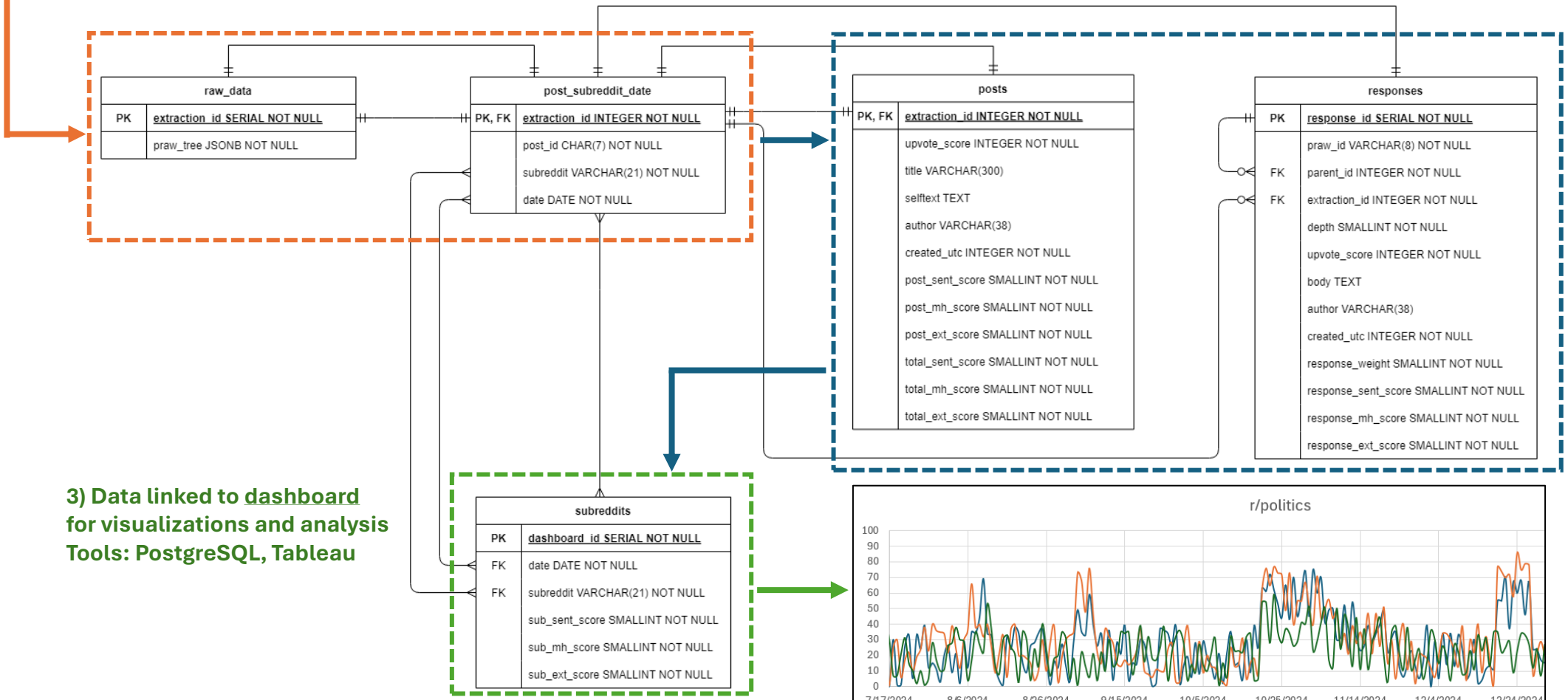
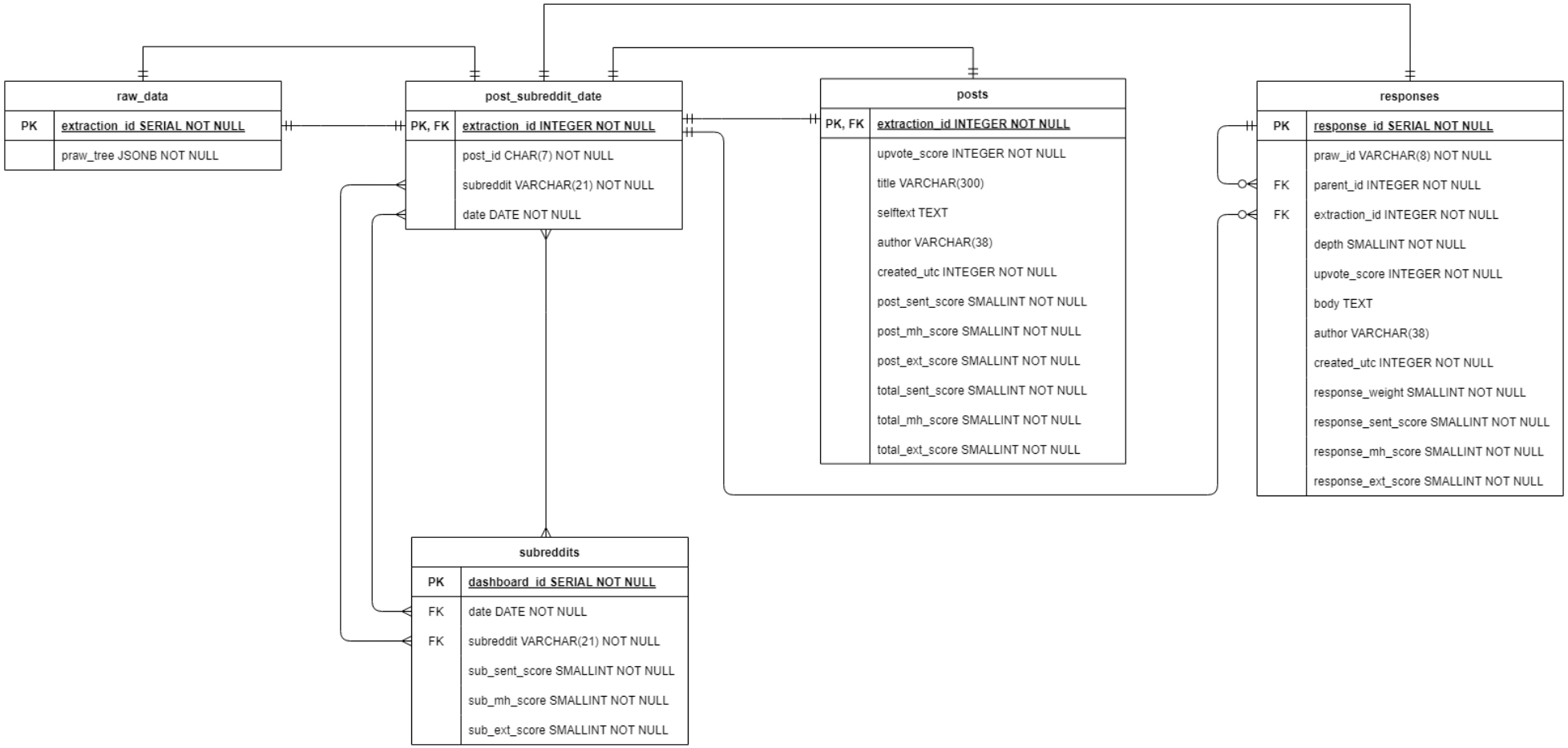
1) Extract Reddit data daily and load to database.
Tools: Python, PostgreSQL, & Dagster Orchestration

2) Transform Reddit data into flattened, useable format for text-analysis, submission-weighting, and aggregation.
Tools: dbt, Python, PostgreSQL, NLP & Machine Learning

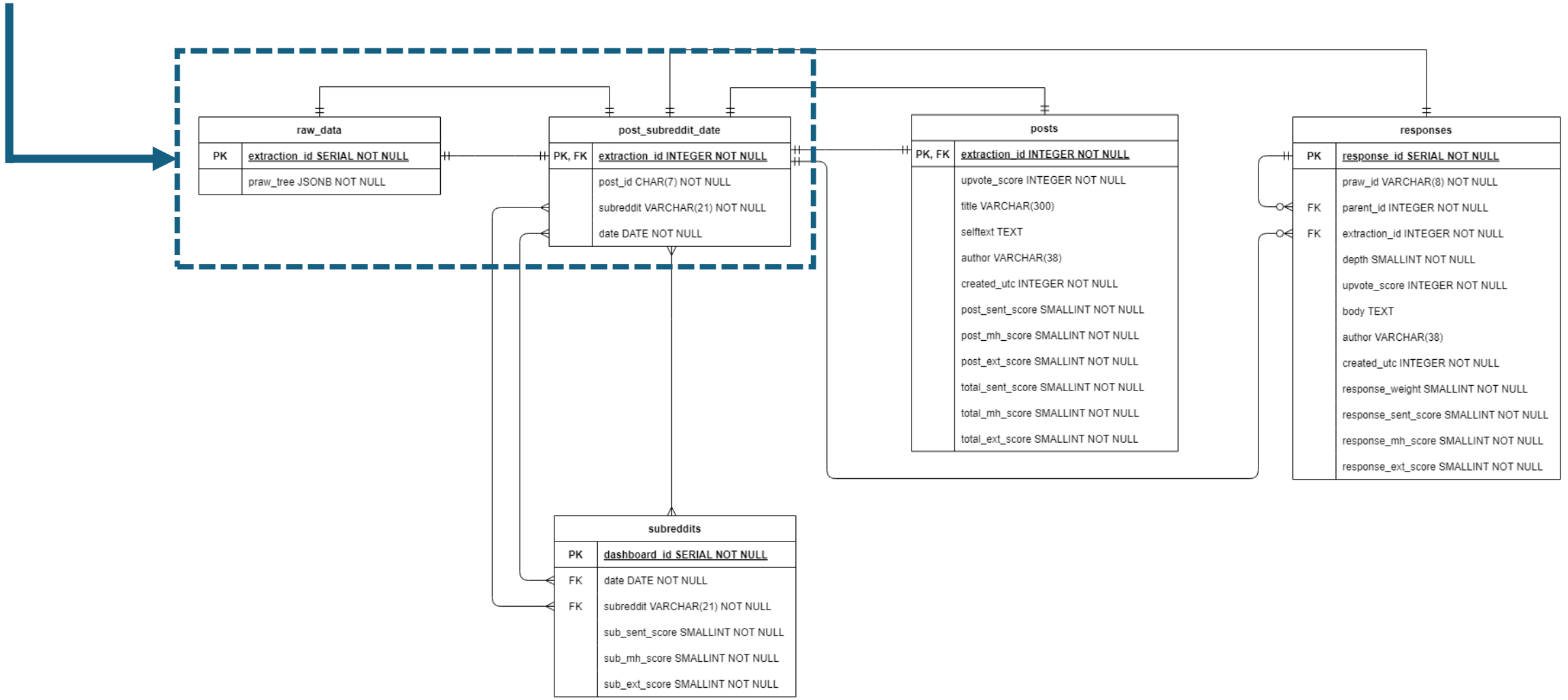


3) Data linked to dashboard for visualizations and analysis
Tools: PostgreSQL, Tableau

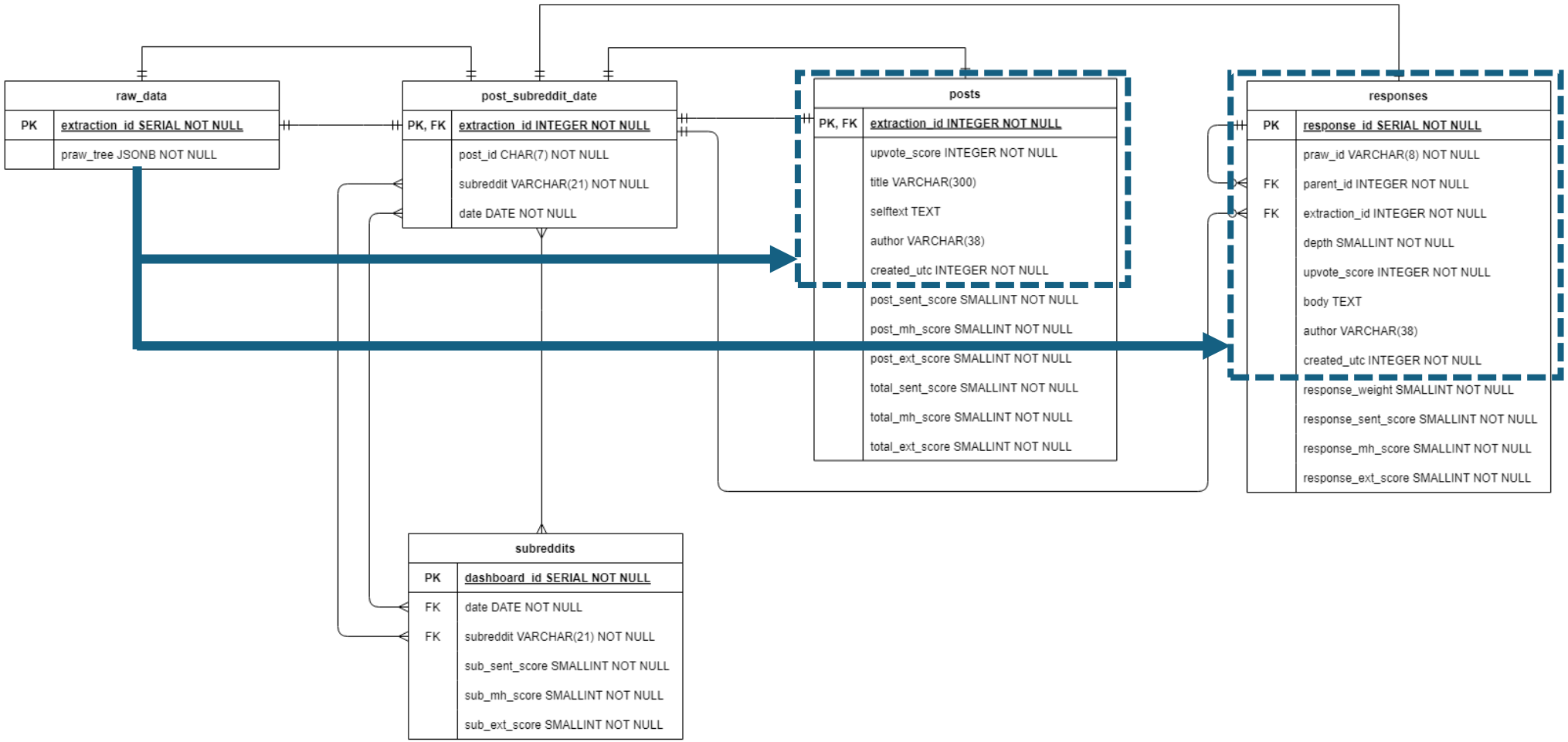
Entity Relationship Diagram (ERD)



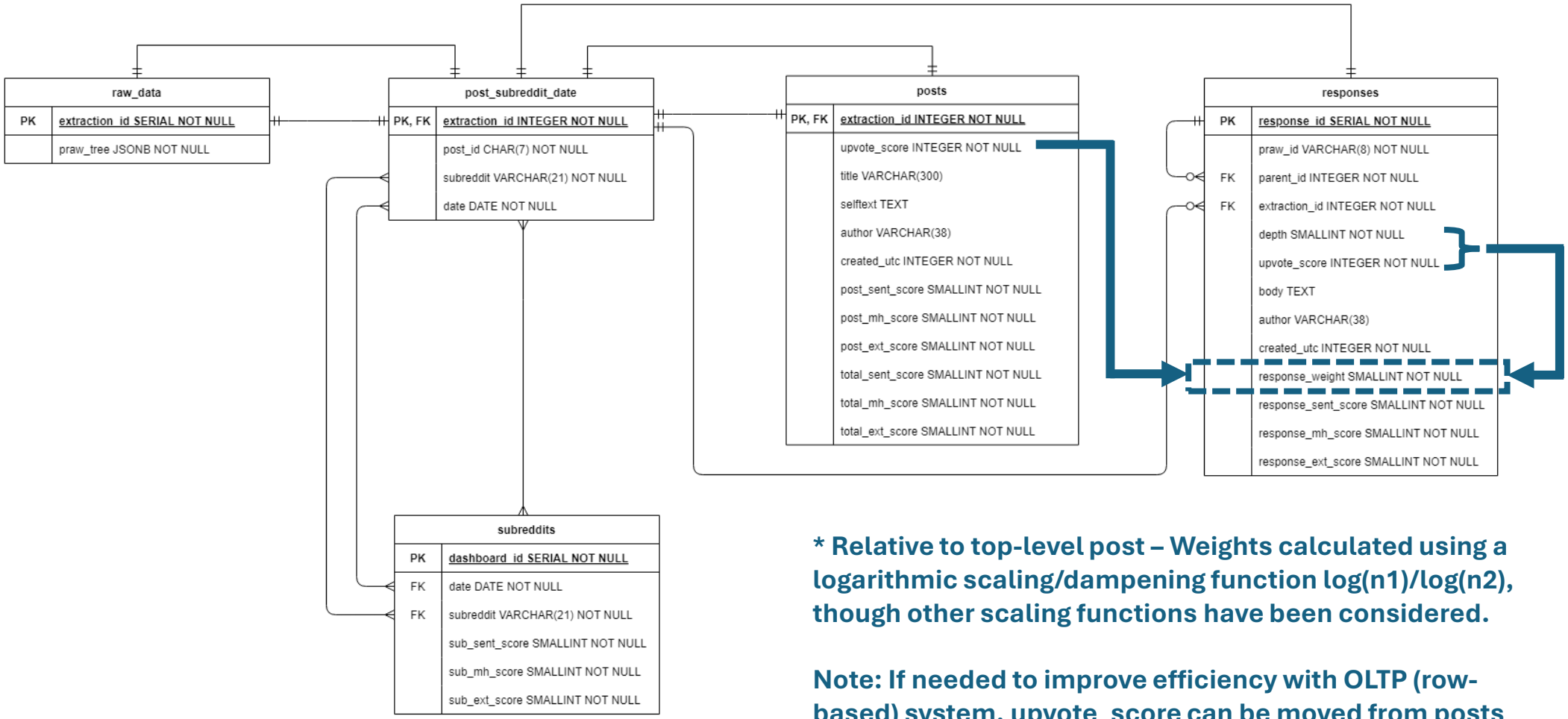
Step 1: Using Python, PRAW, and Orchestration, extract top 10 posts (while limiting comments based on defined extraction parameters) from 25 pre-defined subreddits daily.



Step 2: Flatten JSONB from raw_data into 2 tables:
1) Top-level posts
2) Responses (i.e., both comments and replies)



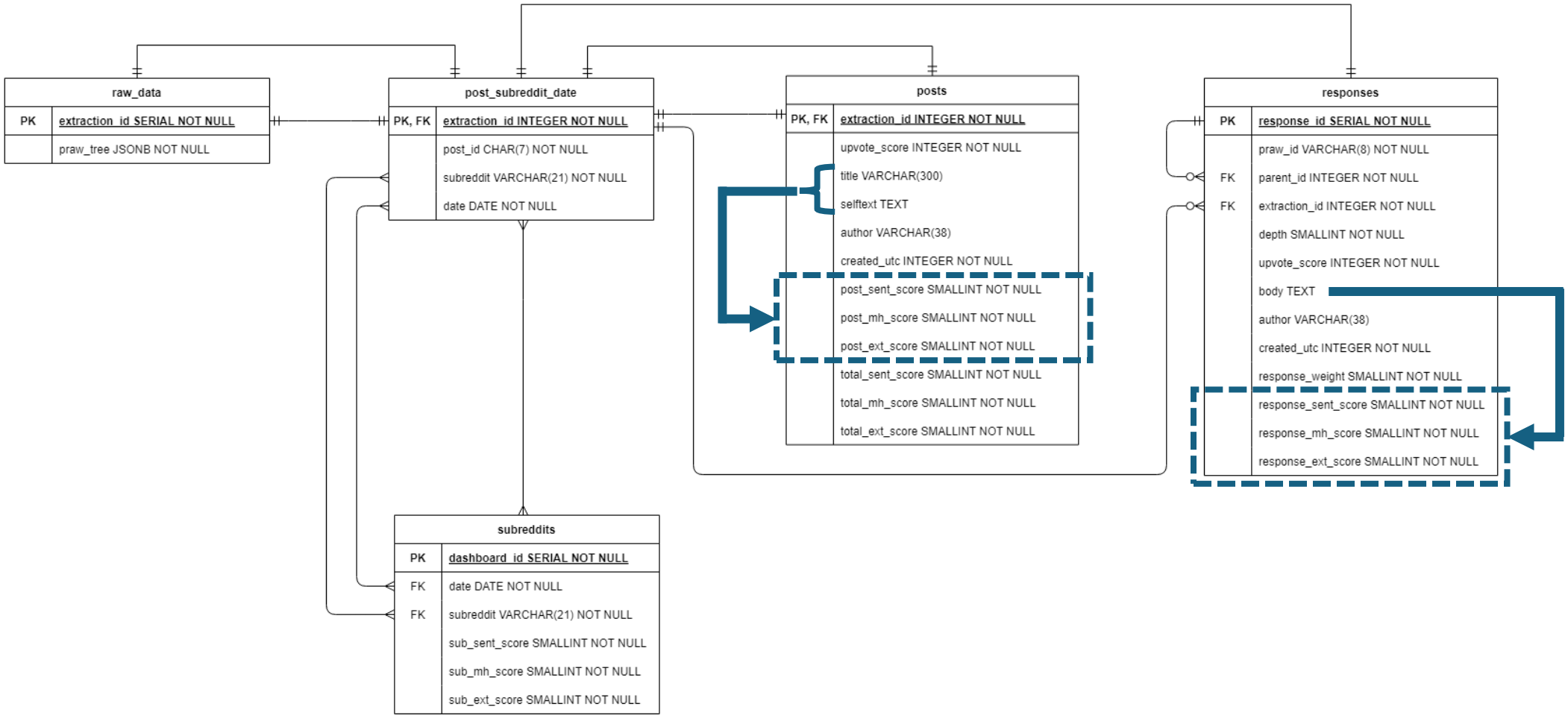
Step 3: Use upvote_score (relative to the top-level post*) and depth to calculate response_weight, which will later be used during post-level text-score aggregation in Step 5.



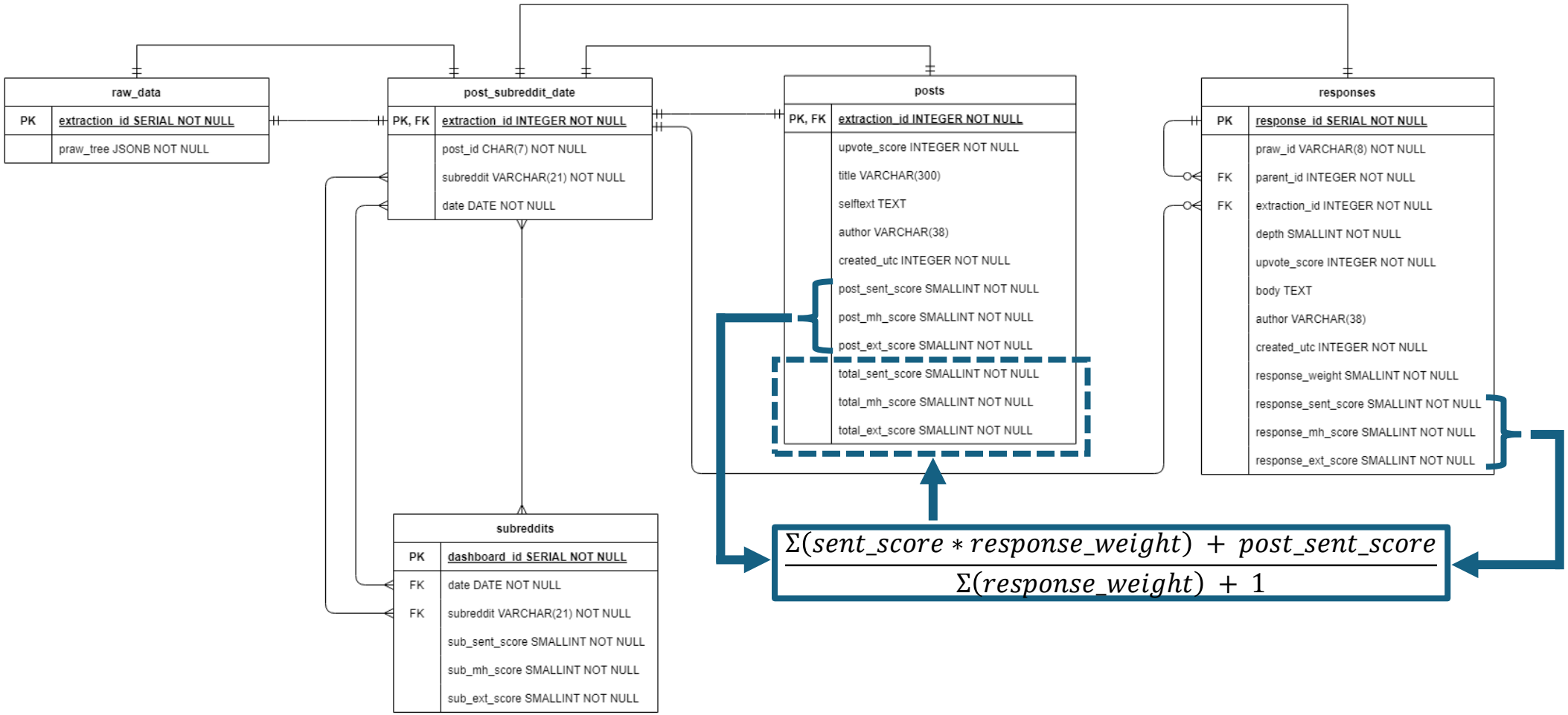
*** Relative to top-level post – Weights calculated using a logarithmic scaling/dampening function log(n1)/log(n2), though other scaling functions have been considered.**

Note: If needed to improve efficiency with OLTP (row-based) system, upvote_score can be moved from posts table to post_subreddit_date association/join table.

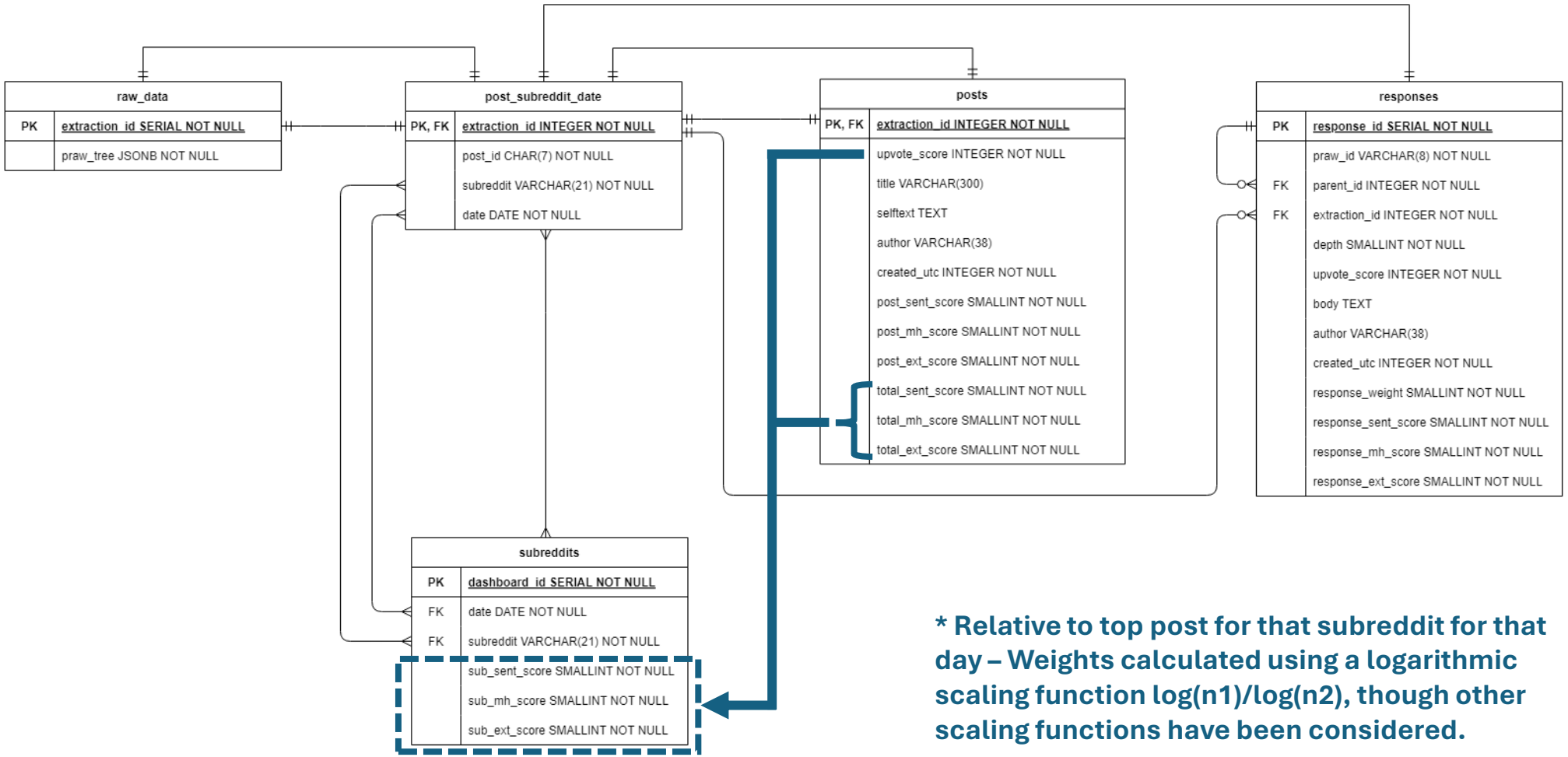
Step 4: Process responses.body and CONCAT(posts.title, posts.selftext) through NLP, ML, and/or LLM to score text in multiple areas (sentiment, mental health, extremism, etc.)



Step 5: Post-level text-score aggregation:
Aggregate all submission text scores (i.e., for all posts, comments, and replies) into a total score for each post.



Step 6: Subreddit-level text-score aggregation:
Aggregate all post text scores into a total score for each subreddit for that day, using upvote_score to weight posts relatively*



*** Relative to top post for that subreddit for that day – Weights calculated using a logarithmic scaling function log(n1)/log(n2), though other scaling functions have been considered.**

Step 7: Results “subreddits” table:
Data available for plotting, dashboard connection, trend analysis, csv export, etc.

subreddits	
PK	dashboard_id SERIAL NOT NULL
FK	date DATE NOT NULL
FK	subreddit VARCHAR(21) NOT NULL
	sub_sent_score SMALLINT NOT NULL
	sub_mh_score SMALLINT NOT NULL
	sub_ext_score SMALLINT NOT NULL