Towards Unified Semantics for RDF Stream Query Processing

Daniele Dell'Aglio[1], Jean-Paul Calbimonte[2], Emanuele Della Valle[1], Oscar Corcho[3]

[1] DEIB, Politecnico of Milano, Italy.
daniele.dellaglio@polimi.it,emanuele.dellavalle@polimi.it
[2] Faculty of Computer Science and Communication Systems, EPFL, Switzerland.
jean-paul.calbimonte@epfl.ch
[3] Ontology Engineering Group, Universidad Politécnica de Madrid, Spain.
ocorcho@fi.upm.es

In recent years, several RDF Stream Processing (RSP) systems have emerged, which allow querying RDF streams using extensions of SPARQL that include operators that take into account the streaming nature of these dynamic data sources [2, 4, 8]. These systems are heterogeneous in terms of syntax and capabilities (due to the choice of operators and syntax selected to extend SPARQL). In addition, they implement different evaluation semantics for a set of constructs that may look similar in principle (for example, they may handle time window operators differently). These engines have different assumptions on how the query processing and delivery of results take place, which makes it difficult to describe, compare, understand and evaluate their behavior.

One of the goals of the W3C RSP Group is to propose a common model for representing and querying RDF streams. The emergence of such a model and its accompanying query language is expected to take the most representative, significant and important features of previous efforts, but will also require a careful design and definition of its semantics. Our main interest in this context, is to help laying down the foundations of formal semantics for the standardized RSP query model, such that we consider beforehand the notions of correctness, continuous evaluation, evaluation time, and operational semantics, to name a few.

We have presented and made available to the community a series of works on this area, which can be summarized as follows. First, in [5], we analyze some of the main RSP query systems and show that their operational semantics substantially differ, making it difficult or impossible to compare them or reliably assess correctness in query evaluation.

In [6], we present CSRBench, an extension of SRBench [9] that considers correctness of evaluation results. It uses the concept of an Oracle[1] that computes a set of possible correct answers considering an extension of the SECRET query model from the database community. Using this oracle, we have analyzed different existing RSP systems and identified issues related to the correctness of query results, and hidden assumptions in their operational semantics.

[1] Cf. https://github.com/dellaglio/csrbench-oracle

Based on these findings, we have worked on a model that helps characterizing these engines and explain the query results that they produce in a continuous evaluation setting. Furthermore, we have proposed the RSP-QL semantics [7], a unifying formal model for representing and processing RDF streams, that reflects the different semantics of existing RSP systems. RSP-QL extends the SPARQL model and also takes into account two existing models coming from the streaming data world: CQL [1] and SECRET [3].

We believe that the RSP-QL model, which already explains the heterogeneous semantics of existing RSP systems, can be used as a basis for the new standardized query model that the RSP Group is pursuing. In this respect, this contribution can strengthen the existing work in progress that is currently being carried out by the group.

Finally, we would like to emphasize the importance of data streams in almost every conceivable use case scenario, be it on the industry or in research. The ubiquity of Big data problems that include large and very dynamic flows of data coming form heterogeneous sources, calls for solutions that consider semantic interpretation of data, while supporting rapidly changing information. The Semantic Web community can provide answers to these challenges, but it needs to continue shifting towards paradigms that include reactive, continuous and event-driven processing.

References

1. Arasu, A., Babu, S., Widom, J.: The CQL continuous query language : semantic foundations. The VLDB Journal 15(2), 121–142 (2006)
2. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: C-SPARQL: A continuous query language for RDF data streams. IJSC 4(1), 3–25 (2010)
3. Botan, I., Derakhshan, R., Dindar, N., Haas, L., Miller, R.J., Tatbul, N.: Secret: A model for analysis of the execution semantics of stream processing systems. PVLDB 3(1), 232–243 (2010), http://dl.acm.org/citation.cfm?id=1920874
4. Calbimonte, J.P., Jeung, H., Corcho, O., Aberer, K.: Enabling Query Technologies for the Semantic Sensor Web. IJSWIS 8(1), 43–63 (2012)
5. Dell'Aglio, D., Balduini, M., Valle, E.D.: On the need to include functional testing in rdf stream engine benchmarks. In: BeRSys 2013 (2013)
6. Dell'Aglio, D., Calbimonte, J., Balduini, M., Corcho, Ó., Valle, E.D.: On correctness in RDF stream processor benchmarking. In: ISWC 2013, Sydney, Australia.
7. Dell'Aglio, D., Della Valle, E., Calbimonte, J.P., Corcho, O.: RSP-QL Semantics: a Unifying Query Model to Explain Heterogeneity of RDF Stream Processing Systems. IJSWIS (to appear) 10(4) (2015)
8. Le-Phuoc, D., Dao-Tran, M., Xavier Parreira, J., Hauswirth, M.: A native and adaptive approach for unified processing of linked streams and linked data. In: ISWC. pp. 370–388 (2011)
9. Zhang, Y., Duc, P., Corcho, O., Calbimonte, J.P.: SRBench: A Streaming RDF/SPARQL Benchmark. In: ISWC. pp. 641–657 (2012)

Authors

Daniele Dell'Aglio is a PhD student at the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB) of the Politecnico di Milano since November 2012. His research activities focus on data stream management and reasoning; his major research topic is the optimization of the continuous query answering process in presence of volatile and heterogeneous data.

Jean-Paul Calbimonte is a postdoctoral research fellow in the Distributed Information Systems Laboratory at EPFL, Switzerland. His work focuses on Web data integration, streaming data processing for the Internet of Things, and the Semantic Sensor Web. He also helps coordinating the W3C Community Group on RDF Stream Processing (RSP).

Emanuele Della Valle is assistant professor at DEIB - Politecnico di Milano. He performs researches that are justified and guided by business needs in smart city, social media analytics, and, previously, in health care and life science. In more than a decade of research, his research interests covered Semantic Web, Big Data, Stream Management Systems, Search Engines, Rank-aware Databases and Service Oriented Architectures. His major research contributions are the Stream Reasoning concept (an approach to master the velocity and variety dimensions of Big Data), and its embodiment in the Continuous SPARQL query language and execution environment.

Oscar Corcho is Associate Professor at Departamento de Inteligencia Artificial (Facultad de Informática , Universidad Politécnica de Madrid) , and he belongs to the Ontology Engineering Group. His research activities are focused on Semantic e-Science and Real World Internet, although he also works in the more general areas of Semantic Web and Ontological Engineering. In these areas, he has participated in a number of international, EU, and Spanish R&D projects.