Knowledge Processing with Big Data and Semantic Web Technologies

Ali Hasnain[1], Naoise Dunne[1], and Stefan Decker[1]

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

Abstract. Knowledge Processing and Acquisition are important activities for Data Scientists in a world rich with large heterogeneous data sources. Big Data not only provide an important set of tools to acquire and capture knowledge for the working data scientist, but also for the Knowledge Acquisition researcher, enabling the processing of large data sources. 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 Big Data processing, as well as the amalgamation with Linked Data and Semantic Web technologies to form a Knowledge Pipeline. 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, the Social Sciences and the Humanities and Archives, 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 be combined. In particular promising seem to be Big 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

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

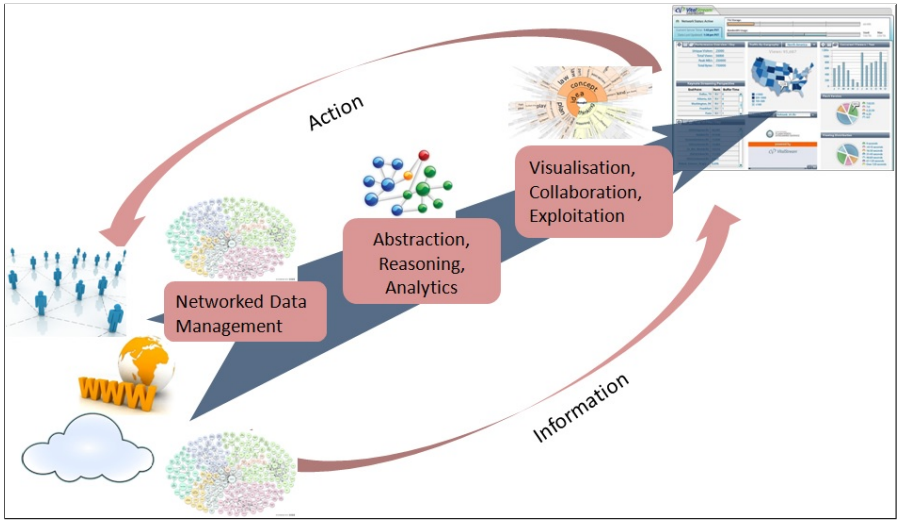

Fig. 1: Knowledge Pipeline

bring heterogeneous data sources together to exploit and make sense of different datasets and making it easier to process semantically heterogeneous data. The amalgamation of technologies from these different fields enables a new process for knowledge discovery, which can be supported by computational infrastructure that empowers data scientists and enables them to make new discoveries. We call this infrastructure a Knowledge Pipeline (see fig 1). The Knowledge Pipeline bundles and organises available technology in a user friendly and integrated way, that enables a scientist and practitioners to explore data, as well as rapidly integrate, transform, filter and process data from distributed data sources, and finally to analyse data and discover new knowledge. It keeps provenance and records the data processing work flow, so enabling the scientist to rapidly replay, analyse and change the processing, lifting the burden from configuring infrastructure or writing code and instead enabling the data scientists to concentrate on the ideas that transform data into knowledge. As indicated the knowledge pipeline is applicable across the area of data intensive sciences or humanities.

In this tutorial we present the anatomy of a Knowledge Pipeline, which covers the process to build, execute, curate, analyse to publish and visualise data. The tutorial will cover the necessary computational infrastructure - enabling scientists to define computational intensive tasks and schedule and execute them on a cluster or cloud infrastructure, to curate and integrate available data sources (e.g., with Linked Data technology). It will further cover data curation (e.g., entity reconciliation) tasks, analytics tasks (e.g., Machine Learning) and Visualisation and Publishing tasks. We aim to provide researchers in knowledge acquisition, an insight and awareness about Big Data technologies, which are becoming increasingly important for knowledge processing.

2 Related Events

Marco et al. [12] 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.

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:

Infrastructure: In this session we will discuss the concerns, available tools and current trends in the creation of the distributed infrastructures for processing of Big Linked Data Knowledge pipeline.
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 Big Data architectures. We will then explore existing popular schedulers Hadoop [18], Yarn and Mesos [8] 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 [16] and emergence of Reactive Pipelines. Finally we will review our own Big Linked Data Knowledge pipeline and how we met these concerns.

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

Data Curation: In this session we will discuss bringing data from other sources into Linked Data platforms. We will review techniques to transforming structured data into linked data. We will demonstrate two popular mapping tools D2RQ [1] and TARQL [9] and how to use vocabularies/ ontologies to build better linked data graphs. We will then discuss some popular linked data representations such as N Triples, JSON-LD and HDT, help the audience understand when to use which representation and show how easy it is to move from one representation to another. Finally we will demonstrate this kind of mapping using the mapping tool in the Linked Data Knowledge pipeline.

Entity Reconciliation: In this session we will discuss the topic of entity reconciliation, basic definitions and different approaches to reconcile an entity. We will also provide an overview of different tools used for entity reconciliation process. This includes SILK [17] and LIMES [13]. We'll also cover the concept of Graph Summarization [2] during this session.

Analyze: In this session we will discuss how to query linked data at scale and use well known graph and machine learning algorithms to make interesting findings in linked data. We will start with the SPARQL query language, and current trends in scaling SPARQL such as federated queries [15,14]. We will then explore link linked data as a heterogeneous graph, we can also use popular graph traversal languages and distributed datasets with linked data. We will demonstrate the use of Gremlin as traversal language and GraphX as a distributed data structure. We will discuss the kinds of algorithms that can be used by these techniques focusing on Clustering - Luvene K-Means, Dijkstra, Page Rank. If we have time we can talk about representation as tensors for traditional machine learning.

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 [11], Genome Wheel-GenomeSnip- Fragmenting the Genomic Wheel to augment discovery in cancer research [10] and FedViz- Formulating Federated SPARQL Query through Visualization.

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 dataset in traditional csv tables, we will ask the audience to create a mapping using a vocabulary, transform the data to linked data, then run a well known algorithm over the data, and visualise this data in our tool. Audience will be asked to schedule a job to run the same mapping and query over a much larger dataset using the Linked Data Knowledge pipeline. They can then share discoveries as they explore this larger dataset through the pipeline's visualisation tool.

3.1 Aims and Learning Objectives

Our learning objectives are the following:

1. Provide basic knowledge regarding the fundamentals of Big Data and related technologies.
2. Elaborate how semantic web technologies are useful in Creating Big Data Knowledge Pipeline.
3. Elaborate to access and benefit from semantic data on the Web.
4. Elaborate how to make use of Big 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 five 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 Big Data. Due to the emerging interest in Big 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, a computer with Java version 7, a working copy of the Git version management tool 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 [3], Big Data, Semantic Models [20] [11], Data Catalouging and Linking [5] [7],Semantic Matching and Relatedness, Link Discovery, Visual Query Formulation [19], Data Provenance [6] and Data Integration [4].

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.

Stefan Decker.
Insight Centre for Data Analytics
National University of Ireland, Galway
Email: stefan.decker@insight-centre.org
List of publications: https://goo.gl/ZgHPjo
Homepage: http://www.stefandecker.org

Stefan Decker is a professor at the National University of Ireland, Galway, and the Director of the Insight Centre for Data Analytics NUIG (formerly Digital Enterprise Research Institute). Previously he worked at ISI, University of Southern California (2 years, Research Assistent Professor and Computer Scientist), Stanford University, Computer Science Department (Database Group) (3 Years, PostDoc and Research Associate), and Institute AIFB, University of Karlsruhe (now KIT Karlsruhe) (4 years, PhD Student and Junior Researcher).
Stefan's main research field is the Semantic Web. Stefan's Google Scholar profile claims more than 15800 citations with h-index of 58.

References

1. Bizer, C., Seaborne, A.: D2rq-treating non-rdf databases as virtual rdf graphs. In: Proceedings of the 3rd international semantic web conference (ISWC2004). vol. 2004 (2004)
2. Campinas, S., Perry, T.E., Ceccarelli, D., Delbru, R., Tummarello, G.: Introducing rdf graph summary with application to assisted sparql formulation. In: Database and Expert Systems Applications (DEXA), 2012 23rd International Workshop on. pp. 261–266. IEEE (2012)
3. 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)
4. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences lod cloud. In: OEDW at EKAW (2012)
5. 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)
6. 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)
7. 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)
8. 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)
9. Kalampokis, E., Nikolov, A., Haase, P., Cyganiak, R., Stasiewicz, A., Karamanou, A., Zotou, M., Zeginis, D., Tambouris, E., Tarabanis, K.: Exploiting linked data cubes with opencube toolkit. In: International Semantic Web Conference (ISWC) (2014)
10. 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)
11. 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)
12. Marko Grobelnik, Blaz Fortuna, D.M.: Introduction to big data (2013)
13. Ngomo, A.C.N., Auer, S.: Limes-a time-efficient approach for large-scale link discovery on the web of data. integration 15, 3 (2011)
14. 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)
15. 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)
16. 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)

17. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk-a link discovery framework for the web of data. LDOW 538 (2009)
18. White, T.: Hadoop: the definitive guide: the definitive guide. " O'Reilly Media, Inc." (2009)
19. e Zainab, S.S., Hasnain, A., Saleem, M., Mehmood, Q., Zehra, D., Decker, S.: Fedviz: A visual interface for sparql queries formulation and execution
20. Zeginis, D., et al.: A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web (2013)