Progress Report

HP: Big Data Analytics
Group 13

Oregon State University
CS 461
2016-2017

Prepared By:
Nic Desilets, James Stallkamp,
and Nathaniel Whitlock

December 7, 2016

Abstract

This document serves as the outline for each of our engagement in our senior design project with HP. The time line which has been laid out is an amalgamation of each of our experiences, therefore we composed this document as a group. Individual experiences are discussed in the video presentation.

Add more?

CONTENTS

1 INTRODUCTION

The main purpose of this document is lay out where we are in the project as Fall term of 2016 comes to an end and recap what we have experienced and learned so far. The following section will reiterate the purpose of our project and also the goals that we hope to ultimately accomplish. Our experiences in the project to date are summarized and broken down into a week by week digest of problems encountered, progress made, and any intermediate goals between weeks. Lastly, a retrospective has been put together of our positive experiences, outstanding issues, and future actions that need to be taken.

2 PURPOSE AND GOALS

The purpose of this project is to build a toolkit to measure system performance metrics from the processing of a SQL query. This will ideally be accomplished by writing a wrapper query that calls out to Oracle native dynamic performance views in order to gather data about a query which just processed.

The goal is to provide a performance report which can be further used for query performance tuning by DBA's and engineers. Additionally, we hope to use this toolkit to further investigate partition design performance, as well as compare any other configurations that we might run against each other. The performance output from these experiments can be used to comparatively perform run-time analysis for each configuration.

3 CURRENT STATE

As a group we are still in pre-development preliminary stages. Though we have started experimenting there has been no formal documentation of development efforts. Much of the writing done during this term has been a planing process to help us identify the requirements needed to fulfill our obligation to our client. Additionally, it has helped to identify deign decision which need to be contemplated before moving forward.

Therefore, the current state of the project relies mostly on planning based documentation. We hope to use what we have learned to guide us on the development process. The documentation will not only help us in further endeavors, but will also serve as a contract with our client that we can refer to later if we encounter any scope issues or complaints about not providing the requested final product.

4 TIMELINE OF EVENTS

4.1 Weeks 0-3

During the first couple weeks of the course we were not yet aside to groups. In lecture we discussed the overall structure of the course, including some of the assignments we would be given, then we individually voted for the top five projects we were interested in.

Once week three came around we finally had our first official meeting with the client. This meeting went pretty well, however, the client had not reserved a conference room for the meeting and the room we were in was booked half way through our meeting. We were forced to find and end of isle cubicle where we spent the remainder of the time. Our project mentors are Kirby Sand, he has a background in chemical engineering and does a lot of the data analytic work in the team, and Andy Weiss, who is an Oracle database administrator. As a team they support the database and provide summary statistics of WebPress performance data.

During the meeting we were exposed to information about the inner workings of an Oracle instance. We also spent quite a bit of time talking about the data dictionary and dynamic performance views. In an Oracle instance, the data dictionary contains definitions for every table, view and index. Dynamic performance views were described as predefined tables

loaded with information pertaining to performance metrics from executing a query. Finally, we discussed the software we would need to install on our personal computers and scheduled the remainder of the meetings that we would have for the remainder of the term.

4.2 Week 4

Before we came together for the meeting this week we attempted to install Ubuntu and CentOS on Orcale Virtualbox in hopes to be prepared for the meeting. However, we ran into issues getting Oracle 12c installed on both Linux distributions. After talking with Andy about it, he recommended that we all install Oracle Linux 7.

During this week we as a team received feedback on our Problem Statement draft. It was somewhat refreshing, but also concerning, that we did not receive much feedback. Nonetheless, effort moving forward with the assignment was placed on taking a higher level approach to the problem. In doing so, we could make the problem more appealing to a generalized audience. Other efforts were spent ensuring the formatting was correct so that we did not lose points.

4.3 Week 5

This week we were still struggling with getting a Linux server and SQL developer installed on our local machines. The issues we faced were both hardware limitations and missing software dependencies. We spent a significant amount of time working with Andy, our database mentor, to Nathaniel switched to a laptop with more memory in order to get the oracle server to perform better. James also had to borrow a laptop from his family in order to have a personal laptop to use. Despite these difficulties we did manage to individually work through our issues and finished installing the software packages by the end of the week.

4.4 Week 6

Most of the effort this week was focused on working on the Requirement Document. During the course of completing this document, our team had a difficult time working on the assignment in concert. Rather, the assignment was completed in individual installments. One of the most difficult parts of this assignment was in following the IEEE standards outline. Upon reading through the IEEE standard initially it became clear that there would be inherent redundancy in our final product strictly based on the required contents of the document. This proved to be true.

Some of the major issues that we ran into during this assignment was trying to overwrite the IEEEtran.cls file. we were well aware of the fact that we were using a class file to style to final document, however, we were not aware of the fact that there were numerous option flags which could be specified in order to adjust the styling. This paper included a lot more structural hierarchy than our other assignments, when we needed to use section subclass tags we ran into issues with mixed Arabic and roman numeral headings. TO resolve this, Nathaniel looked up some examples on line and tried to forcefully change some of the display setting. Lesson learned from this mistake, don't try to override class file settings.

After talking with Jon Dodge about this, it turned out the IEEEtran.cls file that we had was a customized version that did not have the entire class file code. Once downloaded a proper version of the class file, and removed all of the override tags, the class file rendered our document in a visually appealing format.

Moving forward we started reading through an Oracle white paper titled, "Oracle Optimizer Explain the Explain Plan". This paper focused on outlining how a query is broken down into steps, how a cost is associated with each step, and how the optimizer goes with the plan with minimal cost. This was an essential concept in understanding how and why a query is processed in the way that it is.

4.5 Week 7

During our meeting with the client this week we discussed the Oracle Explain Plan document. From reading this paper we learned about how the step by step plan of execution of a query is generated, as well as some of important steps one might see in a plan. Some of these steps include: access methods, join methods, and sorting methods. By looking at an Explain plan, one can deduce potential issues with the optimizer choices in order to further investigate the query execution.

Kirby provided us with a PL/SQL script which could be modified to generate a specified number of rows of data which represent performance output from a WebPress. The goal of this script was to load the HP test server with data, as well as allow us to a smaller collection of data on our person machines to run experiments with.

4.6 Week 8

In our meeting with our mentors this week we got a better understanding of V$ tables, and were able to start querying metrics on specific queries. This week we did not have to spend time working on the design document in our team meeting, however we spend a significant amount of time individually. This week we all researched into our areas of the tech review. We split into three groups, James researched IDE's, Nic researched already existing tool kits, and Nathaniel researched Parallelism and Partitioning.

4.7 Week 9

During our meeting this week more emphasis was put on trying to create performance statistics based on that data available in the native dynamic performance views. We sat around at a conference table trying to Google what was available to us in order to create a plan for our query.

In terms of our class assignments, the design document had a lot of time spent on it during this week. My contribution to this document was really difficult to put together. As with the requirements document, there were many aspects of this document that seemed repetitive, however, each part focused on a different design domain. Talking to Job Dodge about how to approach the document was essential for us in understanding how to approach the assignment.

There were no major problems during this week. We were however told by the client that they were accepted as a speaker to an Oracle convention in Texas from February 27th through March 2nd. This is somewhat good news, but it also means that the development pressure for us has just increased quite a bit.

4.8 Week 10

During the meeting this week we continued on the development track, but this time switching focus to the topic of tracing. Tracing is the process of following the execution procedure for a query where each step is logged out to a file. This file can then be made more human readable by use of other Oracle native tools like tkprof. Running a tracefile through tkprof turns the raw file into a more human readable format. We hope to use this output to gain understanding of the procedural steps taken to process a query. Additional performance metrics can be extrapolated from this data set.

Some of the issues we faced during this period include not being able to get started on the Progress report in a timely manner, as well as getting the design document done in time to get signatures from the client.

5 RETROSPECTIVE

Positives	Deltas	Actions
Meshed with client	Development process	Begin building SQL performance toolkit
Increased understanding of Oracle system	Temp	temp
Temp	Temp	temp
Finished writing each document	Temp	temp

6 CONCLUSION

In summary, the majority of this term was spent on becoming familiarized with the scope of the problem that we will be working on, as well as generating formal documentation. There was also a decent amount of time spent on working through issues in installing our software packages. Though there have been minor experiments in development, they have been mostly focused on gaining a better understanding of what data we have available for us to access. It has been nice having our mentors at HP who can lead us through some of the more involved topics, at least in the sense of giving us direction or material to reference.

Moving forward, we will try to take what we have put together in planning documents and do our best to implement them. There will undoubtedly be more complications along the way, but the time we have spent working on documentation should point us in the right direction.