# Problem Statement

*October 9, 2017*

Alex Schultz
CS461, Fall 2017

Abstract: This document serves to define and describe the problem Hewlett-Packard (HP) prescribed, as well as propose the general solution to said problem. A brief overview of the company's current environment in which HP's Capstone Team ("team") will work will shed light on available or plausible methods. HP's problem is that their Big Data System, which gathers data used for troubleshooting product issues and doing business analytics, processes and stores 350 Gigabytes (GB) of data per day, but they do not know how to efficiently manage the data. They desire to effectively use compression to reduce the storage footprint while maintaining or even improving query performance using an Oracle database. "This project will take a fundamental look at storage size and query performance while implementing a variety of compression techniques" [1]. The team will measure CPU (Central Processing Unit) usage, Disk I/O (Input/Output), Access Path, Memory, Time, etc., to determine the most performant data storage techniques [1].

# 1 Introduction

"PageWide Web Press... a printing division within HP Corvallis, has a large-scale database extensively used for troubleshooting product issues and doing business analytics" [1]. "[Their] products produce 350GB of data per day and create database tables with billions of rows" [1]. The database is currently run by one machine that has 72 cores and 3TB of Random Access Memory (RAM), but some queries take up to three hours to finish. The only current data de-duplication method in place is storing words as numbers.

# 2 Current Proposed Solutions

## 2.1 Oracle's Advanced Compression

One method to consider when conducting research is Oracle's Advanced Compression (OAC) technique. OAC "provides a comprehensive set of compression capabilities to help improve performance and reduce storage costs" [2]. It allows organizations to reduce their overall database storage footprint by enabling compression for all types of data, and it also improves query performance with Advanced Row Compression (ARC). Using OAC allows compressed data to be read in blocks directly without uncompressing the blocks, which helps improve performance due to the reduction in I/O as well as the reduction in system calls related to the I/O operations [2]. Furthermore, the "buffer cache becomes more efficient by storing more data without having to add memory" [2].

## 2.2 File System and Storage Array Compression

Other compression techniques to research are File System Compression and Storage Array Compression. File System Compression (FSC) is a technique that compresses each file as it is written to disk [3]. Although FSC has been known to perform poorly, it should still be investigated to use selectively on highly compressible files that are written once and then accessed infrequently [3]. Storage Array Compression (SAC) is worth considering because HP currently uses SAN file storage rather than local hard disks, and other companies, such as Sun Microsystems, were successful in implementation.

# 3 Initial Meeting

After the team met, the scope of the project was clearly defined; this is a research-focused project, which means there is not going to be a specific process or workflow to the next 9 months. HP stated it is going to be a "follow-your-nose" type project, in which the team will research, implement, test, repeat. The team is finished with the project as soon as they provide HP with a satisfactory or significant improvement regarding the following: maximizing CPU efficiency, reducing Disk I/O, improving storage, data compression, and data decompression. Metrics are difficult to measure in a research-based project, but they will be happy if the team is making progress in research and have at least tried to implement three strategies.

Considerations discussed in the meeting include:

- What happens to compression as block size increases?
- Is there a difference in performance if rows are entered in specific ways?
  - What's the effect of reordering row input (entering as ABC vs. CBA)?
  - What's the effect of entering rows randomly (ABC vs. CAB)?
- What's the effect of storing data as INTs, FLOATs, or NUMs?
- Can one normalize data efficiently?

# 4 Future Work

The team collectively decided to try finishing all writing assignments as soon as possible so research may begin as quickly as possible. Over the next few weeks, writing is the focus of the project, along with getting set up with a test developmental environment. Once the writing portion of the team's project is complete, research will begin. Implementation follows a collectively selected proposed method. If performance increases, the team will move on to the next method to discover an even more efficient solution.

# 5 References

[1] http://eecs.oregonstate.edu/capstone/cs/capstone.cgi?project=413
[2] http://www.oracle.com/technetwork/database/options/compression/overview/index.html
[3] http://searchitchannel.techtarget.com/feature/Top-five-data-storage-compression-methods