**MSDS 7330 Project Proposal**

**Project Title:** An Introduction to Massively Parallel Processing Databases

**Project Team Members:**

* Jayson Barker
* Brandon Croom
* Neha Dixit

**Project Description:**

The project is framed in the context of educating our peers on the topic of Massively Parallel Processing (MPP) databases. It falls under the context of a wider view into the topics of databases and database capabilities that are currently not presented in the MSDS 7330.

The goal of the project is to provide education to peers on the topic of MPP databases. Peers should come away from the project with an understanding of the following topics:

* What is an MPP Database?
  + Parallel processing
    - What is it?
    - What are the components?
* Why use an MPP Database?
  + Scalability
  + Teradata
    - First to market with parallel processing at scale
    - Been in business for many years
  + I/O load balancing
  + Optimizer benefits
* Types of MPP Databases?
  + Teradata
* How does an MPP Database differ from a traditional database?
  + Teradata
    - Amps
    - Spool
    - Optimizer Quirks (collect statistics, volatile tables vs. temp tables)
    - Skew
* Typical MPP Database Architecture
  + Server
    - Physical
    - Cloud
  + Processors / AMPs
  + Interpreter
* Constraints of MPP Databases
  + Proprietary hardware and/or software
    - Locked into a single vendor - costly
  + Software can be behind the curve compared to open source
    - (SQL Assistant – specific to Teradata – very bare bones SQL interpreter compared to TOAD / Microsoft SQL Management Studio)
* Optimization of MPP Databases
  + Teradata
    - Compression
      * Block level
      * Multi-value compression
      * Algorithmic level compression
    - Indexing
    - Leveraging Optimizer
* Examples of MPP Databases in Industry
  + Blue Cross Blue Shield of Texas and Illinois
    - Heavy Teradata shop; uses it to house billion row data warehouse called: Enterprise Data Warehouse (EDW)
      * Each department builds “data marts” from this EDW that have pre-defined joins / views that limit the number of joins that need to be re-created by end users / analysts

**Project Deliverables**

* A PowerPoint presentation – useful for students to have as a quick reference guide
* A paper – useful for students who want to dive deeper than a presentation
* A video – useful for students who want to obtain additional information
* A reference section – useful for students to perform additional research