CS 6400 Team Project Proposal

September 20, 2018

Patel, Bhakti Purshottam *MS in ECE* bpatel87@gatech.edu

- Previous courseworks related to databases: Undergraduate database class
- Practical experience related to databases: Basic programming experience with SQL and Hadoop
- More about me: I am a graduate student from school of ECE. My research interest is in machine learning and data analysis.

Florez, Nathan *MS in Computer Science* nflorez3@gatech.ed

- Previous courseworks related to databases: Undergraduate intro to database systems class
- Practical experience related to databases: Worked as software engineer at Bank of America and regularly worked with Informix SQL database for five years
- More about me: Received my degree as Bachelor of Science in Computer Science at University of Delaware in 2013 and worked at Bank of America following that up until now. Currently I am pursuing my masters in Computer Science with specialization in machine learning.

Lee, Seung Hoon *MS in Computer Science* shoonhye@gatech.edu

- Previous courseworks related to databases: Undergraduate *Intro To Database* class
- Practical experience related to databases: Data Analyst intern for 3 months at Maxim Integrated
- More about me: I received my BS in Chemical Engineering at UC Berkeley. After that, I worked as Chemical Process Engineer and Commissioning Engineer for 4.5 years before I joined Georgia Tech MSCS program.

(i) the scope and the goal –the goal- what are you trying to accomplish in this project

Comparing the performance of DBMS (PostgreSQL, MongoDB, MySQL) when processing text, image, and speech data. For example, we can compare query time, database size, quality of query result, and difficulty of building databases for dealing with image, text, and speech data.

(ii) problem definition – be more precise and define the scope of what you plan to do and what you do NOT intend to do

We do not intend to build a particular application program. Instead, we will analyze and evaluate the performance of existing widely-used Database Management System softwares such as PostgreSQL (Object-relational), MongoDB (Not Only SQL), and MySQL (Relational) when dealing with different types of data such as image, text and speech.

We will evaluate each of DBMS quantitatively and qualitatively for a particular scope of data.

- Quantitative : query time, database size
- Qualitative : difficulty of building DB, whether it has a robust theoretical background (i.e. relational algebra and calculus), flexibility

(iii) description of design, approach, or unique features,

Design & Approach :

1. Pick three categorically different data - Image, text, speech (or something else)
2. Build databases for each data using PostgreSQL, MongoDB, and MySQL
3. Evaluate each one of DBMS in dealing with each one of data type (image / text / speech)

   Quantitative evaluation:

   - Query time: “UPDATE”, “INSERT”, “SELECT” query time with varying amount of data queried. So, it also can measure scalability of queries
   - Database size: measure database storage size of PostgreSQL, MongoDB, and MySQL after building database of image / text / speech database
   - Scalability:
     - How query time increases as the size of data to query increases
     - How database size increases as the size of data to store increases

Qualitative evaluation:

- Quality of query result
  - How image data is queried
  - How speech data is queried
  - How lengthy text data is queried
- Difficulty of building DB
  - Does DBMS have a tool to easily build DB (i.e. MySQL workbench)
- Flexibility
  - Does DBMS have a dynamic schema functionality (i.e. a user can add new columns or fields on MongoDB without affecting existing rows or application performance)
- Theoretical background about DBMS
  - Does DBMS have a strong theoretical background (i.e. SQL relational algebra and calculus)

4. Compare and find out why one DBMS works better than the other in terms of the performance criteria

Unique features of our Project :

1. It will be interesting to observe which DBMS works well for a particular data type. For example, it could be possible that MySQL works better with text data than MongoDB does with text data.
2. Analysis of PostgreSQL performance is not studied well, so this project would be a useful reference for pros and cons of PostgreSQL.

(iv) What data will you use and where does it come from

Three types of data will be used:

1. Image data - MNIST handwritten image, CIFAR image
2. Speech data - TIMIT Dataset
3. Normal text data - Amazon review dataset

NOTE: Those datasets are open to the public.

(v) what / how will be implemented and what technologies/libraries/tools will be used

Database program: DMBSs (PostgreSQL, MongoDB, MySQL) will be used to create database.

Query program: Python or any other language which can digest SQL query

(vi) how the work will be divided among team members and a rough schedule.

Work will be divided based on the data type as follows:

- Image dataset: Bhakti, Nathan
- Speech dataset: Nathan,Brian
- Text dataset: Brian, Bhakti

Tentative Schedule:

9/20 ~ 10/4 : Data acquisition
10/4 ~ 10/18 : Database design and building using the data
10/18 ~ 11/1: Evaluate performance
11/1 ~ 11/13: Write interim report and modify evaluation criteria or data as needed
11/14 ~ : TBD

(vii) what functionality you plan to implement and how it will be demonstrated.

Since we are not developing any application program, we will analyze and evaluate each of DBMSs performance as we mentioned. The following format would be a format to demonstrate to compare the performance :

For image Database:

	MySQL	MongoDB	PostgreSQL
Query_Time			

(SELECT)			
Query_Time (UPDATE)			
Query_Time (INSERT)			
DB size			
Scalability			
Query quality			
etc...			

For text Database:

	MySQL	MongoDB	PostgreSQL
Query_Time (SELECT)			
Query_Time (UPDATE)			
Query_Time (INSERT)			
DB size			
Scalability			
Query quality			
etc...			

For speech Database:

	MySQL	MongoDB	PostgreSQL
Query_Time (SELECT)			
Query_Time (UPDATE)			
Query_Time (INSERT)			
DB size			
Scalability			
Query quality			
etc...			