Senior Capstone Project Group 38

Nicholas Skinner

October 11, 2018

ABSTRACT

This project consists of researching the implementation and optimization of a SQL Query Engine for HP to decide if it is a technology that is right for them. This will involve our team working with our sponsor to familiarize ourselves with the concepts of Big data, compliance systems, how to compare technologies, what methodologies are in place to optimize performance, and how a distributed query engine really works. Our progress for the project is reliant on our interaction with our Sponsor, and maintaining contact to be able to theorize and test ideas. The end research from this project should be culminated in a research paper that expands on the concepts mentioned earlier by applying them to an implementation of a distributed query engine.

PROBLEM STATEMENT

A. *Description of Problem*

Big data is a concept that consists of large datasets that may or may not be analyzed for emerging patterns or trends. In a more direct definition, Big data consists of the three Vs: Volume, Velocity, and Variety. To be useful, big data needs to be all three of these Vs to varying degrees. With these Vs, however, come performance issues due to existing structures, only recently are technologies like Hadoop coming forward that help alleviate the stress of these requirements on the engine.

The sponsor is interested in moving workload of big data from an Oracle based relational database to a distributed system, and they would like to have a more thorough understanding of the differences between 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.

Our sponsor is also interested in the performance benchmarks of the Impala SQL engine, this would consist of the evaluation of categories set by the project sponsor. These benchmarks are intended to allow for a direct comparison of this SQL engine to other similar engines.

As an available expansion to the existing project, the sponsor is also considering several ways to further optimize their database environment, this includes, but is not limited to the supplier distribution of Hadoop, the storage file format, as well as the different technology suppliers.

B. *Proposed Solution*

My proposed approach for this project consists of a four-phase Agile plan: The first phase of this project should consist of research about the requirements and restrictions of the research, what are our sponsors big data needs, and what are they looking to understand from our paper? Our team should collaborate with our sponsor to get answers to these questions to gain a better idea of the scope of the project. This phase can also be revisited if project scale needs to be adjusted, or time constraints become an issue.

The second phase should consist of research about the given concepts: What is a relational database, what is a distributed system? How are these technologies measured? How are they implemented? All questions that should be addressed and answered before researching the query engine should take place in this phase.

The Third Phase should consist of research and testing the SQL engine, this is where we compare the technologies, the implementations, and the benchmarks of the system and document our research and conclusions. This step will be done in tandem with our sponsor to get a realistic idea of how big data, in this case 30 TB, will impact the performance of the engine. The sponsor has also indicated interest in guiding us through some of these sections giving us test ideas for queries that share likeness to queries that they would be actively performing in their production environment.

The fourth and final phase in our cycle will be writing and formatting the research paper. This will consist of all the benchmarking, research, and information about the technologies. As this is an agile structure, we can cycle back to other phases in case some part of our research needs more documentation, more observations, or simply more research. Our sponsor will assist in overseeing and giving descriptions of what they will want on the paper, how thorough our analysis should be, as well as help us determine what information may or may not be relevant to the audience that we are writing for.

*C. **Performance Metrics***

The sponsor has a few base requirements for the project that can be checked off as measurements of progress, questions answered, meetings had, documentation created, etc., but we will also be working in tandem with the sponsor on queries and potential test cases which will help guide our progress through benchmark testing.

In an email conversation with the sponsor they also had this to say about states that the project could lead to, and what is acceptable for them:

Success will be measured by how far we are able to expand our knowledge about how distributed computing works. The goal would be to prove or disprove if a distributed computing environment would be a worthwhile investment for our use cases.

The Project will reach completion when a white paper that possesses the following is delivered: An Executive summary of the technology stacks that are used in the Impala query engine, a list of tools and techniques that are used to measure and optimize query performance, and thorough documentation of the configuration and setup that had been used to benchmark the system performance, and the results of said benchmarks.