Data Analytics Research Workshop

Denilson Barbosa (University of Alberta)

May 27–29, 2011

1 Overview of the Field

Business intelligence (BI) is the commercial term for using information within organizations to make informed decisions, and to run operations effectively based on available data. It has numerous application areas in critical domains including health, energy and infrastructure planning. Broadly speaking, BI is directly related to the areas of computer science, modeling, analysis and forecasting. As a research field, it encompasses data and knowledge management, management of digital media, modeling of processes and policies, data quality, data privacy and security, data integration, data exchange, data cleaning, inconsistency management, information retrieval, data mining, analytics, and decision support.

The goal of the workshop was to provide an in-depth review of the research underway in the NSERC Business Intelligence Network (BIN), a national network currently completing the second of a five year research program. BIN comprises 15 PIs at 7 universities (University of Alberta, University of British Columbia, Carleton University, Dalhousie University, University of Ottawa, University of Toronto and University of Waterloo). Investigators are working in partnership with researchers at a number of organizations including SAP, IBM, iAnywhere, Palomino and Zerofootprint. Over 55 graduate students and post-docs are currently involved in the network.

The industry partners in BIN, all of which have strong presence in the Canadian Information Technology industry, are currently developing expertise around these technologies and incorporating them into their products. In this setting, our network is involved in creating the next generation of software tools that will drive data analytics and management is all areas, and is therefore strongly related the Canadian government's research priorities.

2 Recent Developments and Open Problems

BI is an area rich in open problems, both from a conceptual point of view as well as from the various application areas in which BI is used. Currently, BI solutions have been proposed for addressing large-scale challenges in modern societies, in areas related to health, energy and the environment, urban planning, etc. One common theme in such applications is that they are too complex to be fully modeled in advance, and they produce an abundance of data which are automatically gathered in a multitude of independent systems. This is in part due to the commoditization of computer hardware, which has opened up the possibility of applying massive computing infrastructure to tackle such applications and workloads.

The NSERC BIN network was formed with the goal of developing knowledge as well as expertise within Canada in the area of BI. Broadly speaking, BIN is thematically organized into four main research areas: Strategy and Policy Management, Capitalizing on Document Assets, Adaptive Data Cleaning for BI, and Business-driven Data Integration.

The Strategy and Policy Management theme aims at providing higher-level modelling tools to BI, thus helping bridge the gap between the worlds of business and data. To accomplish this objective, BIN is developing novel concepts and tools for modeling social, organizational and intentional settings (e.g., virtual organizations, organizational structures and actors, strategic and tactical business objectives, government laws, policies and regulations, etc.) as well as linking such models to actual computer systems where the data resides.

The research under the Business-driven Data Integration theme aims at completing the link from the higher-level view of BI into the operational level of data management. Among other goals, this research aims at changing the current paradigm of warehousing all available data into a single repository before it can be processed, integrated, cleaned, understood, and only then used in BI models. Instead, the vision proposed by BIN is to derive and automate all the necessary steps to feed the data into a BI model from the higher-level descriptions of the business and their goals.

The work under Capitalizing on Document Assets aims primarily at enabling the use of information residing in documents to be used seamlessly with or without structured data (residing in databases) in support of decision making. More precisely, this work is aimed at automatically extracting, organizing, cleaning and integrating data expressed in textual form, and originating from both formal documents (e.g., regulations, legal and technical documents, patents, customer reviews, etc.) and informal documetns (e.g., customer comments and complaints, blogs, emails, news articles, etc.).

Also in the spirit of providing higher-level BI tools, the research under the Adaptive Data Cleaning area is building a general framework that includes specification languages for data quality and cleaning; it also includes the development of tools that, based on these languages, allow for the specifications of models of data quality that can be used to assess data, and also to specify and apply data cleaning solutions. Within such a model it will be possible to express, in terms of quality measures, desirable quality properties that can be checked and evaluated with the concrete data at hand.

3 Presentation Highlights

The meeting had two sets of presentations. The project presentations gave an overview of the state of ongoing research, introducing recent results and setting the tone for the network-wide discussion:

A1 One challenge for a network such as BIN is the need for relevant use cases in which real data and expertise is used to can guide the development and evaluation of this kind of research. Dr. Topaloglous presented his efforts and results towards bringing a real use case to BIN researchers, originating from running the Rouge Valley Health System.

A2 Prof. Yu presented, in a sense, a coherent and concise tutorial on the vision for the Next generation BI technologies using strategic business modelling, advanced adaptive software technology and enterprise architecture modelling to create BI-enabled Adaptive Enterprise Architecture. He also pointed out future research avenues, focusing on applying the solutions developed to date to one or more industries, building a business case and creating prototypes for concept demonstration.

A3 Prof. Tompa described ongoing work on compiling high-level (enterprise-level) policies into actionable database constraints, to be defined over the actual production database systems. His approach consists in (1) mapping business policies into constraints on database states and state transitions, (2) capturing policies as constraint diagrams, and (3) produce efficient routines for checking such constraints automatically.

A4 One issue within BI is that business processes must adapt to the content imposed by the specifics of the relevant data, the users, and the stakeholders. Prof. McIlraith discussed an AI approach to alleviate such problems based on (1) developing business process modelling formalisms that support specification of abstract business processes and (2) developing computational machinery to customize, verify, and optimize business processes and data with respect to stakeholder needs.

A5 The a need for dynamic location-aware methods for constructing data cubes that allow scalable, faster access to data was the theme of Prof. Viktor's address. She outlined ongoing work that relies on finding

the minimal set of attributes that correlate with a specific user and her locations of interest. Her goal is to have both the data cube creation and the subsequent data mining model construction location-aware. One challenge to accomplish this goal is to unambiguously identify data items for seemingly disparate sources, which is a problem researched within BIN as well.

A6 Prof. Aboulnaga presented his architecture for "pay-as-you-go" data integration, which enables "situational applications" to access structured data from diverse data sources on the Web at low cost. Furthermore, he outlined an incipient research project geared towards providing scalable support for the complex analytics workloads in a cloud computing environment, with emphasis on performance optimization.

A7 The next presentation described the work of Prof. Amyot and his team, on managing patient flows in health care organizations driven by BI principles. Their approach was to validate the expressive power of current tools by actually modelling the process of cardiovascular care and other patient flows with the tools being developed by the BIN team. One particular difficulty that was discussed was the need for more (real) data in order to effectively conduct this research.

A8 An architecture for data integration driven by (high-level) conceptual models was described by Prof. Kiringa. This work represents substantial steps towards achieving a core goal of the BIN research program: bridging the gap between conceptual and database models. Their approach resulted in the mappings being compiled, which leads to substantially better performance, and also led to clean semantics for MDX.

A9 Prof. Pottinger discussed system where users can coordinate data between two databases where changes in a base database B should be reflected in the contingent database C. Her system is based on the idea of data coordination, which she explained. Another project discussed concerned the design and study of a conceptual language and framework for top-down creation and population of warehouses.

A10 The next presentation, by Prof. Miller, outlined how to use business context and business requirements to guide the construction of database mappings. Two key novel aspects of the approach consists in augmenting basic mapping rules with causality knowledge derived from the business schema, and dealing with incompleteness by using preference rules that define constraints over the set of preferred business schema instances.

A11 The TARgeTEd Social Event Summarization (TARTESES) system was described next by Prof. Lakshmanan. The proposed system would be able to detect events/signals about a specific target within the realm of online social networks and report important events across all social media. Moreover, instead of simply filtering raw data of interest, the system should be able to report its results in a summarized form, and also provide a comparison of the summaries against crowd opinion from different communities.

A12 Prof. Milios described ongoing work that combines interaction and visualization with state-of-the-art machine learning methods and natural language processing tools to enable users to explore and analyze the information enclosed in large document collections. Among the challenges discussed are choosing the appropriate set of text features to be used in the text analysis. The presentation also included results of ongoing work.

A13 One research problem address by Prof. Ng and his team is that of how to perform next-level natural language summarization, abstraction (generate new text). Their focus is on informal conversational data (e.g., emails, blogs, reviews, meeting notes, etc.). The presentation covered recent results in using natural language processing tools to provide abstractive summaries of conversations, in which a new text (the summary) is synthesized from the summary. This technique has been shown superior to extractive summarization (in which fragments of the original text are used as the summary).

A14 Machine reading, which aims at extracting valuable data, information, knowledge automatically from text, is the central theme of the work presented by Prof. Barbosa. He described recent results from this team on focused information extraction from blogs concerning the identification of direct and

indirect relations among recognizable entities in the text corpus. Directions for future work include using relation extraction for question answering and narrative-based summarization of the information extracted.

A15 In the work of Prof. Bertossi and his team, the vision for a general theory of context for data cleaning, in particular for applications in data management, is that (1) a logical theory T is the one that has to be "put in context"; (2) the context is another logical theory, C; and (3) the connection between T and C is established through connection (possibly shared) predicates and mappings. The presentation also discerned about different notions of context and how they are used in practice.

A16 Prof. Özsu presented results and ongoing work of two BIN-related projects. The first, led by Prof. Ilyas uses uncertain data management techniques (possible worlds, consistent query answering) to address the data cleaning problem (entity de-duplication, and removing integrity constraint violations). The other project, led by the presenter, aims at addressing the issue of how to perform multi-query processing over uncertain data streams.

The second round of presentations were led by the industrial partners in BIN, with the goal of bringing perspectives and practical challenges of actual customers and providers of BI solutions deployed today. These discussion were organized by themes:

I1 **Social computing**: as we move to empower the individuals we need to interact with them within their social context. Most decisions in business are made collectively, this ensures that all alternatives are explored, that experience is leveraged and that all parties are bought into the decision. Challenges in this area include: involving others in our decision processes by engaging them with insight around information, being able to draw insight from the Opinion being expressed internally and externally to the organization, and everaging the wisdom of the crowd to the improve quality of information.

I2 **Cloud computing applied to BI**, with the goal of enabling Analytics as a Service. The need for such solutions stem from the fact that cloud deployment is increasingly relevant. The challenges in enabling this infrastructure for BI are: (1) providing cloud-optimized solutions supporting multi-tenancy, resource sharing, scaling in/out and up/down, sharding, workload optimization, license/cost optimization, business and regulatory constraints; (2) understanding whether the cloud infrastructure itself introduces new analytical workflows or makes existing ones no longer intractable; and exploiting new data/access models enabled by this infrastructure.

I3 **Compliance**: as BI providers empower users, they also have to consider the risks represented by distributed decision making. With hundreds of regulatory bodies placing constraints on business, often multiple overlapping any decisions, it is essential that we empower our users to be aware of their risks and to understand if, when and why they are out of compliance. Key observations on the issue are: (1) compliance remains the first priority from the IT management perspective, even as the bodies of regulations continue to expand. Reducing the cost of managing compliance is essential to IT success; (2) Business Users need to be notified when they are in or out of compliance, and how they can mitigate their risks; (3) organizations as a whole need to better understand and more accurately measure risks so they can make informed decisions.

I4 **Consumable Analytics**, which is the view that BI should focus is on the business end user. Key observations in this regard from current experience are that new analytics demand new user experiences, and visualizations, and that novel interactive visualizations and discovery tools will lead to a new set of query-generation gestures. Moreover, the goal of this effort is to provide consumable for all (i.e., attention should be paid to accessibility, globalization, etc.) and everywhere (i.e., not just the office).

I5 **Geospatial and Temporal Analytics**: as the number and diversity of data sources increase, we need finer tools to glean intelligence from such data. This includes ways to address: context, disambiguation, time and space.

I6 **Mobility**: decisions are being made everyday in the organization, and providing users with information to empower them, that information must be present when those decisions are being made. Since the

decision makers all have mobile technology, and rely on it as a portal to resources, we must leverage these devices to deliver rich, effective and timely information. Enabling mobility in BI solutions will require solving pressing issues w.r.t. geospatial and temporal analytics.

I7 **Big Data**: 90% of the data in the world today has been created in the last two years. Everyday, we create 2.5 quintillion bytes of data. Such "Big Data" data sets are too large (and often too unstructured), so much so that they defy conventional analytic techniques and/or can not produce outcomes in tolerable time frames. Effective BI will not be achieved unless we solve this problem.

I8 **User-driven integration and mapping of data**: given the complexity of the BI applications, eliminating the human from the loop will not be feasible for many years, if not decades. Instead, we should focus on making the BI process (particularly the data integration and mapping) more accessible to knowledge workers. Challenges towards this goal include: (1) making the user experience more intuitive; (2) supporting multiple heterogeneous data sources; (3) performance.

I9 **Unstructured data**: a vast amount of information and corporate memory is hidden inside documents on file shares, email servers, web sites, etc. We must convert this to a usable form and make it consumable. Discover knowledge, not find documents. Challenges include: (1) solutions must work on Web-scale, with big data; (2) solutions must work in real-time, on live streams, enabling interactive exploration; (3) solutions must turn knowledge workers into consumers: make sense, make connections, reason, discover, visualize; (4) there must be notions of quality and reliability of the facts extracted: measure, feedback, improve.

I10 **Real time and Streaming Business Analytics**. BI-relevant data not only comes in large quantities, they are also time-sensitive. Effective BI solutions must be capable of more than just event handling, they also need complex aggregate and rule processing on the fly. Further, the relationship between between data volatility and query performance must be understood. Some implications for the modern database architectures include: data partitioning; incremental updates for aggregates; and concurrent writes & consistent reads.

4 Scientific Progress Made

This meeting had two main goals: (1) congregating a large and representative fraction of the network for technical discussions, and (2) enabling the network to identify further collaboration opportunities. In both regards the meeting was an astounding success. The audience represented a substantial subset of the network, and ranged from graduate students to members of the board of directors.

Towards a better understanding of "Core BI". The meeting was the first network-wide opportunity for a deeper discussion of foundational aspects and state-of-the-art in core themes within the network research program. Such discussion was particularly useful in the context of the Strategy and Policy Management Theme, which is concerned with developing usable models and tools for business intelligence end users. Thus, in many ways the success of the network hinges on these models as well as on successfully integrating all other pieces of the research with them. Prof. Yu's presentation started this discussion and set the tone for the subsequent presentations and discussions, resulting in a very productive debate. Dr. Topaloglou's remarks, derived from real-life experiences in deploying a BI solution, were also instrumental in shaping the debate and focusing the research team.

On a similar tone, the meeting allowed the network to gather a broader perspective on the state of state of business intelligence research from other research groups, in the keynote presentation by Dr. Dayal from HP Labs. In particular, the keynote offered a glimpse of new tools and techniques, as well as several success stories of applying these tools to real-life problems in various industries.

BI curriculum development. This deeper understanding of BI enabled further a discussion within BIN into developing reference material on BI, to be used as the basis of introductory texts at the undergraduate level for the field. The need for such a text has been originally articulated as a longstanding and major need from the industrial partners.

Strengthening collaboration. During the meeting several research opportunities were identified, covering all research themes and groups. Novel directions for collaborative research identified during the meeting include: (1) sharing data/expertise from the health care use case (P1, P3, P10, I3); (2) combining efforts in data integration (P6, P9, I8); integrating information extraction techniques (P12, P13, P14, I4, I9); exploring social computing (P11, I1); linking high-level business concepts with operational data management tools (P3, P7, P8, I8); addressing system scalability issues (P6, I2, I7, I10); incorporating mobile and spatio-temporal data into BI solutions (P5, I1, I5, I6).

5 Acknowledgements

The entire BIN team, and particularly the meeting organizers, would like to acknowledge all the support provided by the Banff International Research Station and its staff, who provided BIN with an ideal setting in which to have a focused and productive meeting. Thank you.