Progress Report

Dylan Davis, Trevor Hammock, Alex Schultz
Group 18
Oregon State University
Corvallis, Oregon

December 4, 2017

Abstract

This document serves to inform all interested parties of the progress made within the last 11 weeks. From initialization to current work, a week-by-week synopsis details what the team contributed. Additionally, project purposes, goals, problems, and solutions are briefly discussed.

Contents

1 Project Purposes and Goals

PageWide Web Press, a printing division within HP, produces and supports industrial digital web presses. All of the web presses currently deployed in the market generate product data, which the PageWide Web Press team receives on a daily basis. The data is eventually stored into a Oracle Database and is used to perform business analytics and resolve issues with the web presses.

The large amount of data created some issues for the division; on average, they receive around 350GB of product data per day, which generates Database tables with billions of rows. This massive influx of data has stressed the division's storage and performance capabilities of their current hardware. The stress on the hardware significantly increases the amount time it takes to process and analyze that data, and the division wants to find a more efficient data storage solution. Furthermore, the amount of data that is received everyday has slowly been increasing over time. The current rate of 350GB of data per day will increase, so preparation and efficiency are essential.

If nothing is done about the data collection process, the division may need to make a costly investment into better hardware that satisfies their growing storage and performance needs. Alternatively, they may have to invest into big data software platforms, like Hadoop, that are more optimized and scalable for their given workloads. This project will address the aforementioned issues by using system enabled compression to reduce the amount of physical space that the data occupies. The overall goal for this project is to research system enabled compression options that are available to Oracle Databases and see how these options affect physical storage and query performance. Knowledge obtained from research will be used to accurately predict and measure the way in which data is inserted and stored in the Oracle database.

2 Week by Week Synopsis

2.1 Week 1

The team was initially tasked with submitting project preferences. After investigating all of the available projects, the top 5 most interesting projects were selected. Project preferences and proposals were submitted, and the group was formed.

2.2 Week 2

Meeting with both the team members and the client were the primary goals for week two. A method of communication was established when the team met, and the client provided brief introduction into the problem at hand. Information gathered from the client meeting was used to compose the problem statement.

2.3 Week 3

The third week of Fall term was really when the project began. The client advised the team on environment needs and how it will relate to their environment. Once needs were established, the team then started configuring environments in order to run the necessary experiments. Additionally, the group was tasked with learning about database terminology, data blocks, and compression options available to Oracle Databases. Moreover, the team completed some group related tasks that would aid in the development of the project, including creating a source control repository, designating a dedicated note taker for client meetings, and scheduling meetings with the teaching assistant (TA).

2.4 Week 4

The client charged the team with researching Oracle compression options during week four, and, after accomplishing the assignment, decided to focus on table compression and left the other options as stretch goals if time permits. After agreeing on usable compression options, the user and table nomenclature was established. Standardizing environment variables allows the team to work cohesively without concern over differences when experimenting or analyzing performance. Finally, the client required research into Oracle table compression options and sought to determine which options were free and/or required licensing. Given all of the previous tasks, the team had enough to being writing the requirements document for this project.

2.5 Week 5

During the fifth week the client and team finalized the project timeline, which serves as a general guide for when milestones should be completed. A discussion regarding the relatively free compression options followed the timeline. By the end of week 5, the team finally configured individual environments and also found ways to automate database initialization to help streamline research and testing. The team also met with the TA, Andrew, for the first time.

2.6 Week 6

Progress occurred in week six that eventually snowballed into completing all tasks assigned by the client. Data block research and learning how to dump data blocks was extremely crucial to the project because the ability to find where and how the data is stored, coupled with the ability to unpack its contents, provided a solid understanding of how the data gets compressed. With hard work, the team figured it out and started researching into the details of Oracle trace files to understand what is stored inside them and how they are formatted. This step segued into research into reverse engineering the data blocks, which allows the team to conclude if the underlying data was compressed. Additionally, the final draft for the requirements document was submitted.

2.7 Week 7

A breakthrough in research occurred in week seven. The team managed to find a very informative article that explained the structure of the trace files, how data gets compressed, and how to examine compression within the trace file. To verify this information, the client tasked individual members of the team with testing different components of the compression process. Alex was tasked with testing column reordering between the data blocks; Dylan with finding any differences between blocks with different compression options; and Trevor with testing column reordering within an individual block. Individual rough drafts of the technology review document were also created.

2.8 Week 8

In the eighth week, the team presented results from individual tasks to the client. Pleased with the results, yet still curious, the client assigned more tasks to obtain a more detailed understanding of Oracle's table compression algorithm. Alex was tasked with further examining column order between the data blocks; Dylan with compressing a small set of hand-written data and rebuilding it using the trace file, which would shed some light into how the compression algorithm functions; and Trevor with testing column reordering again, but with pre-sorted data to help remove confounding variables from the previous experiment. Each member submitted a rough draft of the technology review and began working on the final draft.

2.9 Week 9

Once again, the team presented results from the individual tasks. The client then decided to shift focus toward data cardinality and data length to acquire more knowledge of the underlying compression algorithm. Alex was tasked with investigating extent trimming, which affects how data blocks get allocated; Trevor was tasked with testing column tokens and also verifying if the compression algorithm was data de-duplication only. Additionally, the team registered for the engineering exposition on top of finishing and submitting the final drafts for the technology review.

2.10 Week 10

Individual research from week nine was presented to the client. By week ten the team had a pretty solid understanding of how data is compressed, so it was time to start designing some experiments. After discussing potential experiments with the client, the team managed to construct a few. More individual research was assigned: Alex was determined to figure out when and why columns are reordered; Dylan continued to study data blocks to verify what was actually being stored; and Trevor performed more tests with column tokens using interleaved columns. The team began and finished the design document, and began working on the progress report and presentation.

2.11 Week 11

The Fall progress report and presentation will be submitted. The group will also meet with the client to present their individual research and continue to work on the experiments.

3 Current progress

Project members, initially, did not understand the low-level inner-workings of data blocks, which are the smallest units of storage in an Oracle database. After finding an immensely helpful article by Jonathan Lewis, the team was able to experiment with data block analysis and obtained a fundamental understanding of how they work and what information they contain. Continuing from that research and experimentation, the team successfully reverse engineered data blocks to understand exactly what happens to the uncompressed and compressed data. The ability to reverse engineer data blocks led the team to reverse engineer compression algorithms - to understand precisely what the compression algorithms are doing to the data at the lowest level. Experimentation was the driver that led to a firm understanding of what to expect when inserting large amounts of data. The team may now deduce what happens to column order and the size of the data depending on the way in which the data is inserted.

4 Problems and Solutions

The only problem experienced in the past 10 weeks is that two meetings had to be rescheduled, but quick and open communication immediately rectified those issues. These specific meetings were rescheduled from Wednesday to Friday, so the team never went a week without meeting. Despite the research-oriented project, each milestone was met punctually.

5 Retrospective

Positives	**Deltas**	**Actions**
Obtained a solid understanding of data blocks	Begin project deliverables sooner	Assign roles and hold each other accountable
Found an article detailing data block dump files and compression	Meet more often	Meet after talking with the TA since each member will already be there
Began experimentation	Communicate more	Hold each other accountable if there is a lack of communication

6 Conclusion

Fall term was used to familiarize the team with basic concepts and environment configuration. Fortunately the team was able to begin research and has made an incredible amount of progress thus far. Looking ahead, the team will experiment more and develop CPU, I/O, and memory tracking scripts to analyze performance.