Proposal for Designing a Database to Store CommBank's Twitter Data

Introduction

In the ever-evolving digital landscape, social media has become a critical component of customer engagement and sentiment analysis. The Commonwealth Bank of Australia (CBA), in collaboration with InsightSpark, seeks to leverage Twitter data to enhance their customer engagement strategies, monitor public sentiment, and improve marketing efforts. A robust and scalable database is essential to store, process, and analyze this large volume of unstructured data in an efficient and secure manner.

The task at hand involves designing a database to store data derived from the CommBank Twitter account, including tweet details, user replies, retweets, mentions, and related metadata. This structured storage will facilitate effective data querying, processing, and analysis for business insights.

Proposed Database Design

Overview of the Database Structure

The proposed database will consist of several related tables that store key information about tweets, user interactions, and engagement metrics. The database design will adhere to normalization principles, ensuring data consistency, eliminating redundancy, and improving query efficiency.

The core tables include:

- Tweets
- User Interactions (Replies, Retweets, and Mentions)
- User Profiles
- Engagement Metrics (Likes, Retweets, Comments)
- Sentiment Analysis
- Content Categories

Tables and Fields

- Tweets Table
  - tweet_id (Primary Key): A unique identifier for each tweet.
  - tweet_text: The text content of the tweet.
  - tweet_time: Timestamp of when the tweet was posted.
  - hashtags: List of hashtags used in the tweet.
  - mentions: List of users mentioned in the tweet.
  - media_type: Type of media (e.g., image, video, link).

- sentiment_score: Overall sentiment score assigned to the tweet (positive, negative, neutral).

User Interactions Table

- interaction_id (Primary Key): A unique identifier for each interaction.
- tweet_id (Foreign Key): Links to the Tweets table.
- user_id (Foreign Key): Links to the User Profiles table.
- interaction_type: Type of interaction (e.g., reply, retweet, mention).
- interaction_time: Timestamp of when the interaction occurred.
- interaction_text: The content of the reply or retweet (if applicable).

User Profiles Table

- user_id (Primary Key): A unique identifier for each user.
- username: The Twitter username of the user.
- location: Geographical location of the user (if available).
- followers_count: Number of followers the user has.
- following_count: Number of users the user is following.
- account_creation_time: Timestamp when the user's Twitter account was created.

Engagement Metrics Table

- metric_id (Primary Key): A unique identifier for each engagement metric.
- tweet_id (Foreign Key): Links to the Tweets table.
- likes_count: Number of likes on the tweet.
- retweets_count: Number of retweets.
- comments_count: Number of replies.
- views_count: Number of views (if available).
- engagement_rate: Engagement rate calculated from the total engagement metrics.

Sentiment Analysis Table

- analysis_id (Primary Key): A unique identifier for each sentiment analysis record.
- tweet_id (Foreign Key): Links to the Tweets table.

- positive_score: The score representing the positive sentiment of the tweet.
- negative_score: The score representing the negative sentiment of the tweet.
- neutral_score: The score representing the neutral sentiment of the tweet.
- overall_sentiment: Overall sentiment category (Positive, Negative, Neutral).

Content Categories Table

- category_id (Primary Key): A unique identifier for each content category.
- tweet_id (Foreign Key): Links to the Tweets table.
- category_type: Category assigned to the content (e.g., promotional, informational, customer service).
- category_score: Score indicating the relevance or importance of the category.

Database Relationships

- One-to-Many Relationship between the Tweets table and the User Interactions table. Each tweet can have multiple user interactions (replies, retweets, mentions), but each interaction refers to only one tweet.
- One-to-Many Relationship between the User Profiles table and the User Interactions table. Each user can have multiple interactions (replies, retweets, mentions), but each interaction is linked to only one user.
- One-to-One Relationship between the Tweets table and the Sentiment Analysis table. Each tweet will have exactly one sentiment record.
- One-to-Many Relationship between the Tweets table and the Content Categories table. Each tweet can belong to multiple content categories, but each category is linked to only one tweet.
- One-to-One Relationship between the Tweets table and the Engagement Metrics table. Each tweet has exactly one set of engagement metrics.

Primary Keys and Foreign Keys

- The primary keys will ensure each record within the tables is unique, making it easier to reference and query specific data points.
- Foreign keys will maintain referential integrity, ensuring that user profiles and interactions are always linked to valid tweets.

Normalization and Data Integrity

The database design will follow normalization principles to reduce data redundancy and ensure data integrity. By separating the data into relevant tables and defining clear relationships, the

database will ensure consistency and avoid duplication of data. For example, user data is stored in a separate table to avoid repeating user details across interactions, while tweet content and engagement metrics are linked but stored in their respective tables.

Scalability and Performance

The database is designed to handle large volumes of data, as Twitter generates a substantial amount of content daily. Indexing critical columns like tweet ID, user ID, and engagement metrics will optimize query performance, allowing for quick retrieval of data. Additionally, partitioning the database by date or tweet ID can be considered to manage and scale large datasets efficiently.

Conclusion

The proposed database design provides a comprehensive and scalable solution for storing and managing CommBank's Twitter data. By structuring the data into multiple related tables, we ensure efficient storage, maintainability, and scalability. This database will empower data scientists to query and analyze the data efficiently, ultimately providing valuable insights into customer sentiment, engagement patterns, and social media performance. The proposed database design will be a critical tool in helping CommBank derive actionable insights from their Twitter interactions, enabling more effective marketing and customer engagement strategies.