MSDS 7330
Research Project Logistics and Project Topics

Daniel W. Engels

I. Document Summary

This document describes the basic logistics required to successfully complete the term project and term paper for MSDS 7330.

One of the major goals of this course is to introduce the students to solving open ended problems. To this end, the term project is a major component of the course and of the course grade. In performing the term project, you will draw upon your knowledge and experience from the course lectures, readings, your knowledge and experience, and elsewhere to actually perform the work and solve the problem. This document is intended to guide you in the logistics process required for the term project and paper.

The basic topics covered in this document are: the term paper logistics timeline, the term project proposal, background knowledge for the work, the high standards of work expected of us, and some example project topics. This course is meant to be fun and informative, so make sure to choose a project topic that you are interested in and teammates (if any) that share your interest. Interesting topics are the ones that will be fun and the most rewarding to you personally.

II. Research Logistics Timeline

The term project is a major work product upon which your grade is based. Be studious and reach project milestones on time, and you will do well in the course. There are three written reports due in conjunction with the project. Two oral presentations are required. The written reports are the term project proposal, a first draft, and the final report. The oral presentations, are an early presentation given during week 8 class and a final presentation given during the last two weeks of classes. The presentations are in addition to any preprepared video performed as part of the project.

The term project deliverables are given below.

Project proposal. (10% of research project grade)
First presentation. (10% of research project grade)
Draft paper. (10% of research project grade)
Final presentation. (10% of research project grade)
Final paper. (40% of research project grade)
Peer grade. (20% of research project grade)

III. Teams

Teams of three (or four with justification) are encouraged for the term project (teams of two are acceptable, teams of one are forbidden). Teams typically accomplish more and learn more than do persons working alone, so start looking for teammates as soon as possible!

The database and file organization fields are broad and they draw upon multiple diverse knowledge domains. The shear diversity of knowledge and expertise required to research and develop technologies, create applications, and deploy a potentially large scale database management system fosters collaborative work in database research and development.

We all learned how to play nice with others in kindergarten, but we weren't forced to learn how to work with others until we got to college and took great courses like this one. Project teams are often significantly more productive, and the resulting work of higher quality, than if you were working in isolation. Working with others is a learned skill often best learned through actually working with others, and it is mandatory in most jobs in the real world outside of the university. The more you work with others, the better your interpersonal and team skills will become (or you will realize you belong in a job with no living human contact whatsoever, like a morgue technician working on the graveyard shift). Besides, working with others is a lot more fun than working alone. Just think about how much more fun it is to play Doom or Halo or World of Warcraft with a couple of your buddies. Blasting the bad guys isn't nearly as much fun when you're doing it by yourself!

IV. Term Project Proposal

Project proposals of at least one page and up to two pages are due as specified above and in the course syllabus. Use the Word template provided by the instructor. Turn in a Word version of your proposal on 2DS. The instructor will read the proposals carefully over the following few days and get back to you by 2DS with any questions or feedback. The instructor will be looking mainly at three aspects of your proposal: 1) Is it topical? This is to say, does it relate in some way to databases and/or file organization, which it had better, or else. 2) Is it sufficiently focused? Research looking at broad open ended questions typically lasts several years, costs millions of dollars, and has no guarantees of results or outputs. You have two months, no budget, and must guarantee two outputs, your research proposal and your final report. And, 3) Is it too ambitious (or not ambitious enough)? You have two months. You won't win the Touring Award based upon your class project, but you can take the first step with it.

Where do you get a term project idea? That's easy. Choose one of the topic areas that I've given for the class or come up with one of your own. Good class projects can vary dramatically in size, complexity, scope and topic. The only real limitation is that the project must relate in some way to the course. You are strongly encouraged to identify a project related to your work, so that you may have a ready platform on which to apply your knowledge and put your skills to work immediately. Note that purely simple implementation problems, such as installing a large distributed database system, while interesting and potentially a source of great learning, are not acceptable projects. However, evaluation problems, such as installing a large distributed database system and then measuring its performance under a broad range of loads and conditions, are acceptable projects.

Ideally, your project will be targeted towards an open problem with the goal that your results will be novel and publishable in a peer-reviewed academic journal or conference. More realistically, for small groups, it is more practical that your project will be a defined problem that may have been solved before, but one which you are evaluating or solving in a new way.

Note that the list of topics either outline a problem domain or specify a specific problem to be solved. It is your responsibility to take your chosen topic and ensure that a reasonable problem is being addressed by you and your team.

Your term project proposal must contain the following items:

- *Project title.* A detailed project title is better than a vague title. A properly formed title will help to focus your energies on the actual problem being addressed in your project.
- *Names and email addresses of all investigators for the project.* There must be at least two investigators.
- *Clear statement of the problem.* A one to two sentence statement of the problem followed by a one paragraph clarification of the problem. The paragraph should identify clearly the research question you are addressing, and proper motivation for the importance of the problem must be provided.
- *Clear statement of your research methodology.* A multiple paragraph explanation of how you will approach and develop a solution to the research problem under study. Note that you need to identify multiple intermediate milestones. Research, like product development, is not an all or nothing proposition. If you make it all or nothing, you are likely to lose everything. When you strip the problem to its essence, the first step becomes clear. And, that first step is usually a baby step. Identify that baby step, and then the next one and the next one and so on. It is possible to get an ‘A’ on the research project without completing every step you identify, and it is certainly easier to get a good grade when all the steps you’ve identified are baby steps.
- *A statement of previous work related to the problem.* This is a preliminary inquiry into what research has or has not been performed to solve your chosen problem. Your final report will contain an extensive discussion of previous and related research. You should have at least five citations of previous or related work on your topic area.
- *A statement of your research plan and schedule.* The timeline and major milestones that you will achieve in completing your project must be explicitly spelled out. You need to timeline your project to convince yourself (and me) that you can complete the project before the end of the semester and that your project isn’t trivial.
- *A list of resources needed to accomplish your work, with special emphasis on important pieces you don’t yet have access to.* Be as clear and precise as you can in your requirements, and we will work towards getting what you need as quickly as possible. We don’t have a budget, so requests for electron scanning microscopes and new signal generators may not be honored unless they can be borrowed from somewhere else on campus. Even requests for simple things like dedicated computers are unlikely to be met due to the shared nature of the computing environment at SMU. If your request cannot be accommodated for any reason, we will notify you as soon as we find out.
- *Any other questions or clarifications you need from us.*

If your project requires specific types of data sets, attempt to find one or more and include the data set name and/or its location within your proposal. Some data sets may be found at `http://www.statsci.org/datasets.html` and `http://it.stlawu.edu/ rlock/datasurf.html`. The SMU library has access to a broad range of data sets as well.

I am available to answer your questions as you prepare your project proposal. While this is a short document, coming up with a research problem is not always done quickly. A quick look at related and previous work on your chosen topic will give you an idea of how novel your project work needs to be and how hard it will be to achieve something new.

V. First Draft

Your Draft paper is expected to provide an early draft of the work. The completed sections should include an early draft of the Introduction, all background sections needed to understand the work (e.g., if your work involves cancer data and the Cassandra NoSQL database, you should have sections on Cancer, Cassandra, and the Data in your Draft), at least a sketch for the sections involved in your solution approach and results, analysis, and conclusions.

VI. Final Paper

Your final paper is expected to sufficiently describe the problem being addressed, related work, the solution methodology, any results, an analysis of the results and provide some relevant conclusions. When in doubt about whether to include some material or not, include the material. Your final paper should be as self contained as possible. Your final paper should be at least four pages and no more than five pages using the Word template provided by the instructor.

Please don’t wait for me to get back to you before starting your work. Get started as soon as possible! That means today! You have roughly two months to start and complete the term project and all the documentation. This is ample time if your proposal is focused and you start early, but not otherwise. Plan on spending at least one week writing the final paper.

VII. Knowledge Background

In a one semester course, one can cover only a fraction of the database topics, current or otherwise. There are sure to be term projects where the background material needed for the work is not covered in the course or is not covered in sufficient detail or a timely manner. And, even for topics that we do cover in detail, there will certainly be other relevant related work that you will need to be familiar with to finish your project.

A large part of doing data science work is figuring out what has already been done and where the knowledge holes exist. So, you should research your problem's related literature and any other available information as extensively as you can. Keep an eye out for useful software, such as MySQL and MongoDB, or research methodology or tools that you can leverage. This will save you tremendous amounts of time in performing your project and doing the final paper.

VIII. Work Product Standards

Aim high in a focused way, and do the best that you can! The best research papers are sure to be publishable in top ACM or IEEE conferences or appear as articles in journals. In fact, the goal of every research project in this class is to produce a body of work that gets published in an ACM or IEEE journal or transactions. If research does not get published and, therefore, publicized, it is of no use to the community at large. I have this high level of expectation even for those papers that are destined to remain unseen due to corporate or other distribution limitations, so expect to perform well regardless of your final target audience.

I have great confidence that you will far surpass my already high expectations with wonderful work that will further the state-of-the-art in networking.

IX. Example Project Topics

The following is a list of possible project topic ideas, many of which have been done in excellence. You are not required to choose a project topic from this list – in fact, you are encouraged to come up with a project topic of your own choosing!

1) Being able to compare performance of different DBMSs and different storage and access techniques is vital for the database community. To this purpose several synthetic benchmark has been designed and adopted over time (see TPC-C, TPC-H etc...). Wikipedia open source application, and publicly available data (several TB!!), provide a great starting point to develop a benchmark based on real-world data. The project consists in using real-world data, queries and access patterns to design one of the first benchmarks based on real-world data.
2) Amazon RDS is a database service provided within the EC2 cloud. An interesting project consists in investigating performance and scalability characteristics of Amazon RDS. Also since RDS services run in a virtualized environment, studying the ”stability” and ”isolation” of the performance offered is interesting.
3) A similar study may be performed on Amazon SimpleDB which is a NoSQL database.
4) Hosted database services such as Amazon SimpleDB are starting to become popular. It is still unclear what is the performance impact of running applications on a local (non-hosted) platform, such as a local enterprise datacenter, while having the data hosted ”in the cloud”. An interesting project aim at investigating the performance impact for different classes of applications e.g., OLAP, OLTP, Web.
5) Performance monitoring is an important portion of data-center and database management. An interesting project consists in developing a monitoring interface for MySQL or MongoDB, capable of monitoring multiple nodes, reporting both DBMS internal statistics, and OS-level statistics (CPU, RAM, DIsk), potentially automating the detection of saturation of resources.
6) Being able to predict cpu/mem/disk load of database machines can enable ”consolidation”, i.e., the co-location of multiple DB within a smaller set of physical servers. The project would consist in investigating machine-learning and other predictive techniques.
7) Flash memories are very promising technologies, providing lower latency for random operations. However, they have a series of unusual restrictions and performance. An interesting project investigates the performance impact of using flash memories for DB applications.
8) Often database assume data to be stored on a local disk, however data stored on network file systems can allow for easier administration, and is rather common in enterprises. The project will investigate the impact of local-vs-networked storage on query performance.
9) Twitter provides a fire hose of data. Automatically filtering, aggregating, analyzing such data can allow a way to harness the full value of the data, extracting valuable information. The idea of this project is investigating stream processing technology to operate on social streams.
10) Client-side database. Build a Javascript library that client-side Web applications can use to access a database; the idea is to avoid the painful way in which current client-side application have to use the XMLHttpRequest interface to access server-side objects asynchronously. This layer should cache objects on the client side whenever possible, but be backed by a shared, server-side database system.
11) As a related project, HTML5 browsers (including WebKit, used by Safari and Chrome), include a client-side SQL API in JavaScript. This project would involve investigating how to user such a database to improve client performance, offload work from the server, etc.
12) Preventing denial-of-service attacks on database systems. Databases are a vulnerable point in many Web sites, because it is often possible for attackers to make some simple request that causes the Web site to issue

queries asking the database to do a lot of work. By issuing a large number of such requests, an attacker can effectively issue a denial of service attack against the Web site by disabling the database. The goal of this project would be to develop a set of techniques to counter this problem, for example, one approach might be to modify the database scheduler so that it doesn't run the same expensive queries over and over.

13) Scientific community data management requirements significantly differ from regular web/enterprise ones. To this purpose a specialized DB has been developed: SciDB. Studying performance of SciDB on dedicated servers vs. on virtualized environment such is an intriguing topic. Another interesting investigation would cover the impact on SciDB performance of storing the data over the network (e.g., network file system). A third interesting project would explore the performance of clustering algorithms on SciDB vs. MapReduce.
14) Asynchronous Database Access. Client software interacts with standard SQL databases via a blocking interface like ODBC or JDBC; the client sends SQL, waits for the database to process the query, and receives an answer. A non-blocking interface would allow a single client thread to issue many parallel queries from the same thread, with potential for some impressive performance gains. This project would investigate how this would work (do the queries have to be in different transactions? what kind of modification would need to be made to the database) and would look at the possible performance gains in some typical database benchmarks or applications.
15) Extend SimpleDB by Edward Sciore. SimpleDB is a very simple database written in Java. There are a number of ways you might extend it. For example, you could add support for optimistic concurrency control and compare its performance to the basic concurrency control scheme. You could also port SimpleDB to a tiny embedded device for use as a data store. There are a number of other possible projects of this type; we would be happy to discuss these in more detail.
16) One of the dangers with building a system that collects personally identifiable information, such as electronic medical records, is that it collects relatively sensitive personal information. Protecting this information from casual browsers, insurance companies, or other undesired users is important. However, it is also important to be able to combine different users data together to do things like intelligent disease cluster identification or anomaly detection. The goal of this project would be to find a way to securely perform certain types of aggregate queries over data without exposing personally identifiable information.
17) Teach the class about a topic not covered by one of the units, such as Hadoop, Spark, Cassandra, or any of the numerous other topics we have not covered. Prepare a 90 minute video presentation that may be provided asynchronously to the class.

Above all, have fun with the project!
DWE