Coping with the Long Tail of Data Variety

Edward Curry and Andre Freitas

Insight Centre for Data Analytics
National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway
firstname.lastname@deri.org

1 Summary

The talk will discuss current challenges, approaches and future directions for coping with data variety. The discussion will be grounded on exemplar use cases from leaders in industry and in large-scale scientific projects such as IBM Watson, CrowdFlower, BBC, Press Association, ProteinDataBank, Data.gov.uk, Chemspider among others. The use cases were collected in interviews with Big Data industry and academic experts in the context of the BIG Project and provide a glimpse of the state of the art techniques which are currently being used to cope with data variety and the future directions and emerging trends for this field.

2 Contributors' names and short CVs

Edward Curry is a Research Scientist and leads the Green and Sustainable IT research group (dgsit.deri.ie) at the Digital Enterprise Research Institute (www.deri.ie). His research projects include studies of sustainable IT, energy intelligence, semantic information management, and collaborative data management. Edward has worked extensively with industry and government advising on the adoption patterns, practicalities, and benefits of new technologies. Edward has published over 70 scientific articles in journal, books, and international conferences. He has given invited talks at Berkeley, Stanford, and MIT. In 2010 he was a guest speaker at the MIT Sloan CIO Symposium to an audience of 600+ CIOs and senior IT executives. He currently participates in a project for the European Commission to define a research strategy for the Big Data economy within Europe. He has a PhD from the National University of Ireland, Galway (www.nuigalway.ie) and serves as an Adjunct Lecturer within the University. Specialties: Autonomic Computing, Collaborative Data Management, Cyber-Physical Systems, Energy Intelligence, Event-based Systems, Linked Enterprise Data, Semantic Information Management, Sustainable IT, Technology Management, Web of Things

Andre Freitas is a PhD Candidate at the Digital Enterprise Research Institute (DERI) at the National University of Ireland, Galway and is a partner and co-founder at Amtera Semantic Technologies. Andr holds a BSc. in Computer Science from the Federal University of Rio de Janeiro (UFRJ), Brazil (2005). His

main research areas include Semantic Search, Vocabulary-independent Query Mechanisms, Natural Language Query Mechanisms over Linked Data, Distributional Semantics, Semantic Relatedness, Approximate Reasoning and Provenance. Before joining DERI, Andr worked as a research assistant (trainee) at Siemens Corporate Research, Princeton, USA. Andr worked as a software engineer, product designer and project manager in different industries including Oil & Gas Exploration, IT Security, Medical, Healthcare, Banking, Mining and Telecom.

3 Type of the presentation proposed

Research contribution.

4 Extended Abstract

The database landscape is changing rapidly, influenced by the need to cope with data environments with growing data variety, larger schema size and schema dynamicity. The emergence of new data sources such as open datasets on the Web, scientific data, sensor networks, data from mobile applications, social network data, together with the natural growth of datasets inside organizations [3], brings the demand for data management strategies which can operate under the semantic and format heterogeneity characteristics of this new data environment.

Propelled by the growth of the Web and on the number of available computational devices, data management requirements are shifting towards the need to cope with decentralised data generation [2]. Decentralised data generation is intrinsically associated with the long tail of data variety, where data is spread across different domains and the frequency of access for each data item is comparatively low. Additionally, as data coverage increases as users have more tools for decentralised data generation, the connectedness between data items also tends to increase [2].

In a 2010 survey, Brodie & Liu [3] report that database environments in Fortune 100 companies typically consist of tens of thousands of information systems with hundreds of databases per business area, where 90% of them are relational, having an annual growth of 100s of databases per year. Typically, each database has between 100-200 tables, each table containing between 50-200 attributes. The number of views is typically three times the number of tables. In this data environment, the cost of integrating data across databases accounts for 40% of the software project costs. A comparative analysis shows the trend towards more complex and heterogeneous environments: while in 1985 a database would consist of two tables managed with a schema-based design, in 2010 there are 100-1000s of tables which are designed manually in an evidence gathering fashion, with 60-75% of these tables being schema-less.

The increasing availability of data brings the opportunity of directly impacting the fundamental process of sense making and knowledge creation for organisations and individuals, allowing them to reflect the complexity of the

real world into their datasets. This represents a natural evolution of information systems in the direction of a more complete and fine-grained representation of the reality. However, the natural evolution towards the persistence of more complex data inside databases disrupts existing data management practices which assume semantic and syntactic homogeneity.

There is a lack of more systematic analyses of the *data variety* phenomena in Big Data, comparing to the existing discourse with regard to *volume* and *velocity*. This talk will contribute to expanding the understanding of data variety, core techniques and future requirements for coping with it, grounded on the experience of the key players in the area. The talk will target both the technical (practitioners and academics) and the business audiences.

The talk will discuss current challenges, approaches and future directions for coping with data variety. The discussion will be grounded on exemplar use cases from leaders in industry and in large-scale scientific projects such as IBM Watson, CrowdFlower, BBC, Press Association, ProteinDataBank, Data.gov.uk, Chemspider among others.

The use cases were collected in interviews with Big Data experts and practitioners in the context of the BIG Project[1] and provide a glimpse of the state of the art techniques which are currently being used to cope with data variety and the future directions and emerging trends for this field. The uses cases provide a wide spectrum of data variety examples, covering unstructured/structured data, different data quality requirements and distinct economic models.

The talk will also provide a deeper analysis of the dimensions of the data variety problem as perceived by the leading experts in this field. An in-depth understanding of the data variety phenomena will allow us to provide a more precise definition for data variety for Big Data.

Additionally, existing projects provide clear directions on the key processes and technologies which are able to cope with data variety. While some of these elements are implemented as mature processes and functionalities across different organizations, other dimensions emerged as recurrent gaps, defining a roadmap for coping with data variety.

The critical dimensions for coping with data variety are briefly described below:

- *Improved Data Processing Techniques:* covering technologies across next generation data processing scenarios that enable the transformation of raw data into quality (useful) data. In this dimension data curation practices and infrastructures emerge as fundamental elements.
- *Improved Human-Data Interaction (HDI) Models:* enabling domain experts and casual users to query, explore, transform and curate data with a focus on the interactivity and ease of action.
- *Emerging Model Incentives and Rewards Model:* emerging economic models and ecosystems for data curation, Pre-competitive partnerships (e.g. Pistoia Alliance), Public-private partnerships.

[1] www.big-project.eu

References

1. Data centres: their use, value and impact, JISC Technical Report 2011.
2. E. Eifrem, NOSQL overview and intro to graph databases with Neo4j, 2010.
3. M. L. Brodie and J. T. Liu. The power and limits of relational technology in the age of information ecosystems. On The Move Federated Conferences, 2010.
4. http://www.big-project.eu/