Evaluation of Column-Family NoSQL Databases on Hadoop

A dissertation submitted to The University of Manchester for the degree of Master of Science (MSc.) in the Faculty of Engineering and Physical Sciences

2014

Vivek Vijayan
School of Computer Science
Abstract

The need for insights from the massive amount of data generated in today’s world has seen an increase in usage of distributed computing platforms such as Hadoop. Distributed computing enables greater compute power and memory via parallel processing paradigms, but in turn introduces a number of added technical issues. Nonetheless distributed platforms such as Hadoop are able to process large volumes of data in batch in a cost-efficient framework. Hadoop offers both large-scale potential parallelism and economic computing solution using commodity hardware.

With the advent of Hadoop and similar distributed computing platforms, analytical data processing has undergone a huge leap. The robustness of Relational DataBase Management Systems (RDBMS) is well known and for many years have been the standard database model. In contrast, some applications do not require compliance to the ACID (Atomic, ... etc) properties of the relational model, hence there has been an increasing direction of travel towards simpler database designs which are mostly non-relational in nature – hence the Not-only SQL (NoSQL) revolution. A group of databases which are formed under the NoSQL umbrella share common features such as greater speed and scalability. These databases overcome the limitations of adhering to a tightly bound schema and the necessity to join data across multiple tables.

The aim of this work is to evaluate and compare the performance of NoSQL databases based on query performance with Online Transaction Processing (OLTP) and Online Analytical Processing (OLAP) based queries. Also performance testing is carried out using Yahoo Cloud Serving Benchmark (YCSB) tool.

The evaluation is carried out on Hive and two column-family databases, HBase and Cassandra. The column-family databases are designed for applications where the queries do not generally require complete tuple information. Hence this is one of the focuses of the query testing of a column-family database.

The results of this work conclude that Cassandra is comparatively efficient than HBase for the test queries and also provide an easy interface which is SQL-like. HBase performs well, though not up to the mark of Cassandra, but both the databases share similar properties. Hive is not efficient to run queries but it provides additional query support that HBase and Cassandra lack, such as the aggregation functionalities.