HP Big Data Analytics: Effects of Compression on Query Performance

Trevor Hammock
October 2017
CS 461
Fall Term

Problem Statement

Abstract

HP's PageWide Web Press division develops and troubleshoots their industrial web presses. These web presses send over 350GB a day of business analytics and product issues. This has caused a storage/performance dilemma for their team's back-end Database. To solve this my group will research and test various compression options available for Oracle Databases. Once we have gathered the necessary data, we will then propose and implement a solution for the current Database. After that has taken place we will then compile all of our findings into a report and submit that to a Oracle Performance Tuning conference in the Spring

HP Big Data Analytics: Effects of Compression on Query Performance

1 Problem Definition

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. On a routine basis the PageWide Web Press team receives business analytics and product issues from all of the web presses in market which eventually gets stored into a Oracle Database. The team then uses this information to fix and/or enhance their web press products.

However their products, at the time this paper was written, on average produce over 350GB of data per day which tends to create Database tables with over 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, the server hosting the data will exceed it's storage capabilities and will also struggle to lookup/process said data. Most likely resulting in the team having to spend exorbitant amounts of money to upgrade the hardware for their storage needs.

2 Proposed Solution

For my senior capstone project, my group members and I will examine the effectiveness of data compression on the Database. If implemented properly, data compression should significantly reduce the amount of space that the data currently occupies and maybe even improve query performance at the same time (due to less I/O operations and better caching). With this proposed solution, the storage and processing requirements for the data should scale down and help delay a very costly hardware upgrade.

3 Performance Metrics

For this project my group members and I will use the following performance metrics throughout our project:

First we will acquire some background research into the various compression options for our client's current Database environment. Once we know this we will then be able to figure out what we can/can't do, the pros/cons associated with a given approach, and the expected outcome.

Second we will then use that information to design and run a set experiments. These experiments will validate the outcomes we came up with for the various compression options in our research. The experiments will measure a moderate set of factors which will include overall size reduction, CPU usage, Disk I/O, time, etc. We will then analyze the results and come to a conclusion about which approaches performed better than others.

Third we will then use the results from the experiments to propose a solution for our client's storage issue. Once that is done we will then implement that

solution to the underlying database. After that is completed, we will then perform some benchmarks to see what exactly improved or regressed from our given implementation.

Finally we will then compile all of our findings into a written analysis to serve as a guide for others facing similar issues. It will then be presented at an Oracle Performance Tuning conference in the Spring.