CS Capstone Problem Statement

October 19, 2018

Impala performance tuning on HDFS

Prepared for

Hewlett Packard

Andy Weiss ____________________ Signature ________ Date

Prepared by

Group 38

Caitlyn Cook ____________________ Signature ________ Date

Iliana Javier ____________________ Signature ________ Date

Nicholas Skinner ____________________ Signature ________ Date

Amy Tang ____________________ Signature ________ Date

Abstract

The Big Data team at Hewlett Packard (henceforth HP) currently uses an Oracle database to store industrial IOT data gathered from large printing presses. While past projects have improved the performance of this database, there are still opportunities for better performance. To address this, the Big Data team is considering expanding their database by implementing a distributed file system. Distributed computing is an option to improve the storage capabilities and performance of their database, although it is a significantly different model than the centralized system currently in use. This project aims to research the behavior of SQL queries on Apache Hadoop, a distributed computing provider, using the Impala SQL engine. The goal is to compare this new option to the performance of their current Oracle database. At the end of the project, the sponsor will have an improved understanding of distributed computing and related performance capabilities, and will be able to make an informed decision about whether Hadoop and a distributed file system are a worthwhile investment for their purposes.

Contents

1 Description of Problem

The current database used by the team at HP is a 30 TB Oracle database. Although previous efforts have improved its performance, queries on such a large database are still not fast, which is common for large quantities of data. The Big Data team is looking to expand their database further, and would like to investigate distributed computing as a potential solution allowing for larger quantities of data while maintaining performance. Most of the data stored in this database is static, and as a result, an ACID compliant database may not be necessary and causes additional overhead.

As the sponsor is interested in moving workload, they would like to have a more thorough understanding of the differences between data platforms to try to match the query engine they choose with their data needs. For this research project, the sponsor has specifically chosen the Impala SQL engine, which is a Massively Parallel Processing SQL query engine written in Java and C++. Impala also incorporates technologies from Hadoop to ease the burden of the volume of data that is sent from machine to machine through the use of a distributed file system.

The sponsor is interested in receiving thorough documentation of the configuration and setup used to test benchmarks of system performance. The documentation created from this research is intended to look similar to an instruction set, as well as having an executive summary on the research conclusions of the effectiveness of the Hadoop system. If time allows, the sponsor is interested in further researching stretch goals that consist of comparing other distributed database systems, comparable file formats, or even Hadoop suppliers.

2 Proposed Solution

Distributed computing has become popular as a way to address the problem of scale, by removing the burden from a single computer and spread it among several coordinated computers. Traditionally, distributed computing suffers from a problem with consistency, which prevents it from being fully ACID compliant. However, the data used by our sponsor is primarily static, which gives it a slightly different set of needs and circumstances. The goal of ACID compliance is to insure the integrity of the database state across a lifetime of transactions; however, since our sponsors data is primarily static and will not be experiencing many transactions, ACID compliance is less of a concern. This is the motivation behind considering a distributed computation system for this database, as distributed systems avoid some of the overhead associated with ensuring ACID compliance.

The aim of this project is to thoroughly investigate Apache Hadoop as a potential option to increase the capabilities of the Big Data team at HP. There are three primary goals for this project; each will be discussed in more detail below. The first goal is a generally improved understanding of the function of distributed query engines. The second is an understanding of available tools and methods to monitor or improve query performance. The third is an understanding of metrics that represent useful comparisons between related technologies.

The first goal is focused on a general understanding of how a distributed system works. We will examine the functions and behaviors of such a system, particularly as it relates to the business needs and goals of the HP Big Data team. We will compare its behavior to the current database used by our sponsor, and investigate the differences between centralized and distributed computing, as well as differences between other solutions for distributed computing.

The second goal is an understanding of the available performance-related mechanisms. We will gain an understanding of the tools or methods available to diagnose and address performance concerns, so that performance can be further improved should the system be adopted. We will identify methods to investigate query performance so that intensive queries can be targeted for optimization, improving the overall efficiency of the system.

The third goal is an understanding of the metrics that can be used to compare related technologies. We will investigate multiple related technologies, and determine specific points of comparison that can be used to support a decision among the available options. While our investigation will be focused on Apache Hadoop through the Cloudera vendor, which the sponsor is currently considering, we will also compare this specific distribution to other similar alternatives, such as Hortonworks and MapR.

3 Performance Metrics

As this is a research-based project, it is difficult to define specific performance metrics from the beginning. Though we cannot aim for a specific cost reduction or speed increase measurement, we do have a concrete goal to aim for. The sponsor has requested a document containing an introduction to and writeup on the Impala SQL engine and Apache Hadoop system, which is to include an executive summary of the technologies involved as well as the available tools to measure system and query performance. This written document will provide the knowledge requested in each of the three goals discussed above, as well as answering any questions raised in the course of our research or by our sponsor. This document is intended to be high-level, to communicate our findings to interested persons within HP.

The sponsor for the project has also requested that the research be open to flexibility in direction, interested in being guided by curiosity from discoveries about the new technologies and solutions given by the file system. This will cause the required contents of the report to shift over the course of the project as the sponsor guides research towards directions relevant to the needs of their business.