Requirements Document

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

November 4, 2017

Abstract

This document describes the requirements for completion of the HP data compression capstone project. Because this is a research based project, the requirements are given as research milestones describing when the group plans to start and complete the different areas of research that will be covered over the course of the year. Also within this document are descriptions of the purpose and scope of the project, definitions of relevant terms, and an overview of the various pieces of software that will be used to run and analyze experiments. References to important sources of information pertaining to the project are provided as well as stretch goals to be explored if the primary goals of the project are fulfilled.

Contents

1 Introduction

1.1 Purpose

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.

However, this data is currently creating some issues for the division. On average, the division receives around 350GB of product data per day, which generates Database tables with over 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 could turn into 450 to 500 GB of data very quickly.

If nothing is done about the data collection process, the division may need make a costly investment into better hardware that can meet 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.

1.2 Scope

The current scope of the project is to investigate various Oracle database compression options and use them to implement a solution for PageWide Web Press' storage crisis. The solution will not alter or restructure the data itself, but rather change how the Database system stores said data into the physical storage medium. It is not a requirement to actively look into or investigate query performance options for this solution (like Oracle's in-memory compression for example [1]).

Essentially, the goal of this project is to address the division's data issues by using compression to reduce the amount of space that the data occupies. If implemented properly, compression should help delay a costly hardware/software upgrade for their storage needs and hopefully improve the query performance for their database system.

1.3 Definitions, acronyms, and abbreviations

Term	Definition
Block	The smallest unit of storage in an Oracle Database
Central Processing Unit (CPU)	Physical processor that executes instructions
Command Line (CLI)	A application which provides a means of interacting with a computer program by entering lines of
Database	Software that stores data in an organized fashion
Database Administrator (DBA)	A person responsible for the efficiency of a system's database including security, scalability, reliability, and performance
Hadoop	Software framework optimized and designed for processing big data
Hewlett-Packard (HP)	Multinational IT company that is infamous for their printing products
IDE (Integrated Development Environment)	A software application that provides comprehensive facilities to computer programmers for software development
Input/Output (I/O)	Typically associated with the rate of how data is read or written to a physical storage medium
Oracle	Multinational IT company that specializes in enterprise solutions such as Database software and cloud systems
Oracle Database	A relational database system that is produced and supported by Oracle
Oracle Linux	Distribution of Linux that is maintained by Oracle. It is optimized platform for many of their products, such as Oracle Database
PageWide Web Press	A printing division within HP that develops and supports HP's industrial digital web presses
Query	A request that processes and/or returns data from within a database
Oracle Recovery Manager (RMAN)	An Oracle product that is used to backup and recover Oracle Databases
Secure Shell (SSH)	Cryptographic network protocol for operating network services securely over an unsecured network.
Stakeholder	Someone who is not a DBA or developer but is otherwise involved

1.4 References

T. Chien and S. Wertheimer, Recovery Manager (RMAN) Performance Tuning Best Practices, Jun-2011. [Online]. Available: http://www.oracle.com/technetwork/database/focus-areas/availability/rman-perf-tuning-bp-452204.pdf. [Accessed: 03-Nov-2017]

A. Dadhich, K. Patel, F. Khan, and S. Pamu, Advanced Network Compression, Dec-2013. [Online]. Available: http://www.oracle.com/technetwork/database/enterprise-edition/advancednetworkcompression-2141325.pdf. [Accessed: 03-Nov-2017]

Oracle Advanced Compression with Oracle Database 12 c Release 2, Sep-2017. [Online]. Available: http://www.oracle.com/technetwork/database/options/compression/advanced-compression-wp-12c-1896128.pdf. [Accessed: 03-Nov-2017]

A. Rivenes, M. Colgan, and V. Marwah, Oracle Database In - Memory with Oracle Database 12c Release 2 Technical Overview, Aug-2017. [Online]. Available: http://www.oracle.com/technetwork/database/in-memory/overview/twp-oracle-database-in-memory-2245633.pdf. [Accessed: 03-Nov-2017]

1.5 Overview

The second section of this requirements document gives an overview of how the project interacts with other systems, intended functions, characteristics, given constraints, and assumptions. This section is written for the individual stakeholders and examines the project at a much higher level. The third section of this requirements document details the individual requirements that will be implemented. This section possess a lot of technical aspects/terminology and is intended for the implementers for this project. Together these sections form a comprehensive overview for this document.

2 Overall Description

2.1 Product perspective

This activity is not a standalone piece of software but rather it runs on top of a pre-existing system. This project will emulate a Database environment that closely matches the client. The environment will based on a Oracle Linux distribution which will host an Oracle Database instance, and it will serve as the testing grounds for the project's tests and analysis. This will allow the individuals working on this project to make conclusions about Oracle's compression for their Databases.

2.1.1 System Interfaces

The primary system interface will be the operating system for Oracle Databases, Oracle Linux 7. The Databases that we will create and test will run on this distribution of Linux. For system analysis, system calls will be made to gather any relevant information/data.

2.1.2 User Interfaces

The user interfaces will be the Linux command line and SQL Developer IDE. The command line and the IDE will both be used to run queries on the Database to interact with it. Additionally the command line will be used to alter system settings or parameters, such as starting up the Database instance.

2.1.3 Hardware Interfaces

No hardware interfaces are involved; the task is software based, and a virtual machine will be used to emulate the client's system.

2.1.4 Software Interfaces

Oracle 12.2 Database instance on top of Oracle Linux 7 environment. The following system software is required for this project:

- Oracle Linux 7
  - Mnemonic: OL7
  - Version: 7.4
  - Source: Oracle
- Oracle Database 12c Enterprise Edition
  - Mnemonic: 12c
  - Version: 12.2.0.1.0
  - Source: Oracle
- Oracle SQL Developer
  - Mnemonic: N/A
  - Version: 17.3.0
  - Source: Oracle
- Linux Command Line Interface
  - Mnemonic: CLI
  - Version: 3.10.0
  - Source: Linux

2.1.5 Communications

We will be remotely connecting and running commands/scripts to the emulated system. The network configuration will be done through the routine Oracle Database hosting procedures. Once the network configuration is setup, SQL Developer and SSH terminals will be used to remotely interact with the system.

2.1.6 Memory Constraints

The Database instance and all of the tables containing our data will need to fit within the allocated space for the virtual machine (60GB; 4GB RAM).

2.1.7 Operations

There will be multiple modes of operations, including sys level access and basic user access. Sys access will be used to manage the underlying system such as modifying how the data in the Database gets stored. User access will be used for generating and testing the data. Most of the periods of operation will be run in the background unless it is explicitly forced to run. Should time allow, and the backup/recovery options are explored, then explicit backups will be made for testing purposes.

2.1.8 Site Adaptation Requirements

There are no site specific requirements for this project.

2.2 Product functions

There are no explicit functions for this project.

2.3 User characteristics

Users interacting with the system and performing the analyses will be Database Administrators (DBA's) that manage Databases for their systems or product stacks. The users should be proficient in SQL, Oracle Database specifics, and Unix/Linux system administration.

2.4 Constraints

Only compression options available in Oracle Database 12.2 will be considered. System characteristics and properties will also be analyzed; for example, the effect block size has on the varying compression algorithms. Compression options with stricter licensing or hardware dependencies will be ignored; in other words, if it's not free, it's not a viable option. The data, and how it is structured, will be left intact; however, data stored in the physical storage medium will change. Compression options that show little to no benefits will be ignored in the final implementation. Data properties and characteristics will also be examined, which will include:

- How the columns are ordered
- The cardinality
- Repetition within the datasets

2.5 Assumptions and dependencies

It is assumed that all of Databases used in this project will be installed on a distribution of Oracle Linux 7. Additionally, Oracle Databases used in this project will be version 12.2, and the system performing the benchmarks for this project can record system statistics, such as CPU usage, within a reasonable time resolution.

2.6 Apportioning of requirements

If time permits, the following stretch goals may be investigated and potentially implemented:

- Enabling Oracle's new Network Compression [2] and see the effect it has network bandwidth.
- Researching and/or playing with Oracle's Advanced Index Compression [3] to investigate additional space/performance savings.
- Oracle's Backup Compression (RMAN) [4] to address future storage issues with their regularly scheduled backups.

3 Requirements

All tests and analyses will be performed on systems/data that mirror PageWide Web Press' current environment. The members conducting this project will learn about and reverse engineer data blocks in Oracle to understand how the data gets stored. Oracle compression algorithms will be tested and reverse engineered (to the best that can be inferred) to understand how it interacts with a sample set of data. Once the background research is complete, the project facilitators will then design a set of experiments to prove/test the findings they found during the research phase. Experiments include creating and using a set of scripts that will track critical system information such as CPU, I/O, and/or Memory usage. Once all of the data from the experiments is collected, it will be compiled into a white paper that will summarize the research and findings. Finally, the white paper will be presented at an Oracle Database conference in the Spring entitled Hotsos. The remainder of the project will be reserved for design and possibly integrating a solution into PageWide Web Press's current environment.

Gantt Chart outlining the tasks and tentative due dates:

	Task Name	Start Date	End Date	Duration	% Complete	Status
1	**Set up environment**	**10/18/17**	**11/01/17**	**11d**	**100%**	**In Progress**
2	Install Oracle Linux 7.4 in Virtual Box	10/18/17	11/01/17	11d	100%	Completed
3	Install Oracle 12.2.0.1 in the virtual machine	10/18/17	11/01/17	11d	100%	Completed
4	Configure SQL Developer and install tables	10/18/17	11/01/17	11d	100%	In Progress
5	**Research and reverse engineer data blocks and compression algorithms**	**11/01/17**	**11/29/17**	**21d**		**In Progress**
6	Obtain solid understanding of data blocks and how they work	11/01/17	11/15/17	11d		In Progress
7	Reverse engineer data block	11/01/17	11/15/17	11d		Not Started
8	Reverse engineer compression algorithms	11/01/17	11/29/17	21d		Not Started
9	**Design Experiments and develop tracking scripts**	**11/29/17**	**03/09/18**	**73d**		Not Started
10	Design experiments	11/29/17	01/17/18	36d		Not Started
11	Develop CPU, I/O, Memory usage tracking scripts	01/17/18	01/24/18	6d		Not Started
12	**Analyze and present results**	**01/24/18**	**03/09/18**	**33d**		Not Started
13	Analyze results	01/24/18	01/24/18	1d		Not Started
14	Finalize paper so HP can present at Oracle conference	02/15/18	03/09/18	17d		Not Started