**DS 785: Capstone Course Project Proposal**

**Jason Lloyd, Fall 2017**

**Project Description:** This project will cover in detail methods to optimize the build times and end-user response times for a SQL Data Warehouse.

**Rationale for undertaking this project:** Many data warehouses use a SQL-Based DBMS as the underlying platform for storing and retrieving data. In an environment with ever increasing data volumes and end-user requests, it is imperative for a DBA (Data Base Analyst) to ensure short build times to maximize uptime as well as minimize response times for end-users running predefined queries. As a DBA, I have extensive experience in this topic and would have much to add to this topic, but would like to formally investigate this through interviews and discussions with other DBAs in similar solutions, as well as research related Academic papers related to this topic.

**Proposed Project Title**: *Techniques to Minimize Build Times and End-User Response Times in a SQL-Based Data Warehouse*

**Proposed Project Purpose:** This project will be an exploratory case study, looking in detail at techniques related to SQL build times and response times and discussing the underlying theory to minimize both.

**Project Objectives:** Objectives are as follows:

1. Quantitatively characterize the use of SQL joins and selections, and how they are affected by primary keys and indexes, as well as factors limiting the performance of these operations (Disk access rate, table size, and join conditions/number of rows and tables joined)
2. Based on the information in objective 1, research and discuss techniques for minimizing recurring build times, including the use of incremental data builds and the MERGE statement.
3. Based on the information in objective 1, research and discuss techniques for minimizing response times for predefined queries, including the use of reporting tables and proper indexing
4. Characterize the trade-off between objectives 2 and 3 since minimizing response times will increase build times, and how this is truly a trade-off between space and time.

**Project Timeline with Activities:** Proposed timeline is as follows, with Activity Updates as waterfall events:

|  |  |  |  |
| --- | --- | --- | --- |
| **Week** | **Task** | **Deliverable** | **Date** |
| 1 | Idea Submission | **Idea Submission** | **9/11/2017** |
| 2 | Proposal/Timeline | **Proposal/Timeline** | **9/18/2017** |
| 3 | Full Paper Outline and Introduction |  | 9/25/2017 |
| 4 | Schedule Interview times | **Activity Update 1** | **10/2/2017** |
| 5 | Objective 1: Research SQL, Finish section 1 |  | 10/9/2017 |
| 6 | Interview Ted Davis | **Interview 1** | **10/13/2017** |
| 6 | Summarize Interview 1 | **Activity Update 2** | **10/16/2017** |
| 7 | Objective 2: Characterize Build Times |  | 10/23/2017 |
| 8 | Objective 2: Finish section 2 | **Activity Update 3** | **10/30/2017** |
| 9 | Project Post | **Project Post** | **11/1/2017** |
| 9 | Project Follow Up 1 | **Project Follow Up 1** | **11/2/2017** |
| 9 | Project Follow Up 2 | **Project Follow Up 2** | **11/3/2017** |
| 9 | Objective 3: Characterize End User Performance |  | 11/6/2017 |
| 10 | Objective 4: Characterize Time/Space Tradeoff |  | 11/13/2017 |
| 11 | Interview Jason Pettit | **Interview 2** | **11/20/2017** |
| 11 | Summarize Interview 2 | **Activity Update 4** | **11/20/2017** |
| 12 | Objective 3 & 4: Finish section 3 |  | 11/27/2017 |
| 13 | Complete Conclusion section | **Activity Update 5** | **12/4/2017** |
| 14 | Add any relevant graphs, charts, notes, etc. |  | 12/11/2017 |
| 15 | *(Buffer time- Paper should be complete)* |  | 12/18/2017 |
| 16 | Final Review and Submit Final Paper | **Final Paper** | **12/22/2017** |

**Interviewees:** The first interview is planned to be Ted Davis of Werner Electric in Minnesota. He is employed at a sister company to mine, performing the same duties and using the same tools I use, and faces the same challenges I face with SQL and Data Warehousing. The second interview planned is Jason Pettit of Kohls department stores. He works with high-performance tools to capture data. He has been interviewed in the past about Kohl’s marketing strategies; however, networking with Jason may find a better counterpart at Kohls who knows/uses SQL there, and I may interview that person instead.

**Application of Data Science Concepts:** This project will build upon concepts learned in Data Science 715 (Data Warehousing), namely the use of Fact and Dimension tables, indexes, the ETL process, and SQL. However, this will also cover more advanced topics such as the underlying set theory of SQL, the use of advanced SQL statement and structures, the use of reporting tables, and the time-space tradeoff inherent in reporting tables. Related theory from Data Science 730 (High Performance Computing) will be used since optimizing build times requires the use of parallelization (distributing SQL commands across server cores) as well as a staged/waterfall approach. This is highly relevant to SQL because cloud computing services with scalable processing and storage, such as Microsoft Azure, are anticipated to be the future of Data Warehousing, and SQL installations are commonly done on virtual servers for scalability and parallelization reasons.

**Description of Final Document:** The final document will result in a detailed paper approx. 50-100 pages in length covering in detail techniques and structures for improving SQL data warehouse performance. This will have three sections.

The first section it will discuss set theory that SQL is based on, and how joins work, as well as the mathematics behind joins, indexes, and primary keys; this topic will be the foundation for the last two topics.

The second section, it will cover minimizing time for recurring builds, including topics such as the use of ETL and Primary keys, the methodology of a parallelized waterfall-style build (as well as server limitations based on the number of cores and type of storage devices), the use of incremental vs full data builds, handling incomplete builds and/or long downtime, generating metadata on the build process for further analysis, as well as other related topics (including comparison to immediately replicated databases, as well as the strengths and weaknesses of ‘live’ vs replicated databases).

The third section will cover minimizing response times of reports and queries using stored procedures, properly created indexes, and the use of reporting tables and other techniques; this topic will also cover how the use of reporting tables directly conflicts with minimizing build times, how to properly handle added and removed data in a denormalized reporting table, and the theory around the time vs space tradeoff inherent with reporting tables.