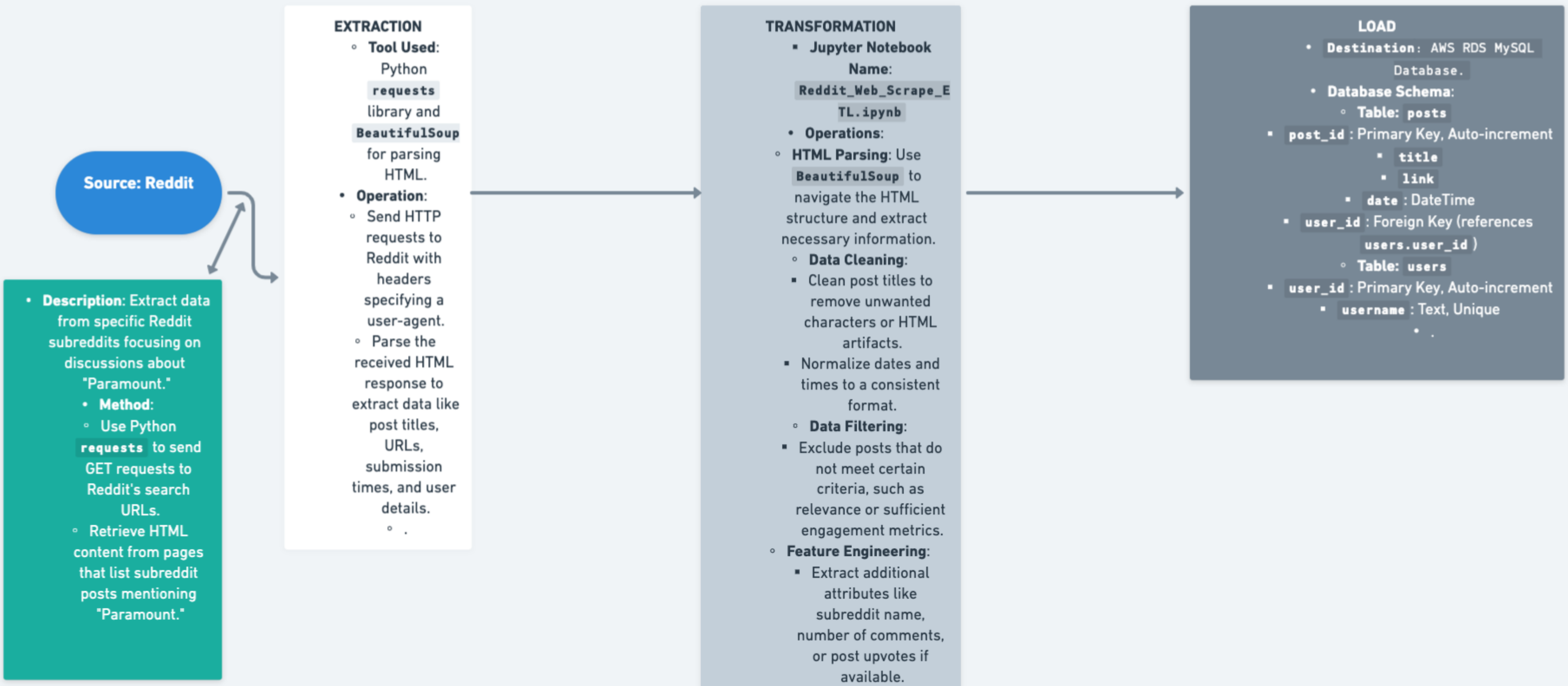

Source: Reddit

- **Description**: Extract data from specific Reddit subreddits focusing on discussions about "Paramount."
  - **Method**:
    - Use Python `requests` to send GET requests to Reddit's search URLs.
    - Retrieve HTML content from pages that list subreddit posts mentioning "Paramount."

EXTRACTION

- **Tool Used**: Python `requests` library and `BeautifulSoup` for parsing HTML.
- **Operation**:
  - Send HTTP requests to Reddit with headers specifying a user-agent.
  - Parse the received HTML response to extract data like post titles, URLs, submission times, and user details.
  - .

TRANSFORMATION

- **Jupyter Notebook Name**: `Reddit_Web_Scrape_ETL.ipynb`
- **Operations**:
  - **HTML Parsing**: Use `BeautifulSoup` to navigate the HTML structure and extract necessary information.
  - **Data Cleaning**:
    - Clean post titles to remove unwanted characters or HTML artifacts.
    - Normalize dates and times to a consistent format.
  - **Data Filtering**:
    - Exclude posts that do not meet certain criteria, such as relevance or sufficient engagement metrics.
  - **Feature Engineering**:
    - Extract additional attributes like subreddit name, number of comments, or post upvotes if available.

LOAD

- `Destination: AWS RDS MySQL Database.`
- **Database Schema**:
  - **Table**: `posts`
    - `post_id` : Primary Key, Auto-increment
    - `title`
    - `link`
    - `date` : DateTime
    - `user_id` : Foreign Key (references `users.user_id`)
  - **Table**: `users`
    - `user_id` : Primary Key, Auto-increment
    - `username` : Text, Unique
    - .

Made with Whimsical