

Why big data matters to companies in retail and media

A straightforward guide for business folk
February, 2012



Introduction

The term “big data” is very much in vogue at the moment. In this white paper, we explore what big data means, what opportunities it presents to companies in the retail and media sectors and outline what companies need to do to take advantage of big data.

The purpose of this white paper is to provide an overview of the opportunities, challenges and success factors around big data. There is a lot to explore in each of the areas that we introduce: we plan to do this in subsequent white papers and blog posts, all of which will be made available on the [Keplar website](#).

A little history

The idea that important decisions in companies should be data-driven is much older than big data. Toyota’s pioneering use of data to drive efficiency in their manufacturing process helped them to steal a march on their American competitors back in the 1970s. Fast forward to Nineties Britain, and Tesco’s pioneering use of customer data collected via their club card scheme (run by Dunnhumby) helped them to establish themselves as the largest supermarket in the UK and in the top ten retailers globally.



By contrast, the history of big data is much more recent. The development of the internet has meant that people now conduct a whole range of activities online – be those activities directly on the Web, or on internet-connected digital platforms such as Xbox or Android. For retailers, this means that customers discover, research, discuss, recommend and buy products online. For media companies, this means that customers explore, discover, discuss, share, buy and consume media online.

When a customer performs these activities on a digital platform, detailed information about that customer’s behaviour can be captured by the retailer or media company operating the service. This data gives the provider unprecedented visibility into the customer’s behaviour, and this data can be analysed, for example to:

- Learn more about the customer, including who she is, how she engages with the product, company or brand, how she feels about the product and what role she plays in evangelising it to others
- Identify ways to better tailor the product and service to that customer or customer segment, improving customer loyalty
- Identify ways to improve the product for all users, by comparing the way that this customer used it with other customers. Are there particular workflows that customers struggled with or abandoned?
- Identify new products / services to offer that customer, or a segment of customers made up of people like her
- Grow customer lifetime value, and hence profit

One of the challenges of doing the above, however, is that the volume of data generated when a consumer engages online is enormous. Whereas in the offline world a company might collect a few million lines of data across all of their customers (e.g. a department store might store a line of data every time a consumer buys a product), several hundred or even thousand lines of data may be generated every time a user goes to a website, browses a selection of products, checks in on a mobile application, or messages his friends via social networks.

This explosion of data driven by online is what is referred to as “big data”. Fortunately, a whole raft of technologies and strategies have been developed that enable companies to store, crunch, make sense of big data and thus use it to drive business decisions. In this white paper, we’ll provide an overview of some of those key technologies, as well as outlining the steps that a company will need to take in order to employ them to drive commercial value. In subsequent posts on the [Keplar blog](#), we will explore some of these steps in more practical ways, including giving concrete implementation examples.

The promise of big data

There are two areas in particular where big data has a potentially transformative impact: consumer analytics and product analytics.

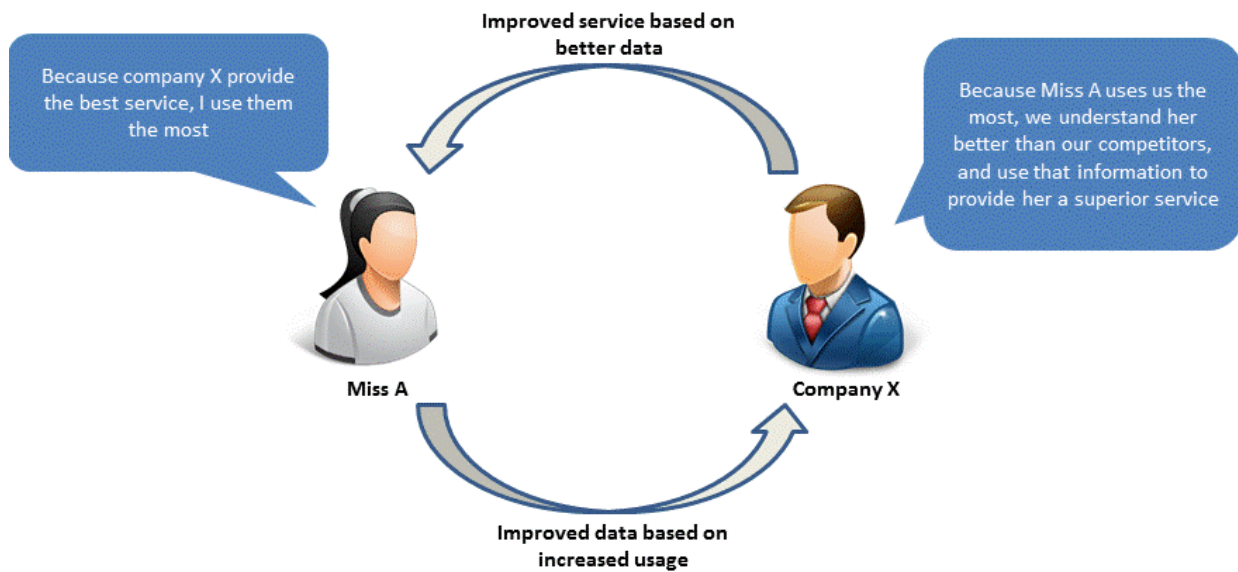
Consumer analytics

One of the most important KPIs for a company – be it a retail or media business – to focus on improving is customer lifetime value. This means understanding:

- Who are your customers?
- What do they buy from you?
- When do they buy it?
- Why do they buy it from you?
- What do they value in their service?
- What can you do to increase the value that you provide them, and the amount that they are willing to spend you?

Tesco and Dunhumby were pioneers in the use of customer lifetime value analysis, growing customer lifetime value by improving the range of products offered to customers, pricing items intelligently to reflect what customers were willing to pay, optimising store layouts and products offered to match local tastes, carefully targeting personalised promotions and developing services in-store (e.g. Tesco Finance) and out of store (e.g. tesco.com) to grow customer value.

A focus on customer lifetime value creates a win-win dynamic where companies build loyalty in their customers by providing them with improved service, which customers subsequently use more frequently, revealing more about their preferences and hence providing the company with more data with which to improve their service. This dynamic benefits both the customer and the company. It is an approach that is applicable not just to retailers, but also to media companies, financial services companies and technology companies.



At the heart of customer lifetime value-led approaches are customer databases (including CRM systems and data warehouses), customer metrics and business intelligence tools. Big data offers retailers the opportunity to capture not just individual sales and telephone engagements with customers (i.e. the type of interactions and transactions prevalent in the offline world), but also:

- How regularly does the customer visit the company's website?
- What information do they consume on the website?
- What products do they consider purchasing?
- Which adverts and promotions have they been shown? Which did they respond to and which did they ignore?
- How often do they visit company stores? (If the retailer has e.g. a location-based mobile app)
- What has the customer said about the company, products and services on social networks? What have they said about rival products and services? How do their views differ from their connections?
- How does the customer engage with the company on mobile application and partner services?

In the case of a media company, the difference is even starker. In the offline world, media companies might record how many unit sales of CDs, DVDs, books, newspapers or similar are sold by channel and geography, and slice that sales data by format, content category and other key variables. By contrast, in the online world a media company can often "drill down" all the way to the level of an individual customer, and understand:

- How do consumers actually engage with that media (be it reading books, listening to music, reading newspapers, watching TV or movies)? When do they consume media? How regularly? And on what platforms (e.g. mobile versus PC versus tablet)?
- What do they think of specific content items? (Especially if they rate individual games, books, music or similar and share those ratings, e.g. by social networks)
- How do their ratings correspond to their consumption habits?

- What triggers them to consume new media? (Recommendations from friends, magazines, adverts, promotions, live events?)
- How do their consumption patterns vary by age group, affluence and geography? How are they shifting over time? What is the impact of the shift on the media company's financials?
- How does the profitability of the media company split by consumer segment? How is that changing over time? What do the mid-term shifts in consumption behaviour mean for future profits? Are new channels and services additive, or do they cannibalise existing services?
- What levers does the media company have at its disposal to impact media consumption patterns for different consumers? How should it use those to maximise profit?

As you can see, big data opens the door on a wide range of questions about actual customer behaviour, and the ways in which either a retailer or media company might look to influence that behaviour to grow customer value.

Thanks to online, retailers now have unprecedented visibility into the browsing and buying behaviours of their individual customers. Similarly, media companies can have unprecedented visibility into the media consumption behaviours of their individual consumers. Used effectively, this knowledge is a tremendous resource to drive operational and strategic decision-making, to maximise customer lifetime value and hence profit.

Product analytics

Using customer data to maximise customer lifetime value is a well-established, well-understood use of big data in the retail sector – and increasingly in the media sector too. Less well understood, however, is the idea that data can also be used to drive digital product development.

Certainly, the importance of “getting a digital product right” is widely understood. You cannot understand Apple's success without seeing how beautifully constructed and easy to use Apple products are. Similarly, you cannot understand the success of Spotify without observing how the core Spotify app is superior to many of the alternative streaming or download services.

However, the role of data in driving the product development decisions which underpin high-quality products is much less well understood. To be fair, e-commerce folk do understand that in the case of their online shops (which are really just a type of digital product), conversion rates can help to understand how effective their shopping experience is; similarly, product managers understand that A/B and multi-variant testing can be used to decide on button appearance or choice of copy. However, both these examples scratch at the surface of how big data can drive product development decisions.

To illustrate the potential here, we need to revisit what makes digital products tick. Broadly speaking, a digital product is designed with a specific set of use-cases in mind. A digital music player, for example, needs to cater to an array of different use-cases, including a consumer:

- Browsing their selection of music
- Discovering new music
- Identifying and listening to a new track that they do not yet own

- Recommending and sharing music with friends
- Pausing the music playback to answer the telephone

The product designer starts with a preset notion of how her digital music player *should* be used by a consumer to accomplish each of these use-cases. But almost always, there is a mismatch between her theories and the actual, real-world use of her app: perhaps the user is doing something that the designer did not explicitly consider; perhaps the user is not performing a specific action at all; perhaps the user is accomplishing something “the wrong way”. Through fully understanding this mismatch, the product designer can build a new, better iteration of her music player.

This process of product analytics, then, aims to answer the following questions about the digital product:

- How is the digital product being used?
- Where is the product working well – for example, users are working through specific workflows regularly and without difficulty?
- Where is there scope for improvement?
- What is the impact of those improvements – including to usage patterns, and to the customer lifetime value of those users?

Product analytics involves mining behavioural data to understand these usage patterns, comparing those patterns against the set workflows and then analysing how past product development decisions have changed those usage patterns. It works best when performed in conjunction with qualitative and quantitative testing with real users, so that the product development team can understand not just how consumers use the product, but also *why* they use it in that particular way and *how* they feel about it.

We will explore the potential for product analytics powered by big data in further Keplar blog posts.

The problem with traditional databases and data warehouses



There are many challenges to using big data to carry out either consumer or product analytics. One challenge with both, however, is that you cannot analyse big data volumes using traditional databases or data warehousing technologies:

1. Traditional databases do not scale to house billions of lines of data

2. Traditional databases cannot effectively house unstructured or semi-structured data

Let us start with scale. Certainly, you can analyse thousands, sometimes tens of thousands, of lines of data using Excel. However, once you have hundreds of thousands of lines of data, Excel will start to struggle, and you will need to look at storing that data in a traditional database such as MySQL or SQL Server, potentially querying this database using a business intelligence toolkit such as MicroStrategy or Business Objects.

As the volume of data then increases to hundreds of millions of lines of data, however, your traditional database will start to stagger. At this stage, you may look to Oracle or Teradata for solutions, but these companies are essentially offering up supercharged versions of the same relational database technologies. These supercharged relational databases are very expensive, and whilst they can be made to handle billions of lines of data, at some point even they will start to struggle.

The issue of support for unstructured or semi-structured data is also an important one. Certainly, an enormous amount of work goes into designing database schemas, and database schemas are fantastic things: they make it easy for anyone with knowledge of the schema to quickly and easily write queries to extract the relevant data point from the database.

However, database schemas have some limitations. In a world where the number of data sources is manageable and all of them have an acceptable structure, developing a schema to house all of that data is achievable (although not always easy: data warehousing projects often take years and cost millions). But in a world in which there are potentially hundreds of different data sources and many of them are not tidily structured (think about pages of internet content or interactions on social networks), going to the expense of designing and deploying a schema before even storing the data is a significant upfront expense, and one that might unravel if new data does not fit the old schema. Worse, some types of unstructured data will never sit prettily in a tidy schema.

Fortunately, big data technologies offer a whole new approach to solving these two problems around scale and data structure. Indeed, these technologies (pioneered by Google) were architected to work with “web-scale” big data from the start, rather than taking a solution that was architected for smaller data sets and trying to stretch it to handle big data. The most important technology to highlight, here, is Hadoop.

Introducing Hadoop

Hadoop is a technology that makes it easy to store enormous volumes of data, and to then run queries against that data to drive consumer or product analytics. We intend to write some more detailed posts on Hadoop in the future, however it is worth outlining some of the key features of Hadoop in this white paper:

Hadoop is built from the ground up with scalability in mind

Hadoop is built using a parallel architecture that divides big analytics tasks into small units which can each be performed on commodity machines, rather than a handful of expensive data warehousing appliances. A lot of the intelligence that comes “for free” with Hadoop is around dividing data and analytic tasks up into discrete units, assigning them to different servers, instructing the individual servers which analytic function to perform, collecting the results and then aggregating them back together.

It is the parallelism of the Hadoop architecture that solves the problem of storing and crunching big data. By dividing comprehensive analyses on big data sets into thousands of smaller analyses which are run in parallel on smaller data sets, Hadoop makes it possible to crunch “impossibly large” data sets. As a result of the Hadoop’s use of parallel architecture, if you double the number of machines in a Hadoop cluster, the time taken to perform an analysis will (very roughly) halve. This is especially appealing in a world where cloud computing services such as Amazon’s EC2 make it easy to rent multiple servers just for the time required to perform an analysis task.

Data is stored in Hadoop without an explicit or pre-defined schema

Data is stored in Hadoop without an explicit data schema, really as a sequence of files. This makes storage very easy: it is just a case of “dumping” the data into Hadoop, and no data transformations are required. Of course, this makes the subsequent analysis more difficult: the analyst has to define the schema at the time of querying, including establishing all of the relationships between the different entities. This does mean, however, that:

1. The upfront costs of getting the data ready for storage are much lower
2. Schemas can evolve over time, as the set of analyses which the company wants to conduct on a given data set changes

In the pre-Hadoop world, companies which wanted to implement data warehouses had to invest significant upfront costs in designing their data warehouse, and then implementing ETL (“extract, transform, load”) processes to grab the data from operational systems and transform it into a suitable schema for the data warehouse. With Hadoop, all of this upfront work is not necessary (although sometimes it is still desirable): it is possible to dump data into a Hadoop cluster as-is, and figure out how to interpret it later.

Conversely, a disadvantage of Hadoop is that extracting a single line of data is much more difficult than in a traditional database. Executing a Hadoop query nearly always involves crunching through a huge volume of data, because there’s no schema or pre-processing to direct the analyst to the specific line of data quickly. However, for a whole raft of data analyses, crunching a huge volume of data is exactly what we want to do – and Hadoop is well-optimised to make that possible.

Not having a schema is a big advantage when it comes to unstructured data

Many of the big data sources consist in their entirety of unstructured or semi-structured data: examples include conversations between users in social media, different versions of e-commerce product descriptions or web log files. This data does not naturally lend itself to a rigid schema – but because this data can be stored in Hadoop without a schema, Hadoop can store unstructured and semi-structured data as easily as it can structured data.

Other relevant technologies for big data

Hadoop is not the only technology which a retailer or media company can employ to leverage its big data for commercial decision making. There are others of note, including:

Technology	Description
Hive, Pig	Tools that make it possible to query Hadoop data without writing MapReduce jobs in Java. The development of these tools has opened up Hadoop to a whole range of analysts and data scientists who previously would have lacked the requisite programming skills
NoSQL “databases”	Data stores of key-value pairs that are generally architected to allow for very fast access. These tend to be used in operational systems (rather than offline analytic systems) when the programming system does not require the query flexibility afforded by SQL interface, and speed of access (or a more flexible data structure) is desirable. Examples include MongoDB, CouchDB, Cassandra, Redis, HBase
Cloud computing	Services like Amazon’s S3 and Elastic MapReduce allow companies to rent computing resources as and when they need them to “crunch” big data, without having to invest in large-scale infrastructure to store and process it themselves
In-memory data analysis tools	<p>Tools like Hadoop make crunching large volumes of data possible, but they do not make it fast. Often the key to a successful analysis session is putting the data scientist in a position where he or she can respond to each answer that the data presents (in the form of a visualisation) with another question (in the form of a query).</p> <p>In-memory analytics engines such as Tableau make this kind of rapid data-interrogation possible by keeping the store of data under analysis in-memory. This will not be possible with a complete big data set, but it may be possible with an aggregate cut of the data. In future blog posts, we will describe how to use in-memory analytics tools such as Tableau in conjunction with big data analysis tools such as Hadoop</p>

We will explore these technologies and their value to media companies and retailers in subsequent blog posts. We have focused on exploring Hadoop in this white paper because it addresses – and illustrates – the fundamental challenge of storing and querying enormous data sets.

How do I start using big data to drive business value?

Having the right technology in place is necessary, but not sufficient, to using big data to drive business value in either consumer or product analytics. In addition to the technology, a data strategy, a set of business processes and a suitably skilled team of people are also pre-requisites.

Using data (whether big or little) to drive business value has never been easy. Developing an effective data strategy requires answering the following questions among others:

- What are the questions which we want the data to help us answer?
- What are the business decisions which those answers will drive, and what is the potential value of answering those questions correctly?
- What are the key unknowns that need to be tested?
- How should we use existing data or collect new data to answer those questions?

To answer these questions requires a firm understanding of the particular company in question, the challenges it faces and the goals it has set itself, as much as understanding the data and being a statistics expert. And leveraging big data is even harder: as well as the challenges associated with the everyday analytics, there are a set of additional challenges associated with the new technologies, many of which require a level of technical understanding and coding ability which is outside the average business analyst's comfort zone.

Effectively using big data to drive business value, then, means getting the following right:

Data strategy

A data strategy encompasses:

- What is it the company needs to understand to operate more effectively? How does this breakdown in the short, mid and long term? How does it break down by business unit and team?
- What is the corresponding roadmap of questions which the company is looking to answer, and what is the value associated with each of those answers?
- What resources can be allocated towards answering each question (including people, time and money)?
- Who is responsible for delivering each answer?

Business processes

Successful delivery of the data strategy will require multiple stakeholders to work together including:

- Business managers (who have to make decisions based on the answers)
- Business analysts or data scientists
- IT

A properly defined set of processes will make it more likely that answers are delivered on time, to budget, and in a way that is understandable to the decision-makers. These processes should encompass both the analysis and the socialising of the results within the business

People and culture

It is not enough to produce analyses to answer business questions. Team members need to understand what the analysis says, be comfortable interpreting it and using data in general to drive decision-making. That requires a basic level of "data literacy" in staff (all the way up to senior management) and a culture that expects and rewards data-driven thinking.

Conclusion

In this white paper, we have summarised how media companies and retailers can use “big data” as a key part of their efforts in consumer and product analytics. We described some of the technical challenges associated with handling big data, and outlined how technologies like Hadoop address those challenges. Finally, we highlighted that technology in itself is not enough, and that data-driven organisations that leverage big data need to employ the right data strategy, business processes, people and culture to maximize the value of big data analytics.

We understand that this white paper raises as many questions as it answers. Through 2012 we will publish a further white papers and blog posts which explore, among other topics, other technologies for big data analytics, organisational design for executing on data strategies and some concrete examples of big data analytics problems.

Interested in using big data and data analytics to drive decision making at your company? For help and support, don't hesitate to get in touch:

Tel: +44 (0)20 3589 6116

Email: contact@keplarllp.com

Address: The Roma Building
32-38 Scrutton Street
London EC2A 4RQ
United Kingdom