Preliminary Design Document

HP: Big Data Analytics
Group 13

Oregon State University
CS 461
2016-2017

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

November 17, 2016

Abstract

This document is focuses on a detailed brake down of ”how” the individual components of our project will be implemented. Each of the three essential pieces: SQL development environment; parallelization and partitioning; and reporting tools are covered with fine granularity.

Elaborate more here in the future...

CONTENTS

1 INTRODUCTION

This should include the full *HOW* of your system. This includes your API, your timeline, necessary testing information, etc. This is important for you as a guide. Remember, a week of debugging can save you 20 minutes of planning! Plan the work, work the plan. For research projects, this will look a little different. This will be more on the order of ”experiment design”, where you detail the junction points/milestones. You will also detail what will dictate your path from then on. You should provide an expected path through these points, any algorithms you will be using or designing, etc.

2 SQL DEVELOPMENT ENVIRONMENT

Miles Stones for getting IDE set up.

- Install local Oracle Instance
  - Download and install Virtual Box.
  - Download and mount Oracle 12c x64 iso.
  - Run Oracle 12c preinstaller.
  - Set up Oracle 12 server on Virtual Box.
- Install SQL developer
  - Download and install SQL developer.
  - Connect to local server.
  - See next step for connecting to HP test server.
- Connect to HP blade server(on campus)
  - Must be on HP campus to connect to test server.
  - Once connected to HP server ready to run experiments.

3 PARALLELIZATION AND PARTITIONING

List of Potential Ideas

- Method of modifying setting for both techniques
  - Degree of Parallelism (DOP)
  - Number of independent processes available to PX coordinator
  - Cardinality and optimizer statistics
- Method of visually evaluating the performance difference
  - Important to poster/expo
  - Discuss how the raw data will be made flexible for multiple visualization techniques
  - Discuss what visualizations will be used (or most reasonable for our purpose)
  - Possibly talk about how we will know when we have met the ideal treshold for system performance or data return
- Partition Designs
  - Evaluate the ideal case for the use of each partition design in order to develop an intuition for selecting the ideal choice
  - Break down some of the reporting analytic queries in order to see how the data is being stored (ie. do they need the entire row, or just some values?)

  - Develop use case for each partition design in terms of the reporting queries that HP runs
- Using toolkit to adjust settings, enable parallel queries with a ”hint”
  - Talk about what a hint is
  - Figure out how it is possible to configure px settings through SQL statements

4 Reporting Tools

List of Potential Ideas

- Collection of Raw Data
  - List the main v$ views that we will query for session statistics (ie. v$active_session_history)
  - Identify how we can correctly calculate statistics of interest (DB time, I/O, etc...)
  - How we would focus development of homebrew toolkit to expand on what is available within the other options
- Milestones

5 Conclusion