Processing Life Science Data at Scale - using Semantic Web Technologies

Ali Hasnain[1], Naoise Dunne[1], and Dietrich Rebholz-Schuhmann[1]

Insight Center for Data Analytics, National University of Ireland, Galway
firstname.lastname@insight-centre.org

Abstract. The life sciences domain has been one of the early adopters of linked data and, a considerable portion of the Linked Open Data cloud is comprised of datasets from Life Sciences Linked Open Data (LSLOD). The deluge of biomedical data in the last few years, partially caused by the advent of high-throughput gene sequencing technologies, has been a primary motivation for these efforts. This success has lead to the growth in size of data sets and to the need for integrating multiples of these data-sets. This growth requires large scale distributed infrastructure and specific techniques for managing large linked data graphs. Especially in combination with Semantic Web and Linked Data technologies these promises to enable the processing of large as well as semantically heterogeneous data sources and the capturing of new knowledge from those. In this tutorial we present the state of the art in large data processing, as well as the amalgamation with Linked Data and Semantic Web technologies for better knowledge discovery and targeted applications. We aim to provide useful information for the Knowledge Acquisition research community as well as the working Data Scientist.

1 Motivation

A study from EMC[1] in 2014 predicts the doubling of the available data in the "Digital Universe" every two years between now and 2020. This rapid growth is a challenge for society – how to put the available data to use effectively? This is a challenge in many areas including Medicine and Life Sciences as well as areas in Engineering and Science, which depend on data such as Energy and Materials research. Advanced technologies and sensors as well as harnessing existing archives of data or the emerging Open Data phenomenon produce an ever-increasing amount of data, which needs to be interpreted and examined. To turn data into knowledge, data scientists need to effectively process, filter, interpret cluster and learn from the available data. This process currently is largely unsupported – data scientists are spending time and money on processing data, configuring infrastructure, writing code etc., which is a large loss of productivity and unexploited opportunities, if data scientists with the necessarily skills are available at all. To meet this challenges technologies from different areas need to

[1] http://www.emc.com/leadership/digital-universe/index.htm

be combined. In particular promising seem to be Large Scale Data technologies - e.g., job scheduling and cloud infrastructures, which help to execute computationally demanding tasks. Linked Data and Semantic Web technologies, coming from a different direction, help to bring heterogeneous data sources together to exploit and make sense of different datasets and making it easier to process semantically heterogeneous data.
In this tutorial we present the anatomy of large scale linked data infrastructure, which covers: the distributed infrastructure to consume, store and query large volumes of heterogeneous linked data; using indexes and graph aggregation to better understand large linked data graphs, query federation to mix internal and external data-sources [13,12], and linked data visualisation tools for health care and life sciences. The tutorial will cover the necessary computational and storage infrastructure - enabling scientists to define and execute computational intensive tasks on large distributed clusters or cloud infrastructure, to index, summarise and aggregate available linked data sources (e.g., with Linked Data technology). It will further cover federation tasks allowing scientists to query beyond their own infrastructure and innovations in Linked Data Visualisation for HCLS. We aim to provide researchers in health care and life science, an insight and awareness about Larg Scale Data technologies for linked data, which are becoming increasingly important for knowledge discovery in the HCLS domain.

2 Related Events

Marco et al. [11] presented a tutorial on Big Data Management and cover the current state of the Big Data area of research and applications. Their emphasis was on the research aspects of the Big Data presenting the characteristic algorithms running under the hood of the Big Data platforms and applications. A Co-Located event namely Big 2014[2] was hosted along with *23rd International World Wide Web Conference- 2014*[3] that covered different aspects related to Big Data. The topics cover included the use of semantic metadata and ontologies for Big Data and methods to establish semantic interoperability between data sources. Tutorial name *"Big Data Stream Mining"* was organized at IEEE BigData 2014 Conference[4], that introduce mining big data streams. The topic includes the classification, regression, clustering, frequent pattern mining and data stream mining on distributed engines. Another tutorial *"Big Data Benchmarking"* was organized as a part of same conference that introduce the set of issues involved in defining big data benchmarks that consider performance as well as price and performance. Presenters presented topics, including heterogeneous data, e.g. structured, semi- structured, unstructured, graphs, and streams; large-scale and evolving system configurations; varying system loads; processing pipelines that progressively transform data; workloads that include queries as well as data mining and machine learning operations and algorithms. In regard

[2] `http://big2014.org/` (l.a: 25-06-2015)
[3] `http://www2014.kr/`
[4] `http://cci.drexel.edu/bigdata/bigdata2014/tutorial.htm`

to query federation, Hartig and Ozsu [1] presented a tutorial on Linked Data query processing and discussed the basics of the Linked Data, SPARQL query and query optimization strategies. Moreover tutorials on RDF Stream Processing (RSP) have been presented both in ISWC 2013 as well as 2014 and in ESWC 2014.

3 Detailed Description

In this section we describe the contents of tutorial, the aims and learning objectives, presentation style and tutorial format, and the prior knowledge required by the attendees. Our tutorial will consist of following sessions:

Classical Query Federation: In this session we will discuss the concepts around SPARQL query Federation to access multiple heterogeneous biological datasets to draw meaningful biological co relations. Real Life sciences Dataset e.g Drugbank, Dailymed will be queried to elaborate SPARQL Federation.

Scalable Infrastructure: In this session we will discuss the concerns, available tools and current trends in the creation of the distributed infrastructures for processing of large Linked Datasets at scale.
We will begin with what considerations are needed when building and running these infrastructures, highlighting the rationale for using containers, the need for schedulers to manage both jobs and resources, and the importance of managing failure in Large Scale Data architectures. We will then explore existing popular schedulers Hadoop [15], Yarn and Mesos [7] and the use of Docker containers.
We will then focus on the specific concerns of Linked Data Pipelines, and the trade-offs vs other architectures. The evolution from Batch to Real Time to Lambda Architectures [14] and emergence of Reactive Pipelines. Finally we will review our own Big Linked Data Knowledge pipeline and how we met these concerns.

Graph Aggregation In this session we will discuss how we can generate useful aggregations of the distributed graph to better understand the structure of the underlying data. This helps us to better interact with the graph [18], helps our understanding of large linked data sets where some of the data schema is missing or mixed with other schema and finally how these techniques can be used to extend the sparql language to include aggregation operations[10] such as to analyse biological networks.

Visualisation: In this session we will demo different visualization tools and application build to visualize Big RDF Data. Applications include ReVeaLD- a Real-time Visual Explorer and Aggregator of Linked Data [9], Genome Wheel-GenomeSnip- Fragmenting the Genomic Wheel to augment discovery in cancer research [8] and FedViz- A Visual Interface for SPARQL Queries Formulation and Execution.

Hands On Session: In this session we will give the audience some practical exposure to the tools and technologies discussed in earlier sessions. Using a sample rdf dataset, we will ask the audience to explore using a visualisation tool to create SPARQL query. We will guide the audience to create a graph summary of this rdf dataset, and show how this can help group results in a meaningful way.

3.1 Aims and Learning Objectives

Our learning objectives are the following:

1. Provide basic knowledge regarding the fundamentals of Large Scale Data in Life Sciences and related technologies.
2. Elaborate how semantic web technologies are useful for managing Large Scale Data.
3. Elaborate to access and benefit from semantic data on the Web.
4. Elaborate how to make use of Large Scale Data and introduce some of the current applications based on Semantic Web technologies

3.2 Presentation Style and Format

Our presentations will be based on animations, running examples, Live Demos, hands-on exercises and visualisation. Audience will be given chance to to ask questions throughout the presentations. The first four sessions will last 35 minutes each. Last section including the practical exposure to the tools and technologies will be conducted at the end of the tutorial and will last for about 1 hour. If time permits, we will discuss some of the open problems in dealing with Large Scale Data. Due to the emerging interest in Large Scale Data and its applications, we expect that most of the semantic web community will be interested in our tutorial and are thus expecting between 25 and 35 attendees.

3.3 Pre-requisite

The audience are required to have basic knowledge of the SPARQL query and Linked Data.

4 Length

We plan it as a half day tutorial.

5 Technical Requirements

Participants of the hands on session require a familiarity with SPARQL, they will need a text editor and will require a HTML 5 enabled web browser.

6 Presenters

Ali Hasnain (primary contact)
Insight Centre for Data Analytics
National University of Ireland, Galway
Email: ali.hasnain@insight-centre.org
List of publications: http://goo.gl/4ptdSU

Ali Hasnain is Research Assistant and PhD candidate at Insight Centre for Data Analytics (NUIG), (formerly: Digital Enterprise Research Institute (DERI), Galway Ireland). Before joining DERI, Hasnain completed a Master Degree in "Engineering and Management of Information Systems" from Royal Institute of Technology, KTH, Stockholm Sweden. He received another Master Degree from the same University in "Project Management and Operational Development". Along with the research activities he had been involved in Teaching and Research and Development activities at KTH and couple of years of industrial experience.

Current research interests include: Linked Open Data [2], Big Data, Semantic Models [17] [9], Data Catalouging and Linking [4] [6],Semantic Matching and Relatedness, Link Discovery, Visual Query Formulation [16], Data Provenance [5] and Data Integration [3].

Naoise Dunne
Insight Centre for Data Analytics
National University of Ireland, Galway
Email: naoise.dunne@insight-centre.org

Naoise Dunne has over 19 years work experience in architect, developer and management roles delivering big data enterprise solutions. At Insight NUI Galway (formerly DERI), he carries the position of a Research Fellow since October 2011 and has worked on Distributed Infrastructure architectures, he has delivered a number of large projects both commercial and research. His work focuses on distributed computing architectures.
Prior to working with Insight, Naoise worked as a developer and coach with advanced agile teams, sat on O2 Europe's technical steering group as a system architect, and was CTO of a small Internet start-up (30 staff).
Naoise's current research interests concern distributed infrastructures, distributed graph systems, distributed linked data indexing and compression, reactive streaming, and data visualisation.
Grant history:

1. RETIS, EI Commercialisation Fund 2011.
2. Smarter Data EI Commercialisation Fund 2013.
3. Safety Monitor EI Commercialisation Fund 2014.

Dietrich Rebholz-Schuhmann.
Insight Centre for Data Analytics
National University of Ireland, Galway
Email: rebholz@insight-centre.org
List of publications: https://goo.gl/TZuFMA

Prof Rebholz-Schuhmann is a qualified doctor with a PhD in immunology. Now a professor of computer science at NUI-Galway and director of the Insight Centre for Data Analytics (Galway site), he specializes in the fields of Semantic Web and Linked Data for biomedical data anlysis. Dr. Dietrich Rebholz-Schuhmann holds a master in medicine (Univ. Duesseldorf, 1988), a Ph.D. in immunology (Univ. Duesseldorf, 1989) and a master in computer science (Univ. Passau, 1993). He worked as a senior scientist at the gsf (Munich) in the field of image analysis and 3D visualization. From 1998 to 2003 he headed a team at LION bioscience AG, Heidelberg working on novel text mining solutions. From 2003 to 2012 he was research group leader at the European Bioinformatics Institute, Hinxton (Uk) doing research in biomedical literature analysis. Prior to his position at Insight (Galway), he was senior researcher at the University of Zürich in the department of computational linguistics leading the Mantra Project.
Dietrich's main research interests are biomedical informatics, literature analysis, ontologies, Semantic Web and Information representation. Dietrichn's Google Scholar profile claims more than 2963 citations with h-index of 34.

References

1. Hartig, O., Ozsu, M.T.: Linked data query processing. In: Data Engineering (ICDE), 2014 IEEE 30th International Conference on. pp. 1286–1289. IEEE (2014)
2. Hasnain, A., Al-Bakri, M., Costabello, L., Cong, Z., Davis, I., Heath, T.: Spamming in linked data. In: Third International Workshop on Consuming Linked Data (COLD2012) (2012)
3. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences lod cloud. In: OEDW at EKAW (2012)
4. Hasnain, A., Kamdar, M.R., Hasapis, P., Zeginis, D., Warren Jr, C.N., et al.: Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery. In: International Semantic Web Conference (In-Use Track), October 2014 (2014)
5. Hasnain, A., Mehmood, Q., e Zainab, S.S., Decker, S.: A provenance assisted roadmap for life sciences linked open data cloud. In: Knowledge Engineering and Semantic Web, pp. 72–86. Springer (2015)
6. Hasnain, A., Zainab, S.S.E., Kamdar, M.R., Mehmood, Q., Warren Jr, C., et al.: A roadmap for navigating the life scinces linked open data cloud. In: International Semantic Technology (JIST2014) conference (2014)
7. Hindman, B., Konwinski, A., Zaharia, M., Ghodsi, A., Joseph, A.D., Katz, R.H., Shenker, S., Stoica, I.: Mesos: A platform for fine-grained resource sharing in the data center. In: NSDI. vol. 11, pp. 22–22 (2011)
8. Kamdar, M., Iqbal, A., Saleem, M., Deus, H., Decker, S.: Genomesnip: Fragmenting the genomic wheel to augment discovery in cancer research. In: Conference on Semantics in Healthcare and Life Sciences (CSHALS) (2014)

9. Kamdar, M.R., Zeginis, D., Hasnain, A., Decker, S., Deus, H.F.: ReVeaLD: A user-driven domain-specific interactive search platform for biomedical research. Journal of Biomedical Informatics 47(0), 112 – 130 (2014)
10. Maali, F., Campinas, S., Decker, S.: Gagg: A graph aggregation operator. In: The Semantic Web. Latest Advances and New Domains, pp. 491–504. Springer (2015)
11. Marko Grobelnik, Blaz Fortuna, D.M.: Introduction to big data (2013)
12. Rakhmawati, N.A., Umbrich, J., Karnstedt, M., Hasnain, A., Hausenblas, M.: Querying over federated sparql endpoints—a state of the art survey. arXiv preprint arXiv:1306.1723 (2013)
13. Saleem, M., Khan, Y., Hasnain, A., Ermilov, I., Ngomo, A.C.N.: A fine-grained evaluation of sparql endpoint federation systems. Semantic Web Journal (2014)
14. Thorpe, S.R., Battestilli, L., Karmous-Edwards, G., Hutanu, A., MacLaren, J., Mambretti, J., Moore, J.H., Sundar, K.S., Xin, Y., Takefusa, A., et al.: G-lambda and enlightened: wrapped in middleware co-allocating compute and network resources across japan and the us. In: Proceedings of the first international conference on Networks for grid applications. p. 5. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2007)
15. White, T.: Hadoop: the definitive guide: the definitive guide. " O'Reilly Media, Inc." (2009)
16. e Zainab, S.S., Hasnain, A., Saleem, M., Mehmood, Q., Zehra, D., Decker, S.: Fedviz: A visual interface for sparql queries formulation and execution
17. Zeginis, D., et al.: A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web (2013)
18. Zhang, N., Tian, Y., Patel, J.M.: Discovery-driven graph summarization. In: Data Engineering (ICDE), 2010 IEEE 26th International Conference on. pp. 880–891. IEEE (2010)