Problem Statement

Dylan Davis
Oregon State University
Corvallis, Oregon

October 16, 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. Because this is a research project, 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. It will also cover the expectations of Hewlett Packard, Inc going into this project as well as their desired outcomes by the end.

1   Introduction

Hewlett Packard, Inc maintains a large scale database for their industrial printing data which receives 350 gigabytes of new data a day from all of the distributed HP industrial printers. They use an Oracle database architecture to run this database. Currently, accessing and analyzing this data can take a very long time due to the massive amount of data and how it is stored. For this research project, we hope to find a more efficient way of utilizing the Oracle database system and its features to improve performance. If a more efficient configuration of the Oracle database is found, HP will be able to implement the necessary changes and make accessing and using their printing data that much easier.

2   Proposed Solution

To find the most efficient way to utilize the 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 dummy database using the same Oracle database architecture used by HP. The different variables that will be tested in the experiments will include the size of the blocks the data will be stored as, 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 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 measured and recorded. Performance will be based on several factors, including CPU efficiency, speed, storage space taken, and disk I/O.

The different block sizes to be tested for this research project are 8K, 16K, and 32K blocks. The ways in which the daata is stored include how the data is represented in the data tables as well as how the rows within the tables are ordered. Research will need to be done on the available Oracle compression algorithms so that a working understanding of how each one works can be gained and the positives and negatives of each can be studied.

3   Project Goals

The outcome for this project will be a research white paper about Oracle database performance. This contents of this paper will be all of the relevent findings from the experiments performed over the course of this research project as well as any conclusions drawn from them. Hewlett Packard will also present the results of the research done for this project at HotSOS 2018, a national Oracle conference. Ideally, this project will provide information that HP can use practically to improve the performance of their Oracle database.