The Three V's of Big Data: Failure to Manage Them is not an Option

The term 'Big Data' was originally conceived by Gartner analyst Doug Laney 2001 in terms of the challenges of managing data growth through three dimensions: Volume, Variety, and Velocity. Based on these three dimensions it is possible to derive a different set of ideas about what big data actually is.

Discussion of 'big data' – Volume - would have us think of a giant database, holding yottabytes of data (10 to the 80 binary digits). To get a true feeling for size, we should start with the extremes; Eric Schmidt has estimated that the size of the internet at 5 million terabytes (5 Billion Gigabytes) of which Google itself has only captured a small percentage. In relative terms, the world's largest commercial data bases of Amazon, YouTube, ChoicePoint, Google, and AT&T, are estimated to be in the 100 petabytes range. Government agencies such as TSA also own vast data sets, other mega database institutions include the United States Customs Service, the IRS, The World Data Center for Climate, The National Energy Research Scientific Computing Center and the CIA.

The first question executives and managers should now want to ask is why would we want to create database as big as these organizations? Additional questions include: if we had such a database what would we actually do with it? How can I determine the return on this investment? And can we manage such a large system ourselves? In addition, and worryingly, the CEO and board will most likely ask: 'will we be building another legacy system that we will live to regret'? The first step a thoughtful BD advocate will need to take is to develop carefully considered answers to these questions.

Discussion of 'big data' – Variety – links the size of the data with the heterogeneous data types it contains, cleverly structured so that it can be queried and mined, to potentially dispense wisdom and intelligence on the fly. This of course assumes that the answers are capable of being extracted from the data in the first place, that the variety of data is rich enough, that the history of each part of the puzzle is long enough, and that analytics can bring this together in a useful manner.

This time, executives need to ask themselves: What types of data do we actually need? How do we capture it? And how do we maintain that data set as it changes? In fact how will we know what data we need to store in order to answer yet unknown questions, five years from now?

The third of Laney's dimensions is 'Velocity,' which in reality is the vector of the speed at which data can be both transmitted and at which it can be processed. Transmission speeds vary. In terms of where the data is going from and to, as well as the limitations imposed by the physical media. Rather more importantly for many companies is data 'latency', the lag of receiving the data from the source. This is important in domains such as high frequency trading. For example, the speed of 100,000,000 'buy-sell' trades London to NY is 0.04 seconds round trip, 0.07 seconds London to Singapore, and 0.1 seconds Singapore to NY. With technology even at the speed of light the Big Data world is not flat and latency is a major factor on competitiveness. The second component of velocity is processing speed. Establishing a large data base is predicated on the ability to process the data through a powerful enough hardware architecture or 'appliance[1]'. Large clusters can have multiple systems with 216 CPU cores, 864GB main memory, 648TB raw storage, and 40+GB transmission connectivity.

For this dimension executives and managers need to ask: How much is this speed going to cost? Is this going to be yet another recurring long term expense? How long is it going to take us to process this data on our servers, move it to where it is needed, and then execute an action based upon the analysis? From a cost perspective, major 'big data' software vendors are now requiring that their clients buy both hardware and software licenses, as well as the

[1] The Oracle Big Data Appliance available from February comes in a full rack configuration of 18 Oracle Sun servers with 216 CPU cores, 864GB of main memory, 648 TB of raw disk storage, and 40GB/sec of InfinBand connectivity for connecting multiple racks together. Oracle did not release pricing information for the product.

hardware itself and the maintenance contracts that go along with them, CFO's and their CIO colleagues need to consider this issue carefully before acquiring another overhead.

By undertaking a thorough analysis of the three V's in this way executives and managers will develop a more complete and integrated consideration of the data, its storage, processing, and its subsequently use. And if all goes well this will lead to agile cost effective analysis solutions, the alternative is to be left behind in the competitive data 'arms race'. Failure, as they say at NASA is not a realistic option.