CSE 304 Term Project

List of Projects

1. Given an unstructured dataset, this project will focus on creating an engine to convert the unstructured data to a given database. This includes understanding the dataset, creating necessary schemas, and inserting data to the given database. This project will also require identifying 5 most useful queries on the given dataset, writing the queries in the database, and measure their performance.
   Example: XML dataset of Stack Overflow to any SQL/NoSQL database

2. Twitter provides a fire hose of data. Automatically filtering, aggregating, analysing such data can allow harnessing the full value of the data, extracting valuable information. This project will investigate stream processing technology to operate on social streams, or extract structure from them and store them in a database system.
   Example: Publicly available Twitter dataset to any SQL/NoSQL database

3. Most graph processing systems end up focusing on simple algorithms (Shortest Paths, PageRank) that are not really used by real graph analysts. This project will study Stanford Network Analysis Project (http://snap.stanford.edu/snap/description.html) to identify common patterns of accesses made by different network algorithms, and design a wrapper over a graph database that can supports a subset of these operations.
   Example: SNAP road networks dataset to Neo4J database

4. Given a dataset, this project will focus on comparing the performance of different SQL/NoSQL/Hadoop based database systems. This includes understanding the dataset, creating necessary schemas, and inserting data to the given databases. This project will also require identifying 3 most useful queries on the given dataset, writing the queries in both the databases, compare their performance, and finally understand the performance differences.
   Example: Compare the performance of Oracle and HBase with respect to XML dataset of Stack Overflow

5. Given a dataset, this project will focus on comparing the performance of different object relational mapping tools (e.g., Hibernate, LINQ, NHibernate etc.) with the traditional database systems. This includes understanding the dataset, creating necessary schemas, and inserting data to the given database. This project will also require identifying 4 most useful queries on the given dataset, writing the queries in both the database and object

relational tools, compare their performance, and finally understand the performance differences.
Example: Compare the performance of using only Oracle and Hibernate on top of Oracle with respect to XML dataset of Stack Overflow

6. This project will focus on implementing any traditional system that requires a database on top of a NoSQL/Hadoop based database. This include creating different input forms and insert/update/delete data to the database using your preferred language.
Example: Real estate management system with MongoDB

List of Databases

1. Relational DB – Oracle, MySQL, SQL Server, PostgreSQL
2. Column-oriented database – HBase
3. Document-oriented database – MongoDB, CouchDB
4. Graph database – Neo4J
5. Key-value database – Riak
6. In-memory database – Redis
7. ORM tools – Hibernate, LINQ, NHibernate

List of Datasets

The ClueWeb09 Dataset
The ClueWeb09 dataset was created to support research on information retrieval and related human language technologies. It consists of about 1 billion web pages in ten languages that were collected in January and February 2009. The dataset is used by several tracks of the TREC conference.
http://lemurproject.org/clueweb09/

Large Health Datasets
https://www.ehdp.com/vitalnet/datasets.htm

U.S. Patents Data
These data comprise detailed information on almost 3 million U.S. patents granted between January 1963 and December 1999, all citations made to these patents between 1975 and 1999

(over 16 million), and a reasonably broad match of patents to Compustat (the data set of all firms traded in the U.S. stock market).
http://www.nber.org/patents/

Freebase Data
Freebase was a large collaborative knowledge base consisting of data composed mainly by its community members. It was an online collection of structured data harvested from many sources, including individual, user-submitted wiki contributions.
https://developers.google.com/freebase/

UN Data
http://data.un.org/Explorer.aspx

Wikipedia XML Data
A complete copy of all Wikimedia wikis, in the form of wikitext source and metadata embedded in XML.
https://dumps.wikimedia.org/backup-index.html

DBPedia Datasets
http://wiki.dbpedia.org/downloads-2016-04

Stack Overflow Data
https://archive.org/download/stackexchange
https://archive.org/details/stackexchange

US Census Data
https://factfinder.census.gov/faces/nav/jsf/pages/download_center.xhtml

Graph Data
http://snap.stanford.edu/data/index.html

General Links
http://www.valleyprogramming.com/blog/big-data-datasets-large-examples-boulder-colorado-hadoop-mongodb
http://www.bigfastblog.com/how-to-get-experience-working-with-large-datasets

Please note that you are free to use any dataset which is significantly large in volume.