Impala performance tuning on HDFS

Chia-Yu Tang

Abstract

In this project, we devote to extending and improving the current database by implementing the distributed technology for HP Inc. Currently, our sponsor have large amount of data that were various measurements collected from sensors and components in the printing machine, and were used to fix errors and make improvement. With the accumulated volume of data increases, the effective management and storage become more important for data usage needs. Apache Hadoop is the distributed filed system we are focusing on. Moreover, we will look at how SQL on Apache Hadoop works, and explore does the distributed technology can be a worthwhile investment compares to the original technology.

Problem Statement

A. *Description of Problem*

• The sponsor manages a 30TB Oracle Database and most of the data is static. Static data structure stored a fixed number of data items, but there are lack of efficiency and flexibility compared to dynamic data. Consider the usage of massive data sets, distribution systems provides an effective way to collect, distribute, store and manage data in real time. Therefore, our sponsor would like to move their relational database workload to a distributed system, and get rid of the traditional ACID (Atomicity, Consistency, Isolation, Durability) compliant relational database.

Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has ability to deal with any kind of data, or big data, and enormous processing power by providing massive storage. Our sponsor would also like to explore whether the query engine on Hadoop database can have better performance than on the Oracle database. The query engine our sponsor has chose is Impala SQL engine, which an open source massively parallel processing query engine for Hadoop and is shipped by Cloudera. Except for Cloudera, we can take other suppliers as reference as expanding the project. Impala provides fast, interactive SQL queries directly on Hadoop data stored in HDFS. In addition, Impala uses the same metadata and SQL syntax from Apache Hive. The HDFS file formats that Impala support are various, such as Parquet, ORC, TEXT. Among them, our sponsor are looking at ORC or Parquet as the file format in this project. To test the performance effectively, we will figure out which tool can provide the best diagnose and present the result that we want.

To bring the further research and discussion of the database optimization, our sponsor would like to have well documentation of the configuration and setup used to benchmark system performance. If time allowed, we can broaden our research by comparing other different distributed system, file formats or suppliers based on our preliminary result.

B. *Proposed Solution*

• Fist of all, we should fully understand of the database background that out sponsor is using. This includes the architecture of the Oracle database, how does it work, and how does it cooperate with machines. For the usage aspect, it is important to know our sponsors requirement and needs, what is the purpose of the data and how are they being used. Next, we can focus on the research on tools that relate in this project, especially, when it comes to big data. For example, how does Hadoop distribute system work, what are influences that it may bring to the current stack, what is Impala SQL, and how does Impala SQL work for Hadoop. Our sponsor also suggests us to think about tools or ways for testing and diagnosing the performance. After that, we will have better image of this project and come up with ways map out the following plan that can prove and support our idea of implementation.

In the aspect of testing Impala SQL engine, we will compare the differences in systems, technologies, implementation, and combination, and documented each of the testing process for the later analysis of query performance. Take consider to big data is important when we do testing. We should always focus on the impact and result on a 30TB huge database.

Finally, we work on the research paper, which includes our knowledge and understanding of the project, and the record from the beginning discussion to the testing. Our sponsor will give us opinion and assist us modifying the paper so that we can present our project perfectly.

C. Performance Metrics

- The sponsor has provides clear description and requirement in the project description site and the email, as well as problems that the sponsor faced and would like to solve. This project is focusing on improving the query performance. Our paper will eventually deliver: how the Impala query engine works, including executive summary of the technology stack(s) used, tools and techniques used to measure and optimize query performance and thorough documentation of the configuration and setup used to benchmark system performance along with the results. 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 sponsor has also mentioned that the research-based project will work differently than development project. We will schedule a meeting with our sponsor every week to make sure that we are in progress rather than spend most of time designing and experimenting the product. Our primary contact information is email. In addition to check our project, the sponsor will lead and assist us until the project have done.