Post Connect

Big data
Buzzword, headache, distraction or opportunity?

The world is experiencing an explosion of data creation. But what is being done with this excess of data? Marketers are finding it's not enough to just push messages out. In order to hit its mark, data needs to be targeted, measured, matched and refined.

By Lisa Schutz and Graham Plant

Beyond the hype – lots of data is not the problem

Large amounts of data are a consequence of several key trends that show no signs of abating. Channel proliferation – particularly online – means that data is being created at a spectacular rate. Significantly, cheap data storage (dropping at a similarly spectacular rate) means that storage of all this data without too much thought about pruning has also been occurring.

Yes, there is a lot of data around. We are experiencing an explosion in data creation and availability greater than we have seen before. IBM recently reported that "every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals, to name a few."

Lots of data of itself should be neither a headache nor a distraction. What makes excess of data a buzzword, and potentially an opportunity, is the other key trend that goes hand in hand with channel proliferation – the need for highly segmented marketing activity to drive better returns on marketing spend.

> "We are increasingly in a big data world, and that has happened for several reasons. There's more ways to collect data, the cost of storing data is continuing to plummet, and there are new, relatively low-cost and scalable solutions for playing with data. It's opening up some interesting and exciting possibilities."
>
> **Joe Megibow,**
> Vice President and General Manager, Expedia US

The overall challenge is to be more accountable for marketing spend – using data. The current global financial pressures have challenged marketing departments and shifted spend towards parts of the budget that can be held accountable – direct and digital are the channels that more easily demonstrate their impact on the bottom line. The challenge to optimally target spend requires a robust measurement and feedback loop. It's not enough to push messages out – the messages have to be targeted to the audience, the responses measured and the process repeated incorporating past learnings.

Why does return on investment increase with data?

A personalised customer experience delivers value to consumers and therefore to the organisation. In the information age, data is our way of automating the personal experience. So, just as the supply of data is increasing, our appetite as businesses to use it is increasing, too. The case for data usage is strong – and big data is inevitable – so what is the problem again?

In the marketing world, it's not big data that's the problem, it's using the data to generate insight that leads to better actions that presents the challenge. As we will show in this paper, what stands between marketers and insight is integration.

If you don't believe that insight created from data is a point of competitive advantage, then look over your shoulder. Do you know if your competitors are spending more on analysts and data integration projects than on infrastructure? They probably are if the mix of job ads in this area is any indication.

> "Data is the next Intel Inside"
>
> **Tim O'Reilly**

This paper is focused on the piece of the puzzle of extracting maximum value from your data – data integration. Yes, analytics delivers the answers and, yes, data management is crucial, but the evidence suggests it's the integration piece that holds more projects up and is also a critical success factor in achieving success in analytics and data management.

Just to confirm, what is not a problem is managing big data itself. While some vendors might claim the problem is about hardware or software, we would suggest that is not the headache that it may first appear to be.

While not wanting to understate the challenges of handling a dramatic increase of data coming into your business, it is a challenge that every business has faced before. Admittedly this time, the volume and speed of data being created is much, much greater than we have experienced – but there are solutions.

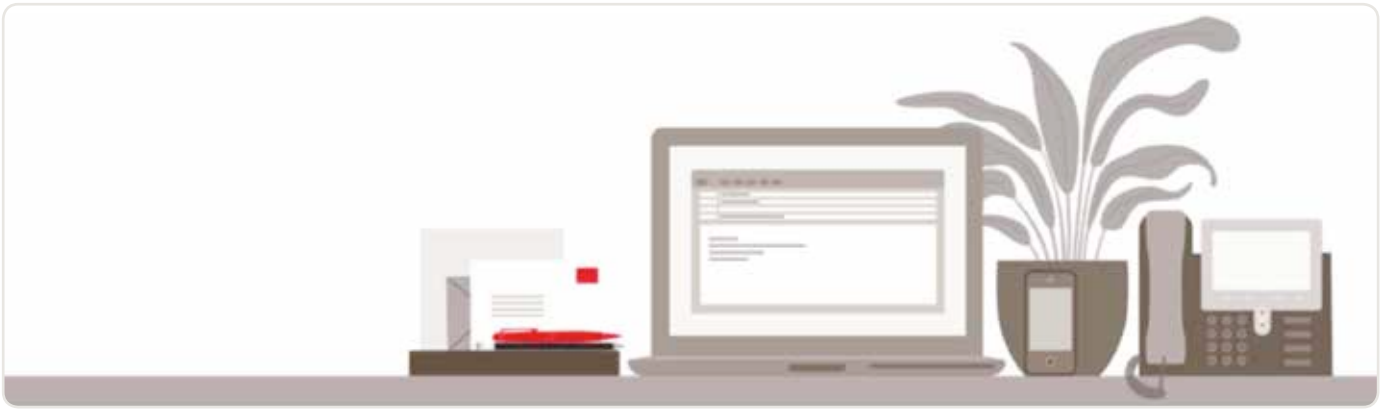
What is not a problem – the process and disciplines of data management

The disciplines and processes necessary to manage big data are not markedly different to the way we do things today and days gone by – especially structured data.

The stages of the data life-cycle are constants. The key steps are collect, store, integrate, generate insight and deploy. Then do it all again.

So, if the process is the same but the supply of raw data is growing. Big data needs a lot of processing to generate insight. And more data without good processing does not create insight – it's worthless.

There are three 'Vs' at work that make it challenging (mind you, not as challenging as doing many of the tasks we do on a standard smart phone compared to on a laptop from 10 years ago!). So what are the three 'Vs' – the Volume of data, the Variety of data and its Velocity – particularly as we add online channels.



Volume of data

Firstly, big data according to whom? In Wikipedia, big data is described as *"...a collection of data sets so large and complex that it becomes awkward to work with using on-hand database management tools."*

From this statement, we can determine that Big Data is a relative term and is largely determined by the capabilities and infrastructure of the business capturing, managing and using data for business decisions. For some businesses, the prospect of hundreds of gigabytes of data pouring into their business seems overwhelming. For many others, this may seem insignificant but tens or hundreds of terabytes would seem frightening.

Variety of data

If you are a technocrat, you might think that the problem of big data is the volumes of unstructured text/visual/voice data that are not in a field format accessible to decision systems and modelling tools. That is an issue, but it's soluble. And frankly, that is not why the phrase is ricocheting around the world.

Velocity of data

The velocity of data, as in the speed of its creation, transmission and consumption, is significantly more than we have experienced. But once again, there is some exceptional technology available that can help businesses manage this.

Yes, we are being deliberately provocative. But even the velocity of data can be handled technically, so it is more a business challenge in the sense that someone has to think about what all that data means, and be prepared to apply the level of effort and technology to administer it. IT is soluble. But the challenge for marketers is not buying infrastructure when they need insight.

> "How well your company integrates its data will be the difference between sales success and failure, customer retention and loss. Properly executed, integration is a critical component of any successful customer-centric business – an essential asset."
>
> **Lou Guercia,** Scribe Software

Single viewpoint – integration is a significant challenge

In order to generate insight, it's necessary to get a common view of the customer across channel and preferably across time. This means linking data from many disparate sources, including online and digital channels, structure and unstructured data – in essence, by every means that customers interact with you and you can record their responses. This is not easy for a number of reasons.

1. Data can be messy

Firstly, data sets often don't link up easily. Names and addresses are often used to link data sets. But formats change, data entry errors occur. So, you will have match rates, and your analysts will apply lots of clever business rules to improve the linkages, but confidence in accuracy is always a concern.

In a March 2012 survey conducted by Columbia Business School and New York American Marketing Association titled "Marketing ROI in the Era of Big Data" 42% of businesses surveyed cited. One of the major challenges related to the use of big data was not being able to link data together at the level of individual customers. A Winterberry study of businesses in the same year reported that the "ability to integrate multiple 'siloed' data repositories" was the most critically important issue in realising value from underlying data. This was followed closely by the "ability to integrate new third-party data sources" and the "ability to closely manage proprietary data assets".

2. More channels and data sources = more variables

This rapid growth of data from a variety of channels creates a significant challenge for businesses. As customer information is distributed across a variety of applications, the unification and integration of customer information from heterogeneous and dispersed applications are big challenges. Forrester Research reported that although 92% of companies said that having an integrated customer application is critical or important, only 2% have managed to achieve this.

3. Integrating physical and digital

We will discuss this a bit later, but something that is still a challenge for even the best industry participants is the integration of physical and digital channels. We live on Twitter and Facebook; we pick up catalogues and letters from our physical mailbox; and we watch TV. At the moment, the digerati would have you think physical location doesn't matter. We would respond with "*au contraire*".

But digital is the future

"The world is digital now!"

"My customers engage with me via email."

"Social media is more important in connecting with my customers."

"Our customers are online!"

Something to bear in mind is that even the most active online users have a postal address. In fact, we're prepared to go as far as saying more people have a postal address than have a Facebook account, Twitter account, email address or mobile telephone.

Where we live defines much about who we are, our values, our social status, our level of affluence, where we work and our lifestyle preferences. If you apply for a bank loan or insurance, one of the first items required to assess you is your address. Think of all the considerations you made in choosing where you live today and where else you would consider living and where you would not.

If you meet someone at a social function, do you ask them where they live, whether they have an email address or what their Twitter ID is? And which one gives you the most insight into this new acquaintance?

If you are planning to leverage the latest Census data being released by the Australian Bureau of Statistics (ABS), guess what? It's address centric or geographically captured and recorded.

In a survey conducted in February 2012 by Scribe Software on the "State of Customer Data Integration" it was identified that only 15% of businesses surveyed reported full integration across their various customer-facing systems; 50% reported partial integration; and 35% indicated they had just commenced their customer data integration.

A true view of each customer

Gaining a true view through customer data integration is the first step to really leverage your data assets, which is why this an important business issue. The key driver for customer data integration is to enable a "true" view of the customer. With a true, single viewpoint you can:

- effectively manage and use customer data by providing a timely and accurate understanding of customer needs and behaviours
- improve cross-selling and up-selling opportunities by understanding prospective customers and the emerging needs of existing customers
- remove duplication and misleading customer information and provide a single version of truth across the various business units of an organisation
- achieve effective campaign management
- comply with legislation, regulations and privacy requirements
- optimise operational, maintenance and enhancement costs by having a central integrated environment.

This goes with that

One of the major impediments to achieving a consolidated view of customer data is the ability to connect disparate elements of data to a customer. Databases are logical storage repositories of data and are (generally) constructed on clear rules to determine what data goes where in the database and what attributes it must have before it can be populated.

When working with data, the simplest and most precise way of matching and then linking is to use data with exact matching fields, or records that share the same key. When a unique key identifies one (and only one) record in a database, it can simply be matched to a secondary database where the same unique key value is recorded and returns an absolute match.

There are only two possible values – match or no match. By using this approach, there is absolute certainty in the correctness of the match and any ambiguity is removed.

However, when databases do not share a primary or unique key, it is necessary to revert to fields of the same types of data (without identifiers), such as addresses, names, telephone numbers or other descriptors, to identify potential matches. When using data elements that may have different structures and varying levels of accuracy, it is possible that there could be hundreds, if not thousands, of potential matched records, which results in three possible values – match, no match or possible matches.

"Data can only be useful if you understand where it is, what is means to your organisation and how it relates to other data in your organisation."

"Imperatives for the new CIO: How to manage data and apply analytics for efficiency and change".
CIO Insights, SAS Institute, 2010

Deterministic and probabilistic matching

The examples of data matching cited on the previous page identifies the two main types: deterministic and probabilistic. The most logical approach to data matching is to start with a deterministic approach and then for the non-matches adopt a probabilistic approach with clear rules and suitable checks and balances to ensure mismatches don't occur.

Companies who demand accuracy in matching for legal, financial or security reasons will only use a deterministic or rules-based approach. It offers clear and precise rules and forces accurate outcomes. A primary key that is applied across databases in the same way every time and forces high levels of accurate matching and appending is "gold" for the data integrator.

However, as you move into areas with data that are increasingly varied in format and are original in source, there seem to be fewer and fewer keys for linking data together.

Physical address might be boring – but it's consistent

A key attribute applied to each customer record in the vast majority of customer databases is the geographical address. The address is one of the most robust and best-structured data elements to work with and creates an avenue for standardising data.

When Australia Post first released the Personal Address File (PAF), it established a set of industry standards for matching, storing and managing customer data that was address centric. Today, most businesses in Australia have a customer database that has been cleansed and validated against the PAF and has embedded within it Australia Post's unique household identifier – the Delivery Point Identifier (DPID).

By using the Australia Post DPID applied to each address, businesses can more easily link disparate datasets containing addresses. Major corporations and marketers spend millions of dollars per annum on data hygiene and Customer Data Integration, and simplifying this process presents significant benefits for any business.

The DPID as a Primary Match Key

The DPID can only be appended using software accredited by the Address Matching Approval System (AMAS). Developers of the software must achieve a minimum match of 99.6% to be accredited and allowed to commercialise their software.

PAF Postal Address File
DPID Delivery Point Identifier
AMAS Address Matching Approval System
CCD Census Collector District

For every record matched, the DPID is returned as an eight-digit random code unique to the address record on the PAF.

The process of achieving AMAS accreditation requires best of breed probabilistic matching to a comprehensive and precise set of business rules and requirements. Once this is completed and the DPID is appended, the DPID can then be used as a Primary Match Key, adopting deterministic techniques to match disparate sources of data.

Obviously, this requires that each of the databases seeking to be matched is address based and has been processed through the AMAS-accredited software. When dealing with unstructured data and non-address-related data, such as transactional information, it is necessary for the data integrator to identify a way of matching this to geography.

The times are a-changing, and the complexity of matching the huge variety of new data sources will test every business. Understanding how to link data back to an address will enable you to integrate your customer data. Whether it is modelled to a Census Collector District (CCD), postcode, state or store catchment area, each of these address attributes can be appended to a DPID and provide consistency in the linkage of unstructured data to structured data.

Congratulations – we've come a long way! Australia Post used to process files where 40% of the records were inaccurate. Error rates now sit at around 5% with the application of its Personal Address File.

From barcoding of mail to data quality

When Australia Post released the PAF, many customer databases were so poor and devoid of standards that the initial files processed through the AMAS designed to release the PAF were achieving matches of around 60%. That means a whopping 40% of customer records were inaccurate. Today, most customer databases that have been regular users of the PAF via AMAS-accredited software achieve around 95% accuracy in their matching. Functional standards, clear business rules and application of data management principles supported by a comprehensive and accurate base of customer centric data in the PAF enabled this dramatic transformation.

Beyond 5% accuracy

Australia Post recognises that the PAF should be made more accessible to assist with the needs and challenges of a big data world. Working through our reseller networks and directly, we are releasing a set of enriched offers designed to package the suite of data that marketers and Customer Relationship Management teams need to integrate and enrich their data assets to generate the insights and superior marketing returns on investment.

Summary

Focus at the start on your goal – it improves the chances you'll get there!

When reviewing your own business and the need for integrating your customer data, it is important to first ask:

"What business problem are we solving by integrating?

"How much data do we need to integrate?"

"Is the integration of data required in real-time or can it be programmed?"

Answering these questions will assist you to define the benefits to your business that customer data integration can deliver.

One final word – don't be daunted

Wherever you are in your insight generation journey, our final advice would be to urge you to not be daunted by buzzwords such big data. Sure, there are plenty of technical challenges, but common sense and your understanding of your goal will get you where you need to go.

And, should you find yourself struggling with lower match rates than you would like, or you would like to enrich your data assets to help you target your campaigns more effectively, then talk either to us or one of our data resellers. Consider Australia Post – we have actually been in the data integration business for longer than anyone.

To find out more about smarter ways to connect with your customers, contact PostConnect on 1800 353 883 or visit www.postconnect.com.au.

About the authors

Lisa Schutz, General Manager PostConnect, leads Australia Post's multi-channel communications business. Lisa started out in strategy consulting and has focused throughout her career on the strategic use of data. Lisa co-founded the analytics firm Equigen Consulting (bought by Veda Advantage). She's since held leadership positions focusing on data innovation at Veda, MasterCard and InFact Decisions and has been a key advocate for data-driven approaches to media, marketing, telecommunications and IPTV.

Graham Plant is a marketing strategist and sales professional, who is passionate about improving sales and marketing effectiveness through customer insights and engagement. With more than 20 years' experience in driving better business outcomes by applying strong expertise in data-driven, long-term strategic planning, Graham has held various executive and general management positions with some of Australia's market leaders, including PMP Print, Australia Post, Telstra, Salmat Computing Services, Ausdoc and KPMG.