CSCI 620/Section 02/Mior, Introduction to Big Data, Spring 2235

Assignment 5 – Data integration

Description

Let's assume the following sources:

- S1: NetworkingPosts(Id, Title, Score), which stores posts with the tag "networking"
- S2: NonNetworkingPosts(Id, Title, Score), which stores posts that do not have the tag "networking"
- S3: NetworkingUser(Id, DisplayName, Reputation), which stores users who have commented on at least one post with the tag "networking"
- S4: NonNetworkingUsers(Id, DisplayName, Reputation), which stores users who have never commented on any post with the tag "networking"
- S5: CommentedOn(UserId, PostId), which stores all users who have commented on posts

Let's assume the following global schema:

- AllPosts(Id, Title, Score, Tag), which contains each post with its main tag. (For this assignment, the main tag is either "networking" or "other".)
- AllUsers(Id, DisplayName, Reputation), which contains each user.
- AllComments(UserId, PostId), which stores users commenting on posts.

Your tasks

1. Implement the previous sources in your relational database from assignment 2 using both non-materialized and materialized views. Provide your code. **(20 points)**

2. Using GAV mappings, describe the global schema based solely on the sources. These will be non-materialized views using the views in Q1. You should have two versions of these views using the materialized and non-materialized views from Q1. **(20 points)**

3. Provide queries for answering the following queries over global schema. **(5 points per query)**

    3.1. Users with reputation more than 100 who have commented on at least 10 posts.
    3.2. Users whose display name starts with "john-" and who have never commented on any post with the tag "networking".

4. Provide the resulting queries after expanding Q3.1 and Q3.2 using the GAV mappings defined in Q2 over the sources. You should simply substitute the definitions of the views into the queries. Run the resulting queries. Report your timings when you use non-materialized vs. materialized views from Q1.
   (25 points)

5. Optimize the queries from Q4 by eliminating redundant joins and sources where possible. Run the resulting queries. Report your timings when you use non-materialized vs. materialized views from Q1.
   (25 points)