Demo: YABench - Yet Another RDF Stream Processing Benchmark

Maxim Kolchin[1] and Peter Wetz[2]

[1] ITMO University, Russia,
kolchinmax@niuitmo.ru
[2] Vienna University of Technology, Austria,
peter.wetz@tuwien.ac.at

1 Introduction

Since the global diffusion of the world wide web, continuously flowing and rapidly changing data becomes a more and more prevalent paradigm. Often, we want to analyze data also by taking its temporal dimension into account. We are interested in what happens right now to draw conclusions about the current or future state of a system. Application domains for which making sense of frequently changing data flows include (i) data fusion in smart city contexts, (ii) environmental monitoring, (iii) public transport, (iv) and health care.

Data stream management systems (DSMSs) and accompanying query languages have already been addressed since 2003 [2]. DSMSs have their roots in traditional Database Management Systems (DBMSs), but extend them to deal with frequent data changes. They also introduce continuous queries which – in contrast to DBMSs – are executed continuously to provide updated results based on newly arriving data [8].

Recently the semantic web community gained interest in addressing the challenge of dealing with dynamic data streams. The combination of streaming data approaches with the power of semantic web technologies has the potential to yield promising results and allowing for novel methods in data stream analytics. A major advantage is that semantic web principles enable the fusion of different data sets by means of federated queries. This allows to use external data (e.g., *GeoNames*, *DBpedia*) in queries. Moreover, it is possible to conduct reasoning, i.e., inference of logical consequences via rules based on asserted facts. Finally, semantic approaches are expected to facilitate dealing with heterogeneous data sets, incomplete data, and complex domain models. Considerable work has already been done to cover these features in RSP (RDF Stream Processing) engines, however, especially reasoning and coping with heterogeneity are still in their infancy. A consecutive step will be to cover reasoning and heterogeneity also in benchmarking.

To realize the vision of semantic data stream management systems, there is a need to make a transition from *one time semantics* to *continuous semantics*. As a consequence, the new paradigm of continuous semantics gives rise to new research challenges: (i) How to conduct reasoning on streams? (ii) How to deal

with noisy and incomplete data? (iii) How to design a query language for semantic streams? (iv) How to enable parallel and distributed processing? (v) How to benchmark proposed approaches and engines?

In this paper and accompanying workshop we want to present preliminary results of our work dealing with the latter research challenge, i.e., benchmarking for semantic stream reasoning engines, called YABench (Yet Another RDF Stream Processing Benchmark). Based on above explanations several research prototypes capable of processing semantic data streams have already been proposed. C-SPARQL [3, 5, 4], CQELS [11], and $\text{SPARQL}_{\text{Stream}}$ [7, 6] represent the most prominent efforts related to the DSMS paradigm. On the other hand, EP-SPARQL [1], INSTANS [14, 13], and Sparkwave [10] are more closely related to complex event processing. Even though all of these approaches try to solve similar challenges, they differ in various important aspects: Among others, they employ different underlying systems, query rewriting mechanisms, execution strategies, and query semantics. Due to the plethora of proposals in this still very young field, it is crucial to provide benchmarks allowing for a fair and rigorous comparison along different dimensions (e.g., performance, correctness). YABench aims at substantially contributing to the field of RDF Stream Processor Benchmarking by providing a decoupled and flexible architecture for defining and conducting benchmarks.

2 Related Work

LSBench (A benchmark for Linked Stream Data processing engines) [12] first tackled the issue of comparing available RSP engines (C-SPARQL, CQELS, and JTALIS). Their evaluation framework uncovers conceptual and technical differences of such engines. The researchers identify performance shortcomings, but also conduct functionality and correctness tests.

SRBench (Streaming RDF/SPARQL Benchmark) [15] provides a set of queries covering important aspects of RSP, such as joining static data with streaming data, or performing ontology-based reasoning. They provide a functional evaluation of C-SPARQL, CQELS, and $\text{SPARQL}_{\text{Stream}}$, leading them to the conclusion that the capability of the engines is still limited.

CSRBench (Correctness checking Benchmark for Streaming RDF/SPARQL) [9] deals with the issue of checking the correctness of stream query results. Moreover they test if engines comply to their own operational semantics. This benchmark is complementary to others which deal with functional (SRBench), and performance-based (LSBench) evaluations. They find that none of the tested engines (C-SPARQL, CQELS, and $\text{SPARQL}_{\text{Stream}}$) passes all tests and provide explanations on why certain engines fail at specific queries.

3 YABench

YABench is a novel benchmark (work in progress) for RDF Stream Processing engines which builds upon findings of previous benchmarks in this domain. The

benchmark moreover relies on early feedback from the community (primarily from the W3C RSP Community Group[3]) and is currently hosted on GitHub[4].

The primary goal is to provide a decoupled and flexible architecture for creating, executing, and analyzing benchmarks on different streaming engines. Consequently, this will be the first attempt to facilitate joint evaluation of functional, correctness, and performance testing. YABench provides the following components to fulfill this vision: (i) a configurable stream data generator, (ii) a set of test queries, (iii) integration of engines to run the tests on, (iv) an oracle, which implements different operator semantics, (v) and a reporting application to visualize test results.

To this end we enable a parametrized definition of streaming scenarios based on the LinkedSensorData data set (cf. [15]). This generator is configurable, meaning that it enables to test an engines' scalability through creation of big test data sets. After having generated test stream data, an evaluation based on predefined queries can be run on supported engines. Next, an oracle, which implements different operational semantics, is executed over the same set of test stream data. The oracle is responsible to calculate precision, recall, and f-measure via comparing expected results with actual results received from the engines, therefore extending the approach undertaken in CSRBench. Moreover, YABench provides performance indicators, such as delay of results delivery. These measurements, for instance, allow to identify correlations between window parameters (e.g., size, frequency) and performance indicators (e.g., query processing delay, memory consumption) of an engine.

The engines' capability to produce correct results under high load within a short time span is essential to the success of RDF stream processing applications. Therefore YABench provides means to measure throughput, response time, hardware consumption, and scalability with respect to the correctness of results.

To conclude we would like to participate in the RDF Stream Processing Workshop at ESWC 2015 to present preliminary results and the status quo of YABench. Since the benchmark is still work in progress, we expect direct feedback and fruitful discussions to positively influence our work. Subsequently YABench – as a benchmarking framework – will also have a positive impact on the future developments of the RDF Stream Processing group's efforts, i.e., among other things, the definition of a common query syntax for continuous RDF streams.

References

1. Anicic, D., Fodor, P., Rudolph, S., Stojanovic, N.: EP-SPARQL: a unified language for event processing and stream reasoning. In: Proc.of the 20th Int. Conf. on World Wide Web, WWW 2011, Hyderabad, India, 2011. pp. 635–644 (2011)

[3] https://www.w3.org/community/rsp/ [Accessed 16th March, 2015]
[4] https://github.com/YABench [Accessed 16th March, 2015]

2. Arasu, A., Babcock, B., Babu, S., Datar, M., Ito, K., Motwani, R., Nishizawa, I., Srivastava, U., Thomas, D., Varma, R., Widom, J.: STREAM: the stanford stream data manager. IEEE Data Eng. Bull. 26(1), 19–26 (2003)
3. Barbieri, D.F., Braga, D., Ceri, S., Grossniklaus, M.: An execution environment for C-SPARQL queries. In: EDBT 2010, 13th Int. Conf. on Extending Database Technology, Lausanne, Switzerland, 2010, Proceedings. pp. 441–452 (2010)
4. Barbieri, D.F., Braga, D., Ceri, S., Valle, E.D., Grossniklaus, M.: C-SPARQL: SPARQL for continuous querying. In: Proc. of the 18th Int. Conf. on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009. pp. 1061–1062 (2009)
5. Barbieri, D.F., Braga, D., Ceri, S., Valle, E.D., Grossniklaus, M.: Querying RDF streams with C-SPARQL. SIGMOD Record 39(1), 20–26 (2010)
6. Calbimonte, J., Corcho, Ó., Gray, A.J.G.: Enabling ontology-based access to streaming data sources. In: The Semantic Web - ISWC 2010 - 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. pp. 96–111 (2010)
7. Calbimonte, J., Jeung, H., Corcho, Ó., Aberer, K.: Enabling query technologies for the semantic sensor web. Int. J. Semantic Web Inf. Syst. 8(1), 43–63 (2012)
8. Cugola, G., Margara, A.: Processing flows of information: From data stream to complex event processing. ACM Comput. Surv. 44(3), 15 (2012)
9. Dell'Aglio, D., Calbimonte, J.P., Balduini, M., Corcho, O., Della Valle, E.: On correctness in RDF stream processor benchmarking. In: The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II, pp. 326–342. Springer (2013)
10. Komazec, S., Cerri, D., Fensel, D.: Sparkwave: continuous schema-enhanced pattern matching over RDF data streams. In: Proc. of the Sixth ACM Int. Conf. on Distributed Event-Based Systems, DEBS 2012, Berlin, Germany. pp. 58–68 (2012)
11. Le-Phuoc, D., Dao-Tran, M., Parreira, J.X., Hauswirth, M.: A native and adaptive approach for unified processing of linked streams and linked data. In: The Semantic Web - ISWC 2011 - 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I. pp. 370–388 (2011)
12. Le-Phuoc, D., Dao-Tran, M., Pham, M., Boncz, P., Eiter, T., Fink, M.: Linked stream data processing engines: facts and figures. In: The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part II, pp. 300–312. Springer (2012)
13. Rinne, M., Nuutila, E.: Constructing event processing systems of layered and heterogeneous events with SPARQL. In: On the Move to Meaningful Internet Systems: OTM 2014 Conferences, Amantea, Italy, 2014, Proceedings. pp. 682–699 (2014)
14. Rinne, M., Nuutila, E., Törmä, S.: INSTANS: high-performance event processing with standard RDF and SPARQL. In: Proc. of the ISWC 2012 Posters & Demonstrations Track, Boston, USA, November 11-15, 2012 (2012)
15. Zhang, Y., Duc, P.M., Corcho, O., Calbimonte, J.P.: SRBench: a streaming RDF/SPARQL benchmark. In: The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I, pp. 641–657. Springer (2012)