Comparação entre sistemas de banco de dados NoSQL com modelos chave-valor e orientado a documentos em operações sobre arquivos de genótipos[1]

Arthur Lorenzi Almeida[2,3], Vinícius Junqueira Schettino[2], Leojayme Rodrigues Manso Silva[2], Wagner Arbex[4, 5]

[1]O presente trabalho foi realizado com o apoio do CNPq, Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brasil: Parte do projeto Detecção de Assinaturas de Seleção em Bovinos a Partir de Dados Gerados por NGS, liderado por Marcos Vinícius Barbosa da Silva;
[2]Graduando em Sistemas de Informação - UFJF.
[3]Bolsista do CNPq. e-mail: lorenzi@ice.ufjf.b
[4]Analista, Embrapa. e-mail: wagner.arbex@embrapa.br
[5]Orientador

Resumo: O tratamento adequado de arquivos de genótipos por bancos de dados é de grande importância para bionformática e genômica e, notadamente, os recursos providos pelos tradicionais sistemas gerenciadores de bancos de dados relacionais (SGBDRs) não cumprem esta tarefa. Arquivos de genótipos são bases de dados não clássicas, de domínio restrito, alta dimensionalidade e, em geral, ocupam muito espaço etc., fazendo com que SGBDRs não sejam boas soluções. O objetivo deste trabalho foi avaliar o desempenho de dois diferentes sistemas NoSQL, ou "not-only SQL", no caso MongoDB e Tarantool, que representam, respectivamente, as "famílias" de sistemas com modelos de dados orientados a documentos e chave-valor. Foi utilizado o *framework* Yahoo! *Cloud Serving Benchmark* (YCSB) com uma *workload* de 5 mil registros e 56 mil campos de 1 byte, representando o tamanho da população e o número de marcadores por indivíduo, respectivamente. O experimento foi dividido em dois cenários, C1, com as operações de inserção e, C2, com as operações de leitura e atualização. Para C1, foram executadas 5 mil operações de inserção, para os 5 mil registros da *workload*, quando MongoDB foi um pouco superior em relação ao Tarantool. A taxa de operações/s (*throughput*), mostrou que MongoDB foi cerca de 11% mais eficiente. Todavia, em C2, foram executadas 10 mil operações, divididas uniformemente entre leitura e atualização, e o resultado se inverteu, visto que, MongoDB não chegou a 2% do *throughput* de Tarantool. A análise dos resultados permitiu concluir que para armazenamento, leitura e atualização de arquivos de genótipos, o uso de um banco de dados NoSQL, tal como o Tarantool, deve ser considerado, se comparado a sistemas com modelo baseados em documentos, como é o caso do MongoDB. Trabalhos futuros devem partir (i) da hipótese de que sistemas NoSQL com modelos baseados em chave-valor apresentam desempenho superior em relação a outros sistemas NoSQL; e (ii) de cenários que avaliem a escalabilidade.

Palavras–chave: banco de dados, bioinformática, ciência de dados, genótipo, NoSQL