BY
John Lynch

What is Big Data? Busting the myth

"Big Data" is a buzzword that has gathered traction in the last few years, many people have heard of it but few actually know what any of it means.

John Lynch is a CPA working for the Ardagh Group as Financial and Systems Manager. He is currently studying for a Masters in Data Analytics.

Big Data does not have a specific agreed definition. It is used as an umbrella term that includes data science, data mining, predictive analytics, data modelling and business intelligence, among other adjectives.

What do these words actually mean? They are all interchangeable and in layman's terms it is about the bringing together of good old fashioned Maths & Stats with the computing world to be applied to real world scenarios to glean new information from the data presented.

The main reason why the concept of Big Data has really taken off in recent times is primarily that the tools that underpin it have advanced in leaps and bounds. In the past, a project of this nature would involve tedious and careful computer coding plus significant computing power was required to "crunch the numbers". As a result, these tools were only in the remit of large institutions like banks and insurance companies which had the means to finance and run initiatives of this type.

However, modern software has automated a great deal of the intricate coding of the past. Furthermore, computational power has increased dramatically and the cost of processing has fallen considerably. Today's smart phone has more computing power than NASA used for the first moon landing!

The dual improvements in hardware and software have brought down turnaround time and overall project costs dramatically. This has liberalised and democratised these methods and tools and brought them into the reach of a great many people and organisations.

For simplicity, I will use the term "Data Science" in lieu of the term "Big Data" as it is more apt, in my opinion. I will try to briefly set out various aspects and applications of data science and its relevance for the accounting profession.

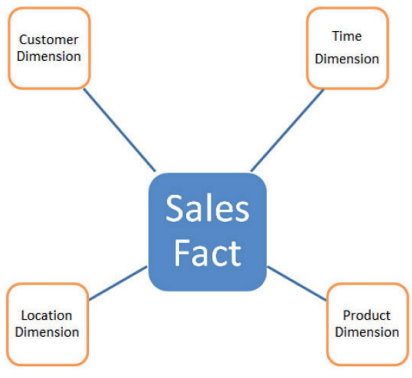
Dimension Modelling

Believe it or not, many of you have already engaged with dimension modelling. My first experience in this field was more than 10 years ago so it is not necessarily a new age technology.

Dimension modelling is a technique where data is extracted from a transaction processing system such as the sales ledger, general ledger, etc. and the data is reorganised and stored purely for the purposes of reporting and analysis. The advantage of this is that the sales ledger system's primary function is to process and record sales ledger transactions, not necessarily to report on sales ledger activity.

Conceptually, dimension modelling incorporates a central fact table with numeric measures (e.g. the sales value, sales volume, discount, VAT, etc.) around which dimensions are organised; these are factors such as customer, product/product type or location, for example.

The representation below is the logical design in a "star" schema with the central fact table and the various dimensions around it:



▶ Continued on Page 26

BY
John Lynch

▶ Continued from Page 25

From the user perspective, at the front end of a modelling system a user can move interchangeably between the dimensions and different levels of abstraction (e.g. country to county to town) and examine the different measures available and so on.

To simplify this, it is comparable to the use of pivot tables in Microsoft Excel. It is possible to click and drag dimensions from row to column, nesting more than one dimension into a row or column, apply a filter on different dimensions. It's this ability to work with multiple dimensions and measures at the same time that is its strength, and often you will hear it referred to as a "cube".

Tools such as this were once only within the reach of larger organisations. However, they are becoming more easily available; anyone with Excel 2016 and the "Power BI" suite of tools now have this modelling capability although without all the bells and whistles of a full blown dimension modelling tool.

Application of data science in practise

Modern day general ledgers now hold transactions at the atomic level. In the past, there may have been a single aggregated journal from a sales ledger for daily, weekly or monthly sales. Nowadays, there are sales transactions at the customer level in the general ledger. This is useful for the GL accountant who desires this level of granularity, but for the auditor and for an audit process that is heavily reliant on sampling techniques, this explosion in transactions presents a challenge.

With data science, it is possible to analyse large volumes quickly and easily and therefore asses the entire population of a particular dataset, even with the new expanded datasets which characterise modern day accounting systems.

Simple tasks can be performed to identify transactions such as:

- unusual processing dates or times – weekends/out of office hours
- postings after period end
- non-standard journals
- postings involving unusual account combinations.

More elaborate tools such as text mining can assist. This is the process of converting words or blocks of words into numeric indices in order to conduct statistical analysis. An example of applying text mining is when scanning journal descriptions for certain red flag descriptions – e.g. "profit adjustment".

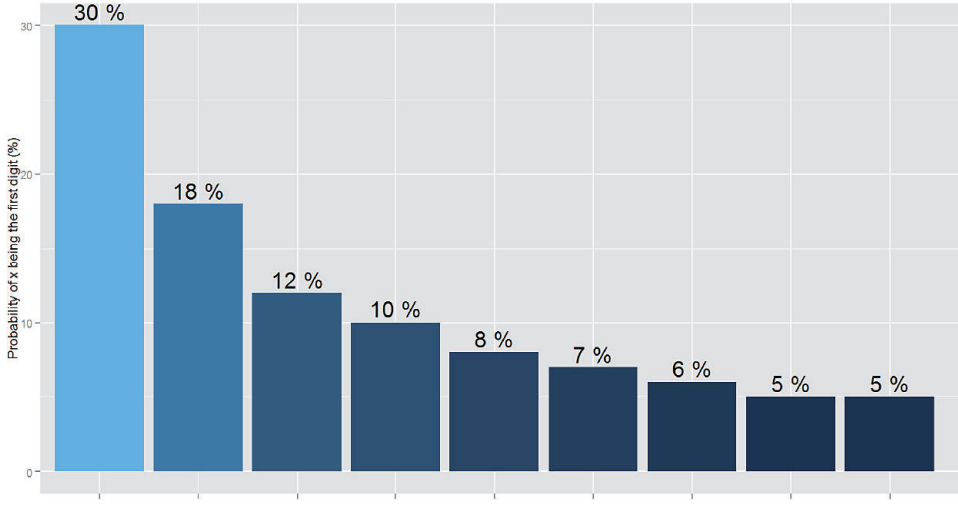
The process of text mining can factor in some of the issues which occur when trying to examine text. Using the example above of "profit adjustment", this could also be presented in a number of different ways including profit adjust/profit adj/profit adjmt. Text mining cannot completely mitigate against the problem of variants on a word that means the same thing but can enhance the process significantly.

Text mining tools can be applied in various other ways, for example:

- To compare supplier details against employee details for fraud detection;
- To compare employee/supplier/customer details for duplicates;
- To conduct a data quality check on postal addresses.

Another tool for audit and fraud detection is "Digit Analysis", which is the examination of the digits in a number. Benford's Law is an algorithm which examines the frequency of the first digit of a number for a range of numbers. Often in the real world the same frequency distribution is observed of the first digit, as per the below representation.

If you were to take a trial balance, extract the first digit of every account balance, the expectation is that the digit "1" will occur 30% of the time, "2" will occur 18% of the time and so on. In practise Benford's law is applied in more complex ways but the core concept remains the same.



Predictive Analytics

Employee churn, for some organisations, can be a human resource management issue. Frequently, the reasons for departure are not known, are misunderstood and/or unrecorded. Contemporary techniques such as exit interviews may or may not reveal the true reason for an employee's departure and often this type of data, if recorded, is held as soft data (anecdotal and unquantifiable).

Predictive analytics can be used as a tool in employee retention programmes, by examining available hard data and generating a mathematical model from it.
A very simplified example is set out below.

To start with, data is extracted from the historical employee repository. Details - called predictor variables – are extracted such as age, salary, department/section, years in current role, education level, number of years of service and so on.

	Predictor Variables							Target
Record No	Age	Salary	Department	Yrs in current role	Total Yrs of Service	Education level	...	Did they Churn
1001	23	31,200	Sales	2	2	Degree		Yes
1002	42	62,400	Finance	4	15	Masters		No
1003	27	41,600	Logistics	3	5	Senior cert		No
1004	55	68,640	Sales	9	22	Degree		Yes
1005	40	62,400	R&D	4	4	PHD		Yes
1006								
1007								
1008								

The table of predictor variables is used as a basis to generate a mathematical model, which in turn is then applied to current employee data to predict potential churn candidates.

This same approach could also be used in other scenarios such as for prediction of customer churn, production line faults etc.

Conclusion

When discussing big data, the "big" refers to the Googles and Amazons of the world which process mind boggling vast amounts of data. However, much of what these companies do is to employ data science techniques, albeit in newer and innovative ways and with specialist tools to handle the scale.

The accounting profession can benefit from similar use of data science techniques; the reduced cost of employing data science has made it more accessible to all, it can improve audit/fraud detection, reporting, and predictive analytics can build models to better manage your business.

The beauty of it is that the data science toolset can be applied to anything, once you have the data!