|  |
| --- |
|  |
| Project Plan for Data Patterns |
| **Program**  Computer Engineering Technology - Computing Science  **Student Research Assistants**  Dante Beltran  Hamza Salman  Kevin Truong  Mark Greenidge  Bochi Li  **Industrial Supervisors**  Gurbachan Singh & Paul Trudel  **Faculty Supervisor**  Abdullah Kadri  7/28/2021 |

# Executive Summary

When it comes to data exploration, there is a high learning curve for researchers and data analysts to turn raw data into useful information. Currently, there are very few options to visualize data, most of which are paid options. What these solutions have in common is that the licensing fees are expensive and not always accessible to the average consumer. The other option is to use statistical languages for analysis such as Python. However, the average user is not always a developer, and therefore will struggle trying to do so. The goal of Data Patterns is to provide a more cost-effective alternative to the paid solutions that exist. The “Data to Intelligence” website will provide a one stop shop for individuals or small organizations, to easily upload and understand their data, while using statistical and visualization tools to find diverse ways to make sense of it. Additionally, all issues, bugs and milestones achieved by the developers, will be documented as the project is developed. This way, users can learn about how we achieved this result and increase their knowledge of data science along the way.

[Executive Summary 3](#_Toc78922848)

[Definition 5](#_Toc78922849)

[Requirements 6](#_Toc78922850)

[Scope 7](#_Toc78922851)

[Specifications 7](#_Toc78922852)

[Assumptions 7](#_Toc78922853)

[Schedule 8](#_Toc78922854)

[Deliverable Breakdown Structure (DBS) & Work Breakdown Structure (WBS) 8](#_Toc78922855)

[Development Model & Sequence of Activities 8](#_Toc78922856)

[Design & System Architecture 9](#_Toc78922857)

[Preliminary Development & Implementation 10](#_Toc78922858)

[Shared Repository 11](#_Toc78922859)

[Work Breakdown Structure 11](#_Toc78922860)

[UML 12](#_Toc78922861)

[Use Cases & Sequence Diagrams 12](#_Toc78922862)

[Gantt Chart 13](#_Toc78922863)

[Resources 14](#_Toc78922864)

[Tools 14](#_Toc78922865)

[Team Members 15](#_Toc78922866)

[Communications 15](#_Toc78922867)

[Distribution of information 15](#_Toc78922868)

[Performance Reporting 15](#_Toc78922869)

[Risks 16](#_Toc78922870)

[Manufacturing & Test/Evaluation 16](#_Toc78922871)

[Intellectual Property 17](#_Toc78922872)

[Safety 17](#_Toc78922873)

# Definition

The intent of this document is to provide an application development project plan. The scope of this document covers the project planning phase and demonstrates how a Work Breakdown Structure (WBS) and associated Resource Breakdown Structure might be incorporated into key project documents. This document also provides a structure for presenting:

* Project deliverables
* Project risks and opportunities
* Estimates
* Project resource information
* Project delivery method
* Configuration and change management

**Need for project**

This project will benefit the data science community in a couple diverse ways. First, the purpose for building this application is to give data analysists, researchers, and master’s students a free tool to both analyze their data, and to help researchers more easily recognize patterns within their data. Researchers spend a lot of time both collecting and analyzing data manually which can be both cumbersome and time-consuming.

As well, many of the enterprise tools for data analysis that currently exist are both expensive and have a high learning curve for the average consumer or research groups with a limited budget. With a tool this is both free to use and user-friendly, the product we are developing hopes to give an alternative solution to these currently existing problems in the data science field.

**Challenges**

The main challenge of the project is learning data science techniques. They are critical to transforming data both in real life data consolidation and digital consolidation. Our team members are not data scientists and have only brushed on the surface of the field of techniques, and this will require time for training to complete project tasks.

Another challenge in the project is the team’s unfamiliarity with the current technologies being utilized. Our team is unfamiliar with Mongo database technologies. As well, the server language, Python, and the D3.js libraries being utilized for visualization are foreign technologies to the development team and will also require research and training to progress the project.

**Opportunities**

By implementing the project, here are some opportunities that will result from its completion:

* The application will enable users to turn raw data into useful information.
* The application will allow users with extremely limited knowledge of data science to utilize the website to full capacity and let them learn about data science along the way.

## Requirements

The requirement for this solution is to create an application that can convert raw data into detailed and informative data dashboards. This requires the ability to easily upload and store data. The ability to efficiently process and analyze data, as well as the ability to generate viable visualizations.

Currently, there are few available software to visualize data, most of which are paid options. Some of the existing solutions are Power BI, SAS Forecasting, and erwin Data Modeler. All these alternatives are very expensive and not always accessible to the average consumer. The other option is to use statistical languages for analysis such as Python which has a very high learning curve most users. As a result, researchers are left with limited options for data analysis tools with very limited capabilities.

Compared to the existing solutions there are some limitations that may come into play:

* Users need to be able to upload raw data files in multiple formats to allow for a wider customer base. The application should allow for mass data and can process it quickly. This will require an efficient solution for uploading and processing data.
* Data would need to be stored safely in a way that it is not accessible from outside of the website applications. This would require encryption on each individual table in our database as one database will be holding many data sets.
* Manage old data and give the users the ability to clear their data after they have finished using the website. The data uploaded by each user would need to be cleared after a set duration of time to ensure stability of the system.
* The website would be required to analyze each individual variable in the provided data to understand the type of data coming in to be able to visualize it properly. This would require the website to be trained on different formats of data and how to process them as the user will not have the ability to create custom visualizations themselves.

These are the criteria which will be used to measure project success:

* Complete application implementation by the end of Fall 2021
* Establish a platform that will allow users to upload their data to a website to gain useful information through visual descriptions using d3.js libraries.
* Provide a repository of issues, substantial changes and major accomplishments made by the developers throughout the project’s life cycle. This hub will be meant for users to dive into how the project was completed.
* Data is automatically purged 24 hours after the data has been uploaded and stored safely.
* The website can perform in depth analysis of data without the need for user interaction.

## Scope

The scope of this project is the research, UI (User Interface) implementation, database integration, CSV data processing, data visualization, user-data interaction, real time data analysis dashboards, dashboard extractability, documentation of each stage, source code for all data analysis and visualizations. Instruction set for recreating visualizations and data models, testing documentation, project overview documents, operation manuals. The website design and layout are not in the scope of this project as it was already provided by the client as a bare bones template.

## Specifications

The success factors of this website must be well defined to satisfy the requirements outlined by the client. For this project to succeed, the website must meet the following criteria:

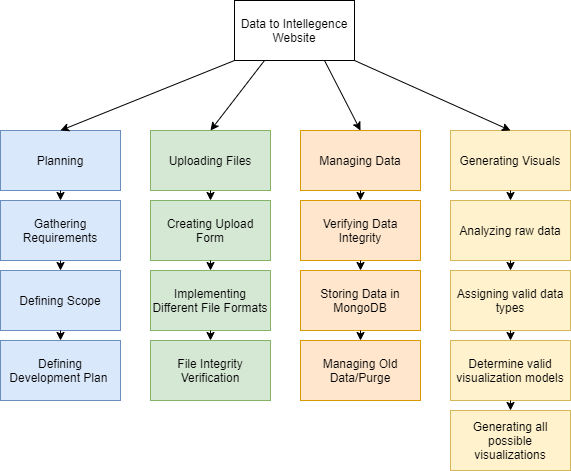
* Be able to develop in Linux Ubuntu environment, version 21.04.
* Be able to utilize Python libraries for development, version 3.8.5.
* Be able to use Flask to deploy JavaScript frameworks, version 2.0.1.
* Be able to run MongoDB database to store unstructured data, version 4.4.5.
* Be able to design and visualize the look and feel using D3.js and DC.js libraries, version 7.0.0.
* Be able to securely obtain data in correct format and verify the integrity of the data.
* Be able to manipulate and massage the data efficiently.
* Be able to dismantle the data exactly 24 hours after uploading, in respect of privacy.
* Be able to implement at least ten different types of visualizations from Python libraries to different types of data.
* Be able to perform business analysis and regression analysis for its scalability.
* Be able to incorporate functionality to extract data into .pdf files or webpages.
* Must allow users to view all code issues/resolutions related to the development process of the project through developer sections from the development team.

## Assumptions

* File types will be restricted to CSV, TSV, TLX, TXT and JSON files. Ultimately, since we are basing the project on the fact that the user has little to no knowledge in regard to json files, it would not make sense to do more than that.
* Another constraint will be the number of visualizations that the application will contain. D3.js is home to hundreds of different visualizations, and the application will limit the amount to 10-20 visualizations based on the project’s timeframe.

# Schedule

## Deliverable Breakdown Structure (DBS) & Work Breakdown Structure (WBS)



## Development Model & Sequence of Activities

The development model being used is the Kanban model. The development team will meet weekly independently of the product owner to discuss development obstacles and general progress, as well as divide up the work amongst the team. Tasks will be tracked through tickets, and progress will be measured through sprints, with new tasks being assigned at the beginning of each sprint. Tickets and documentation will be maintained and tracked through Microsoft Teams.

There are 3 important stages to the development of this project. First, there is the Managing Data stage in which the team will implement the ability to upload CSV files to the server. The team will also implement methods to manage and manipulate data with the Python server, to produce a working example of a Geo map with interactive data on the website. This stage is expected to take approximately a month.

The second stage is the Descriptive Analysis stage. The main deliverables for this stage are the implementation of extra filetypes like XML, JSON, and Pickle files. The development team is also responsible for updating their developer web pages detailing their progress, research, and problems encountered in development. This stage is estimated to take up to 2 months.

The third stage development is the Visualizations stage. This stage will focus on the implementation of 10 different visualizations of data being a key deliverable. This stage will also be devoted to testing the existing solutions for bugs. This stage is expected to take approximately one month.

Business Analysis is the fourth stage of the project. This stage will expand upon the work previously done, and a key deliverable will be the addition of twenty new visualizations to the website.

Each developer will build an in-depth data analysis on a topic of their interest and develop 4 different data visualizations of their choosing dedicated to this topic. This stage is estimated to take approximately six weeks.

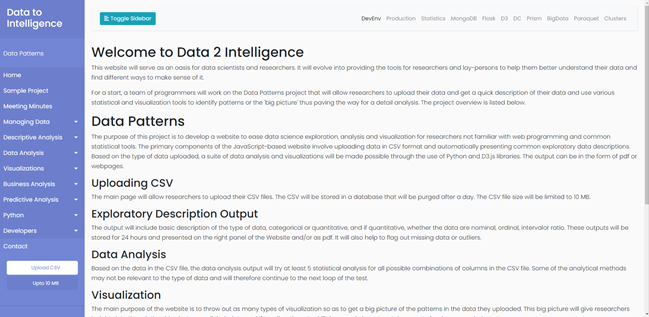
The final stage of development will be dedicated to Review and Finalization. This stage will center largely on stress testing the web application and fixing any outstanding bugs. In addition, documents such as the technical report and training guide will be finished. At the end of this stage, the final product will be handed over to the client for public use. This stage is estimated to be completed within the remaining eight weeks.

## Design & System Architecture

The design is a website which is used both for data scientists to submit their data electronically, as well as a reference for any users looking to replicate this process on their own. Acceptable formats of data are limited to include only CSV file formats at this time. The data received will then be entered into a database. The server would then perform various operations on the data and if the process were successful, the data would then be passed to the D3 and DC libraries to generate visual representations of the data in a new window.

The side menu would be the user’s main method of navigation through the website, with the ability to upload their data using the bottom button of the menu. Users can also view descriptions of diverse types of analysis, different Python libraries, and programmer pages to detail to the user different roadblocks or difficulties faced during this implementation, which will be periodically updated by each developer.

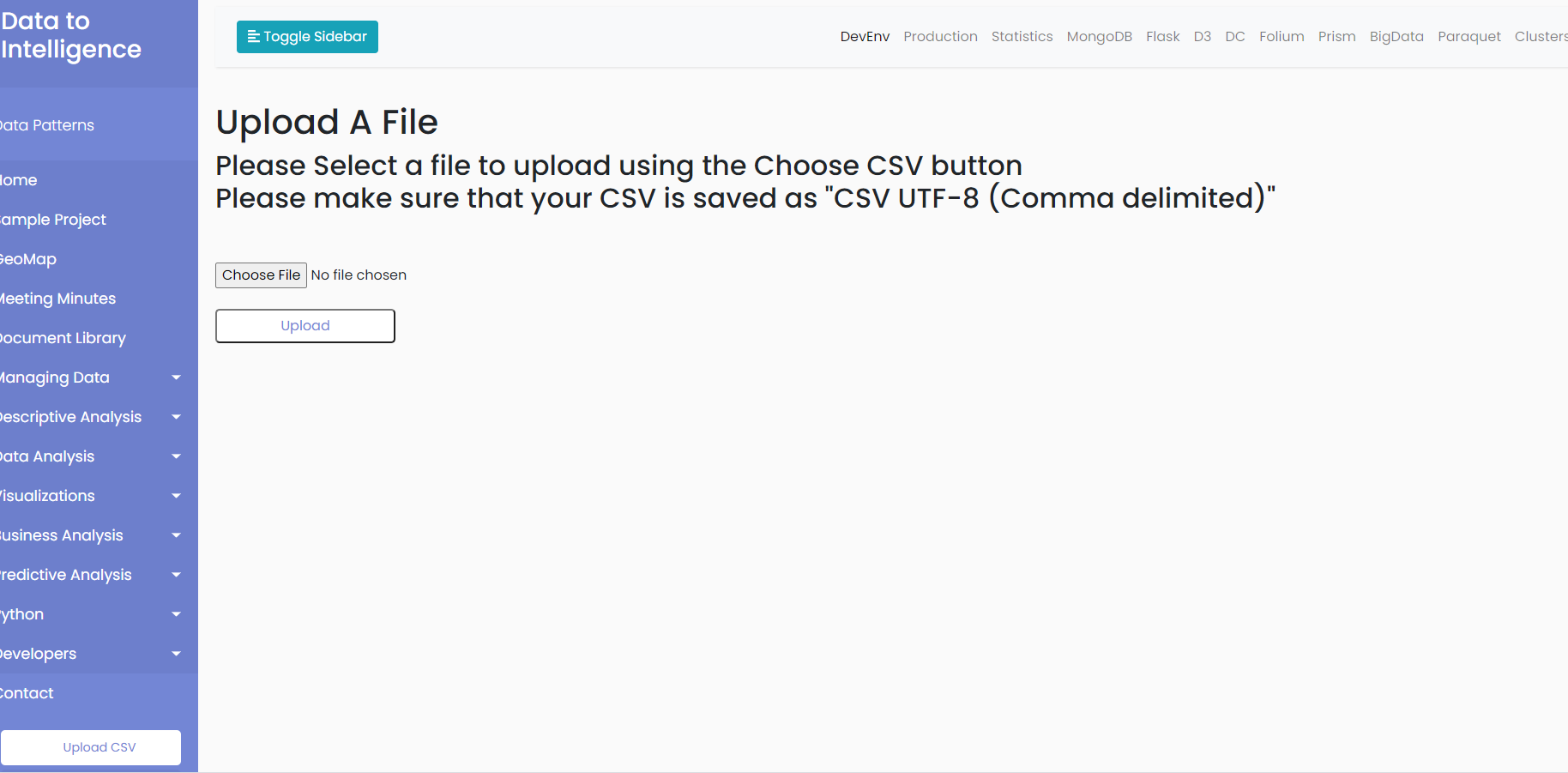
The top menu bar will be the users main resource tab to access throughout the use of the website. This menu allows the user to access all the main resources that is used to build the application, the user can find additional information on what tools and libraries that were vital to the development of the application and become stronger data scientists along the way.

A detailed photo of the website’s homepage can be seen below.

Button to upload data to the database Resource Bar Tab for developers to detail progress

## Preliminary Development & Implementation

For preliminary development and implementation, the application is still in its preliminary stages and thus, there is not much to show. However, some of the development that we have completed is we have added the functionality of uploading files onto the web server.



This page allows you to upload CSV files onto the website, however at the current time the files are simply stored on the web server and do not yet have any functionality paired with them.

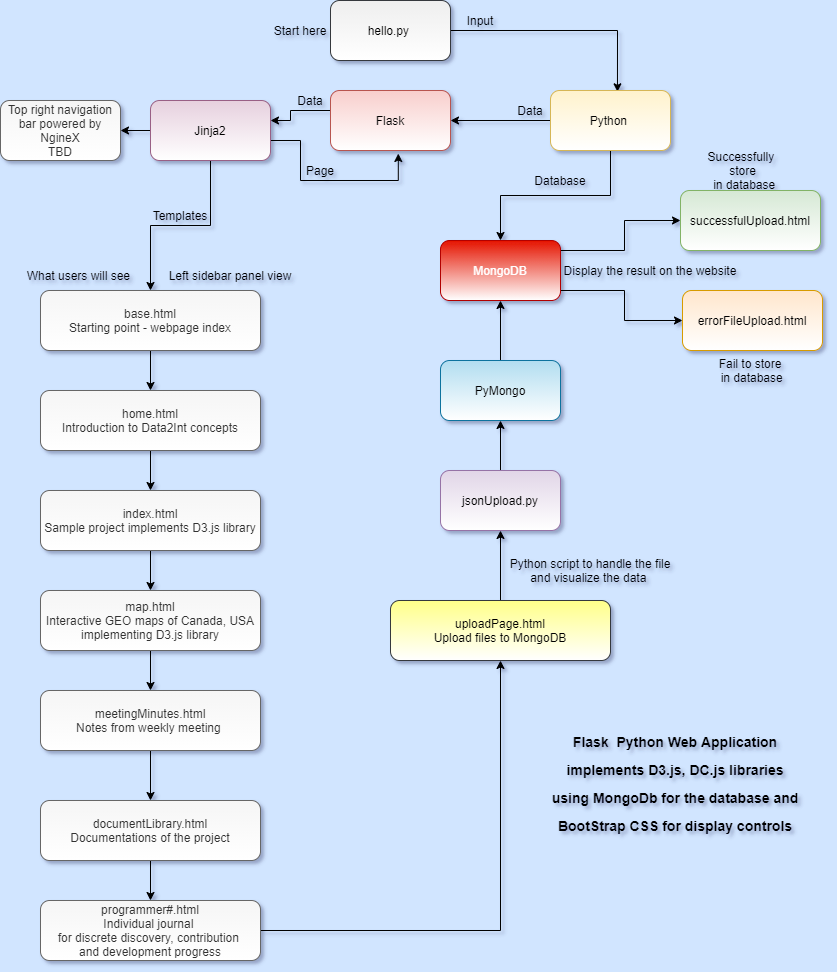
In the future, we plan to preform file integrity verification on each file and once validated, they are stored in the database.

## Shared Repository

When it comes to storing the application on a shared repository. We plan to store completed “releases” of the application on GitHub. After each milestone, we will push each new change onto the shared repository. Each developer will be able to access this repository and can view each commit and revert to any point at any time during the development of the application.

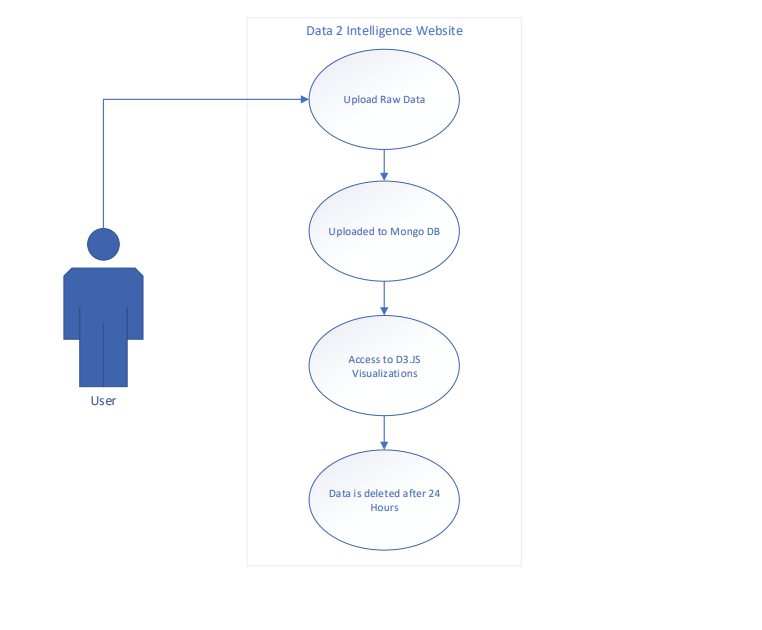
## UML

The following diagram is a UML of the Data patterns System Architecture:



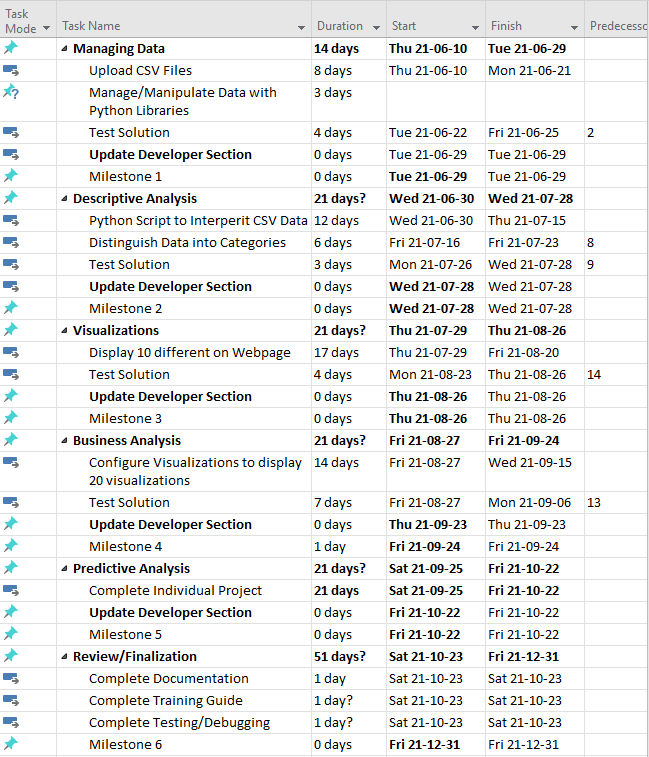
## Use Cases & Sequence Diagrams

The following diagram demonstrates the primary use case of the application from an end user's perspective:



## Gantt Chart

After gathering the requirements from the client, we have defined our goals and set milestones for this project. These milestones are illustrated on the Gantt Chart below:



# Resources

## Tools

This portion of the project plan contains all the tools required for the development of the application:

* Ubuntu version - 21.04
* Python version - 3.8.5
* Flask version - 2.0.1
* D3.js & DC.js version - 7.0.0
* Gunicorn version – 20.1.0
* PyCharm Professional version - 2021.2
* Nginx version - 1.20.0

## Team Members

This section of the project plan defines the various roles and responsibilities of the members of the project team. It also considers each team members level of authority within their scope of responsibility (e.g. approve, support, or conduct).

Staffing is subject to change as roles and responsibilities will vary throughout the life cycle of the project.

|  |  |  |
| --- | --- | --- |
| Member | Roles | Authority |
| Dante Beltran | Team Lead, DEV, TESTER | Approve, Support, and Conduct |
| Kevin Truong | DEV, TESTER | Support, and Conduct |
| Hamza Salman | DEV, TESTER | Approve, Support, and Conduct |
| Mark Greenidge | DEV, TESTER | Support, and Conduct |
| Bochi Li | DEV, TESTER | Support |

# Communications

## Distribution of information

Throughout the process of the project, project members will use ZOOM, Discord, and Microsoft Teams to communicate and communicate. This software has no geographical restrictions. Anyone can connect and join the meeting. All emails and files are sent by the project team leader. This includes project questions, links to weekly meetings, and estimated time for group meetings. At the same time, all members and project leaders can receive messages, and an email will be sent once a week to inform the project leader to participate in the ZOOM meeting. the link to.

## Performance Reporting

Throughout the whole week, in total, there are four recurring meetings that will be held on Zoom. On every Monday and Wednesday, a meeting among us will happen and during this time, we will report to our leader with what we have done, what is our goal for this week and what are we planning to present for our upcoming client meeting. On Thursday, we will have a meeting with our professor to summarize our work in the past week and then we will have a client meeting where we present

# Risks

Uncertainties are quite common in the early phase of development, and it is crucial to identify these risks and take them into consideration. Some risks are worth taking but sometimes, there are some you need to avoid. After gathering all the requirements from the client, we have analyzed and determined that these risks will have a high impact on our project development plan:

* Uploaded files may contain viruses
  + The probability of this risk is extremely high because the website is open to the public and it is necessary to validate all files that are being uploaded to the website because threats are everywhere on the internet
  + This can heavily impact the website and the server. Viruses, malwares such as Trojan can damage, disrupt, steal, monitor or destroy the website in a matter of seconds
  + There are many ways to reduce the risks by using a scanner, rename files on upload, validate files before execution but the risk is still there but it is very minimal
* Multiple file extensions can cause a delay in software development
  + The probability of this risk is medium because allowing multiple file extensions will take more time to analyze files and time is highly valuable to project deliverables.
  + This will put a stress test on the team because validating and analyzing files can take a lot of time.
  + A solution is to convert these files into one file format (JSON for example)
* Bigger files are complex to handle and will take more time to analyze and visualize that then can cause suspend the server
  + The probability of this risk is high because the website relies on the server and freezing the servers can cause a lot of problems
  + When the servers are frozen, the website does not have access to the database which can cause the website to white out (blank pages)
  + We can limit the file size but that also means limiting the amount of data that can be processed

# Manufacturing & Test/Evaluation

* Manufacture Is not required for the development of the Data Patterns Project
* Facilities will not be used throughout the development of the project
* The testing and verification of the project will be compared/aligned with the project criteria

# Intellectual Property

* All rights of the website belong to Gurbachan Singh
* All software used in the development of the application are open source.

# Safety

To ensure safety throughout the development of the project, one of the key safety measures that will take effect on the application is the verification of uploaded files. When a file is uploaded, the application will perform file integrity verification on the file to prevent any malicious files from being added to the database. This solution will enhance the security and safety of the users that browse the website.

Add numbering to headers.