

6 Conclusion, related and future work

In this dissertation, we proposed a solution to handling the challenge of keyword search over Linked Data. As the Web of Data keeps growing, more and more proposals to approach this task should emerge. Combined with the Cloud Computing high computational power availability by relatively low costs, there is now no reason not to experiment with different combinations of tools, techniques and non-traditional architectures.

We provided a mechanism to store large RDF graphs in a distributed way. Using the tensor matrix representation as input, we developed a mechanism that shards the RDF graph and persist data in separate, distributed datastores. Building on top of this datastores, we provided a mechanism to execute scalable keyword-based search mechanism for RDF graphs. The proposed architecture proved to be a viable way to store and scale large RDF datasets.

As it is often the case with public clouds, we ran into limitations that are inherent with how the public cloud is built. Internal network bandwidth is still unpredictable, and virtual CPUs performance can vary wildly in a matter of minutes. But even with those hurdles, we could develop an implementation of the architecture and achieve meaningful results. The implementation is published as open source [23], and the datasets used in this work are publically available [7] [21].

Related work by the same authors of the tensor-based approach seek to improve keyword search over Linked Data, by means of clustering substructures of the RDF graph and scoring the relevance of those groups [24]. Other proposals to the keyword search are explored with different approaches, such as pre-indexing all documents, running distributed full text searches on demand [25], and ranking similar keyword search terms and presenting the results as word-clouds [26].

The proposed architecture is not the only approach possible. A possible variation is transferring the sharding logic directly to the database - e.g., by using

Cassandra [27]. Also, no caching layers were applied to the whole system. We leave those options open as future work.