 m-lab 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 Design Map

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. This is to allow a better understanding of the data flow through the modules and the the overall presentation of the software.

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.

2 Design Considerations

2.1 Assumptions

- 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.

2.2 Constraints

Constraints include

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

2.3 System Environment

Recommended hardware specifications

- Processor - 4 (Quad) Core Intel i3 or better
- Memory – 2GB RAM or better
- Free space – 10MB

Required software

- Python 2 or higher
- MongoDB server in case of hosting a local server
- MongoDB Compass

2.4 Design Methodology

The overall design 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 creates the connection to the database which may be local or hosted online and 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.

2.5 Risks and Volatile Areas

The only possible risk is of compromise of the database credentials. This is possible as the user must provide his/her mongoDB url to connect to an online Mongo service. SiDBa does not store these credentials and they are destroyed as soon as the connection to the online Mongo service is established. This means that the software is completely safe.
The user must make sure to check this if siDBa is integrated in some third-party software as it is an open-source application and the source code can be tampered with by the third party.
In case of using third party applications the best way is to create a collection in your local mongoDB instance. This means that the url does not require the user name and the password keeping the user safe from any compromise.

3. Architecture

3.1 Overview

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. Then he/she has to mention the mongoDB url with the proper credentials and hit the “Create” button. This will work its magic on the csv file and convert it to a mongoDB collection in the user specified database.

3.2 Subsystem, Component, or Module 1…N

The software can be broken down into the following categories

- 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 and 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 Strategy

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.

4 Database Schema

4.1 Tables, Fields and Relationships

MongoDB is a non-relational database design. This means that each entry in the database is treated as an individual object. Hence it is possible to have a number of different entries in each object of the same table. It can also store other objects as its entries. This means that a single entry can contain more than one entry depending upon the type of data in the CSV file.

4.2 Data Migration

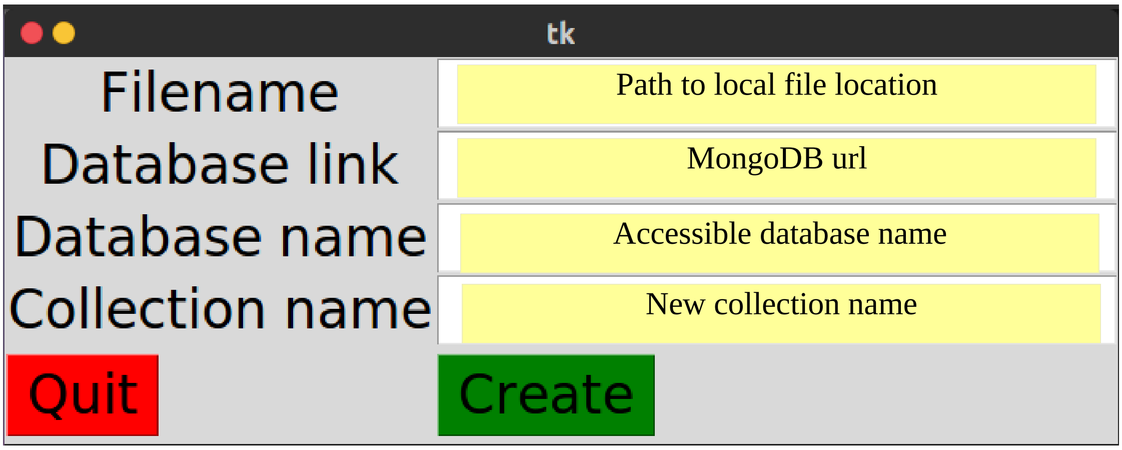
The communication interface is handled by the pymongo library used in Python. It takes care of securing all the connections to the database in case it a local database or hosted online by a service provider.
In-case the software is hosted online as a web page, it will also have a secure https connection 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.

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.
- In-case 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.
- In-case 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.

5 High Level Design

5.1 View / Model element



Filename	Path to local file location
Database link	MongoDB url
Database name	Accessible database name
Collection name	New collection name

Fig: Form layout

6 Low Level Design

6.1 Module 1…N

Module	Functionality
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. It also 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.

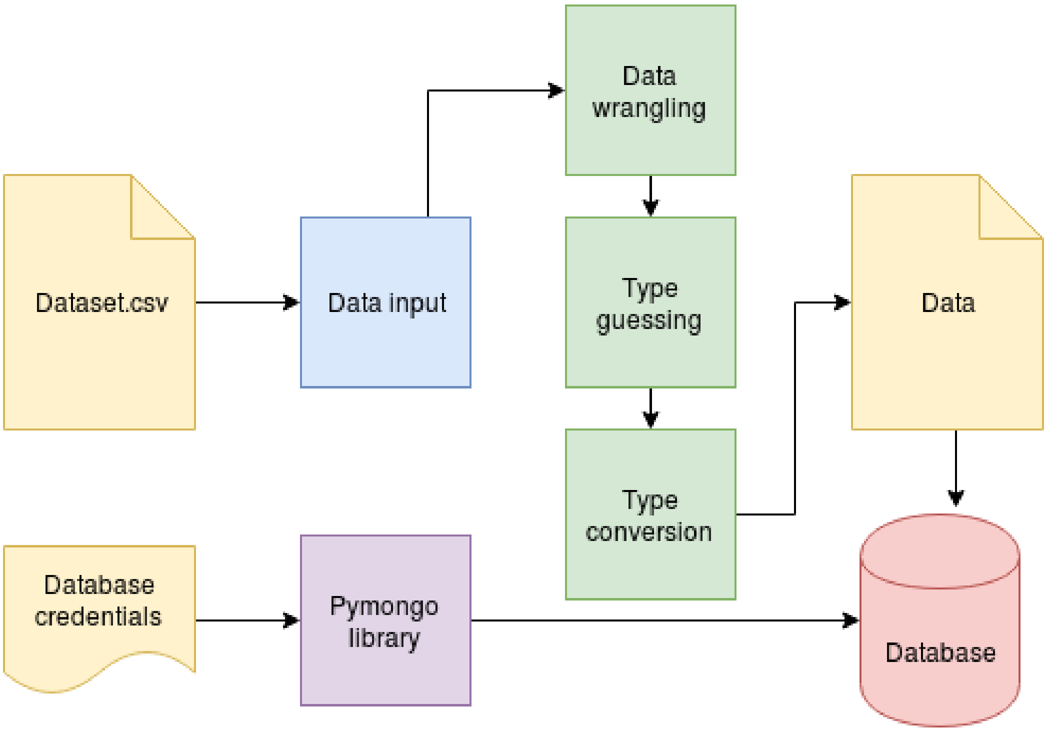


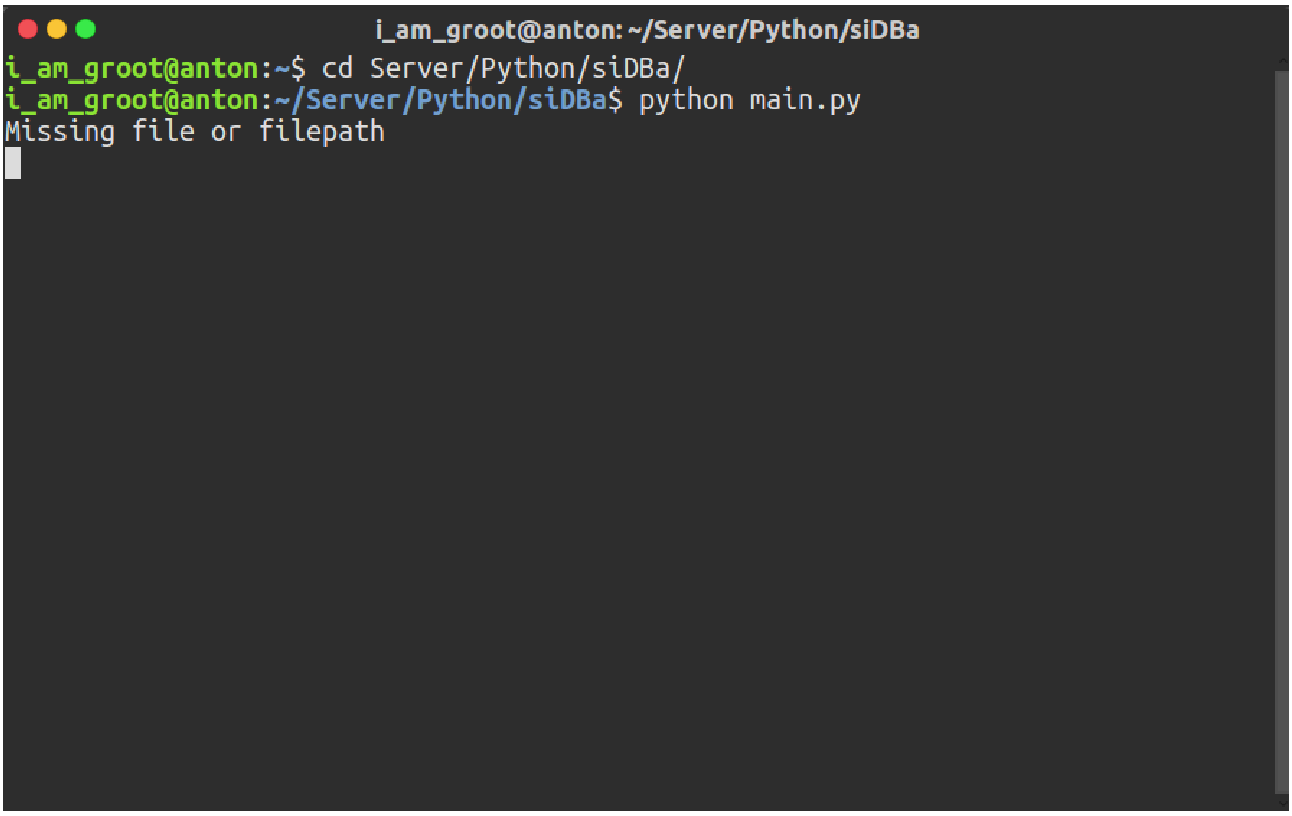
Figure: Module interactions

7 User Interface Design

7.1 Application controls

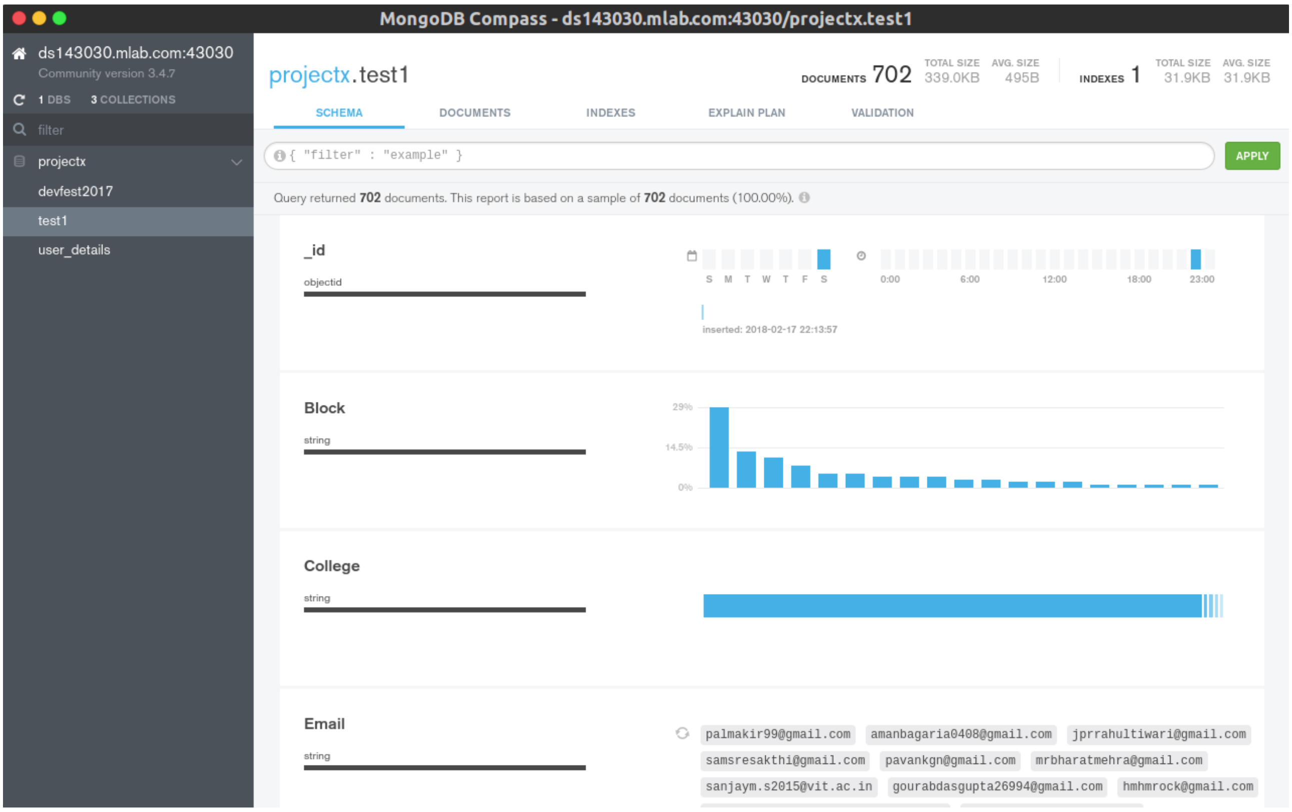
Button	Functionality
Create	Create the required collection in the user specified MongoDB and inserts all the CSV entries into it.
Quit	Saves all the settings and quits the application.

7.2 Screen 1…N



```
i_am_groot@anton:~$ cd Server/Python/siDBa/
i_am_groot@anton:~/Server/Python/siDBa$ python main.py
Missing file or filepath
```

Figure: Admin terminal



Query returned 702 documents. This report is based on a sample of 702 documents (100.00%).

palmakir99@gmail.com amanbagaria0408@gmail.com jprrahultiwari@gmail.com
samsresakthi@gmail.com pavankgn@gmail.com mrbharatmehra@gmail.com
sanjaym.s2015@vit.ac.in gourabdasgupta26994@gmail.com hmhmrock@gmail.com

Fig: MongoDB Compass

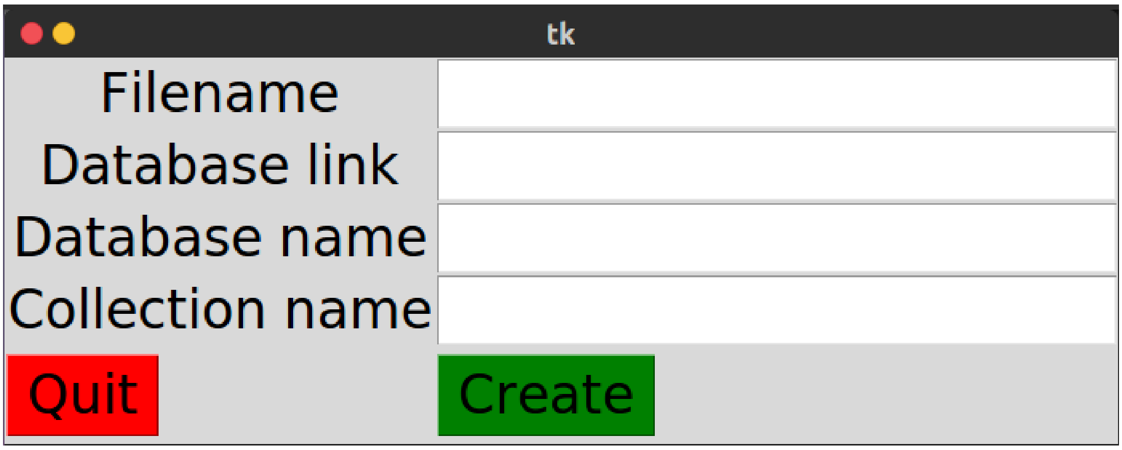


Fig: Software interface