Report on the implementation of 4chan and Reddit Crawler

Mukhil Venkataramanan
SUNY Binghamton
Binghamton, USA
mvenkatarama@binghamton.edu

Gurusaran Venkatachalam Rajarajacholan
SUNY Binghamton
Binghamton, USA
gvenkatachal@binghamton.edu

Sai Prakash
SUNY Binghamton
Binghamton, USA
snalubolu@binghamton.edu

ABSTRACT

This report explains the design and implementation of two data collection system that periodically retrieves data from Reddit's sports subreddits and 4chan's sports board using a custom-built API client. The system will preprocess, clean, and filter the data before storing it in PostgreSQL-timescaleDB. The system uses Faktory, a producer-consumer model to continuously collect data by scheduling crawling jobs periodically almost at even intervals. The report elaborates the Data sources used in the data collection system, the system architecture, implementation of the system, changes in the system implementation since proposal, data collected over time and gives a brief overview of the collected data from preliminary data analysis.

KEYWORDS

Data Collection, Reddit API, 4chan API, Sports Subreddits, Report, Implementation, PostGreSQL, Faktory, Docker, Python

1 INTRODUCTION

The proliferation of data on social media platforms such as Reddit and 4chan provides a rich source of information for real-time analysis and insights. These are two of the most influential online platforms, playing a significant role in shaping public opinions, and trends, especially within niche communities. Collecting data from these platforms, particularly in the sports category, provides us a wealth of user-generated content that is highly unfiltered and diverse.

Reddit, with over 57 million daily active users worldwide, hosts communities (subreddits) on nearly every topic imaginable. Approximately 5.7 percentage of global internet users interact with Reddit, making it a crucial source for understanding popular sentiment, emerging trends, and user engagement in real time. 4chan, while smaller with around 20 million monthly users, is known for its anonymity and rapid dissemination of viral content. It often acts as an incubator for cultural trends and internet memes. Although only a small fraction of the global internet population uses 4chan, its influence on broader internet culture is disproportionately large.

The real-time, organic conversations that take place on these platforms offer invaluable insights into not only what people are discussing but also how ideas and sentiments spread across different demographics.Our goal is to create an efficient pipeline that can handle large volumes of data collected from these platforms in the category of sports and provide actionable insights for various stakeholders, including sports analysts and researchers.

2 DATA SOURCES

Data will be gathered from the following sources:

- **Source:** Reddit
  - **API:** https://www.reddit.com/dev/api/
  - **Subreddits**
    - **Sports:** https://www.reddit.com/r/sports/new.json,
    - https://www.reddit.com/r/sports/comments/[article-id] .json
    - **Cricket:** https://www.reddit.com/r/Cricket/new.json,
    - https://www.reddit.com/r/Cricket/comments/[article-id] .json
    - **Soccer:** https://www.reddit.com/r/Soccer/new.json,
    - https://www.reddit.com/r/Soccer/comments/[article-id] .json
    - **Football:** https://www.reddit.com/r/Football/new.json,
    - https://www.reddit.com/r/Football/comments/[article-id] .json
    - **Tennis:** https://www.reddit.com/r/tennis/new.json,
    - https://www.reddit.com/r/tennis/comments/[article-id] .json
- **Source:** 4chan
  - **API:** https://a.4cdn.org/
  - **4chan Boards**
    - **Sports:** https://a.4cdn.org/sp/archive.json
    - **Threads of Sports:** https://a.4cdn.org/sp/thread/[thread-id].json

3 ARCHITECTURE DIAGRAM

The architecture of the system is shown in Figure 3. This diagram outlines the major components and their interactions within the system.

- **Data Sources**: Reddit API and 4chan API - The system fetches data from both platforms via their respective APIs
- **API client**: This part of the system automates the data collection process through a scheduling mechanism.
- **Data pre-processing**:Once data is collected from Reddit and 4chan, it goes through a pre-processing phase. This step involves cleaning the data, filtering out irrelevant content, and normalizing the data for consistency and further analysis.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

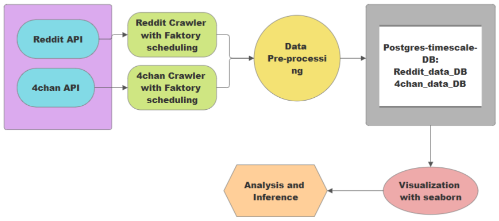
Reddit API
4chan API
Reddit Crawler with Faktory scheduling
4chan Crawler with Faktory scheduling
Data Pre-processi ng
Postgres-timescale-DB: Reddit_data_DB 4chan_data_DB
Visualization with seaborn
Analysis and Inference

Figure 1: System Architecture Diagram

- **Data Storage**: After pre-processing, the clean data is stored in a database. A PostGreSQL Timescale database, well-suited for handling data that has been collected continuously.
- **Data Visualization**: Data from database is retrieved and visualized using a visualization tool. This helps to create graphs, charts, and other forms of data representation to explore trends, insights, sentiment analysis, etc.
- **Analysis and Inference**: Finally, insights and conclusions are drawn based on the collected and visualized data. The analysis phase might involve sentiment analysis, trend detection, and other types of statistical or machine learning models to understand the data better.

4 IMPLEMENTATION

4.1 Technologies Used

Our systems are based on a modular architecture using Docker, Python, PostgreSQL TimescaleDB, and Faktory. The crawlers are responsible for collecting posts, comments, and metadata from the respective platforms.

4.2 System Architecture

The architecture of the data collection system follows a modular approach, allowing easy scaling and maintenance. The core components include:

- **Python Backend**: Both crawlers are implemented in Python. The Reddit crawler leverages Reddit's API for data retrieval, while the 4chan crawler uses HTTP requests to scrape data from specific boards.
- **PostgreSQL TimescaleDB**: The collected data is stored in a PostgreSQL database enhanced with TimescaleDB extensions to efficiently handle time-series data. This ensures that data is stored in a structured format, allowing easy querying and analysis.
- **Faktory**: Faktory, a job management system, is used to handle the continuous scheduling and execution of crawling jobs. Jobs are created at regular intervals (e.g., every five minutes for Reddit) and consumed by workers that execute the crawling tasks.
- **Docker**: All services, including PostgreSQL TimescaleDB and Faktory, are containerized using Docker, providing portability and ease of deployment. This ensures consistent behavior across development and production environments.

4.3 Data Collection Process

The Reddit and 4chan crawlers are designed to run in parallel. They collect data at intervals defined by Faktory, which schedules jobs to ensure regular data collection. Reddit posts are crawled every five minutes, while 4chan boards are crawled at similar intervals.

The system consists of two independent crawlers:

- **Reddit Crawler**: This crawler uses Reddit's API to collect posts, comments, and metadata from predefined subreddits. Each job fetches the latest posts and stores them in the database. Posts are enriched with additional metadata, including the author, score, and number of comments.
- **4chan Crawler**: This crawler uses 4chan APIs mentioned in the open public gitHub repository of 4chan API documentation. Special care is taken to handle errors, and rate limits imposed by 4chan's servers.

For both crawlers, data is inserted into the database through SQL queries. The database schema is designed to store posts, comments, and their associated metadata, indexed by time.

The system continuously runs through Faktory, where producers schedule jobs and consumers fetch and process them. The jobs are distributed across multiple workers, each capable of handling concurrent crawling operations.

5 CHANGES SINCE PROPOSAL

This sections describes the changes in the implementation of the system from the proposal. Several challenges were encountered during implementation.

- **Change of Storage:** We used PostgresSQL - timescaledb as the Database to store the collected data, which is different from MongoDB which was proposed initially. The change is made after conducting analysis on the features of PostgreSQL and MongoDB. PostgreSQL seems to be more suitable for this project of data collection.
- **Change of scheduler:** We initially planned on using cron jobs to schedule the API calls at regualr interval of time, but upon analysis, we get to know that there are limitations in using cron jobs which may cause unwanted overhead in handling the errors while collecting data from 4chan.

6 PROJECT STATUS

Currently, the data collection systems for both Reddit and 4chan are fully operational and running continuously. Both crawlers are built using Python, with data being stored in PostgreSQL TimescaleDB, and the workflow managed through Faktory for job scheduling.

- **Reddit Crawler:** The Reddit crawler is successfully collecting data from multiple subreddits including `sports`, `cricket`, `soccer`, and `football`. The crawler runs periodically every 5 minutes to fetch the latest posts, along with their associated comments.
- **4chan Crawler:** The 4chan crawler is actively collecting data from the `/sp/` (sports) board. It retrieves information from ongoing threads, capturing the board name, thread number, post number, and post content. The collection process is scheduled periodically to gather updates on threads and posts continuously.

- **Challenges:** During the implementation phase, some challenges were encountered, particularly with parsing the collected data and storing it into the database appropriately. These issues have been addressed and database insertions are handled efficiently.

In summary, both crawlers are currently running smoothly, and the data collection process is progressing as planned. Regular checks are being conducted to ensure the systems continue to function correctly and to monitor the volume and quality of the collected data.

7 DATA FIELDS STORED

This section explains which data fields are collected from 4chan and reddit and explains it's significance in analysis.

7.1 Reddit

For the posts table:

- **Post ID:** Unique identifier for each post. This is essential for tracking and avoiding duplicate entries.
- **Title:**: The title gives a brief description of the post's content, making it crucial for text analysis. It also allows for easy post retrieval based on search queries or keywords.
- **Content (selftext)::** The main body of the post, which contains the detailed text that users write. It is important for analyzing the actual discussions happening in a subreddit. By storing the full content, we can extract insights like common topics, user interests, or text sentiment.
- **Author:** The username of the person who made the post. Collecting author information helps to analyze user behavior patterns, track the activity of specific individuals, and correlate authorship with post quality, engagement, or popularity.
- **Score:** The post score represents the total number of upvotes minus downvotes. This is a key metric for understanding community engagement and gauging the overall popularity or reception of a post.
- **Comments:** Number of comments on the post. The comment count indicates the engagement level of the community with a particular post, providing insight into the relevance and discussions generated by the content.
- **Created At:** This is the timestamp indicating when the post was created. This field is essential for time-series analysis, understanding patterns of activity over time, and examining trends within a subreddit.
- **Subreddit:** The specific subreddit where the post was made. This is important for categorizing data and allowing us to compare patterns across different communities.
- **URL:** The direct URL to the post, useful for retrieval, validation, or manual inspection. Having this stored allows for easy referencing of the original content on Reddit.

For the comments table:

- **Post ID:** The foreign key connecting the comment to its parent post. This establishes a relationship between posts and their comments, allowing us to perform queries that analyze comment activity on a per-post basis.
- **Comment ID:** Unique identifier for each comment. Similar to the post ID, this ensures that each comment is uniquely identifiable and prevents duplication.
- **Author:** The username of the commenter. Tracking who is making comments can reveal patterns in engagement and contribute to user-based analysis, such as identifying frequent contributors or potential influencers.
- **Body** The text of the comment. This allows for sentiment analysis and deeper textual analysis of discussions happening in response to posts. It's crucial for understanding the nature of conversations around posts.
- **Score:** The score of the comment, representing upvotes minus downvotes. Just like post score, this helps gauge the community's reaction to individual comments, allowing for an understanding of which comments contribute most to the discussion.
- **Created At:** Timestamp indicating when the comment was made. This field is vital for studying patterns over time, understanding how quickly discussions grow after a post is made, and performing time-based analysis.

7.2 4chan

For posts table:

- **Board:** This field indicates which specific 4chan board (such as /pol/, /b/, or /g/) the post belongs to. Each board has a distinct community, rules, and content focus. Collecting this data allows for filtering and analyzing posts by the board, which helps understand different discussion trends, content types, and user engagement across various topics. It's essential to identify patterns and differences in discourse depending on the context of the board.
- **Thread Number:** A thread on 4chan serves as a container for a series of related posts. The thread number is vital for grouping posts together, as it ensures that all related posts can be analyzed as a single conversation. This data can be used for analyzing the lifespan of a thread, how quickly discussions evolve, or how viral certain threads become. It's also crucial for identifying influential threads within a board.
- **Post Number:** The post number provides a unique identifier for each individual post within a thread. This field is necessary for maintaining the hierarchy of posts, especially when analyzing post reply chains or creating visualizations of user interactions within threads.
- **Data**: The content of the post (or "data") includes the text, images, or other media shared by users. This is the most important field for sentiment analysis, text mining, topic modeling, and other content analysis techniques. Capturing the post data enables a deep dive into what users are discussing, the tone of their interactions, and emerging themes or controversial topics within the boards.

8 DATA COLLECTION OVER TIME

The following plot shows the number of posts collected per day over the past two weeks.

8.1 Reddit Data

This section shows how much data collected from reddit over time.

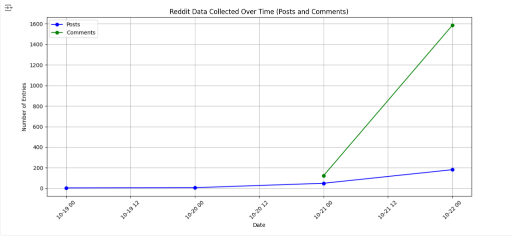

Figure 2: Reddit Data Collected Over time

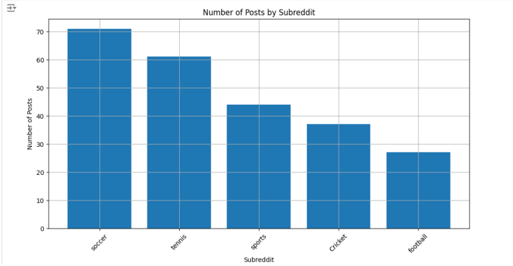

Figure 3: Posts per Subreddit

8.2 4chan Data

This section shows how much data collected from 4chan over time.

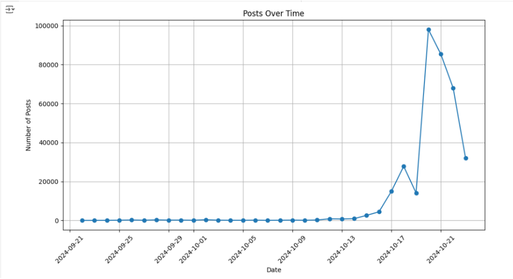

Figure 4: 4chan Data Collected Over time

9 PRELIMINARY DATA EXPLORATION

We have collected a significant volume of data over time. As of the writing of this report, the Reddit crawler has collected over 500 posts and 3500 comments while the 4chan crawler has collected 400k posts. Figure 4 shows the growth in data collected over time.

10 NAPKIN MATH

This section describes the updates projections of data collection volume over the time.

- **Reddit**: We have collected reddit data for 2 days and the collected data sums upto to size 3MB(seems pretty low). Hence roughly, we have collected 1.5MB of data per day and which in-turn means that we will be collecting around 12MB to 15MB of data per week. Since the major political events are nearing, we can have an upper bound of 25MB per week. We are planning to change the subreddits from where the data is being collected, hence there might be a significant change in this napkin math, but doesn't exceed 150MB to 200MB of data per week.
- **4chan**: We have collected 4chan data for 4 days and the collected data sums upto to size 120MB. Hence roughly, we have collected 30MB of data per day and which in-turn means that we are collecting 210MB of data per week. Since the major political events are nearing, we can have an upper bound of 300MB per week.

11 CONCLUSION

In conclusion, the implementation of both the Reddit and 4chan crawlers has been successful, with continuous data collection running smoothly. Despite initial challenges related to concurrency, error handling, and scheduling, the system is now stable. Preliminary exploration of the data suggests that the crawlers are capturing relevant and valuable data, with potential for deeper analysis in the future.

Next steps include further data cleaning, enhancing analysis pipelines, and scaling the system to include more subreddits and boards. With improved data projections, the system is expected to continue collecting valuable data for ongoing and future analyses.

REFERENCES

[1] **4chan API documentation:** *https://github.com/4chan/4chan-API.*
[2] **Official Reddit API documentation:** *https://www.reddit.com/dev/api/.*
[3] **Faktory Official Github Page:** *https://github.com/contribsys/faktory.*