Name: Kevin Gillanders

Title: Popularity prediction using reddit

Document: Functional Specification

Student number: 13410072

Finished: 25/11/2016

Functional Specification Contents

Contents

1. Introduction

1. Overview

This project will involve analysis of data from the site ‘reddit.com’. The factors which popular posts on the site have in common will be examined in the interests of finding what makes a post popular, and this will be used to predict the future popularity of new posts.

The project will be conducted in two phases, both of which will involve the collection, cleaning, storage and analysis of data on posts from Reddit. Initially, older archived posts where the values for all attributes are now static will be examined. This should allow establishment of trends between posts which have high final scores, giving a preliminary idea of what attributes are most important in determining post popularity.

Over two months of archived Reddit content has been obtained. This content will require data cleaning and preprocessing before it will be placed in an SQL database. Due to API limitations placed on batch calls to Reddit there are a few a few variables which are not returned. Posts must be called individually to retrieve these variables. Due to a two second wait placed on all API calls this will be quite time consuming.

For the second phase, the top 100 from the front page will be pulled in real time over the course of several weeks. While the first phase will give an idea of the ‘why’ posts are popular, this will indicate ‘how’ they are popular. As the information gathered during this phase will be dynamic it will show how the popularity of a post develops by looking at its half-life. Some popular posts may be ‘flash in the pan’ popular, with high popularity but short half-life, while others may display more of a slow burn, eventually reaching the same score but with a much longer half-life and longer visibility time. Once this attribute has been established, the other attributes influencing which of these categories a post falls into can be looked at.

This data will be accumulated over the course of weeks, and will be fed into a parallel processing pipeline. As Reddit has an IP based API restriction, multi-threading will be used. The main thread will perform the batch calls to Reddit and then pass the content to a separate thread which will handle data cleaning. This cleaned data will then be passed to another thread which will handle pre-processing and finally the clean preprocessed data will be passed to a thread which will handle database entry. The main thread will also write the ID to a memory location shared by two virtual machines These virtual machine will make individual calls to Reddit to obtain additional information required for analysis. Speeding up this process two-fold. This strategy could be extended if the process is still taking too long.

Following data collection, cleaning and storage for both stages, the data will be analysed to assess trends in popularity and regression techniques will be applied in order to attempt to predict the ‘virality’ of new content.

This project could have applications in advertising and media. Looking at Reddit specifically, specific subreddits likely stand a higher chance of reaching more people and having more widely visible posts. In terms of the sale of advertising space and style of advertising, this sort of information could be invaluable. However, results from this style of analysis could also be more widely applied. Increasingly, media outlets and advertisers alike are seeking to create 'viral' posts; information such as that to be found in this project is the key knowledge required to create such posts and tailor them to specific requirements. For example, if a media outlet or advertiser wishes to either get information to a large number of people quickly, or ensure that a post remains visible to a large number of people for longer, then knowledge regarding the factors determining half-life would be useful.

1. Glossary

- Regression techniques: Regression is a statistical process for estimating the relationships among variables

- Subreddit: A sub forum on the site, can be default which means that when a new user is registered they will automatically be subscribed to receive that subreddit's content

- Score: When a user makes a post on reddit readers then can vote on the post, either giving it a positive or negative score. The collection of these votes is the score

- Half-life: When a post reaches half of its final score

2. General Description

2.1 Product / System Functions

.

This is primarily a research project and so will not have an end product as such. However, a system will be produced in the form of a data analysis pipeline, the general functionality of which could have multiple applications requiring viral posts. Examples of this include advertising or news articles.

Data obtained from Reddit will be stored in a relational SQL database. Each comment has a unique ID which may be used as the primary key in a comment data table. Subreddit information, such as amount of active users and if it is text based, will also be stored in a subreddit data table, using subreddit_id as the primary key. Usernames are also unique so user statistics will also be included as a data table.

The second set of data contain many duplicate entries as each post will appear at multiple time points. As the entire point of this stage of the project is to look at the state of the same post at different timepoints, all of these entries will need to be stored. This will be achieved by making a composite key on post_id and

time_pulled. Items such as ‘title’ and ‘poster’ will only need to be stored once for each post_id, while variables such as current_rank and current_score will be stored for each entry. Using these it will be possible to track the popularity of a post over the course of its life. From this it should also be possible to see what subreddit posts get to the front page most often and which stay active for longest

2.2 User Characteristics and Objectives

The user base for this project would be companies which want to use reddit as an advertising platform or news agencies who want to strategically post an article on reddit so that it will get the maximum amount of views.

The user would require that the data which they are receiving is reliable. While it is possible for this prediction to say a post will receive a high score with a quantifiable amount of certainty, there is no way to guarantee this with absolute confidence, as with all viral content there is an element of luck involved. For example, the initial readers’ response may dictate the entire course of the post as an ambivalent response from the earliest readers, and therefore failure to spread or discuss the content, could eliminate any possibility of a post going viral. Describes the features of the user community, including their expected expertise with software systems and the application domain. Explain the objectives and requirements for the system from the user's perspective. It may include a "wish list" of desirable characteristics, along with more feasible solutions that are in line with the business objectives.

2.3 Operational Scenarios

1. A company which wants to gain more brand recognition decides to run a viral marketing campaign on Reddit. To ensure they generate the maximum amount of discussion may want to consult the analysis generated.
2. A sociology researcher who wants to conduct a study of how users interact with content and is interested in performing their study on Reddit could use the data which will be studied in this project

2.4 Constraints

There is a 2 second time constraint placed on all API calls to reddit. This constraint is enforced on the IP address of the machine making the call. They do, however, allow for batch calls to be made to the system returning several hundred at a time. The batch call, through an API limitation, does not return a variable which is likely important to the analysis: upvote_ratio. To get this variable a more detailed call is required. This is where the 2 second API limit becomes an issue.

To subvert this time constraint a multiple virtual machine solution will be used, where the multiple machines will have a shared piece of memory which contains all the IDs, makes new calls and returns the ratio and ID to a separate data entry program.

3. Functional Requirements

- **Description -** Acquire data from Reddit
- **Criticality -** This step is critical to the project as the data is required for all subsequent steps.
- **Technical issues -** Data must to be cleaned, spurious data must be removed and finally the data then must be placed in a database.
- **Dependencies with other requirements -** All other requirements depend on this step

- **Description -** Perform data cleaning.
- **Criticality –** This is an important step as the data must not contain any accidental irregularities which may skew data analysis
- **Technical issues –** Due to the large quantity of data this may take some time to actually complete and some irregularities may go undetected
- **Dependencies with other requirements –** This step is important for analysis of the data

- **Description -** Store data in SQL database
- **Criticality -** This step is critical to the system as this project deals with large amounts of data and a database will be required to access it quickly
- **Technical issues -** The database schema will need to be planned and built
- **Dependencies with other requirements –** This is dependent on obtaining data to store

- **Description –** Data validation
- **Criticality –** This is a critical step as in ensures the model is working on clean and helpful data.
- **Technical issues -** Due to the large quantity of data this may take some time to actually complete and some irregularities may go undetected
- **Dependencies with other requirements –** This step is important for analysis of the data

- **Description -** Analysis of data to gain meaningful information
- **Criticality -** This is critical to the final deliverable.
- **Dependencies with other requirements -** This is dependant on having clean data and a database so it is accessible

- **Description -** Make prediction
- **Criticality -** This is the end goal and final deliverable
- **Technical issues -** With all prediction models there is a certain inherent risk that predictions have only a low accuracy, possibly due to unforeseen variables affecting the predicted variable. If this occurs, the model will still take into account relevant variables but would require further enhancement to improve accuracy.
- **Dependencies with other requirements -** Dependant on clean data and good analysis

4. System Architecture

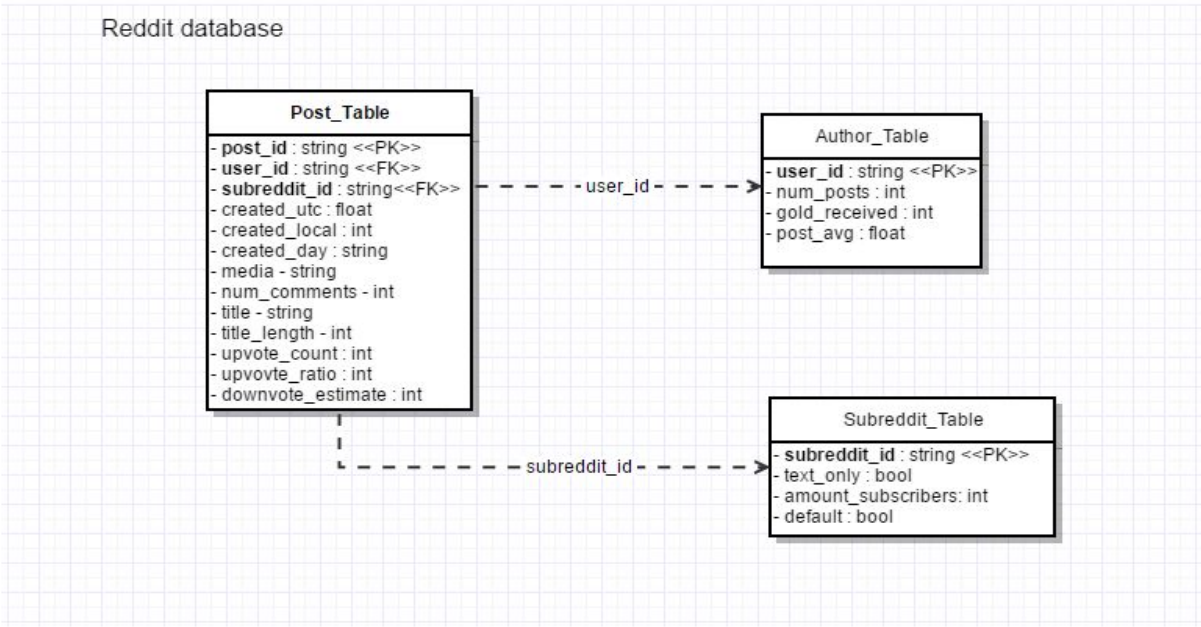
The first diagram shows the planned schema for the first batch of data. All Reddit posts have a unique post ID; this will be used as the primary key for the 'posts' data table. It will also contain any other relevant information relating to the post.

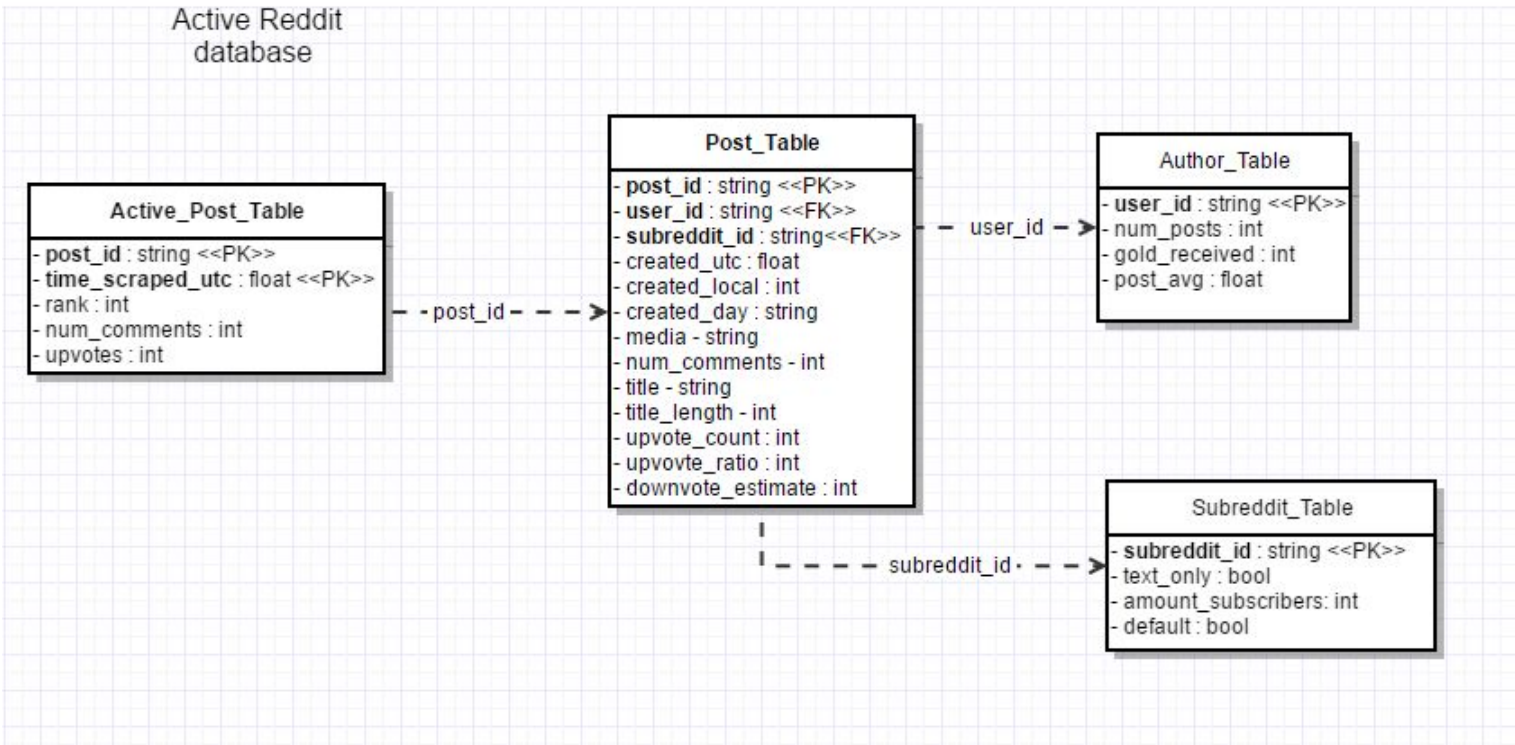
The user name of the author of the post is also unique, so metadata relating to the author will also be collected. This will include the number of posts they have made and their average score. These values will likely be of interest during the data analysis, as the user attributes which contribute to higher scores could contribute a layer of accuracy to predictions of popularity.

Subreddits also have unique IDs; this can be used to store information on each subreddit that appears in the data, information such as amount of subscribers, whether it is a default Subreddit or if it is text only.

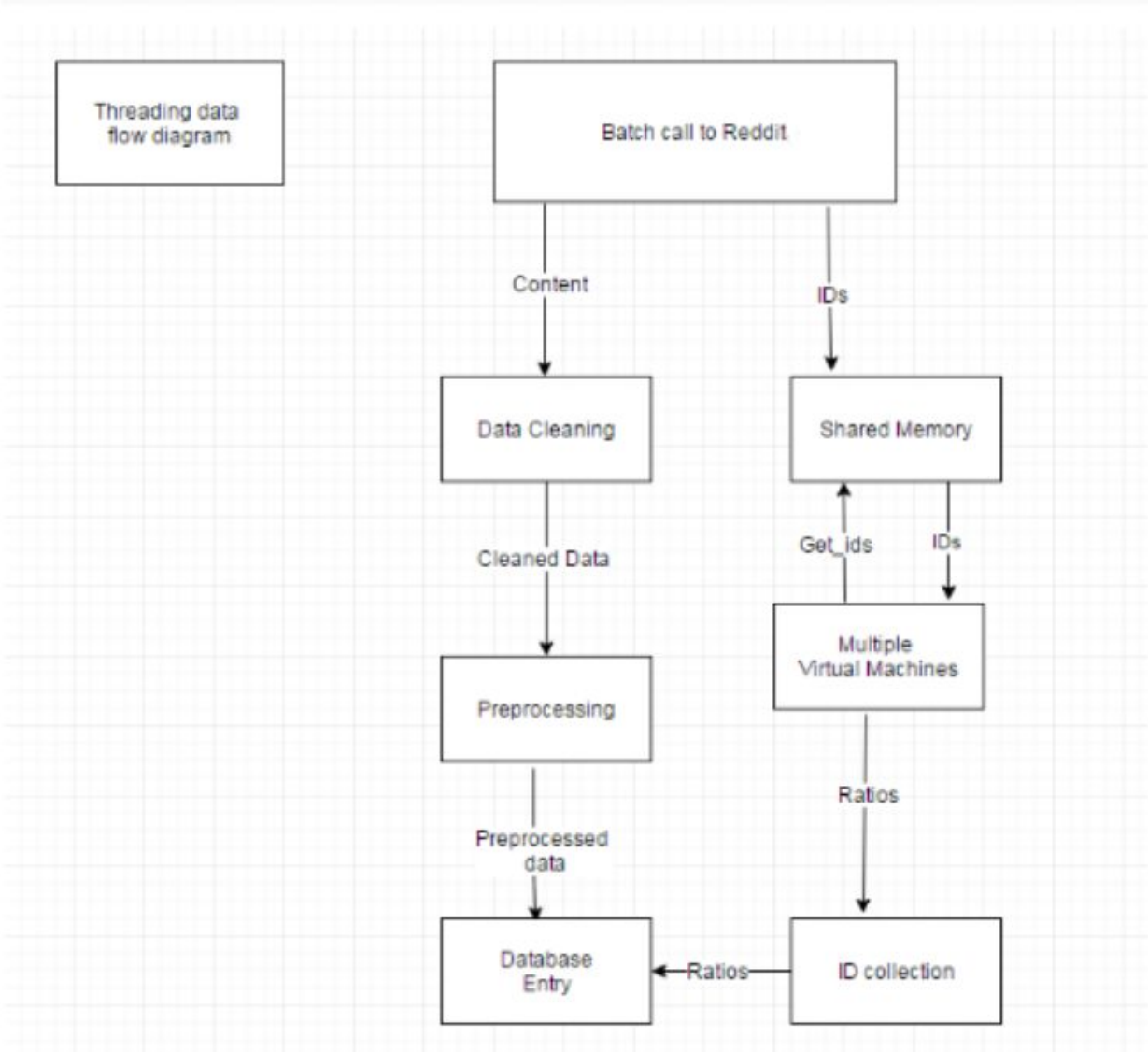
The second database diagram describes the schema for actively scraping Reddit's frontpage, i.e. the second stage of data collection. It has a very similar schema to the first, however it shows an extra data table called Active_Post_Table which has a composite key on post_id and time_scraped. It records the rank of a post out of 100

at the time of scraping and various other attributes liable to change, such as: number of upvotes and number of comments. It does not record attributes, such as the post author, which cannot change as this would be data duplication.





5. High-Level Design



This Data flow diagram outlines how the parallel processing pipeline for the second batch of data will work. As reddit has an IP based API restriction, multi-threading will be used.

The main thread will perform the batch calls to reddit and then pass the content to a separate thread which will handle data cleaning. This cleaned data will then be passed to another thread which will handle pre-processing and, finally, the clean preprocessed data will be passed to a thread which will handle database entry. The main thread will also write the ID to a memory location shared by two virtual machines. These virtual machines will make individual calls to reddit to obtain additional information required for analysis. The pipeline will speed up this process two-fold. This strategy could be extended if the process is still taking too long.

6. Preliminary Schedule

The major tasks in this project are:

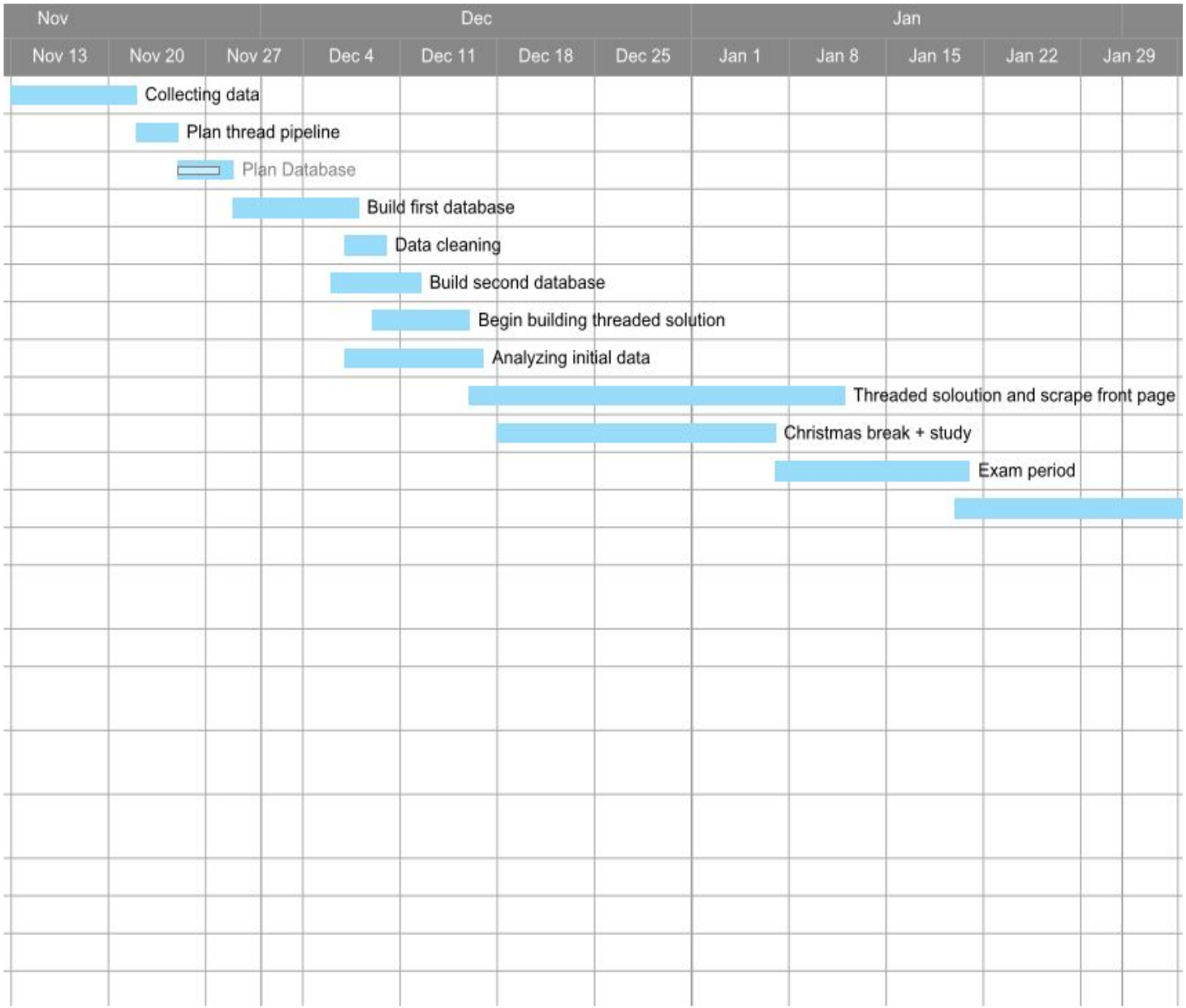
1) **1st set of data**
    - Collecting the data
    - Cleaning the data
    - Building the database
    - Performing in depth data analysis
    - Building the model
    - Writing up results

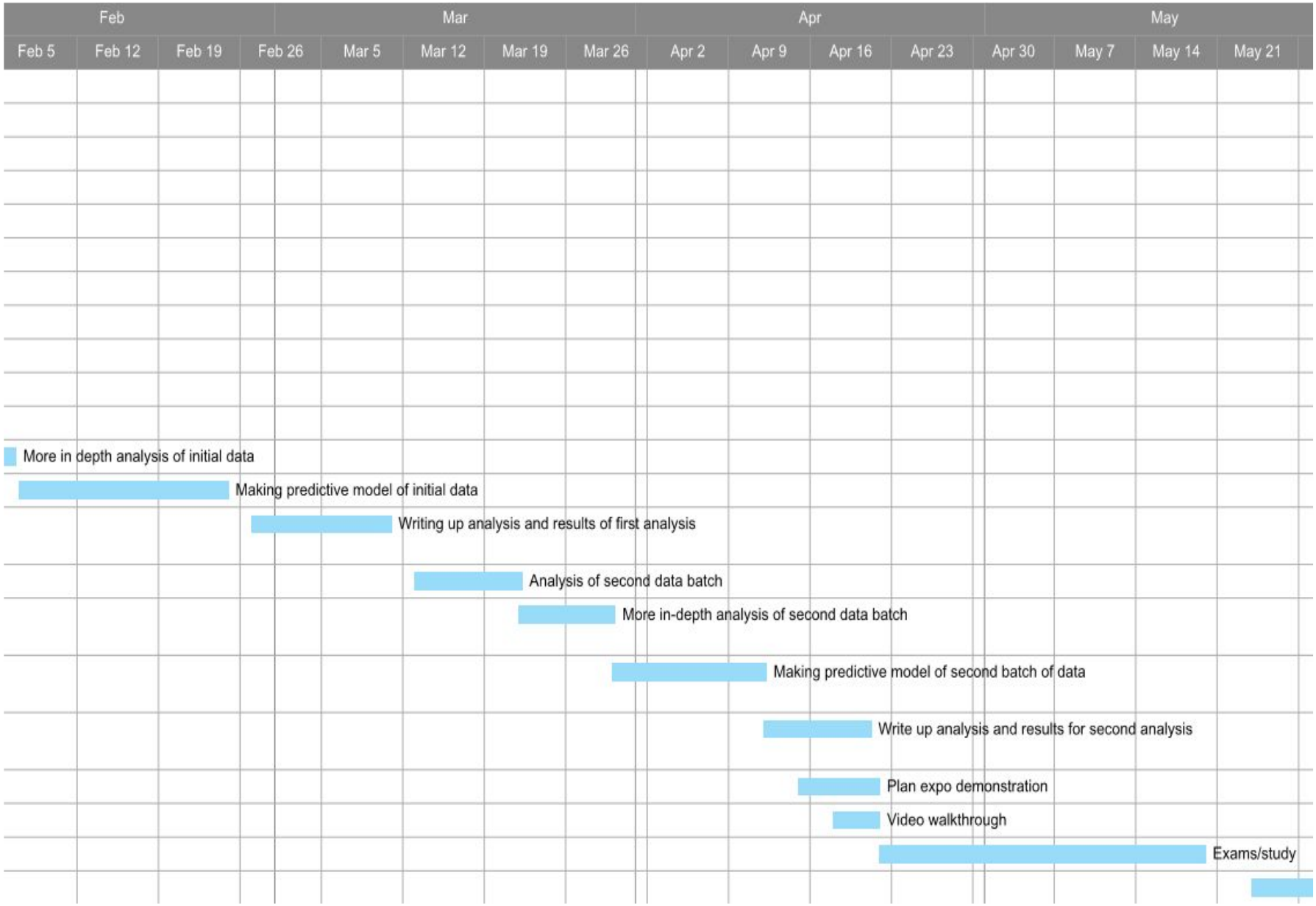
2) 2nd set of data

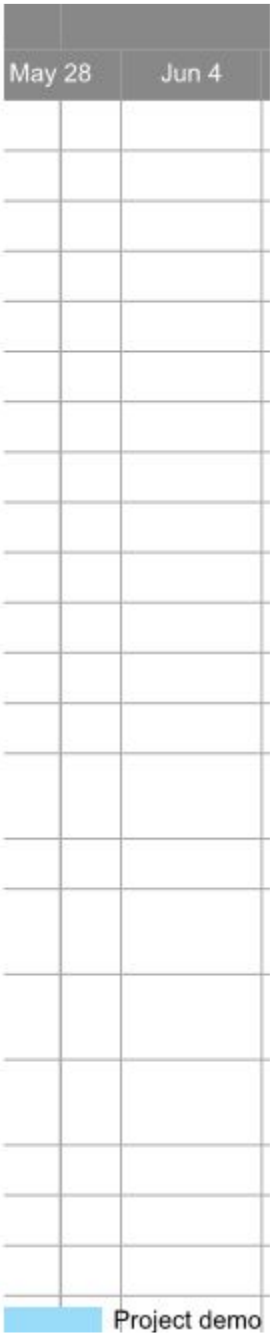
- Building multithreaded pipeline
- Collecting the data
- Clean the data
- Build database
- Perform in depth data analysis
- Building the final model
- Write up results

3) Expo and project demo

- Present results


Nov
Dec
Jan
Nov 13
Nov 20
Nov 27
Dec 4
Dec 11
Dec 18
Dec 25
Jan 1
Jan 8
Jan 15
Jan 22
Jan 29
Collecting data
Plan thread pipeline
Plan Database
Build first database
Data cleaning
Build second database
Begin building threaded solution
Analyzing initial data
Threaded soloution and scrape front page
Christmas break + study
Exam period


Feb
Mar
Apr
May
Feb 5
Feb 12
Feb 19
Feb 26
Mar 5
Mar 12
Mar 19
Mar 26
Apr 2
Apr 9
Apr 16
Apr 23
Apr 30
May 7
May 14
May 21
More in depth analysis of initial data
Making predictive model of initial data
Writing up analysis and results of first analysis
Analysis of second data batch
More in-depth analysis of second data batch
Making predictive model of second batch of data
Write up analysis and results for second analysis
Plan expo demonstration
Video walkthrough
Exams/study


May 28
Jun 4
Project demo