Impala Performance Tuning on Apache Hadoop: Problem Statement

Iliana Javier

October 11, 2018

ABSTRACT

Our project is to compare the Apache Hadoop distributed file system with Hewlett-Packards current Oracle database system. We will conduct a formal write-up of how each systems respective SQL engine functions, which support vendor will best suit the clients needs, and which file format the data will be stored in to achieve optimal results.

PROBLEM STATEMENT

A. Description of Problem

Hewlett-Packard (HP), like many large companies, stores large amount of data in various databases. As database research continues and new products for database implementation and management are released, large companies will want to stay at the forefront of this technology in order to to get the most out of their data. We want to see if any of the current products on the market, particularly those suggested by our client, will be a significant upgrade to HPs current database system- to make data queries faster, implement scalability, and / or increase productivity.

B. Proposed Solution

In order to answer our clients questions about the current distributed, SQL running file systems available, we will research and test a subset of the known products and solutions. This subset contains the Apache Hadoop distributed file system working in junction with the Apache Impala engine, the Orc and parquet file types, and several support vendors. The Apache line of products generally provides extensive documentation for the use and workings of the solution, so our research will consist of us reading this documentation and compiling the information that is relevant to our client in an easy to read format. We will also be learning by experimenting with the actual solutions themselves using large amounts of dummy data to reflect how the database will actually be used on the job in real time.

C. Performance Metrics

This Project will be completed when we have completed a write up of our findings of the Apache Hadoop as it compares with their current database implementation using Oracle. Special attention will be paid to the performance differences between the two SQL engines. This will achieved by performing many efficiency and timing tests as well as simple documentation research to explain the inner workings and features of the solutions to our client. Specific metrics in terms of quantifiable improvement have yet to be discussed with the client and require further review.