Oregon State University
College of Engineering

CS Capstone Research Proposal

April 19, 2019

Impala performance tuning on HDFS

Prepared for

HP Inc.

Andy Weiss _____ _____
Signature *Date*

Prepared by

Group 38

Team Name

Caitlyn Cook _____ _____
Signature *Date*

Iliana Javier _____ _____
Signature *Date*

Nicholas Skinner _____ _____
Signature *Date*

Amy Tang _____ _____
Signature *Date*

Abstract

The goal of this project is to improve the cost effectiveness and scalability of the HP database system by archiving static data into a distributed system. Data from industrial presses are currently stored in a centralized database; however, as more presses are sold and more data is desired about their operation, the rate of data flow has demonstrated exponential growth. To preserve or improve the level of performance in the face of this growth, HP plans to implement a distributed system which can significantly reduce the load on the centralized server. The distributed system will be used to archive static data that does not require a high level of resilience to errors resulting from transactions. The team's end product should be a white paper detailing the impala system, and general guidance on aspects of the system.

Our task is to research the capabilities of Hadoop and Apache Impala to improve the efficiency of the database environment for HP. We will (1) investigate the compatibility of these systems with the existing querying and processing tools in use, and (2) compare the performance and optimization capabilities of the fully centralized system to a hybrid system. (https://tobi.oetiker.ch/lshort/lshort.pdf)

Contents

1 Update Table

Section	Original	New
Abstract		The team's end product should be a white paper detailing the impala system, and general guidance on aspects of the system.
Proposed Methodology	To determine the best software and methods, we will execute queries for experiment on different implementation. By this way, we can compare performance and ability for the usage aspect on clients data.	Removed, no longer in scope of project.
Proposed Methodology	To switch the Oracle centralized system to a hybrid distributed system without changing the original query system, we will obtained knowledge from research that can help us measure and evaluate the capabilities and effectiveness of different distributed system.	Removed, no longer in scope of project.

2 Project Description

HP has spent many years fine tuning and garnering a deep understanding of their Oracle database and the SQL commands specially crafted to query it. As more printing presses are sold, the Big Data team says its current database size of 30 terabytes is growing exponentially. This level of growth means even the most optimized of database queries can result in long running times due simply to the large amount of data tables that need to be scanned, reduced, or joined by the single monolithic server. Yet gathering and processing large amounts of data is essential if HP wishes to continue growing. So the central problem remains: How do we handle big data without the system to process, query, and store it becoming a big cost?

Cost can be thought of in terms of money and time. On the monetary side, it takes sophisticated and expensive hardware to process and store data. Whenever one hard drive runs out of memory space to house its data, new hard drives need to be purchased before any new data can be stored. With data size growing exponentially, the need to buy new hardware becomes more and more frequent, meaning it is not a valid long-term solution. This does not take into account the cost of reserving a physical location for these servers to live or the cost of powering them.

In addition to cost of hardware, the software that runs on the hardware incurs a monetary cost. Oracle databases are a managed system, which require purchasing a license to be able to use. These licensing costs are often recurring, with companies needing to pay by time period, amount of servers or cores in use, and/or by job run. Thus, as the data amount grows, so does the licensing cost. Finally, databases garner a time-cost in the amount of time they take to query

and maintain their dataset. As the size of the dataset increases, so does the amount of time required to run queries. This extra time creates an opportunity cost caused by running on subpar systems, and wastes the time of users.

With all of the aforementioned sources of cost taken into account, it becomes apparent that HPs current database system, housed on a single server, is cost prohibitive and unsustainable with their current rate of growth. This necessitates a new system, one with more processing power, storage space, and that will scale with their data size. The implementation of an entirely new system needs to be proven to save HP both time and money before being put into action. It must save the company money in the short and long term, through a system that requires less hardware and that moves from a provisionally licensed software to an open source one. It must also save the company time, particularly by ensuring the new system runs on SQL and that its SQL engine is as close in implementation to the current system as possible. Should the Big Data team decide to move to a different database with a non-SQL based engine, years worth of work that went into understanding every angle of how SQL logic interact with their data will be made obsolete. As this change would negate literal years of work, providing a work around and preserving the current SQL logic is a top priority.

In the search for such a new system, the Big Data team has brought our team on board to investigate Apache Impala, a distributed database SQL engine.

3 Research Intent

HP believes that moving their 30 Terabyte Oracle 12c (12c being Oracles 12th iteration featuring cloud technologies) database from a purely centralized data platform to a hybrid of a centralized and distributed system will result in a net performance increase. The distributed system that HP is interested in transitioning into is an Apache Impala distributed system. The sponsor believes that this transition will increase the overall performance of their current system by reducing the amount of data that is scanned, but remains curious on what trade offs will come with different technologies. The sponsor also believes that due to both of the platforms being based on the SQL query system, HP will not need to spend capital translating existing data processing assets into another data centric language.

To determine if this is a correct choice of software, research might give HP a fundamental idea of what happens within the Impala system when a query is submitted. More precisely, the research will explore query execution such as: (1) how the query is parsed within the Impala system, (2) how an execution plan is generated by the system, (3) how that plan will be interpreted by the system, and (4) how execution plans are distributed across nodes. Understanding these processes will assist HP in determining if the system will be right. The research report will also contain (5) an explanation of how operations are conducted within the system. The research will also (6) define how the system will perform a more complex task, such as joining tables when executing a query, or how the system uses analytic functions. If the system supports analytic functions, the sponsor is interested in (7) understanding how an analytic function interfaces with the node series brought forth by the distributed system.

Given that the distributed system provides better data read access performance, HP is interested in implementing a hybrid system. This system will consist of two parts: Oracle 12c to host actively updated data and files on a centralized system, and Apache Impala to host static data that has not been updated in over two years in a distributed system. The new system will be required to run existing Oracle 12c query scripts with minimal to no modification of the SQL syntax.

4 PROPOSED METHODOLOGY

This system were studying transitions data older than two years to a distributed system. To determine the performance of the proposed system, we will examine query plans for selected queries and compare the decisions made by each system. We will time the execution of selected queries and compare their performance on each system, using the examined plans as a guide to understand the observed differences.

We will have comprehensive understanding of these entirely different systems, and related technologies. To approach that, we are going to start with reading existing documentation and research of these main topics: Oracle Optimizer, Impala Hadoop, SQL engine execution in different system and Relational Theory.

We found that previous Capstone projects which worked with Big Data used similar methods. One of the projects built a toolkit to measure system performance metrics from the processing of a SQL query. They have a lot of writing in the first phase that help identify the requirements needed to the project. And then these documentation helped them in further works. The other project was doing research on system enabled compression options that work with Oracle Databases and compare the storage and query performance on each of these options. They started with evaluated research that would be used to accurately predict and measure ways that data is inserted and stored in the Oracle database. Once they got comprehensive knowledge of the Oracles table compression, they start designing some experiments and made the following research organized.

In the lifecycle of a SQL query, before it gets to the executing step, it compares the cost time of different query plans with their own optimization, and then picks the optimal plan. Understanding the query plan is a standard place to begin with performance optimization.