Problem Statement

Dylan Davis, Trevor Hammock, Alex Schultz
Oregon State University
Corvallis, Oregon

October 18, 2017

Abstract

The purpose of this document is to give a general description of the goals for the HP Data Compression capstone project, as well as the current plans to accomplish those goals. HP's PageWide Web Press division develops and troubleshoots their industrial web presses. These web presses send over 350GB a day of data, which enables business analytics and product troubleshooting. Thus, HP is faced with a storage/performance dilemma for their team's back-end database. This document will describe the general direction the research will take, including the variables that will be changed and the outcomes that will be measured. This team will research and test various compression options available for Oracle Databases, and, once we have gathered the necessary data, we will then propose and implement a solution, ensuring HP is satisfied with the performance enhancements. Our findings will be presented to the Hotsos Oracle Performance Tuning conference in Spring 2018.

1 Introduction

PageWide Web Press, a printing division within HP Corvallis, is responsible for developing and troubleshooting HP's industrial web presses. These web presses are used by various companies for very large-scale digital printing operations. The PageWide Web Press team regularly receives business analytics and product issue data from all of the web presses in the market, which eventually gets stored into an Oracle Database. The team then uses this information to fix and/or enhance their web press products.

The printers, at the time this paper was written, produce over 350GB of data per day, which tends to create database tables with billions of rows. Additionally, the amount of data generated per day is slowly increasing over time. This is problematic because if no action is taken, then the server hosting the data will exceed its storage capabilities and will also struggle to process said the data. If a more optimial configuration for Oracle Database is found, HP will be able to access, store, retrieve, and analyze their printing data much more efficiently.

2 Proposed Solution

To find the most efficient implementation for HP's Oracle Database system, the research for this project will focus on the ways in which data is compressed and stored. The experiments will be performed on a mock database using the same Oracle Database architecture used by the PageWide Web Press division. Variables include the size of the blocks on which the data will be stored, how the data is stored, and what Oracle compression algorithms will be used. Any feasible combination of inputs for these variables will be tested in order to find the optimal configuration of the database system best suited to HP's needs. Once each configuration is set up, queries will be performed on the data and performance will be benchedmarked. Performance will be based on several factors including CPU efficiency, speed, storage space, and disk I/O; each will be tested on 8K, 16K, and 32K block sizes.

3 Project Goals

We will use the following performance metrics throughout our project:

First, we will research various compression solutions for our client's current database environment. Once we find a feasible solution, we will weigh the pros and cons associated with that approach against the expected outcome. For example, if a solution would take months to implement, we would note it as a possible solution, but ultimately disregard it in the current scope.

Second, we will use our research to design and run a set of experiments, which will measure a moderate set of factors including data size reduction, CPU efficiency, disk I/O, and time. We will then analyze the results and draw conclusions about which solutions performed the best.

Third, we will use the results from the experiments to propose a solution for our client's current issue. Once we have isolated a solution, implementation follows. Benchmarks will reveal improvements or possible regressions.

Finally, we will compile our findings into a written analysis to serve as a guide for others facing similar issues. HP will use this research as the foundation of their presentation during Oracle's conference in Spring 2018.