

Architectural Design Document

Ctrl Alt Defeat

June 26, 2023

1 Architectural Design Strategy

1.1 Approach

We have opted to follow the approach of: Design based on Quality Requirements

1.2 Reasons for approach

- This approach allows us to model our system as a solution to an abstract problem (that being the satisfaction of quality requirements). This allows us to make technology-independent design decisions.
- By formulating our design based on quality requirements, we are able to ensure (make a huge step towards ensuring) that the system does indeed meet its quality requirements first and foremost before any code has been written.
- By analyzing the system in terms of quality requirements, it becomes clearer which architectural strategies and patterns (discussed below) are most suitable for the system's implementation.

1.3 Reasons to not follow other approaches

- Since our application is dashboard centric, which limits edge case interaction from users, we found it unsuitable to take the approach of designing the system via the generation of test cases. Furthermore, designing test cases would not necessarily help us in identifying appropriate architectural strategies and patterns (at least not to the extent that our favoured approach allows)
- There are relatively few object-like structures that can be decomposed into suitable hierarchical or relationship structures (at least not to enough of an extent) in our system. Hence we felt that a decomposition approach was not necessarily warranted (at least not over our favoured approach, which we deemed to be the most suitable for our purposes)

2 Architectural Quality Requirements

While almost every possible quality requirement is something that should be aimed for as a general rule, we have identified five quality requirements most important

for our system. The quality requirements below are listed in order from the most to "least" prioritised.

2.1 Usability

The most essential quality requirement of our system is usability. The target user for our application is not necessarily someone with high proficiency with software systems, sentiment analysis and natural language processing, or IT in general (ex: social media managers, small business and restaurant owners, etc) - consequently, the systems needs to be intuitive and non-overwhelming, both in the user interface as well as in the numerical metrics provided. As such, the following measurable requirements must be met (and can be tested by asking a non-IT person to try use the system)

- With zero developer interference, the user should be able to do the following: create a new profile, change the theme of the website, and create a new domain
- With minimal developer interference, the user must be able to do the following: create a new source to pull data from for analysis, refresh the source for new data, and delete the source
- With zero developer interference (and the user having access to Google), the user must be able to do the following: Interpret each of the statistics on the dashboard, and interpret the chart visualizations of each statistic
- After using the system for half an hour, the user should be able to: switch between viewing a single source and viewing the entire domain, and log out

2.2 Security

As will be explained below, our system will be hosted on a virtual machine on a public IP address - experience and measuring has shown that the network of computers on which the virtual machine lies is subject to up to 17 000 cyber-attacks every 24 hours. Consequently, the security of the system needs to be state of the art in order to prevent would-be attackers from wreaking havoc upon the system. To measure the strength of our security, the following requirements must be met:

- No user that interacts with the system may access the profile or domain/source data of another user
- Should the database be leaked, all password data must be hashed to obfuscate sensitive password information
- The system should be immune (to the most reasonable, state-of-the-art extent possible) to the following attacks: CSRF, SQL injection, and clickjacking
- No user shall have the ability to make any modification to the system (ie: there will be no admin or super user)

2.3 Performance

Performance is an important quality requirement of our system. Our system includes a high degree of data processing and analysing (which can be a time-intensive activity) in addition to the ability to fetch data from external sources (which can also add precious seconds to execution time). Slow performance ruins the user's experience (pertaining to usability) and renders the app more frustrating than useful. Consequently, the following measurable requirements must be realised:

- When a user refreshes a source, the total time taken for the data to be retrieved from an external API, preprocessed, analysed, aggregated, and returned to the user must never exceed 20 seconds.
- When a user performs any other action that requires database interaction (ie: profile changes, accessing existing sentiment data, performing CRUD operations on domains and sources), the total time taken for the data to be queried and returned to the user cannot exceed 5 seconds.

2.4 Scalability

Scalability goes somewhat hand in hand with performance in that the system should not become overly encumbered (resulting in slower performance) when a high number of users are interacting with the system. To measure whether this is attained, the following requirements must be observed:

- When the system has 50 concurrent users, the performance metrics specified above cannot exceed 1.1 times their normal threshold time
- When the system has 100 concurrent users, the performance metrics specified above cannot exceed 1.2 times their normal threshold time
- When the system has 500 concurrent users, the performance metrics specified above cannot exceed 1.5 times their normal threshold time

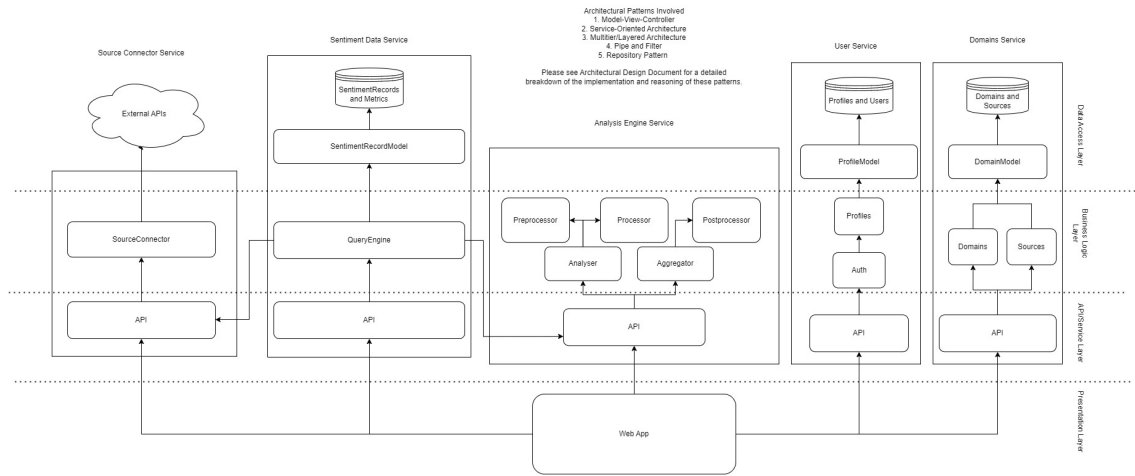
2.5 Modifiability

Ideally, this system will be useful to a fair number of people, who will in turn provide feedback for functionality, sources, and metrics they would like to be able to investigate using the system. Consequently, modifiability of the system is a desirable quality requirement in order to ensure the continuous growth and evolution of the system. Modifiability in this case is to be defined by the following measurables:

- It should not require more than 24 total work hours to add a new type of source for analysis (ex: Twitter or Reddit)
- It should not require more than 15 total work hours to add a new metric/statistic to be included in the analysis
- It should not require more than 10 total work hours to add a new type of visualization for a particular statistic
- Changing database for any service should not exceed more than 10* total work hours (*though this is dependent on the type and amount of data to be transferred)

3 Architectural Design and Patterns

3.1 Diagram



Please see a description of the employed architectural patterns and tactics below.

4 Architectural Strategies

We have adopted aspects of several architectural styles, patterns and strategies to help realise our system and its quality requirements. Tactics additionally are grouped by architectural patterns.

4.1 Service-Oriented Architecture

4.1.1 Description

The overarching software architecture is that of a Service-Oriented Architecture. Our system consists of five different, independently deployable unit, each of which is responsible for providing (a) different services(s).

4.1.2 Quality Requirements Addressed

- **Scalability:** The use of a Service-Oriented Architecture promotes scalability of the application by distributing the computational responsibilities across different deployed units (tactic: distributed processing) - thus helping reduce bottleneck.
- **Performance:** The separation of services allows services that require higher computationally expense to be scaled vertically or horizontally (tactic). Furthermore distributed processing of tasks ultimately improves performance.
- **Modifiability:** A SOA takes separation of concerns (tactic) to an extreme, and allow for individual services to be modified totally independently of one another.

4.2 Multitier/Layered Pattern

4.2.1 Description

A Multitier architecture is achieved by physically separating the modules responsible for presentation, business logic, and data logic. Furthermore, this architecture is Layered in the sense that each layer provides services to the layer below it to use. The layers/tiers are as follows: Presentation layer (frontend UI), Service Layer (public APIs for each service), Business Logic Layer (to handle the application logic), and the Data Access Layer (to handle the database interaction)

4.2.2 Quality Requirements Addressed

- **Modifiability:** Once again, separation of concerns (tactic) is achieved (and is more finely grained) by separating modules physically by responsibility
- **General:** It is good practice to separate a system into layers such that updates to one layer in one unit does not affect others. Furthermore this helps realise the Agile methodology which requires that modules are updated and developed rapidly.

4.3 Model-View-Controller Pattern

4.3.1 Description

The use of Angular as a frontend technology (discussed below) inheritedly makes use of the MVC pattern.

4.3.2 Quality Requirements Addressed

- **Usability:** By separating data, visuals, and logic, is it possible to build a highly responsive application with stunning and intuitive visuals.

4.4 Pipe-and-Filter Pattern

4.4.1 Description

The pipe-and-filter architectural pattern is a means to conceptually reason about the sentiment analysis pipeline. The pipeline is as follows: 1. Data is pulled from an external API 2. The data moves through the Query Engine and is formatted and pushed further along the conceptual pipeline 3. As part of the Analyser, the data is preprocessed using NTK 4. As part of the Analyser, the data is subject to a number of sentiment analysis models to produce metrics 5. As part of the Aggregator, the data is combined in a meaningful and aggregated manner 6. The next step conceptually is that this fully transformed data is returned to the user on the frontend and presented visually

4.4.2 Quality Requirements Addressed

- **Performance:** The use of this conceptual pipeline structure is conducive to parallel processing, an architectural tactic (since this pipeline spans across three

different deployable units) improving the time taken for data to be analysed overall on average

- Scalability: Closely related to performance, the use of a somewhat pipe-and-filter structure allows for pipelining (tactic) in the sense that while data is in one stage of the pipeline, different data may be simultaneously processed in a different component of the pipeline - improves overall scalability and performance when multiple users are invoking the pipeline simultaneously
- Modifiability: By implementing the sentiment analysis pipeline as a series of filters/components, one can make modifications to one part of the pipeline (ex: preprocessing) without affecting any other stage of the pipeline

4.5 Repository Pattern

4.5.1 Description

As part of the Data layer of the application (see Multitier/Layered Pattern) we have opted to implement the Repository pattern such that we have an intermediary between the databases and the rest of the application. This is done for all deployable units that include a database (of which there are three).

4.5.2 Quality Requirements Addressed

- Modifiability: By placing an abstraction between the database and the application logic (tactic), the rest of the system can be totally agnostic of the database technology used, meaning that changes to the database require a minimal amount of code to be changed, namely only the repository/model modules that interact with the database, the rest of the application can remain unchanged.
- Performance: Since we can thus guarantee that interaction with the databases takes place in a consistent manner, we can preemptively create optimized database queries (tactic: database optimization) for each predefined database-interaction method defined in the repository/model modules (since the application will always access the databases via these modules). By using optimal queries will the performance of the application improve.

4.6 Important Notice regarding other identified Quality Requirements

Note that no pattern directly addresses the quality requirement of security. This is because this quality requirement is addressed by the choice of technology (detailed below). Furthermore, the important quality requirement of Usability is only directly addressed via one pattern above - please note however that this is not an oversight. Instead, Usability is so core to our application that it is mainly ingrained into the business logic and technology choices of the system.

4.6.1 Security

- Authentication is token based (tactic). JWT tokens will be used to authenticate a user across multiple deployed and independent services.
- Backend technology choice of Django (discussed below) is designed with security in mind, employing various mechanisms to combat common security vulnerabilities

4.6.2 Usability

- Attention to detail and smooth-ness when designing the UI
- Sentiment analysis model outputs are appropriately transformed on the backend such that they are easy to interpret and visualise.

5 Architectural Constraints

5.1 Client-Defined

- The system needs to consist of at least two separate deployable units.
- The system must be deployed to a private virtual machine with a public IP address (provided by Southern Cross Solutions). This machine is subject to extremely frequent cyber-attacks for which the system must protect against.
- The system must make use of at least one NoSQL document-based database

5.2 Hardware and Operating System Constraints

- The system is to be hosted on a single virtual machine (running on a single physical server). At a later date this may be expanded to make use of multiple virtual machines running on different physical machines.
- The system must be suitable for and run on an Ubuntu-style Linux operating system

6 Technology Choices

6.1 Frontend Technology

Our frontend consists of the use of Angular which is an open source web application development framework that was developed by google. Angular is used to build dynamic and responsive applications.

6.2 Angular

The reasons below are some of the key points as to why Angular is best suited for our application

- Angular is a complete framework which means that it doesn't have to rely on any other frameworks to complete any functionality front end wise, this is good because we already have multiple deployable units and having a singular front end framework would be a great way to declutter the project and not have a steep learning curve with multiple frontend frameworks.
- Angular promotes a Component based architecture which is great in terms of our project because as mentioned above we have multiple deployable units and as Angular promotes component-based architectures this will be good as we will need modularity, code reusability and separation of concerns.
- Angular has a large component for testing, including unit tests, integration tests and end to end tests, angular makes use of the Jasmine and Karma for writing and running tests which allows for higher quality and reliability of applications developed with Angular

6.3 Backend Technology

Our backend consists of five different deployable units, each developed using Django to perform the business logic, data logic, and API functionality. The reasoning behind this technology choice is explained below.

6.3.1 Django

The following reasons are the main driving force this design decision:

- Django is, by design, highly secure 'out of the box' and provides comprehensive protection for all the most common security threats (SQL injection, CSRF, clickjacking, etc). This is of particularly high value to us since (as pointed out in the architectural constraints) the system will be hosted on a virtual machine with a public IP that is subject to frequent cyber-attacks. In order to minimize risk wrought on by inexperience or oversight, a highly secure framework such as Django is a very suitable choice for our purposes.
- Django is an entirely Python development environment. Since our system includes a significant portion of data manipulation (across all services), Python will improve the ease with which we are able to develop the system. Furthermore, Python has a number of fantastic libraries for Natural Language Processing and Machine Learning (ex: NLTK, PyTorch, etc), this language is a great choice to handle the application's sentiment analysis component. Additionally, all group members have high proficiency with Python, allowing us to lean into our strengths as programmers.
- Django is designed for rapid and smooth development, this is suitable and desirable since we are adopting the Agile Software Process. That is, by developing our system using Django, we are able to rapidly prototype and build on different aspects of our system - allowing for maximum flexibility, while speeding up iteration-by-iteration improvement (highly desirable and necessary trait to successfully adopt the Agile process)

However, no technology choice is ever perfect and there are some trade-offs which need to be minimized.

- Django accomplishes the advantages pointed out by points 1 and 3 above by being a 'batteries included' framework. Unfortunately this means that Django is more heavyweight than other Python frameworks such as Flask. A more lightweight framework may be more desirable since we are adopting a service-oriented architecture (ie: not every service will have use of all of Django's built-in features). We do however believe that the advantages described above outweigh this point, since each service in the architecture will leverage at least a few of Django's native features to great effect.
- Django is somewhat opinionated about the structure of the application which once again is not particularly desirable for a service-oriented architecture - however, this point is easily negated as it is entirely possible (and fairly easy) to add (through the use of user-defined Python modules) and remove components to accomplish our architectural structure.

6.4 Database Technology

Our system comprises three separate databases, each running on different deployable units and each serve a different purpose. The description and reasoning of each is below.

6.4.1 Sentiment Records Database

This is a NoSQL document-based MongoDB database that runs on the Data Warehouse deployable unit. It is the main and largest database on the application, and contains all the sentiment records/data that have been analyzed, as well as the sentiment metrics that have been computed for each record.

- A NoSQL database lends itself to this data because of unpredictable size and variety of the sentiment records that are retrieved online.
- Furthermore, the metrics that are computed for each record are closely tied to the record itself (since the record will never be accessed outside the context of checking the sentiment metrics). By leveraging the document structure allowed by MongoDB, it is possible to store both the data and the metrics as a single document/unit.
- Potentially most importantly, a NoSQL database (especially the document-based MongoDB) allows for flexibility and modifiability of the metrics computed on a particular sentiment record. For example, in the future it may be desirable to include an objectivity-subjectivity score on new articles that have been analysed. The flexibility of MongoDB would make this modification extremely simple, as no change needs to be applied to the database schema, or to any of the other existing documents in the database.

6.4.2 Domains and Sources Database

This is a NoSQL document-based MongoDB database that runs on the Domains deployable unit. It is responsible for storing the information and structure of a domain, as well as the sources specified within it (including the information and parameters for those sources)

- Since domains effectively act as 'folders' for sources, MongoDB is once again a suitable choice of technology, as it allows us to store sources as nested objects within a domain (which is the document). This helps accomplish 1. domains physically group their associated sources (and easily catering for a variable number of sources), as opposed to a SQL approach, which would almost certainly require the use of complex joins, and 2. accomplish some form of query/database optimization in that since domains will ALWAYS be accessed in order to retrieve ALL their contained sources, we avoid consistently having to use joins like we would in a SQL database, instead by retrieving a single document (domain) we have all the information we required.
- Furthermore since the parameters for different sources vary both in number and in type/name, by leverage a MongoDB database we do not need to create some static schema that needs to cater for all possible parameters. Instead we can simple create an object that contains key value pairs of the parameter name and its value.

6.4.3 Profiles, Users, and Authentication

This is a SQL Postgre database that runs on the Profiles deployable unit. It is responsible for storing a user's profile information and preferences, linking it to their domains in the above database, and managing user and authentication data.

- The data to be stored by this database is highly structured and well-defined, and consequently a SQL database is highly suitable for this purpose.
- Additionally, Django (backend technology) provides seamless and secure integration with Postgre 'right out of the box' and hence we are able to leverage this predefined interaction in our development. (It is worth noting that this seamless integration is designed with security in mind, providing appropriate and comprehensive protection for database attacks such as SQL Injection)