DS 4300: Assignment 1
Generate Twitter Data

DESCRIPTION

In order to test different data modeling techniques, we need to generate some data. We'll begin by generating some sample twitter-like data and testing performance of a twitter-like system using MySQL – A relational database system. These tests will form the baseline for future comparisons using NoSQL techniques.

Our twitter-like application should support:

a) Users rapidly posting tweets
b) Users following other users
c) Users repeatedly requesting a list of tweets posted by those they follow

IMPLEMENTATION

1. Implement a relational database to manage users and their tweets. Your database should have two tables:

TWEETS – The tweets posted by users

tweet_id	long
user_id	long
tweet_ts	datetime
tweet_text	varchar(140)

FOLLOWERS – Who follows whom

user_id	long
follows_id	long

2. Write a tweet generator that creates a random tweet and inserts it into the database. The tweets can be random worlds from some corpus of your own choosing. Some of the words in the tweet should be hashtagged. Your resultant database should contain at least 1 million tweets.

3. Write a separate program that repeatedly picks a random user, and fetches the most recent N tweets posted by that user's followers.

4. Optional: Write a third program that does nothing but fetch the most recent N tweets containing a specified hashtag.

ANALYSIS (Be Creative – but here are some general guidelines)

1. Document your hardware configuration (CPU speed, number of cores, RAM, Disk etc.) Also document and justify your data model assumptions: number of users, number of tweets per user, distribution of the number of followers per user, number of tweets retrieved per user request, or any other parameters that might affect your analysis.

2. In each system, investigate performance: how fast are you able to perform inserts and timeline requests when they are performed separately vs. concurrently?

3. Discuss the implications of your system parameters on performance. For example, how would a random-vs-skewed distribution of followers impact your results? Does your performance change as your database fills up with records / values? If I had asked you to generate 1 billion tweets, would that have been a bad idea? Is so, why?

4. Twitter generates about 10,000 per second. Could your system keep up? How long would it take to fill up your harddrive? Estimate the incremental storage cost per year based on current harddrive prices on Amazon.

SCORING

You will be graded on your MySQL implementation, the quality of your code, your experimental design, and the thoroughness of your analysis. *We will compare results in class and I may ask several of you to present your analysis to the rest of the class.*