MapReduce and Parallel Databases application for data analysis problems

Dmitry Spikhalskiy
E-mail: dmitry@spikhalskiy.com
Lomonosov Moscow State University, System Programming Department
Advisor: Professor of Computer Science at Lomonosov Moscow State University, Principal research officer at Institute for System Programming, Russian Academy of Sciences, Sergey Kuznetsov, Ph.D.

Modern practices in business management are based on decision making through historical data, so that companies interested in systems that are able to manipulate data, process and analyze them at different detalisation levels. At the same time, exponentially increases the amount of data which you want to store and process through analytical database systems. Main reason of such a situation is growth in the level of automation of production data, increasing the number of sensors or other devices that generate data, the transition to the use of WEB-technologies in interactions with customers [1].

To solve these problems, many companies offer their parallel DBMS with sharing-nothing architecture - a cluster of independent machines with their own local disks and main memory, connected by high-speed network. There is a hypothesis that this architecture scales well, especially if you take into account the cost of hardware. Workload of data analysis usually contains many large scan operations, multidimensional aggregations and joins with a star schema, which is relatively simple parallelized across nodes in the network without shared resources [2]. On the other hand, some experts argue that in order to perform analysis on hundreds or more nodes best of all are systems, based on the MapReduce paradigm, as they originally were developed based on the scaling up of thousands of nodes in the sharing-nothing architecture [3]. Thus, comparison and analysis of the applicability of the existing approaches to storing and processing large amounts of data is an urgent task facing the developers of both storage and to companies that specialize in data analysis.

This paper considers the architecture of different analytical data storages - parallel DBMSs and MapReduce. For both paradigms the complexity of development and support of implemented systems are estimated. As an experimental research of the paper influence of the characteristics and architectural features of the storage on the performance of analytical queries in situations of vertical and horizontal scaling are considered. To solve this problem there have been developed set of analytical functions, and test data that extend the research of "A Comparison of Approaches to Large-Scale Data Analysis" [4], held launches on clusters of Hadoop&HDFS, Vertica, Greenplum, and classical row database systems. As a test configuration we used a cluster of 100 workstations, as well as fully connected network of 3 high-end servers. Practical importance of the paper is reasonable explanation of differences in performance of storages at different volumes of data, tasks, and the cluster size, and described the differences in the areas of applicability of the researched systems.

Bibliography

1. Abouzeid A., Bajda-Pawlikowski K., Abadi D., Silberschatz A., Rasin A. *HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads.* Proceedings of the 35th VLDB Conference, August 24-28, 2009, Lyon, France.

2. Lammel R. *Google's MapReduce programming model — Revisited.* Science of Computer Programming, 2007, Volume 68, Issue 3, pp. 108-137.

3. Dean J., Ghemawat S. *MapReduce: Simplified Data Processing on Large Clusters.* Communications of the ACM – 50th anniversary issue: 1958 – 2008, 2008, Volume 51, Issue 1, pp. 107-113.

4. Pavlo A., Paulson E., Rasin A., Daniel J., Abadi D., D.J. DeWitt, Madden S., Stonebraker M. *A Comparison of Approaches to Large-Scale Data Analysis.* Proceedings of the 35th SIGMOD International Conference on Management of Data, 2009, pp. 165-178.