Managing Big Multidimensional Data

Torben Bach Pedersen

Aalborg University
Department of Computer Science
Center for Data-intensive Systems (Daisy)
tbp@cs.aau.dk
http://people.cs.aau.dk/~tbp

Business Intelligence Versus Big Data

Multidimensional data has traditionally primarily been used for Business Intelligence (BI) applications, so it is interesting to check the similarities and differences between BI and Big Data. We now define the two terms.

Interestingly, the term Business Intelligence was coined in 1958, more than 56 years ago. Here, H. P. Luhn defined BI as “the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal” (Luhn, 1958) More recently, in 2013, Gartner Group defined BI to be “an umbrella term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to improve and optimize decisions and performance” (Gartner Group, 2013). Thus, we can conclude that optimizing your business using data is not a new idea.

The term Big Data is of course more recent, although the first uses of the term seems to date back to the 1990’s. Wikipedia defined Big Data to be “an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process them using traditional data processing applications” (Wikipedia, 2014). So, the data should be so “big” that it becomes “difficult” to do it the traditional way.

When defining what is means to be “big”, the 3 V’s of Volume (very large datasets), Velocity (data arriving very rapidly), and Variety (data of very different format and structure) are well-known. However, up to 6 extra V’s are also mentioned, namely Veracity (how much can we trust data ?), Visibility (data must be visible to the Big Data processes), Variability (the meaning of data changes over time/place/context), Viability (can our data be used for anything useful ?), Value (what real value can this data add to our business ?), and Visualization (complex visualization is needed to fully understand and get value).

Now, we can finally compare BI and Big Data. The following issues are *similar* for the two concepts, and thus not novel for Big Data :

— Collecting, integrating, and analyzing data to gain knowledge
— Large data volumes
— Data (often) arrives at a fast pace

Certainly, the first item is what BI was always all about. Traditional data warehouses have also held very large, and increasing, amount of data, and the concept of real-time data streams has been known for more than 15 years. However, there are also quite a few interesting *differences* between the two, summarized in Table 1. The change is perhaps best illustrated by the picture that in BI were are "sipping data on a straw", whereas with Big Data, data is coming right at us from a garden/fire hose.

	BI	**Big Data**
Data types	Structured (mostly)	Unstructured (also)
Data sources	Mostly internal	Mostly external
History	Essential	(Often) less relevant
Users	Manager/controller	Data scientist
Precision	Exact results	Approximate results
Privacy	Not critical	Critical
Control over data	Almost full control	Little or no control

TAB. 1 – *Differences between BI and Big Data*

Whereas BI data are mostly structured data from relational databases, Big Data contains new data types such as text, multimedia, social network graphs, etc. The BI data sources were almost entirely internal to the organization, e.g., ERP or CRM systems, whereas Big Data such as social network updates originate entirely outside the organization and the external data is often the most important. In BI, keeping the full history of, e.g., a customer is essential, whereas with Big Data, it is mostly not possible, or even relevant to keep the history, as the meaning of many concepts, e.g., hashtags, evolve quickly over time. The main users of BI systems are classical company profiles likes controllers, managers, etc., while Big Data is mostly used by the brand new job type of *data scientists*, which mix deep statistical, computer science, machine learning, and domain skills. In BI systems, computing the exact result, e.g., an amount, down to 4 decimal points is the norm, whereas approximate results may be the only meaningful option for Big Data since the available data is only a sample of the full reality anyway. In the company-internal world of BI, privacy us not a big issue, whereas it becomes critical in the world of Big Data with people often revealing more about themselves on the social media than they want to. Finally, the perhaps most profound difference is the level of control : in the traditional BI world, the organization has more or less full control over the data, whereas for the (external) Big Data, the organization has little or no control.

Références

Gartner Group (2013). Gartner IT Glossary - Business Intelligence (BI). http://www.gartner.com/it-glossary/business-intelligence-bi/.

Luhn, H. P. (1958). A Business Intelligence System. *IBM Journal of Research and Development 2*(4), 314 – 319.

Wikipedia (2014). Big Data. http://en.wikipedia.org/wiki/Big_data.

T.B. Pedersen

Summary

Multidimensional database concepts such as cubes, dimensions with hierarchies, and measures have been a cornerstone of analytical business intelligence tools for decades. However, the standard data models and system implementations (OLAP) for multidimensional databases cannot handle “Big Multidimensional Data”, very large amounts of complex and highly dynamic multidimensional data that occur in a number of emerging domains such as energy, transport, logistics, as well as science. This talk will discuss similarities and differences between traditional Business Intelligence (BI) and Big Data, present examples of Big Multidimensional data with the characteristics of large *volume*, high *velocity* (fast data), and/or high *variety* (complex data) and discuss how to manage Big Multidimensional Data, including modeling, algorithmic, implementation, as well as practical, issues.