Drivers for Big Data

Big Data Analytics is a popular topic. While everyone has heard stories of new Silicon Valley valuation bubbles and critical shortages of data scientists, there are an equal number of concerns: Will it take away my current investment in Business Intelligence or replace my organization? How do I integrate my Data Warehouse and Business Intelligence with Big Data? How do I get started, so I can show some results? What are the skills required? What happens to data governance? How do we deal with data privacy?

Over the past 9 to 12 months, I have conducted many workshops with practitioners in this field. I am always fascinated with the two views that so often clash in the same room—the bright-eyed explorers ready to share their data and the worriers identifying ways this can lead to trouble. A similar divide exists among consumers. As in any new field, implementation of Big Data requires a delicate balance between the two views and a robust architecture that can accommodate divergent concerns.

Unlike many other Big Data Analytics blogs and books that cover the basics and technological underpinnings, this book takes a practitioner's viewpoint. It identifies the use cases for Big Data Analytics, its engineering components, and how Big Data is integrated with business processes and systems. In doing so, it respects the large investments in Data Warehouse and Business Intelligence and shows both evolutionary and revolutionary—as well as hybrid—ways of moving forward to the brave new world of Big Data. It deliberates on serious topics of data privacy and corporate governance and how we must take care in the implementation of Big Data programs to safeguard our data, our customers' privacy, and our products.

So, what is Big Data? There are two common sources of data grouped under the banner of Big Data. First, we have a fair amount of data within the corporation that, thanks to automation and access, is increasingly shared. This includes emails, mainframe logs, blogs, Adobe PDF documents, business process events, and any other structured, unstructured, or semi-structured data available inside the organization. Second, we are seeing a lot more data outside the organization—some available publicly free of cost, some based on paid subscription, and the rest available selectively for specific business partners or customers. This includes information available on social media sites, product literature freely distributed by competitors, corporate customers' organization hierarchies, helpful hints available from third parties, and customer complaints posted on regulatory sites.

Many organizations are trying to incentivize customers to create new data. For example, Foursquare (*www.foursquare.com*) encourages me to document my visits to a set of businesses advertised through Foursquare. It provides me with points for each visit and rewards me with the "Mayor" title if I am the most frequent visitor to a specific business location. For example, every time I visit Tokyo Joe's—my favorite nearby sushi place—I let Foursquare know about my visit and collect award points. Presumably, Foursquare, Tokyo Joe's, and all the competing sushi restaurants can use this information to attract my attention at the next meal opportunity.

Sunil Soares has identified five types of Big Data: web and social media, machine-to-machine (M2M), big transaction data, biometrics, and human generated.[1] Here are some examples of Big Data that I will use in this book:

- Social media text
- Cell phone locations
- Channel click information from set-top box
- Web browsing and search
- Product manuals
- Communications network events
- Call detail records (CDRs)
- Radio Frequency Identification (RFID) tags
- Maps
- Traffic patterns
- Weather data
- Mainframe logs

Why is Big Data different from any other data that we have dealt with in the past? There are "four V's" that characterize this data: Volume, Velocity, Variety,

and Veracity. Some analysts have added other V's to this list, but for the purpose of this book, I will focus on the four V's described here.

Volume

Most organizations were already struggling with the increasing size of their databases as the Big Data tsunami hit the data stores. According to *Fortune* magazine, we created 5 exabytes of digital data in recorded time until 2003. In 2011, the same amount of data was created in two days. By 2013, that time period is expected to shrink to just 10 minutes.[2]

A decade ago, organizations typically counted their data storage for analytics infrastructure in terabytes. They have now graduated to applications requiring storage in petabytes. This data is straining the analytics infrastructure in a number of industries. For a communications service provider (CSP) with 100 million customers, the daily location data could amount to about 50 terabytes, which, if stored for 100 days, would occupy about 5 petabytes. In my discussions with one cable company, I learned that they discard most of their network data at the end of the day because they lack the capacity to store it. However, regulators have asked most CSPs and cable operators to store call detail records and associated usage data. For a 100-million-subscriber CSP, the CDRs could easily exceed 5 billion records a day. As of 2010, AT&T had 193 trillion CDRs in its database.[3]

Velocity

There are two aspects to velocity, one representing the throughput of data and the other representing latency. Let us start with throughput, which represents the data moving in the pipes. The amount of global mobile data is growing at a 78 percent compounded growth rate and is expected to reach 10.8 exabytes per month in 2016[4] as consumers share more pictures and videos. To analyze this data, the corporate analytics infrastructure is seeking bigger pipes and massively parallel processing.

Latency is the other measure of velocity. Analytics used to be a "store and report" environment where reporting typically contained data as of yesterday—popularly represented as "D-1." Now, the analytics is increasingly being embedded in business processes using data-in-motion with reduced latency. For example, Turn (*www.turn.com*) is conducting its analytics in 10 milliseconds to place advertisements in online advertising platforms.[5]

1.3 Variety

In the 1990s, as Data Warehouse technology was rapidly introduced, the initial push was to create meta-models to represent all the data in one standard format.

The data was compiled from a variety of sources and transformed using ETL (*E*xtract, *T*ransform, *L*oad) or ELT (*E*xtract the data and *L*oad it in the warehouse, then *T*ransform it inside the warehouse). The basic premise was narrow variety and structured content. Big Data has significantly expanded our horizons, enabled by new data integration and analytics technologies. A number of call center analytics solutions are seeking analysis of call center conversations and their correlation with emails, trouble tickets, and social media blogs. The source data includes unstructured text, sound, and video in addition to structured data. A number of applications are gathering data from emails, documents, or blogs. For example, Slice provides order analytics for online orders (see *www.slice.com* for details). Its raw data comes from parsing emails and looking for information from a variety of organizations—airline tickets, online bookstore purchases, music download receipts, city parking tickets, or anything you can purchase and pay for that hits your email. How do we normalize this information into a product catalog and analyze purchases?

Another example of enabling technology is IBM's InfoSphere Streams platform, which has dealt with a variety of sources for real-time analytics and decision making, including medical instruments for neonatal analysis, seismic data, CDRs, network events, RFID tags, traffic patterns, weather data, mainframe logs, voice in many languages, and video.

1.4 Veracity

Unlike carefully governed internal data, most Big Data comes from sources outside our control and therefore suffers from significant correctness or accuracy problems. Veracity represents both the credibility of the data source as well as the suitability of the data for the target audience.

Let us start with source credibility. If an organization were to collect product information from third parties and offer it to their contact center employees to support customer queries, the data would have to be screened for source accuracy and credibility. Otherwise, the contact centers could end up recommending competitive offers that might marginalize offerings and reduce revenue opportunities. A lot of social media responses to campaigns could be coming from a small number of disgruntled past employees or persons employed by competition to post negative comments. For example, we assume that "like" on a product signifies satisfied customers. What if the "like" was placed by a third party?[6]

We must also think about audience suitability and how much truth can be shared with a specific audience. The veracity of data created within an organization can be assumed to be at least well intentioned. However, some of the internal

data may not be available for wider communication. For example, if customer service has provided inputs to engineering on product shortcomings as seen at the customer touch points, this data should be shared selectively, on a need-to-know basis. Other data may be shared only with customers who have valid contracts or other prerequisites.

Over the past year, the Information Agenda team has been asked to conduct a number of Big Data Analytics workshops. The three most common questions have been as follows:

1. What is Big Data and what are others doing with it?
2. How do we build a strategic plan for Big Data Analytics in response to a management request?
3. How does Big Data change our analytics organization and architecture?

Most of the material included in this book was collated in response to answering these questions.

This book provides three perspectives on Big Data Analytics.

First, why is Big Data Analytics becoming so important, and what can we do with it? The book projects major trends behind the rise of Big Data and shows typical use cases tackled by Big Data Analytics, where leading organizations are already seeing major benefits.

Second, the book lists major components of Big Data Analytics and introduces an integrated architecture—Advanced Analytics Platform (AAP)—that combines Big Data Analytics with the rest of the analytics infrastructures and integrates with business processes. It shows how these components work together in the AAP to provide an integrated engine that can combine Big Data with traditional Data Warehouse and Business Intelligence to provide an overall solution.

Third, the book provides a glimpse at implementation concerns and how they must be tackled. How do we establish a roadmap and implement key pilot programs to gather momentum and persist to create a game-changing vision? How do we provide governance across this data when the originating data may have varying quality or privacy constraints?

The big elephant in the room is data privacy. I confess I have not taken a position on data privacy, nor have I predicted how the world will deal with it. It is an evolving topic, with many complications, geographical differences, and

unknown consequences. However, I have outlined a number of critical areas to probe further, as well as a number of required components, irrespective of the position taken.

I have relied heavily on my personal work for illustrations of the concepts discussed in this book. As a result, most of the examples are tilted towards CSPs, advertising, and retail industries. This is not to say that these industries are leading the pack or that other industries do not have good Big Data opportunities. To the contrary, we are finding a large number of examples across many industries.

Chapter 2
Drivers for Big Data?

We are increasing the pace for Big Data creation. This chapter examines the forces behind this tsunami of Big Data. There are three contributing factors: consumers, automation, and monetization. More than each of these contributing factors, their interaction is speeding the creation of Big Data. With increasing automation, it is easier to offer Big Data creation and consumption opportunities to the consumers and the monetization process is increasingly providing an efficient marketplace for Big Data.

2.1 Sophisticated Consumers

The increase in information level and the associated tools has created a new breed of sophisticated consumers. These consumers are far more analytic, far savvier at using statistics, and far more connected, using social media to rapidly collect and collate opinion from others. We live in a world full of marketing messages. While most of the marketing is still broadcast using newspaper, magazine, network TV, radio, and display advertising, even in the conventional media, narrow casting is gradually becoming more prominent. This is seen in local advertisement insertions in magazines, insertion of narrow cast commercials using set-top boxes, and use of commuter information to change street display ads. The Internet world can become highly personalized. Search engines, social network sites, and electronic yellow pages insert advertisements specific to an individual or to a micro-segment. Internet cookies are increasingly used to track user behavior and to tailor content based on this behavior.

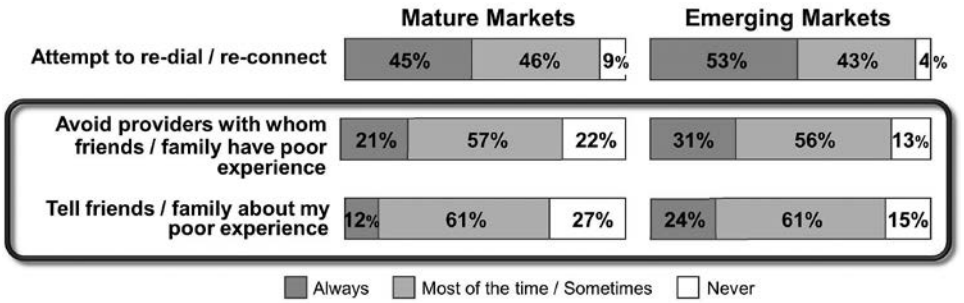
Email and text messages rapidly led toward increased interpersonal interactions. Communication started not only with marketers but also with third parties and friends. Communication expanded to bulletin boards, group chats, and social media, allowing us to converse about our purchase intentions, fears,

expectations, and disappointments with small and large social groups. Unlike email and text, the conversations are on the Web for others to read, either now or later.

So far, we have been dealing only with single forms of communication. The next sets of sources combine information from more than one media. For example, Facebook conversations involve a number of media, including text, sound clips, photos, and video. Second world and alternate reality are becoming interesting avenues for trying out product ideas in a simulated world where product usage can be experimented with.

We often need experts to help us sort out product features and how they relate to our product usage. A large variety of experts are available today to help us with usage, quality, pricing, and value-related information about products. A number of marketers are encouraging advisor or ambassador programs using social media sites. These selected customers get a preview of new products and actively participate in evaluating and promoting new products. At the end of the day, people we know and trust sway our decisions. This is the biggest contribution of social networks. They have brought consumers together such that sharing customer experiences is now far more frequent than ever before.

How would a consumer deal with a poor service quality experience? Figure 2.1 shows typical behaviors in mature and emerging markets as studied by an IBM Global Telecom Consumer Survey conducted with a sample size of 10,177.[7] In this survey, 78 percent of the consumers surveyed in the mature markets said they avoid providers with whom friends or family had bad experience. The percentage was even higher (87 percent) in growth markets. In response to a



Source: 2011 IBM Global Telecom Consumer Survey, Global N = 10177; Mature Countries N = 7875

Figure 2.1: Behaviors in response to poor service quality experience

related question, survey participants said that they inform friends and family about poor experience (73 percent in mature markets and 85 percent in growth markets). These numbers together show a strong influence of social network on purchase behavior. These are highly significant percentages and are now increasingly augmented by social media sites (e.g., the "Like" button placed on Facebook). The same survey also found that the three most preferred sources for recommendation information are Internet, recommendations from family/friends, and social media.

In any group, there are leaders. These are the people who lead a change from one brand to another. Leaders typically have a set of followers. Once a leader switches a brand, it increases the likelihood for the social group members to churn as well. Who are these leaders? Can we identify them? How can we direct our marketing to these leaders?

In any communication, the leaders are always the center of the hub (see Figure 2.2). They are often connected to a larger number of "followers," some of whom could also be leaders. In the figure, the leaders have a lot more communication arrows either originating or terminating to them compared with others.

How do we identify the leaders? IBM Research conducted a series of experiments with CSPs.[8] Call detail records, which carry information about

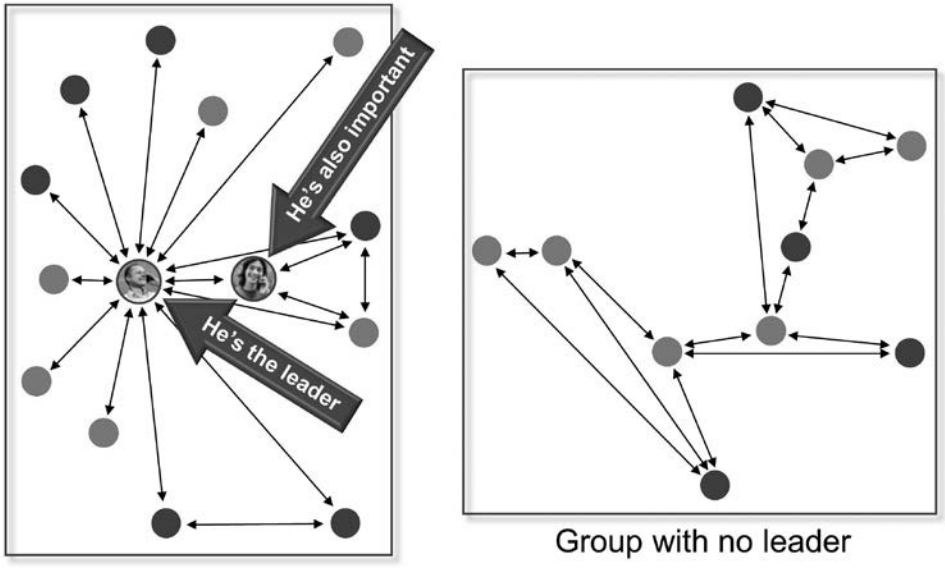


Figure 2.2: Leaders in a communications network

person A calling person B, were analyzed. By synthesizing call information and abstracting communications networks, we discovered webs of communications across individuals. We also used the customer churn information to correlate churn among leaders to subsequent churn among followers. Here are some of the highlights from one of the experiments I helped conduct:

- Leaders were 1.2 times more likely to churn compared with non-leaders.
- There were two types of leaders: disseminating leaders who were connected to their group through outgoing calls, and authority leaders who were connected through a larger proportion of incoming calls.
- When a disseminating leader churned, additional churns were 28.5 times more likely. When an authority leader churned, additional churns were 19.9 times more likely.
- Typically, there was a very limited time between leaders' churn and the followers' churn.

Social groups can be inferred from any type of communication—emails, SMS texts, calls, Facebook friendships, and so on. It is interesting to see strong statistics associated with leaders' influence on the group.

There are many ways to utilize social networks to influence purchase and reuse:

- *Studying consumer experience*—A fair amount of this data is unstructured. By analyzing the text for sentiments, intensity, readership, related blogs, referrals, and other information, we can organize the data into positive and negative influences and their impact on the customer base.
- *Organizing customer experience*—We can provide reviews to a prospective buyer, so they can gauge how others evaluated the product.
- *Influencing social networks*—We can provide marketing material, product changes, company directions, and celebrity endorsements to social networks, so that social media may influence and enhance the buzz.
- *Feedback to products, operations, or marketing*—By using information generated by social media, we can rapidly make changes in the product mix and marketing to improve the offering to customers.

Society has always played a major role in our evaluation process. However, the Internet and social networking have radically altered our access to information. I may choose to "like" a product on Facebook, and my network now has

instant access to this action. If I consider a restaurant worth its money, Yelp can help me broadcast that fact worldwide. If I hate the new cell phone service from a CSP, I can blog to complain about it to everyone.

2.2 Automation

Interactive Voice Response (IVR), kiosks, mobile devices, email, chat, corporate Websites, third-party applications, and social networks have generated a fair amount of event information about the customers. In addition, customer interactions via traditional media such as call centers can now be analyzed and organized. The biggest change is in our ability to modify the customer experience using software policies, procedures, and personalization, making self-service increasingly customer friendly.

Sales and marketing have received their biggest boost in instrumentation from Internet-driven automation over the past 10 years. Browsing, shopping, ordering, and customer service on the Web not only has provided tremendous control to users but also has created an enormous flood of information to the marketing, product, and sales organization in understanding buyer behavior. Each sequence of Web clicks can be collected, collated, and analyzed for customer delight, puzzlement, dysphoria, or outright defection. More information can also be obtained about sequence leading up to a decision.

Self-service has crept in through a variety of means: IVRs, kiosks, handheld devices, and many others. Each of these electronic means of communication acts like a gigantic pool of time-and-motion studies. We have data available on how many steps customers took, how many products they compared, and what attributes they focused on, such as price, features, brand comparisons, recommendations, defects, and so on. Suppliers have gained enormous amounts of data from self-service and electronic sensors connected to products. If I use a two-way set-top box to watch television, the supplier has instant access to my channel-surfing behavior. Did I change the channel when an advertisement started? Did I turn the volume up or down when the jingle started to play? If I use the Internet to shop for a product, my click stream can be analyzed and used to study shopping behavior. How many products did I look at? Did I view the product description or the price when looking at the product? This enriched set of data allows us to analyze customer experience in the minutest detail.

What are the sources of data from such self-service interactions?

- *Product*—As products become increasingly electronic, they provide a lot of valuable data to the supplier regarding product use and product quality. In many cases, suppliers can also collect information about the context in which a product was used. Products can also supply information related to frequency of use, interruptions, usage skipping, and other related aspects.
- *Electronic touch points*—A fair amount of data can be collected from the touch points used for product shopping, purchase, use, or payment. IVR tree traversals can be logged, Web click streams can be collected, and so on.
- *Components*—Sometimes, components may provide additional information. This information could include data about component failures, use, or lack thereof. For example, a wireless CSP can collect data from networks, cell towers, third parties, and handheld devices to understand how all the components together provided a good or bad service to the customer.

2.3 Monetization

From a Big Data Analytics perspective, a "data bazaar" is the biggest enabler to create an external marketplace, where we collect, exchange, and sell customer information. We are seeing a new trend in the marketplace, in which customer experience from one industry is anonymized, packaged, and sold to other industries. Fortunately for us, Internet advertising came to our rescue in providing an incentive to customers through free services and across-the-board opt-ins.

Internet advertising is a remarkably complex field. With over $26 billion in 2010 revenue,[9] the industry is feeding a fair amount of startup and initial public offering (IPO) activity. What is interesting is that this advertising money is enhancing customer experience. Take the case of Yelp, which lets consumers share their experiences regarding restaurants, shopping, nightlife, beauty spas, active life, coffee and tea, and others.[10] Yelp obtains its revenues through advertising on its website; however, most of the traffic is from people who access Yelp to read customer experience posted by others. With all this traffic coming to the Internet, the questions that arise are how is this Internet usage experience captured and packaged and how are advertisements traded among advertisers and publishers.

Big Data Analytics is creating a new market, where customer data from one industry can be collected, categorized, anonymized, and repackaged for sale to others:

- *Location*—As we discussed earlier, location is increasingly available to suppliers. Assuming a product is consumed in conjunction with a mobile device, the location of the consumer becomes an important piece of information that may be available to the supplier.
- *Cookies*—Web browsers carry enormous information using web cookies. Some of this may be directly associated with touch points.
- *Usage data*—A number of data providers have started to collect, synthesize, categorize, and package information for reuse. This includes credit-rating agencies that rate consumers, social networks with blogs published or "Like" clicked, and cable companies with audience information. Some of this data may be available only in summary form or anonymized for the protection of customer privacy.

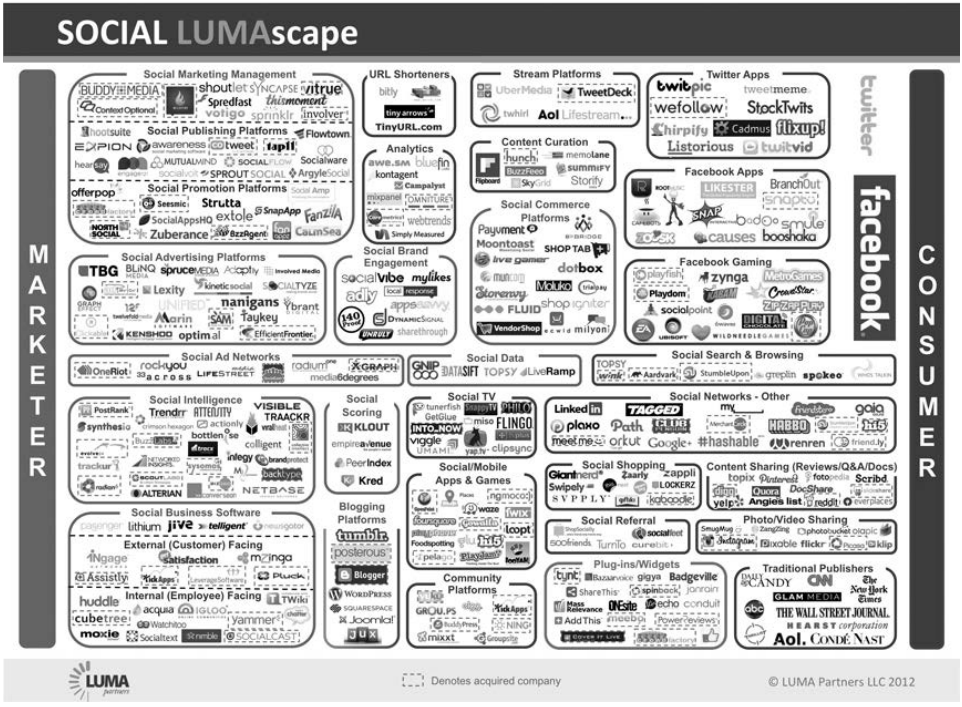


Figure 2.3: LUMA Scape for social media (reprinted with permission)

Terence Kawaja has been studying this market for a number of years and has characterized a number of markets and associated players. "Terence Kawaja has a new way for potential investors to visualize it," says *Wall Street Journal* writer Amir Efrati. "The market involves hundreds of small and large companies that help advertisers reach consumers and help website publishers, mobile-application developers, search engines, and other digital destinations generate revenue through advertising. Kawaja, who runs the investment firm LUMA Partners, spent months putting together six new graphics that show how 1,240 different companies fit into the following categories of online advertising: display, video, search engines, mobile, social, and commerce."[11] I have replicated Kawaja's Social Media LUMA Scape in Figure 2.3. For the rest of the LUMA Scapes, visit Kawaja's website: *www.lumapartners.com*. A number of intermediaries play key roles in developing an advertising inventory, auctioning of the inventory to the ad servers, and facilitating the related payment process, as the advertisements are clicked and related buying decisions are tracked.