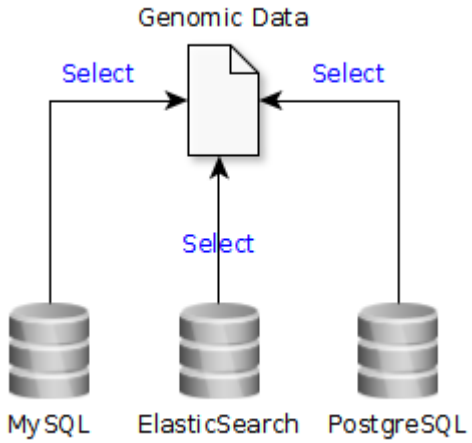
MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

Optimizing queries via search server ElasticSearch: a study applied to large volumes of genomic data

Vinicius Seus[a] (viniciusseus@gmail.com), Alex Camargo[a] (alexcamargoweb@gmail.com), Diego Mengarda[b] (diegormengarda@gmail.com)
[a] *FURG*
[b] *UNIPAMPA*

Graphical Abstract



Abstract

This work aims to use the ElasticSearch server to optimize searches on genomic data made publicly available by the UCI Machine Learning Repository. As a case study, the results obtained were compared with the MySQL and PostgreSQL relational databases. With the proposal presented, a gain of more than 90% was achieved through the use of ElasticSearch technology.

Introduction

ElasticSearch[1] is an open source search server started by Shay Banon project published in 2010. Its main concepts of use include: index, document, document type, nodes, cluster, shard, and replic [Kuc and Rogozinski 2013]. In this technology the records do not use the usual normalization of tables because the tool structure is designed to have superior search performance. Databases like NoSQL and MongoDB also operate in a very similar way.

When it is necessary to analyze large volumes of data, Bioinformatics acts as a multidisciplinary field that integrates knowledge from different areas. Its applicability goes from the analysis of biological data to the construction of tools and methodologies that allow the use of the computer for tasks usually laboratory. An important fact in this issue was the advent of the Human Genome[2] Project (HGP) and the subsequent availability of the data obtained for the entire scientific community [Pennisi 2001]. With this, the search for results in viable processing time has become a great challenge among bioinformatics, especially with regard to genomic data[Alencar 2010].

1 https://www.elastic.co/products/elasticsearch

2 Genome is the name given to the DNA set of all the chromosomes of an ovum or sperm, being constituted of 3.4 billion bases

Materials and Methods

The hardware composes a structure of a computer by: 1 processor with 8 cores 2.2 GHz, RAM of 6 GB and hard disk of 100 GB/SSD. The queries were performed on the same database, replicated in each of the following technologies: MySQL, PostgreSQL and ElasticSearch. In order to organize the applied methodology, the term "query" refers to a "select" in MySQL/PostgreSQL and a "search" request in ElasticSearch. Table 1 shows the results of the experiments for the "splice.data" file for the Molecular Biology (Splice-junction Gene Sequences) Data Set, available from the UCI Machine Learning Repository <https://archive.ics.uci.edu/ml/datasets>. The final, preprocessed file resulted in 1.4 million records, inserted through scripts in each database. The query was performed in the "NameGene" column, a field of type varchar (50).

Table 1. Results for the different technologies

Technology	Index method	Search time (s)
MySQL	*Full-Text Index*	0,697
PostgreSQL	*Full-Text Index*	1,125
ElasticSearch	*Default*	0,058

Table 1 shows that the technology with the highest performance and speed is ElasticSearch. This is mainly due to its more accurate caching system for repeated searches. For example, when performing a query that has already been done previously, the tool already maps the records, organized in the form of documents, which guarantees speed for searching for values in databases.

Conclusions

According to the experiments, it is clear that ElasticSearch achieves very good results in the response time of a query. A performance gain of 91.7% and 94.9% was observed using ElasticSearch technology compared to the MySQL and PostgreSQL relational databases, respectively. As a future work, we intend to investigate how to adapt the searches to the ElasticSearch database with the use of LIKE, commonly used in queries with autocomplete, but not contemplated until the last version tested.

References

ALENCAR, Sérgio. Utilização de ferramentas computacionais para o estudo do impacto funcional e estrutural de nsSNPs em genes codificadores de proteínas. 2010. Doctoral thesis. Universidade Federal de Minas Gerais.

KUC, Rafal; ROGOZINSKI, Marek. ElasticSearch server. Packt Publishing Ltd, 2013.

PENNISI, Elizabeth. The human genome. Science, v. 291, n. 5507, p. 1177, 2001.

WELCH, Terry A. A technique for high-performance data compression. Computer, v. 6, n. 17, p. 8-19, 1984.