BIG DATA ANALYTICS ENTERS A WORLD OF OPEN SOURCE POSSIBILITIES

Connectivity, big data, and the bigger challenge

The concept of a network of smart devices emerged as early as the 1970s. Around 1972 - Prancing Pony - a computer-controlled vending machine selling snack foods on credit at the Stanford Artificial Intelligence Laboratory, became one of the first Internet-connected appliances. There began the saga of pervasive connectivity – where every device is plugged into everything else – creating the defining trend of 2010 to 2020. In fact, the Internet of Things is anticipated to burgeon to about of 26 billion units excluding PCs, smartphones, and tablets by 2020 – and perhaps several categories of these items, that will be connected in 2020, don't even exist at present.

The Internet of Things will cause connectivity to explode, and it will also create value – as much as US$6.2 trillion in annual revenues by 2025, says a global consulting company. But it will also create massive amounts of data – 40 zettabytes by 2020, according to one estimate. And as we all know, the bulk – over 80% – of big data is unstructured, and in motion, existing in a variety of forms and formats both inside and outside company walls.

Gathering this data is a huge challenge, but one that technology today is capable of. It's what comes next – extracting accurate insights in real time and creating foresight from it – that enterprises are yet to nail.

To share is to learn

Several industries, such as financial services, telecom, retail, and insurance, are among the leaders in collating, processing, and analyzing big data into reliable findings. Even more importantly, they have the ability to arrive at these insights in very quick, if not real time. In telecom, big data analytics has helped providers mitigate the high rate of churn by predicting which customers are most likely to leave, enabling operators to target promotional offers more accurately, and even scouring social media conversations to spot telltale signs of defection. On the other hand, insurance companies have managed to speed up claims processing, improve risk management, and price products based on predicted behavior (think auto insurance premiums based on driving patterns), and accelerate report generation using analytics. Then there are retailers, who have learned to exploit the vast customer data at their disposal to identify customer behavior, seasonal trends, replenishment cycles, merchandising requirements, and so forth. Financial services firms, on the other hand, leverage data to quantify risk and provide transparency to regulators – which in turn is a great driver of operational efficiency.

Note how differently each of these industries uses big data. It clearly signals the huge potential for sharing, and cross-pollinating learning between industries, even among those who are analytically progressed.

What's your problem?

One of the biggest lessons in big data analytics is that it is what an enterprise 'does' with its data and analytics software that counts. Defining – sometimes even discovering – the problem is the most important part of the insight generation process. Retailing's success with analytics owes much to the nested question, a series of questions that, with each succeeding question, closes in on the problem. Unfortunately, in their impatience for quick resolution, most enterprises cut straight through to finding the answer to a problem they haven't identified in the first place. For them, the outcome in a best-case scenario is symptomatic relief.

This is exactly what new-age "problem finding" concepts like design thinking seek to address. The overarching goal of design thinking is to get to the root of a known problem or identify one that hasn't been recognized – staying as close to business reality as possible. It does this in a succinct, three-step process of establishing (end user) desirability, (technical) feasibility, and (business) viability.

Establishing desirability is all about understanding user need, and what the end user is trying to accomplish. A good indicator of desirability is the extent of empathy one has for the end user – the more empathetic the creator of the solution is, the more desirable the solution.

Feasibility is essentially a matter of mapping problem-resolution to technical capability. The enterprise knows what problem to solve and how to solve it in theory, but must figure out if there's a technology that will do it in practice.

Viability determines whether a problem that is both desirable and feasible to solve, is economically attractive. Here, business metrics, such as measurable business value, cost versus benefit, payback period, and return on investment, come into play.

> Design Thinking gives enterprises a mechanism to define the "What". Now remains the challenge of solving the "How".

A sea of data and a data lake

Proprietary statistical tools have proved to be of limited utility in crunching massive data of the order of millions of records into insights – and foresight thereon. They're sluggish, cost millions of dollars in capital expenditure, and worst of all, are not very amenable to change or expansion of scope. But now, open source technology has given us a very promising alternative. At its foundation is the notion of a data lake – "…a storage repository that holds a vast amount of raw data in its native format until it is needed." It is this absence of rigidity – on data structure, format, and also end purpose – that differentiates the data lake from any method of storage the world has ever known, and also enables it to overcome all the major limitations of proprietary statistical tools of analysis.

Architecturally, the data lake comprises the Hadoop File System (HDFS) that pools in the data from every source. Because it is so accommodating on structure, the data lake is not constrained to support only a predetermined type of analytical problem solving; indeed, it can take on new analytical use cases endlessly, at virtually no additional cost. Unlike data brought into warehouses and marts, the "open" data in a lake needs no integration effort; using MapReduce and other algorithms, enterprises can quickly be on their way.

Above all, the data lake stores information in a highly granular "microdata" form, unlike licensed off-the-shelf solutions, which aggregate or pre-compute data to expedite analysis but end up compromising fidelity.

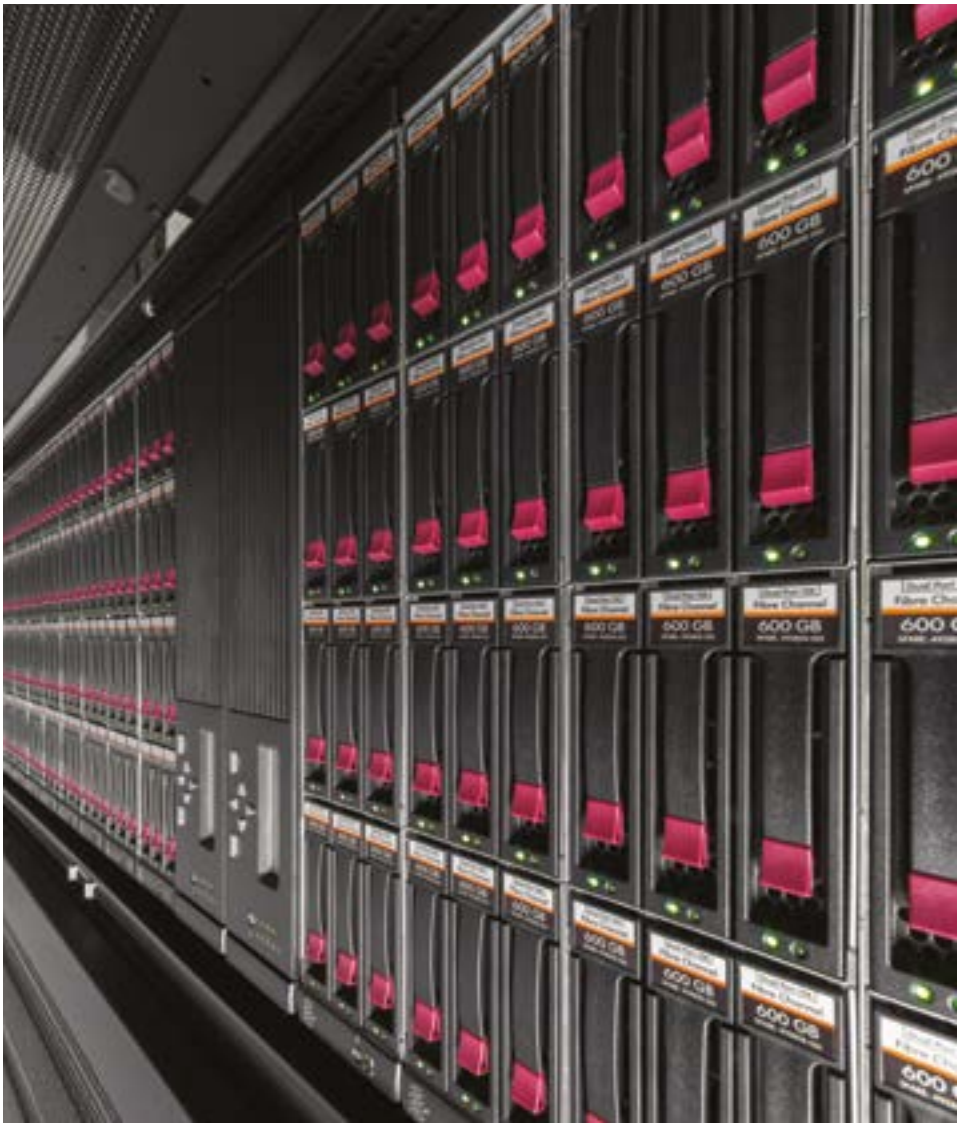
In contrast, the data lake has an almost infinite capacity to store data at the finest level, at the "power of one" so to speak, and refine, and add information at will. This data is fed into open source software, which can run through any number of data layers, and indeed any amount of data, in a very short time. The analysis arrives in real-time, is accurate, and keeps improving as the datasets become larger.

When they want to solve a particular problem, enterprises need only pull the required data from the data lake on to a data foundation. This data – which should ideally be of high-quality and granularity to deliver accurate results – is now stored on commodity hardware, such as Amazon Web Servers, Azure, or custom-built commodity servers.

The analytics or data science layer sits atop the data foundation. Using machine learning, data scientists run various mathematical models of statistical analysis, and make that data science available as packaged, open-source software. Finally, the analytics results are presented in business-consumable form by visualization software like Tableau, or open source components like D3.

Opening up the possibilities

Open source technology has revolutionized data and analytics at every step of the value chain, from data storage to analysis, to visualization. Viewed from a design-thinking perspective, open source makes every aspect desirable, feasible, and viable: enabling sharp insights into problem discovery and solution desirability; making it technically feasible to deliver accurate real-time analysis no matter how big the data; reducing cost of data



storage and processing dramatically to make every project affordable and viable.

As open source throws open immense possibilities, its biggest challenge will be to assure security, access control, and governance of the data lake. There is also the risk that a data lake that is not managed thoughtfully could end up as an aggregate of data silos in one place. Industry watchers caution about the need to train lay users in appreciating key nuances – contextual bias in data capture, incomplete nature of datasets, ways to merge and reconcile different data sources, and so on – which is a Herculean task in every way.

While potential users are quite concerned about these issues, overall they are very excited about the opportunity. Meanwhile, the technology industry is trying to accelerate adoption by making all the open source capabilities discussed here available in a pre-tooled, enterprise-ready, "out of the box" format. A global, office automation firm has deployed such a solution to crunch the time it takes to process two million records to a few seconds, from several dozen minutes earlier. It is now able to make business predictions of 80 percent accuracy. And all of this has come at an investment that is a fraction of the cost of a proprietary statistical analysis tool. Open source technology has enabled it to simply do more with less for more.

Author

Abdul Razack

In a career that spans over two decades, Abdul has been involved in several engineering and consulting roles at Commerce One, Sybase, KPMG Peat Marwick, and SAP. Abdul holds a Master's Degree in Electrical Engineering from Southern Illinois University, and a Bachelor's Degree in Electronics and Communication Engineering from the University of Mysore, India.

If you wish to share your thoughts on this article or seek more information, write to us at Insights@infosys.com