TED UNIVERSITY

Faculty of Engineering

Department of Computer Engineering

CMPE 491 (Senior Project 1) – Project Analysis Report

by

Yağız ÇİMEN – Efe TONTU – Mesut Nadir SEYFELİOĞLU

Supervisor: Assoc. Prof. Dr. Tansel DÖKEROĞLU

Table of Contents

1 Introduction

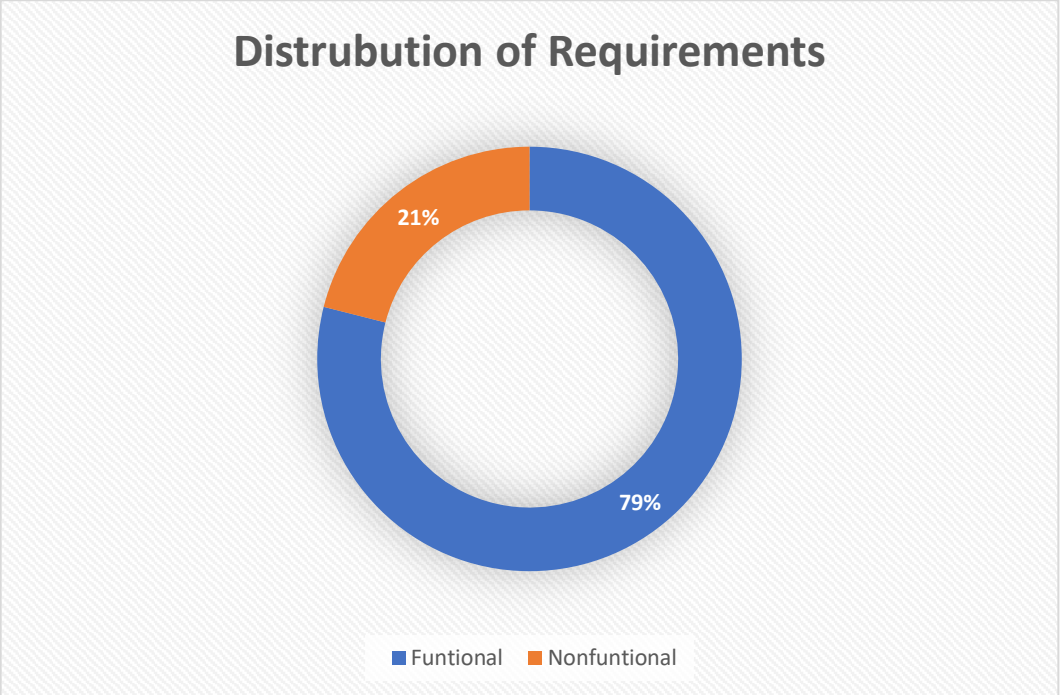
As one of the most hyped terms in the market, Big Data also took the interest as senior projects. On the other hand, there is one more topic that is as popular as Big Data which is social media. This situation leads us to do a project that combines both these important and popular topics. Nowadays most of the software companies It is a very exciting area to work in because big data has lots of demands on the software market and social media has lots of users. Thus, we thought that it would be a very important thing in a couple of years. So, creating a big data system that predicts communities through social media data will change the eCommerce world and advertising work. The main reason for this change is, advertisement companies will know what kind of advertisement will be better for specific groups, and eCommerce companies can easily define their possible customer's thanks to these kinds of systems.

As we know that commercials are very important for e-commerce and companies that work for commerce. Nowadays lots of companies prefer advanced, faster communication platforms like social media to advertise their products and reach out to more customers. We figure out if we can create a system which finds the communities in the social media and gives information to companies about these communities. For this purpose, the best thing to do is creating a system that helps these kinds of companies with big data systems. That’s why as a project group we decided to create a system for eCommerce and advertising companies. This system will search through hashtags, keywords, and comments through social media and it will find user’s communities by their social media sharing. This will help companies who want to advertise their products. By using this application company can easily find their communities that are related to their product. Additionally, these popular communities can be a great way to advertise their products through social media. The latter feature will be for companies who want to see their competitors' results so that they can build their marketing strategy by investigating their competitors' strategy.

2 Proposed System

For our Project, there are two main requirement groups to be considered. For the first phase, we have functional requirements that include software technologies and other hardware components to run stable and correct analysis over big data. Analyzing big data cannot be done by traditional techs. There must be specific tools to compete in the process like the Hadoop environment etc. Analyzing big data requires some hardware specifications with it, like powerful computers and processors. In addition, the plan for deployment is to convert to a mobile application in which companies and regular users can access and run algorithms better. For that purpose, we need to have an application that can work on many operating systems like IOS and Android. Production should be available 7/24, manage risks like information security, back-up, etc.

On the other hand, we need to clarify the nonfunctional requirements. We should have satisfied users. To achieve that goal, we are going to plan our requirements to reach maximum customer happiness. We aim to achieve a user-friendly application, with all the risk management and legal issues.



2.1 Overview

The “e-big” project is a combination of different areas. Such as; marketing, bid data, social media, social media marketing, machine learning and data analysis. Thanks to this combination it may work in any area that consist of data. Our project’s aim is creating the system that finds communities from social media sharing’s and the tags of the sharing. After that project's specifications can change with the customers' wishes. Since our system is working with social media data that means our customers will be the companies that want to sell their products or companies that want to analyze their market through social media. Thanks to this system our customers can also check the demand for their products. For example, since we will take our data frequently, we can see the new comments, also we can check changing demands. This was a simple example of telling what “e-big” project will work on.

Since currently we are working on YouTube data, we decided that checking video tags is the best way for the community analysis. For example, the “e-big” system will check the videoID’s tag and it will show predict the community/category of the video. After that we will look at the comments that is related with the tag and we will report the author_id of this comment, which will help for putting the YouTube user to the according community. The other important question is “what will be our costs?”. Since we will do all the work as software and we won’t need any hardware most probably our cost will be 0$. However, since we are collecting our data from RapidAPI which is free for 500 data per day maybe we will need to buy access to collect more data per day, but it is not a case for now.

2.2 Functional Requirements

2.2.1 Performance

Our project must be efficient and reliable. For that reason, we preferred to use Java python and spark technologies. For big data processing the kind of tools and technologies are proved that by their performance. On the other hand, our team going to compare the algorithm speeds on both java and python.

2.2.2 Reporting

As another requirement we should have reporting. Reporting is one of the most important things in progress. So, documentation the results and problems in order to make process more efficient is quite important.

2.2.3 Supporting Articles

To make our project we need some articles from real world. As a result of our search we have found some articles related to our project which has done by different institution's. You can find the links for article in References part.

2.2.4 Traceability

We should trace the events and their effect to our project in order to make it more efficient. We must know the work principles of each system and algorithm to design better products.

2.2.5 Hadoop Environment

Spark is the most popular big data analysis tool nowadays. Apache Spark contains lots of libraries to make it easier. Big data analysis is a hard thing to do. Spark is much faster than Hadoop analysis system. As a requirement we need to add spark technology to our project

2.2.6 Apache SPARK

Spark is the most popular big data analysis tool nowadays. Apache Spark contains lots of libraries to make it easier. Big data analysis is a hard thing to do. Spark is much faster than Hadoop analysis system. As a requirement we need to add spark technology to our project

2.2.7 Java, Python Technologies

In out project we are going to need JAVA and PYTHON technologies to run all this system. Hadoop HDFS system works on java basically. Also spark libraries can be implemented both in java and python programming languages. Our aim is to implement binary classification and machine learning algorithms to achieve trustful analysis. Also, we want to show the performance comparison between java and python on algorithms.

2.2.8 Datasets

We have some big datasets to analyze. Our project is going to be on social media platforms for that reason we found our datasets from YouTube. We are going to analyze these data sets in order to find and detect the communities in social media.

2.2.9 Operating Systems

Our system must work on both ios and android operating systems. In future we can develop our project as a desktop application. In this case we should make it compellability to windows Linux and mac operating systems.

2.2.10 Maintenance

Our team should run performance and correctness check in a plan in order to balance the maintenance of our project. As a requirement we begin to plan our test dates for project.

2.2.11 Security

Information security is the most important thing for our project. Because we are working on bigdata and real world datas. For that reason, we cannot allow anyone to use our private data for another reasons. For both mobile and desktop applications we should run security checks on each operating system to have a secure application.

2.2.12 Real Time Estimation

Another important thing to be consider is to have an application that can real time estimations. As new datas arrive to our system we should analyze and update the system eventually without storing them for future progress.

2.2.13 Minimum hardware for Big Data Process

We are working on big data so we should have powerful computers and processors to complete operations like search update add delete…

2.2.14 Accessibility

Product must be accessible from everywhere. We cannot foresee the future. There might be customers with different equipment's. So as a requirement we need to add an accessibility from everywhere feature to our project.

2.2.15 Back-up and Restore

Backing-up and restoring big data is not an easy goal to achieve. But working on big data comes with responsibilities. We should arrange a storing system and restoring system to decrease customer unhappiness. Not just satisfaction but losing data can cause more problems for analysis.

2.3 Nonfunctional Requirements

2.3.1 User Friendly

Applying a user-friendly project can lead us to better performance. User friendly interfaces can decrease the distraction for customers and make them easy to develop filters to search communities.

2.3.2 Project Branding

We are going to use Social media platform to brand our project to world. As we are working on social media it is the best platform to make customers hear our voice.

2.3.3 Project Risk Analysis

Risk analysis can occur in every phase of project. Testing implementing. Reporting… In each part we should consider the risks and report them as documents. As a requirement applying Risk analysis simultaneously while developing the project is quite important to have bug-free applications.

2.3.4 Legal and Compliance

As our team explained before in Ethical issues part, as a requirement we should consider the rules of legality to develop right project. We don’t want any of our customers to complain about the features of our project.

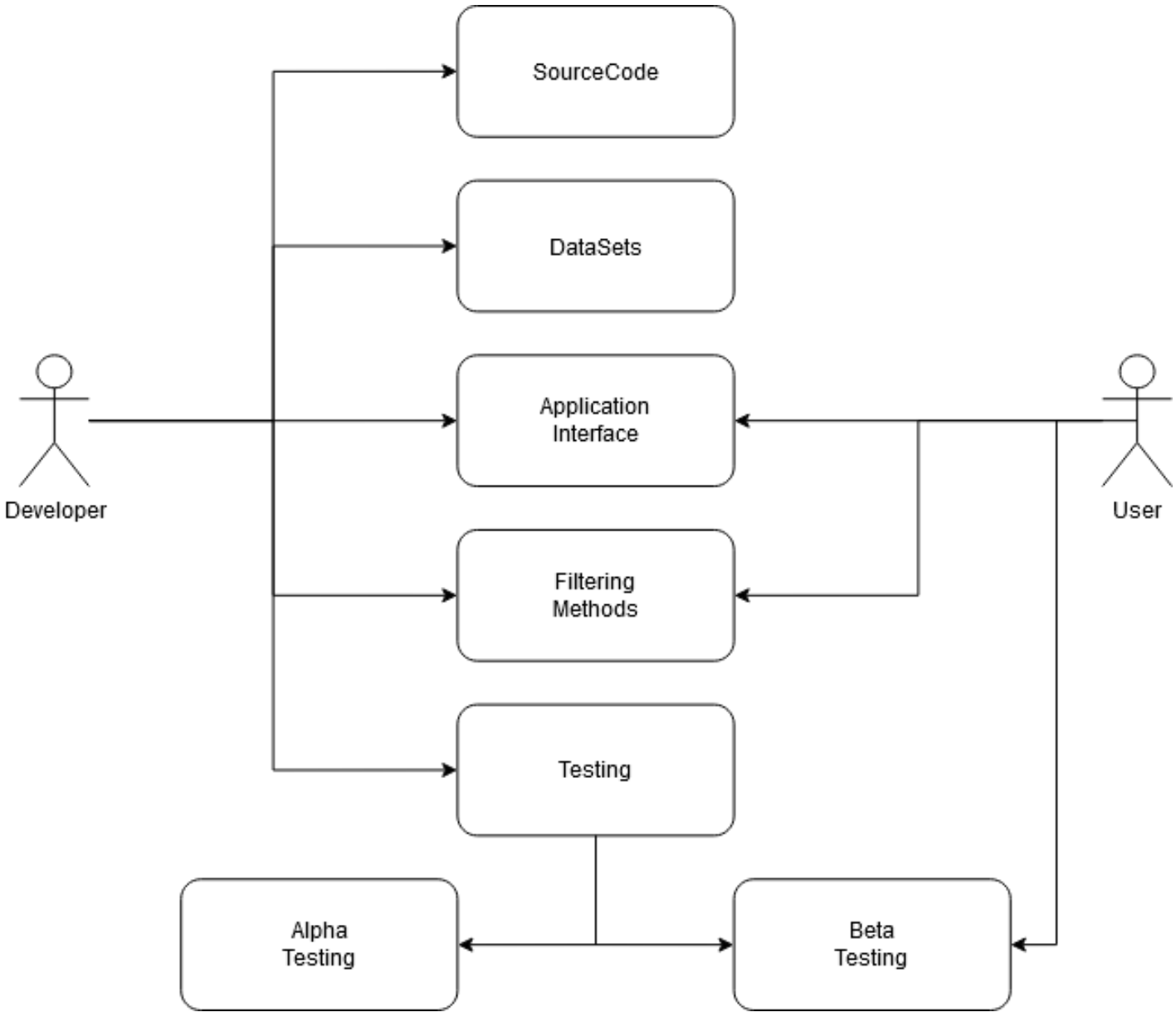
2.4 Pseudo Requirements

The three most significant project constraints -- schedule, cost, and scope -- are sometimes known as the triple constraint or the project management triangle. A project's scope involves the specific goals, deliverables, and tasks that define the boundaries of the project. The schedule (sometimes stated more broadly as time) specifies the timeline according to which those components will be delivered, including the final deadline for completion. Cost (sometimes stated more broadly as resources) involves the financial limitation of resources input to the project and the overall limit for the total amount that can be spent.

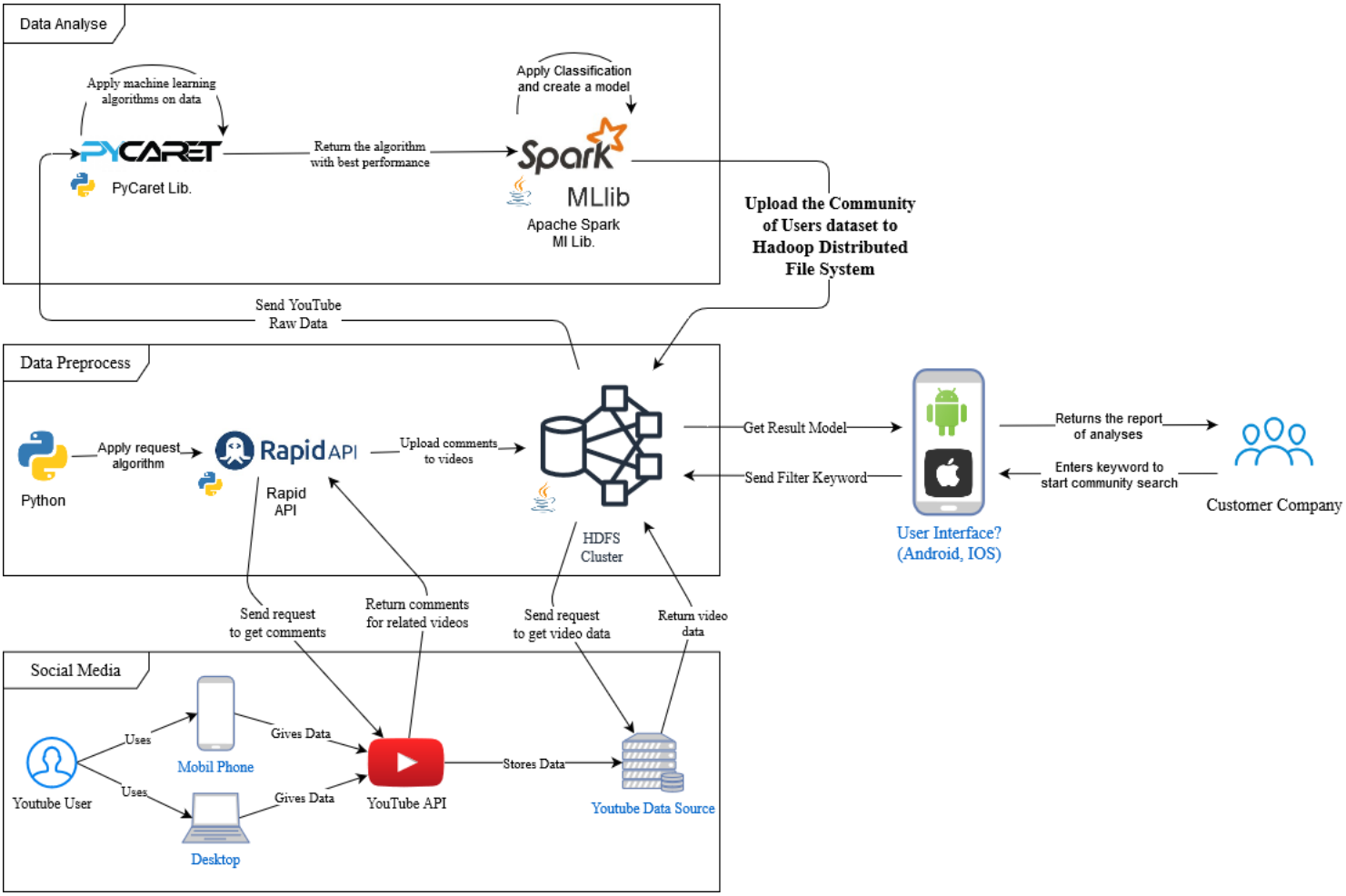
As a result of triple constraint, we have these limitations too. Our schedule is ready until the end of the semester and explained day by day in detail. We are planning to finish implementations at the end of the semester and our deadline is January 2020 for the e-big project. The other constraint is the cost, but this constraint is not that important for us because we are not using any external hardware devices, instead we are using open-source tools. The scope is one of the constraints that we face. Out project can be used in different areas such as social media, marketing, and so on. However, we have YouTube datasets for creating our models and testing them. After creating our models, we will test our system for test datasets that we will create, it will be ready to real life testing's. After all, e-big project will help marketing companies to find communities.

2.5 System Models

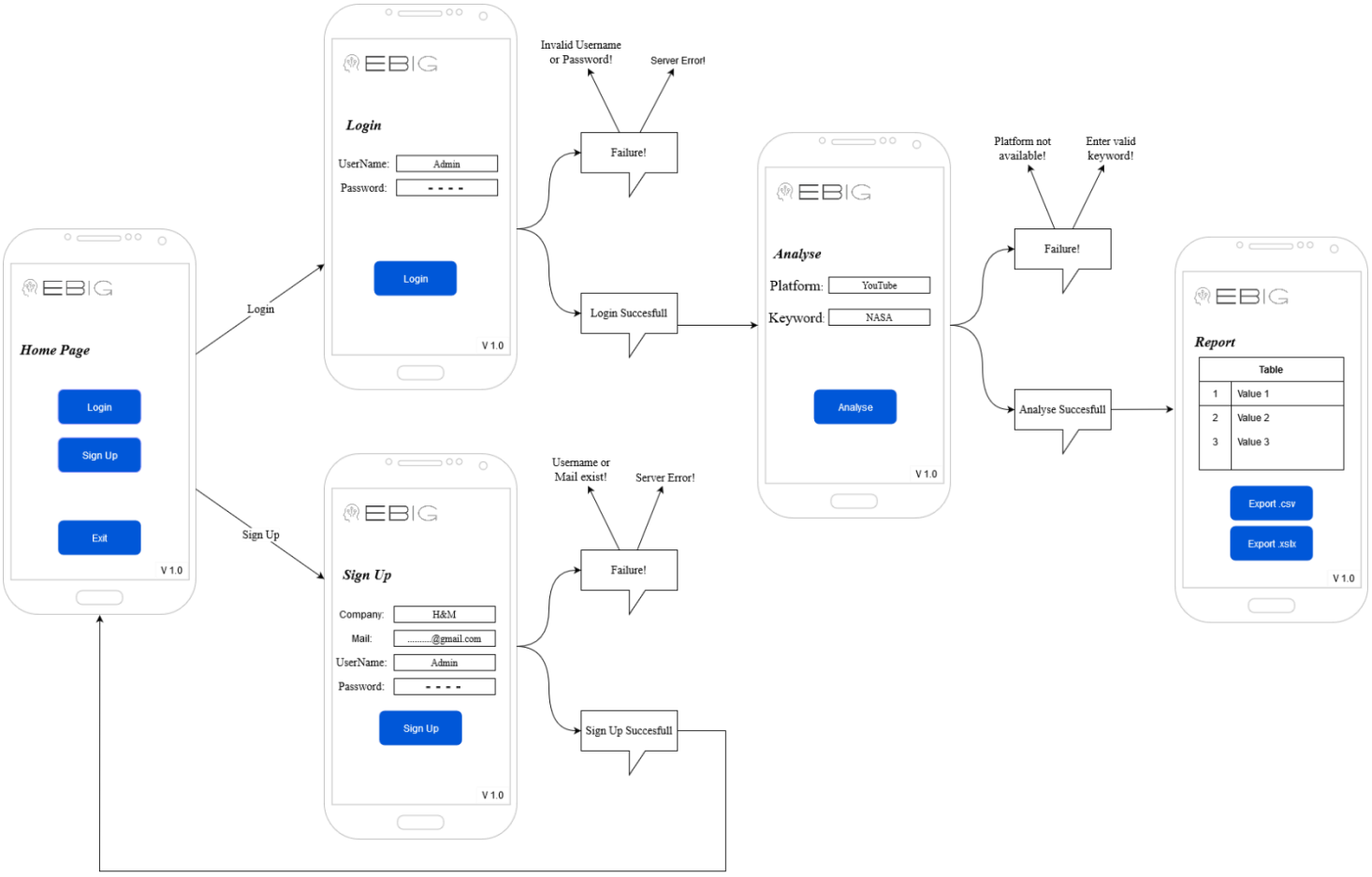
2.5.1 Use Case Model



2.5.2 Context Diagram



2.5.3 User Interface – Navigation



3 Glossary

API: A tool, or library, that assists developers in writing code that interfaces with other software.

Big Data: Big data refers to the large, diverse sets of information that grow at ever-increasing rates. It encompasses the volume of information, the velocity or speed at which it is created and collected, and the variety or scope of the data points being covered. Big data often comes from data mining and arrives in multiple formats.

Communities: A community is a group of people with shared values, behaviors, and artifacts.

Machine Learning: Machine learning is the concept that a computer program can learn and adapt to new data without human intervention. Machine learning is a field of artificial intelligence (AI) that keeps a computer's built-in algorithms current regardless of changes in the worldwide economy.

Triple Constraints: The Triple constraint theory tells us that every project that is being developed or has been developed in the past, operates within the boundaries set by the three constraints of project management (time, scope, cost).

Data Preprocess: Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, lacking in certain behaviors or trends, and is likely to contain many errors.

4 References

1. https://www.pmi.org/-/media/pmi/documents/public/pdf/ethics/pmi-code-of-ethics.pdf?sc_lang_temp=en
2. https://simplicable.com/new/project-constraint
3. http://xajzkjdx.cn/gallery/311-april2020.pdf
4. http://www.alochanachakra.in/gallery/586-acj-june-2256.pdf
5. https://iopscience.iop.org/article/10.1088/1757-899X/925/1/012015/pdf
6. http://ijiepr.iust.ac.ir/article-1-1067-en.pdf
7. https://poseidon01.ssrn.com/delivery.php?ID=352021002125001086000016120113013107036024008001019007007068122100020029097083098069121058118006041127001089097088086093079068030007004001088000083127085096123112010020086054028005121108096009126025124008119105080066007115127096075120029027121069020094&EXT=pdf
8. https://www.hindawi.com/journals/sp/2020/8884926/