CS543 - Paper Review Report # V

Hailu Belay Kahsay - 20155624

Title: **HadoopDB: An Architectural Hybrid of MapReduce and DBMS Technologies for Analytical Workloads**

Author: Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel Abadi, Avi Silberschatz, Alexander Rasin

In the paper it is stated that the amount of data that needs to be stored and processed by analytical database system is exploding due to: [1] the increased automation(more business processes are becoming digitized) [2] the proliferation of sensors and data-producing devices [3] Web-scale interactions with customers and [4] Government compliance demands along with strategic corporate initiatives requiring more historical data to be kept online for analysis. And because of this problem some of the analytical database start-ups deploy their DBMS on a shared-nothing architecture which is a collection of independent (possibly virtual) machines, each with local disk and local main memory, connected together on a high-speed network; these databases are referred as parallel databases in the paper. Teradata, oracle and Microsoft(for its exadata and Madison projects) have used this architecture.

Even though performance and efficiency of parallel databases makes them well suited to perform big data analysis and they can scale well into tens of nodes , scalability becomes an issue when the number of nodes increases(hundreds of nodes) due to the fact that: failures become increasingly common as one adds more nodes to a system, nearly impossible to achieve pure homogeneity at scale. Third, and parallel databases have not been tested at larger scale applications. In order to alleviate this scalability issue, the authors proposed HadoopDB, by combining the scalability advantages of MapReduce with the performance and efficiency advantages of parallel databases to achieve a hybrid system. In the paper the performance, fault tolerance, Ability to run in a heterogeneous environment, flexible query interface have been chosen as the desired design properties.

The paper also talks about the Hadoop architecture that is composed of Data center connectors, Catalog, Data Loader and SQL to MapReduce to SQL (SMS) Planner. The performance evaluation showed that although Vertica's percentage slowdown was larger than Hadoop and HadoopDB, its total query time (even with the failure or the slow node) was still lower than Hadoop or HadoopDB. In addition to this, Vertica's performance in the absence of failures is an order of magnitude faster than Hadoop and HadoopDB (mostly because its column-oriented layout of data is a big win for the small aggregation query).

Finally the paper concludes with the idea that HadoopDB, a hybrid of the parallel DBMS and Hadoop approaches to data analysis, achieving the performance and efficiency of parallel databases, yet still yielding the scalability, fault tolerance, and flexibility of MapReduce-based systems.