NoSQL technology in order to support Semantic Health Search Engine

Romain LELONG[a,b], Lina F. SOUALMIA[a,b,c], Saoussen SAKJI[d], Badisse DAHAMNA[a], Stéfan DARMONI[a,b,c]
[a] *Department of Biomedical Informatics, Rouen University Hospital, France*
[b] *TIBS - LITIS EA 4108, University of Rouen, France*
[c] *French National Institute for Health, INSERM, LIMICS UMR-U1142, Paris, France*
[d] University of Jendouba, Tunisia

Introduction : Not Only SQL (NoSQL) databases have emerged to meet the needs of managing the increasing volume of available (e.g. Electronic Health Records fields [2]) but remain immature and lack of standards [3] which make them difficult to use in Information Retrieval (IR) context. LiSSa provides a French access to 1,191,977 health bibliographic references [1]. This study, presents the implementation of the LiSSa's back-end underlying semantic search engine (SSENoSQL) using Infinispan Technologie.

Method : Infinispan is one of the best open-source In Memory Data Grid (IMDG) as regards to data access performance [4]. The data are modeled as RAM-stored hash tables structures (viz. Maps). The data model of the SSENoSQL (see table 1) has been designed in a generic way by following the Resource Description Framework (RDF) graph model [5].

Table 1. *Maps-based Data model. "⟨X⟩" designate a list of elements of type X. All the data is modeled as entities associated to attributes or to other entities through a relationship triplet* $(id_s, lable, id_t)$.

Map name	*key*	*value*
$Map_T\ (types\ of\ entities)$	id	$type$
$Map_E\ (entities)$	id	$entity$
$Map_A\ (attributes)$	id	$\langle (name, attribute)\rangle$
$Map_R\ (relationships)$	id_s	$\langle (id_s, label, id_t)\rangle$
$Map_J\ (joints)$	$id_s + label + type_t$	$\langle id_t\rangle$
$Map_M\ (model)$	$type$	$model$

Result : A comparison between the execution times of 10 queries on both the SSE_{NoSQL} and the older version of the search engine SSE_{SQL} is performed. We can observe averages run times in the range of 3 seconds for the SSE_{NoSQL} against 12.6 seconds for the SSE_{SQL}.

Table 2. *Average execution times of 10 queries*

query	n	$t_{SQL}\ (ms)$	$t_{NoSQL}\ (ms)$	*gain* (%)	*factor*
$asthma.mc$	4,879	440	139	68.4	3.16
$humans.mc$	448,893	29,538	3,288	88.8	8.98
$asthme/diagnosis.mc$	1,013	916	315	65.6	2.91
$drugs.mt$	141,729	14,137	3,213	77.3	4.40
$(asthma/drug\ therapy.mc\ OU\ (asthma.mc\ ET\ drugs.mt))$	387	4,128	2,237	45.8	1.85
$drugs.mt\ ET\ surgery.mt\ ET\ therapeutics.mt$	11,042	13,962	2,755	80.3	5.07
$drugs.mt\ OU\ surgery.mt\ OU\ therapeutics.mt$	315,619	36,720	4,977	86.5	7.38
$iatrogenic\ disease.sr$	57,555	9,305	3,046	67.3	3.05
$patient.tc$	259,554	9,882	5,310	46.3	1.86
$(surgery.tc\ OU\ procedure.tc)\ ET\ patient.tc$	67,155	7,863	5,349	32.0	1.47

Discussion & Conclusion : Due to the small size of the query set and the difference between hardware systems of each version the search engine, another contextual evaluation will be soon conducted based on a more comprehensive set of queries as well as in Clinical IR context. The use of an IMDG enables a sharp improvement of processing times and also reduce variability of run times with a reduction of the standard deviation from 11,267.9 ms to 1,757.6 ms.

References :

[1] Griffon, N et al. A Search Engine to Access PubMed Monolingual Subsets: Proof of Concept and Evaluation in French. J Med Internet Res, December, Volume 16, Number 12, Pages e271, 2014.
[2] Luo, J et al. Big data application in biomedical research and health care: A literature review. Biomedical informatics insights, 2016, vol. 8, p. 1.
[3] Nayak, A et al. Type of NOSQL databases and its comparison with relational databases. International Journal of Applied Information Systems, 2013, vol. 5, no 4, p. 16-19.
[4] Salhi, H et al. Open Source In-Memory Data Grid Systems: Benchmarking Hazelcast and Infinispan. In : Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering. ACM, 2017. p. 163-164.
[5] WORLD WIDE WEB CONSORTIUM, et al. RDF 1.1 concepts and abstract syntax. 2014.