Big Data Everywhere, and No SQL in Sight

Editorial: from SIGKDD Chairman

Usama M. Fayyad

Chairman, SIGKDD Executive Committee
Executive Chairman, Oasis500 – www.oasis500.com
Chairman & CTO, Blue Kangaroo (ChoozOn Corp)
Usama_fayyad@yahoo.com

ABSTRACT

As founding editor-in-chief of this publication, I used to systematically address our readers via regular editorials. In discussions with the current Editor-in-Chief, Bart Goethals, we decided that it may be a good idea to restore the regular editorial section from me as SIGKDD Chairman/director. So this is a first in what hopefully will be a regularly occurring event. This Editorial contribution discusses the issues facing the fields of data mining and knowledge discovery, including challenges, threats and opportunities. In this Editorial we focus on the emerging area of BigData and the concerns that we are not contributing at the level we should to this important field. We believe we have a unique opportunity with the growth of BigData to play a big role in defining the future and providing the much needed talent for this key area in the organizations of today and the future.

KEYWORDS

BigData, data mining technology, challenges in KDD, threats and opportunities in KDD, Editorial

1. BACKGROUND

As I write these lines, I am observing the end of the year 2012 and anticipating the new year of 2013. As I reflect on the past year in terms of what is happening in our small world of Data Mining and Knowledge Discovery in Databases (KDD), I see many positive developments in our field. The KDD conferences continue to attract record numbers of attendees – 1200 attendees where at the KDD-2012 conference in Beijing. This is an all-time high for the KDD conference and is happening despite the fact that a plethora of major conferences focused on data mining and predictive analytics are mushrooming in number. So this additional supply has not decreased the demand for KDD conferences: in fact the evidence is that demand has increased. There is also unprecedented demand for *Data Scientists* as this new position has risen to the highest in demand by tech companies, in financial services, in telco and in insurance companies. The biggest supply of Data Scientists to fulfill this demand is actually, in my opinion, coming from qualified individuals in our community as these are typically the most data-oriented analysts who are comfortable using computational techniques intensively. This is a huge opportunity for our community to play a great role in further educating and preparing our membership to better benefit from the opportunity that is developing.

My belief is that this increased demand is driven by the growth of BigData as major area of importance in business, government, and even in academia. The term BigData is not well-defined, but is generally used to refer to the type of data that breaks the limits of traditional data storage and management state-of-the-art. Usually these limits are tested on 3 major fronts: the 3-V's: Volume, Velocity, and Variety. I like to talk about the 4-V's of data mining, where the 4th V stands for Value: delivering insights and business value from the data. More on this topic in a future Editorial. For now, I would like to focus on some of the opportunities and threats arising from this developing situation.

2. BREAKDOWN IN THE RELATIONAL DATABASE WORLD

While most relational database researchers and practitioners are still living in denial about what is going on, what I am witnessing is nothing short of a major breakdown. Almost every start-up company that has BigData needs, and this pretty much includes all of the ones I work with or I am familiar with, are going through a transformation in their data infrastructure towards using NoSQL databases. So while we used to see Oracle, Microsoft, MySQL, Postgress, and other RDBMS's, I am now witnessing a veritable exodus from these relational systems towards systems like CASSANDRA, MongoDB, Hadoop FS, and other systems that depart from the traditional relational DB framework. While I believe the role of RDBMS is essential in the world of transaction processing, I strongly believe that we are seeing a decisive, and potentially very dangerous, move away from relational databases as a framework for storing data to drive on-line systems (Operational Data Stores - ODS), to provide real-time information, and to serve business intelligence, targeting, reporting, and tracking needs of companies.

I work closely with many startups either as a board member, advisor, or investor. Without exception, I have seen every one of them make the move away from RDBMS as the ODS towards some NoSQL type of storage I have also seen several moving towards specialized data engines like Data Management Platforms -- DMP's (e.g. nPario, Epsilon, [x+1], etc…) All of these companies are moving their data assets away from traditional RDBMS framework and in the process giving up the convenient and well-understood semantics of the relational world. Of course, the reason for this is that they are looking for an escape from the rigidity of the relational systems, and more importantly seeking higher real-time retrieval and update performance.

It is also my observation that the RDBMS world is strangely unaware of this dramatic movement. The startups, especially the ones with serious market traction, are trend setters and should be creating a serious warning sign to the relational database community.

3. A MAJOR MESS IS COMING

By moving to more flexible data platforms, many of the companies moking the move to NoSQL data systems are taking on a major liability down the road. The very flexibility of these systems contain the very seeds of problematic issues and management down the road. This mess comes from the fact that systems that allow programmers and administrators to create new properties and values "on the fly" are ultimately creating the kind of systems that are nearly impossible to document.

After several weeks or several months of convenient development, the companies using such new frameworks find that down the road the systems they produce are not sufficiently well-documented and will inevitably have a major reverse-engineering job to capture the documentation and semantics of their data stores.

3.1 A DEARTH OF ANALYTICS TOOLS

Despite the risk in data management that companies are undertaking to achieve performance gains and flexibility if data storage schema, there are few rewards to be had on the predictive analytics side. Tools from R, to SAS, to much of the work on open source analytical tools, to the work done by academia is unfortunately not easy to use on these new generation data stores. We as a community have not invested in the new platforms. This means that to use the analytical tools coming out of the analytics communities, these companies have to remap their data to flat files or to traditional database systems. This effectively means that we are not going to see wide adoption of the tools and techniques we work on

3.2 NO MANAGEMENT FRAMEWORKS FOR ANALYTICAL MODELS

Using a predictive analytics tool typically requires a significant amount of data manipulation, entity extraction, and heavy processing to render the data into a format consumable by our algortihms. In a RDMS, much of this logic can be captured effectively in stored queries and in data views that persist over time. As attributes change, these views and queries can be updated to reflect these changes. In the brave new world where new attributes can be created on the fly a serious challenge is faced. Furthermore, the tools used for extracting structure from the unstructured or structured data will also need to be managed. This means that if an analysis is to be readone in a few weeks or a few months, or if new staff comes on board or old staff leaves the company, much of the knowledge required to figure out what is needed and how to do it is no longer available and is extremely difficult to reverse engineer.

4. OPPORTUNITY AHEAD

The current broken situation and the looming mess both represent a unique chance to the field of KDD. We can develop the appropriate systems for managing data models, for tracking input semantics, for performing repeated analysis operations effectively and reliably, and for running such computationally expensive operations efficiently. I have personally dealt with situations where years down the road one is trying to reverse engineer the semantics/contents of a key-value pair. That job is daunting and is often never done.

The other opportunity lies in figuring out how to embed the analytical algorithms so they operate over these new types of NoSQL stores. Much like Mahout represents a very good approach to bring analytics algortithms to the Hadoop grid, we need similar kind of implementation for the new popular data stores. If we do not go to where the data are, we stand to lose relevance and effectiveness. Once we lose that, we have effectively given up on our most important "customers" in our field.

The other major opportunity is to provide solutions for pragmatic problems in model building: solutions for the problem of when a model must be updated, the problem of managing models over time, the problem of aging models and understanding when certain events cause models to cease to function properly. E.g. a model for predicting the likelihood of interest in purchasing a product, e.g. a digital camera, would increase that likelihood with recent activity predictive of the event, until the purchase happens, when suddenly counter to all the statistical predictions, interest drops dramatically just when predictions become very significant statistically speaking.

About the author:

Usama Fayyad is Chairman & CTO (ChoozOn Corp/Blue Kangaroo): mobile vertical search engine for offers via personalization/context. In 2010 he was appointed by King Abdullah II of Jordan to lead Oasis500 as Executive Chairman: top tech startup investment fund/accelerator (500 startups in 6 years) in MENA.

2008 founded Open Insights LLC: data strategy/technology firm to help enterprises develop data strategy & BigData solutions to effectively grow revenues.

2004-2008: Yahoo!'s chief data officer & Executive VP of Yahoo!'s global BigData systems/policies & data scientist group using Big Data for content/ad targeting: growing Yahoo! revenues from targeting by 20x in 4 years while processing 25+ Terabytes of data/day. Founder of Yahoo! Research Labs: the premier scientific research organization to develop the new sciences of the Internet.

2003: co-founded/led DMX Group, a data mining/data strategy company -- acquired by Yahoo! in 2004. In early 2000: co-founder/CEO of Audience Science (digiMine, Inc.) the leader in Behavioral Targeting & ad networks.

1995-2000: led Data Mining & Exploration group at Microsoft Research, built data mining products for Microsoft's server division. From 1989-1996: held a leadership role at NASA's JPL in analysis of Big Data in Science earning him the top research excellence award from Caltech, as well as a U.S. Government medal from NASA.

Fayyad's Ph.D. in engineering is from the University of Michigan, Ann Arbor (1991). He holds BSE's in both EE & CSE (1984); MSE in CSE (1986); and M.Sc. in mathematics (1989). He published over 100 technical articles, holds over 30 patents.

Fellow of AAAI (Association for Advancement of Artificial Intelligence), Fellow of ACM (Association of Computing Machinery), editor two influential books on data mining; Founding editor-in-chief of primary scientific journal in field (Data Mining and Knowledge Discovery) and of SIGKDD Explorations Newsletter, Chairman of ACM SIGKDD which runs the world's premiere data science, big data, and data mining conferences: KDD. He is an active angel investor in U.S. and Middle East specializing in early-stage tech companies.