FROM GUT FEELING TO FACTS

BIG DATA CAN YOU EAT THAT?

By Marianne faro, Principal Consultant and competence lead Analytics team at Itility

Big data is trendy. Research has shown that practically all enterprises have Big Data initiatives on their agenda. But although Big Data is happening and everyone knows you should do something with it, not everyone knows what you should do with it as yet. So sometimes you might hear the phrase 'we're doing Big Data'. As if doing data (is that even possible?) is a goal in itself. It's like a restaurant saying 'we're doing food'.

TRANSLATED FROM ORIGINAL DUTCH ARTICLE:
ITILITY.NL/WHITE-PAPERS/BIG-DATA-CAN-YOU-EAT-THAT/

The concept 'Big Data' refers to a collection of data so big that it's difficult to process within traditional database systems. Big Data focuses on storing and analyzing many different types of data source – both real-time and near real-time – in huge quantities. It involves using the famous 3 Vs – Variety, Velocity and Volume to achieve the most important V: Value. You get Value if you use the insights from your data analysis to deliver added value to the client. Big Data therefore involves far more than tooling alone. It's mainly about using the right skills for translating the data into Value. It seems that balanced collaboration between a data scientist, a data engineer, an IT tooling expert and a domain expert is essential. If you don't set this up deliberately, then you may well end up with a nice tool and fancy pictures, but no new customer value.

Teamwork

The IT tooling side can provide the right Big Data platform – an agile platform that can store lots of data, is scalable, easy to adjust and cost-efficient, and which you can use to search and analyze data at the speed required. On the data engineering side, you ensure that different data sources can be connected easily (through api's, scripts or data links) and that the right queries deliver the right processed data sets to the right people. In addition, the right semantics are developed for making data analysis widely accepted in the organization. On the data scientist side, you use mathematics, statistics, algorithms and visualization techniques to correlate data and make predictions, but also to generate hypotheses that release the usable intelligence. And finally, the domain expert is the "value creator", who filters the hypotheses and decides whether or not they should be translated into new business rules. He knows the business better than anyone.

Pragmatists think they can combine all these roles, although it has proven virtually impossible to do so. Reinforcing this division of roles has proved successful mainly from the perspective of time. In the end, the relationship between data scientist and domain expert is the most important – this is where the real competitive edge lies.

Restaurant

In the past, data analysis was like the vending machine next to the coffee machine, with an array of long-life products to take the edge off your appetite in the afternoon. Input and output is structured: box A1 has a bag of potato chips, and box D3 has a Mars for the same amount of money. But one day a strategic initiative

comes along: to focus on Big Data. In other words: let's start a restaurant. But where should we start? All that was required for the vending machine was a service engineer (who refilled the boxes every month and reprogrammed box D3 once a year, replacing the Mars with a slightly more expensive Snelle Jelle). But a restaurant requires a team with different kinds of skills. As well as needing our own kitchen staff, we have to make smart choices in the buying process, be hospitable and deliver our services on time. And we need someone who can transform the ingredients into a really good dish: the data scientist.

Browsing

Let's go back to Big Data. 'We're doing it' because we think it's valuable and can deliver added value. You want to browse through big amounts of data to gain new insights and create new business rules, through using statistic methods, correlations, visualizations and machine learning. This requires teamwork – both inside your restaurant (chef, sommelier and waiter, all working as a team with different roles) and even more in interacting with your guests. Successful Big Data programs involve their guests right from the start in searching for the most delicious recipe; because Big Data involves searching and finding, searching and selecting, and searching and fine-tuning. It's an ongoing process between the chef, the sommelier and the connoisseur/guest. Or rather between the data scientist (who's capable of searching smartly and quickly in large amounts of data) and the domain expert (who's familiar with recent products and services and can define hypotheses).

Hypotheses

Searching large amounts of uncorrelated data for insights, you don't know exactly what insight you're looking for just yet. You're not quite sure which questions to ask to get added value from the data, and the business rules don't arise directly from the visualizations and queries. Forming a hypothesis (a thesis that can be answered with true or false) can be very helpful. Hypotheses help with thinking out-of-the-box, as you don't have to sketch out the graphics or dashboard you want to see. You just have to put forward propositions based on existing domain knowledge. So generate as many hypotheses as possible, whether they can be validated or not. The next step is to let everyone vote: is this hypothesis true or false? This allows you to create ownership with regard to business rule definitions – everyone wants to think about the hypotheses and everyone wants to know the answers to the hypothesis.

Next, the data scientist delves into the data to find the answer to the hypothesis. He can't do so alone, however. He has to "pick the brains" of the domain expert, as all the data requires interpretation, domain knowledge and an understanding of the underlying preconditions. And with this under your belt, you progress from gut feeling ('this dish is good') to facts ('in 90 percent of tastings, a combination of 15 percent sour with 25 percent salty leads to empty plates').

These facts are important. They can help estimate the fifth 'V': the Viability of the data. Most data scientists assume that only 5 percent of relevant variables are responsible for 95 percent of the value that can be derived from data. Hypothesis validations ensure that this 5 percent is found. They also help to find inputs (familiar and unfamiliar) for creating a model, and to determine which variables are relevant to that model. They show which inputs are variable or constant, and which have predictable value.

The role of the CIO

How can a CIO play a role in this Big-Data exploration, or even better: how can he accelerate this exploration? The main question, of course, is whether he wants to do so and is able to do so. If the answer is 'yes', he can base his management mainly on setting up a Big Data service as if it were a restaurant, i.e. involving more than just tooling. The next step is to search for suitable topics and stakeholders within the organization to launch the Big Data project. The CIO's traditionally independent role is an asset here, provided of course that there is an intrinsic interest in the business process within the IT department. Otherwise, it is useless starting on the project, as you will end up with a fancy kitchen rather than a restaurant.