Relative Evaluation of NoSQL Databases For Manipulating Genotype Data

Vinícius Junqueira Schettino[1], Arthur Lorenzi Almeida[1], Fernanda Nascimento Almeida[1,2], Wagner Arbex[1,2]

[1]*Federal University of Juiz de Fora (UFJF),* [2]*Brazilian Agricultural Research Corporation (Embrapa)*

One of the greatest challenges on bioinformatics research is to manipulate the data. Genotype files, vastly used in this field, are known by their high dimensionality and unbalancing. These aspects are some of the reasons RDBMSs, traditionally signed for tabular information persistence, have not been shown as good infrastructure to analysis that rely on this kind of data. Therefore, this abstract aims to evaluate the relative performance among NoSQL engines on genotype data manipulation. In this text we present the extension of previous studies, encompassing the viewpoint of scalability, as well as including results from three representatives of distinct NoSQL databases families. For our evaluation, we used the Yahoo! Cloud Server Benchmark, a framework designed for asserting NoSQL databases aspects. Three databases were considered, each of them representing one family of NoSQL engines: Tarantool as "Key/Value Based", MongoDB as "Document Based" and OrientDB as "Graph Based". We simulated two populations with 5,000 individuals, with a hypothetical SNP sequence for each individual. One population with 20,000 SNP markers, the other with 56,000. Two scenarios were considered: One with 5,000 insert operations, and another with 10,000 equally divided read and update operations. To assert the scalability, we measured the throughput of each workload for these engines. On the insert scenario, using 20,000 SNP markers, MongoDB was capable of handling 56.8 ops/s on average, followed by Tarantool and OrientDB with 41.8 and 33.7 ops/s, respectively. For the 56,000 SNP markers population, MongoDB, Tarantool and OrientDB handled, on average, 21.7, 14.0 and 12.0 ops/s each. For read and update operations with the 20,000 SNP markers population, Tarantool executed on average 515.1 ops/s, followed by MongoDB with 170 ops/s and OrientDB with 26.0 ops/s. With 56,000 SNP markers, the results were: 307.5, 50.0 and 9.9 ops/s on average for Tarantool, MongoDB and OrientDB, respectively. Comparing these results, Tarantool was capable of keeping an average 33.6% of its performance on insert operations and 59.7% on reading/updating when we increased the SNP markers sequences from 20,000 to 56,000. MongoDB kept on average 38.2% on insert operations and 23.4% on read/update operations, while OrientDB preserved 35.5% of its performance on insert operations and 38.3% on read/update operations. For insert operations, the three engines scaled similarly, though MongoDB did the best absolute throughput for this kind of operation. Tarantool scaled better for read and update operations and presented the best absolute throughput.

Supported by CAPES, CNPq, Embrapa, FAPEMIG and UFJF.