Towards maintainable Web APIs with Linked Event Streams

Search Strategy

De Clercq, Pieter - De Witte, Andreas - Devos, Robin - Veldeman, Sieben

October 25, 2020

Step 1

We started with meeting with our supervisors about the subject. During this meeting, the supervisors explained us what the expectations were and gave us some papers and directions to start from. We also planned future meetings and decided to meet every week on Monday.

We were given the following papers:

1. R&Wbase: Git for triples [1].
2. Comparing a polling and push-based approach for live Open Data interfaces [2].
3. Geospatial partitioning of open transit data [3].

Additionally, we were told to read about the following subjects:

1. Kafka[1]
2. Ostrich[2]
3. Event streams[3]

[1] https://kafka.apache.org/
[2] https://github.com/rdfostrich/ostrich
[3] https://github.com/TREEcg/specification/tree/master/examples/eventstreams

Step 2

After the first meeting, we started reading in into the subject by reading the provided papers and subjects from step 1. These provided us with some basic knowledge about the subject. During the following meeting, we discussed our original search strategy. We started by discussing which technologies seemed interesting for the target of the research. We decided that we would look into the following terms:

1. Kafka
2. Git based
3. MQTT
4. DNS
5. Decentralization
6. Azure service bus
7. Linked data streaming
8. Government Open Data

These terms will lead to the actual papers that are found in the following steps.

Step 3

1. Search “mqtt event streaming” on ResearchGate, resulted in [4].
2. Search “DNS based data publishing” and “DNS based data” on Google Scholar, resulted in [5] and [6].
3. Search “messaging protocols” on ieeexplore, resulted in [7].
4. Search “service bus” on the Azure-Docs GitHub repository, resulted in this overview of the Azure Service Bus[4].
5. Search “DNS update propagation” on ieeexplore, resulted in [8].
6. Search “DNS-over-HTTPS” on ResearchGate, resulted in [9] and [10].

[4] https://github.com/MicrosoftDocs/azure-docs/blob/master/articles/service-bus-messaging/

7. Search “event streams linked data” on Google Search, resulted in [11] and [12].
8. Search “linked data querying” on Google Scholar, resulted in [13].
9. Search “linked data git” on Google Scholar, resulted in [14], [15], [16], [17] and [18].
10. Search “git versioning of datasets” on Google Scholar, resulted in [19].
11. Search “linked data stream” on Google Scholar, resulted in [20].
12. Search “Government Big Open Data” on Web of Science, resulted in [21].
13. Search “open data updates” on Web of Science, resulted in [22].

Step 4

Forward and backward tracking of given and found papers resulted in the following resources.

1. Backward tracking from [9], resulted in [23].
2. Use Google Scholar to research papers citing [1], resulted in [24].
3. Use Google Scholar to research papers citing [13], resulted in [25].
4. Use Google Scholar to research papers citing [26], resulted in [27].
5. Use Google Scholar to research papers cited by found papers (backwards), this did not yield any additional useful results.
6. Research main contributors of Apache Kafka. Research work of the contributors is not very useful, but we found [28].
7. Use Web of Science to research papers cited in [21], resulted in [29].
8. Use Web of Science to research papers cited in [29], resulted in [30].
9. Use Web of Science to research papers cited in [30], resulted in [31].

List of articles of interest

Will certainly appear in the survey paper

- [1] R&wbase proposes a git-based file structure and organisation to allow versioning linked data.
- [3] The paper tries to address the problem of storing too large datasets for (in this case) public transit data. The proposed solution consists of fragmenting the data and making use of linked data, as to limit the amount of data needed to resolve a query.
- [4] This paper describes the application of MQTT as a communication protocol for IoT devices. We can use this to send updates to clients with very low latency and high performance.
- [13] One of the main problems of SPARQL endpoints is the limited scalability. This paper proposes Linked Data Fragments as a way to increase the scalability of SPARQL endpoints in turn for increased query times.
- [20] This paper is about C-SPARQL, an extension on top of SPARQL. C-SPARQL makes it possible to query over a stream of RDF triples. Since it is an extension on top of SPARQL, already existing functionalities of SPARQL can be used. The same goes for the SPARQL syntax. The RDF stream is defined as an ordered sequence of pairs, where each pair is made of an RDF triple and its timestamp. Especially the part about how the RDF stream is defined is usefull.

- [22] This paper discribes an approach to updates caches of open data using an appplication-aware change prioritization aproach.

Will most likely appear in the survey paper

- [7] Exploration of different MoM techniques. Instead of only using MQTT, this paper also presents AMQP which allows unlimited message lengths, as well

as storing messages

- [24] The paper presents a new version controlled way of storing and working with Linked Data. The paper also mentions related work such as R&Wbase: git for triples (by P. Colpaert e.a.) and claims this to be an improvement.
- [9] Instead of running DNS over UDP as is the case today, this paper proposes to use DNS over HTTPS. This could be beneficial, because we can make use of existing caching solutions provided by the HTTP standard.
- [12] Document proposes another way of publishing event streams as linked data, using their own defined entities for events and such. This corresponds to the TREE specification and it could be interesting to comapre.
- [10] A new way to use DNS imposes new privacy and security challenges to tackle. This paper gives an outline of the main concerns.
- [14] The paper is about versioning linked data. this builds on top of the git stack.
- [15] The paper is doing versioning of RDF data using a combination of SPARQL 1.1 and Git.
- [19] The paper is about comparing versions of a git versioned quad store by using Quit diff.
- [21] This paper gives an overview of current data distribution and interpretation mechanisms used by Governments.
- [29] This paper is about the current need and use of intermediaries on government open data.
- [31] This paper gives an overview of the existing state of the art in open data.

Will not likely appear in the survey paper

- [2]
- [8]
- [28]
- [11]

- [16]
- [17]
- [18]
- [25]
- [23]
- [26]
- [27]
- [30]

Will not appear in the survey paper

- [5]
- [6]

References

[1] M. Vander Sande, P. Colpaert, R. Verborgh, S. Coppens, E. Mannens, and R. Van de Walle, “R&wbase: Git for triples.” 2013, p.

[2] B. Van de Vyvere, P. Colpaert, and R. Verborgh, “Comparing a polling and push-based approach for live open data interfaces,” in International conference on web engineering, 2020, pp. 87–101.

[3] H. Delva, J. A. Rojas, P.-J. Vandenberghe, P. Colpaert, and R. Verborgh, “Geospatial partitioning of open transit data,” in International conference on web engineering, 2020, pp. 305–320.

[4] R. Atmoko, R. Riantini, and M. Hasin, “IoT real time data acquisition using MQTT protocol,” Journal of Physics: Conference Series, vol. 853, p. 012003, May 2017, doi: 10.1088/1742-6596/853/1/012003[5].

[5] W. Spektor Daron (Seattle, “DNS-based content routing,” 86892802014 [Online]. Available: https://www.freepatentsonline.com/8689280.html

[6] I. Bogner Etay (Tel-Aviv, “Data distribution using DNS,” 79872912011 [Online]. Available: https://www.freepatentsonline.com/7987291.html

[7] J. E. Luzuriaga, M. Perez, P. Boronat, J. C. Cano, C. Calafate, and P. Manzoni, “A comparative evaluation of AMQP and MQTT protocols over unstable and mobile networks,” in 2015 12th annual IEEE consumer communications and networking conference (CCNC), 2015, vol., pp. 931–936, doi: 10.1109/CCNC.2015.7158101[6].

[8] M. R. Parwez, M. Akbar, S. Haider, and M. S. Javaid, “DNS propagation delay: An effective and robust solution using authoritative response from non-authoritative server,” in 2010 2nd IEEE international conference on information management and engineering, 2010, vol., pp. 150–153, doi: 10.1109/ICIME.2010.5477485[7].

[9] S. Hrushak and C. Pavlenko, “Advantages of DNS-over-HTTPS over DNS,” 2020, p., doi: 10.30837/IVcsitic2020201356[8].

[10] K. Borgolte et al., “How DNS over HTTPS is reshaping privacy, performance, and policy in the internet ecosystem,” 2019.

[11] R. Tommasini, M. Ragab, A. Falcetta, E. D. Valle, and S. Sakr, “Bootstrapping the publication of linked data streams,” in ISWC satellites, 2019.

[12] A. Harth and R. Stühmer, “Publishing event streams as linked data’,” Technical report, Karlsruhe Institute of Technology, http://km. aifb. kit ..., 2011.

[13] R. Verborgh, M. Vander Sande, P. Colpaert, S. Coppens, E. Mannens, and R. Van de Walle, “Web-scale querying through linked data fragments,” in CEUR Workshop Proceedings, 2014, vol. 1184, p.

[14] N. Arndt, P. Naumann, and E. Marx, “Exploring the evolution and provenance of git versioned RDF data,” in MEPDaW/LDQ@ESWC, 2017.

[15] N. Arndt, N. Radtke, and M. Martin, “Distributed collaboration on RDF datasets using git: Towards the quit-store,” 2016, p., doi: 10.1145/2993318.2993328[9].

[16] A. Meroño-Peñuela and R. J. Hoekstra, “SPARQL2Git: Transparent SPARQL and linked data API curation via git,” in Proceedings of the 14th extended semantic web conference (ESWC 2017), poster and demo track, 2017.

[17] D. O. Kubitza, M. Böckmann, and D. Graux, “SemanGit: A linked dataset from git,” in The semantic web – ISWC 2019, 2019, pp. 215–228.

[18] A. Merono Penuela and R. J. Hoekstra, “Grlc makes GitHub taste like linked data APIs,” in The semantic web - ESWC 2016 satellite events, revised selected papers, 2016, vol. 9989 LNCS, pp. 342–353, doi: 10.1007/978-3-319-47602-5_48[10].

[19] N. Arndt and N. Radtke, “Quit diff: Calculating the delta between RDF datasets under version control,” in Proceedings of the 12th international conference on semantic systems, 2016, pp. 185–188, doi: 10.1145/2993318.2993349[11] [Online]. Available: https://doi.org/10.1145/2993318.2993349

[20] D. Barbieri and E. Della Valle, “A proposal for publishing data streams as linked data - a position paper.” 2010, vol. 628, p.

[21] M. Lněnička and J. Komárková, “Developing a government enterprise architecture framework to support the requirements of big and open linked data with the use of cloud computing,” International Journal of Information Management, vol. 46, pp. 124–141, Jun. 2019, doi: 10.1016/j.ijinfomgt.2018.12.003[12].

[22] U. Akhtar et al., “Change-aware scheduling for effectively updating linked open data caches,” IEEE Access, vol. PP, pp. 1–1, Sep. 2018, doi: 10.1109/ACCESS.2018.2871511[13].

[23] P. Hoffman and P. McManus, “DNS queries over HTTPS (DoH),” RFC Editor; Internet Requests for Comments; RFC Editor, RFC 8484, 2018.

[24] M. Graube, S. Hensel, and L. Urbas, “R43ples: Revisions for triples an approach for version control in the semantic web,” CEUR Workshop Proceedings, vol. 1215, p., Jan. 2014.

[25] T. Kuhn, C. Chichester, M. Krauthammer, and M. Dumontier, “Publishing without publishers: A decentralized approach to dissemination, retrieval, and archiving of data,” in The semantic web - ISWC 2015, 2015, pp. 656–672.

[26] R. Verborgh et al., “Querying datasets on the web with high availability,” in The semantic web – ISWC 2014, 2014, pp. 180–196.

[27] P. Folz, H. Skaf-Molli, and P. Molli, “CyCLaDEs: A decentralized cache for linked data fragments,” in ESWC 2015, 2015.

[28] G. Wang, Awesome streaming. 2020 [Online]. Available: https://github.com/guozhangwang/awesome-streaming

[29] F. V. Schalkwyk, M. Willmers, and M. McNaughton, “Viscous open data: The roles of intermediaries in an open data ecosystem,” Information Technology for Development, vol. 22, pp. 68–83, 2016.

[30] Y. Dwivedi et al., “Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modelling,” Information Systems Frontiers, p., Jul. 2016, doi: 10.1007/s10796-016-9675-5[14].

[31] M. Hossain, Y. Dwivedi, and N. Rana, “State of the art in open data research: Insights from existing literature and a research agenda,” Journal of Organizational Computing and Electronic Commerce, vol. 26, p., Dec. 2015, doi: 10.1080/10919392.2015.1124007[15].