Conceptual Database Design

For the given data set we identified the entities like "User", "Tweet", "Media", "HashTag", "Out\_Links" and "Source" and their corresponding relations. The chatgpt data has the above mentioned components. There are following relationship we captured between entities as "User\_Mentions", "Media\_Usage" and "Out\_Links". The main relation is "Tweets" relation which kinds of ties everything together. For the sources we see very high redundancy so keep it out in a different table which have only 843 diffenent entires so keeping in out in a different table minimize the table size also.

While designing the database we keep the read and write operations in mind. Like "Public\_Metrics" which will be keep updated after every impression so keeping those things out from "tweets" table will minimize the unnecesasary updates to the "tweet" table. Once a tweet is saved there are no updates regarding that tweet.

To capture all the "hashtags" in the system, we keep them in a sepaerate table and there is relation "hashtag\_usage" will keep record which tweet uses which hashtags. If someone delete their tweet, it will not delete the hashtag, because they may be used in other tweets. We used the similar design for the media also which is again reused many time.

"User\_mentions" table will keep which user mentioned in which tweet, since it captures the only users who are already in the system, so there is relationship between "users" table and "user\_mentions" table.

For the conversations, its associated with the parent tweet so we keep the conversationId within the "tweets" table and there are no other attributes associated with the converation whitin the dataset which require it to be a seperate entity.

Our database is 3NF normalized and we did not keep any paritial and transtive dependencies. Our design keep in mind the kind of use case we are solving and make those queries as fast as possible with very minimal joins.