Demystifying big data analytics

April, 2014

A White Paper by

Dr. Barry Devlin, 9sight Consulting
barry@9sight.com

Big data analytics is an essential tool in understanding and driving today's combined online and physical business world. However, there exists considerable confusion about how traditional business data (e.g. POS, inventory, shipping, etc.) can be combined with modern digital data sources. Add the reality of a still evolving technological environment, and the outcome is confusion for many aspiring analysts. The aim of this paper is to demystify the situation.

We begin with a brief summary of the technical roadblocks to adoption of big data analytics directly on today's most popular big data storage and processing platform, Hadoop. In contrast, three successful use cases, based on real implementations across different industries, demonstrate what is possible and where real benefits lie. Their success stems from advances in agile context creation and timely, graphical analysis by business users, as well as consolidating data from both traditional and so-called "unstructured" data sources in Hadoop in their raw, original form.

This white paper also offers a useful model of the modern data landscape, breaking it up into three distinct areas: process-mediated data, machine-generated data and human-sourced information. Understanding these three data types is a foundation for advancing big data analytics and choosing an architectural approach to its implementation.

Finally, we describe an approach to creating a big data analytics platform capable of delivering sustained business value.

Contents

"Sweet Analytics, 'tis thou hast ravished me." [1]

For the business analyst, big data can be very seductive. It exists in enormous quantities. It contains an extensive and expanding record of every interaction that makes up people's various daily behaviors. According to all the experts, the previously unnoticed correlations it contains hold the potential for discovering customer preferences, understanding their next actions, and even creating brand new business models. Trailblazing businesses in every industry, especially Internet startups, are already doing this. The future beckons…

However, for the ordinary business analyst in a mainstream enterprise or emerging online business, it's often not as easy as it appears. The technology is still emerging and often requires significant IT skills to create and manage data stores. Data must be combined from a disparate set of sources. A key data source—traditional business transactions and other operational and informational data—is often largely isolated from the big data scene.

Hadoop offers size and agility but needs a new user-friendly face at every stage, from data acquisition through to analysis and use.

Today, big data and Hadoop are synonymous. Hadoop is certainly popular but it's far from friendly. Your typical business analyst is likely to be daunted by the programming skills often needed to explore the data therein. It's a long way from the graphical BI tools, SQL queries and even Excel used by analysts today. Hadoop may be today's favorite big data processing engine, but it definitely needs a user-friendly face at every stage, from data acquisition through to analysis and use. Addressing these issues is key to demystifying big data analytics.

Big data comes in a wide variety of shapes and structures from a plethora of sources. One of the most commonly used is web log data, which takes the form of a simple text string with a predefined order of data items. Comma separated variable (CSV) files are used as a common exchange format for many systems. JavaScript Object Notation (JSON) objects are generated by many devices on the emerging Internet of Things. The structure-agnostic Hadoop reservoir is an ideal place to store such diverse data and process it with agility, but before an analyst can make use of it, somebody—often the analyst herself—has to define what each piece of data means.

And although the Hadoop store is the default destination for all big data, perhaps the most important data of the actual business—the customer and product records, the transactions, and so on—usually reside elsewhere entirely, in the relational databases of the business' operational and informational systems. This data is key to many of the most useful analyses the business user may desire.

These three roadblocks—technical complexity, lack of contextual information, and missing traditional data—are the price paid for Hadoop's strengths in handling large data volumes and its agility with novel types of data. They are serious barriers to adoption of big data analytics by mainstream and emerging companies, as well as ordinary business analysts. Fortunately, a better understanding of today's data types and tools that address the underlying problems are emerging. We explore these aspects later, but first let's examine three business use cases, generalized from existing, successful implementations. We note very distinct value drivers across different industries, but observe some common theme too: the extensive use of data from multiple sources as input to decision making and a graphical, iterative approach to the analytics. We see the emergence of a fact-based enterprise in many aspects of traditional decision support as well as in novel business analytics.

Creating the fact-based enterprise

"People…operate with beliefs and biases. To the extent you can eliminate both and replace them with data, you gain a clear advantage." [2]

There is much talk about the need for business to become data-driven. Indeed, there is research that shows a direct correlation between data-driven decision making and business performance. For example, Erik Brynjolfsson et al. showed in 2011[3] an increase in output and productivity of 5-6% in firms that are data-driven, although profit margin showed less correlation. On the other hand, industry observers consistently note a range of inhibitors to data-driven decision making. These range from purely technological to organizational. We might conclude, therefore, that many enterprises are failing to achieve the levels of business benefit that data-driven decision making can offer.

Considering how these benefits can be achieved, four areas of improvement in big data analytics are needed to deliver a fact-based enterprise:

1. Combining data from traditional and new sources
2. Creating context for data while maintaining agile structure
3. Supporting iterative, speed-of-thought analytics
4. Enabling business-user-friendly analytical interface

Building a fact-based enterprise delivers bottom-line benefits, but will require improvements in big data analytics tools.

We can clearly see one or more of these needs expressed in the following use cases.

Shopping—online gets physical, physical goes online

Online retailers were among the earliest users of big data analytics. Web logs recorded every click and page view of a shopper on the website. Analyzing that data alone gave the retailer a wealth of information about shoppers' behavior: how they arrived on the site, where and how long they browsed, what and when they bought or, more worryingly, when they abandoned a cart with planned purchases. Websites were optimized, customers upsold or cross-sold, product lines redefined on the basis of this data. Most online retailers used Hadoop to store and analyze the data. As early adopters of web technologies, they had the technical skills required to set up and use the environment.

Retailing has increasingly moved from very separate online and physical experiences to the world of clicks and mortar. The data of interest no longer originates in its entirety from the web logs. Point of sale data from stores travels through retail applications and ends up in data warehouses in relational databases. Supply chain management systems store further useful data. Customer movements are tracked in-store via their smartphones. Call centers record phone interactions with customers, storing data in bespoke applications. Social media interactions may also be tracked.

The customer's journey across different channels from interest to purchase and beyond is recorded in multiple places and formats. The challenge is to follow the trail; the value is in truly understanding the behavior and optimizing the value of the customer. Figure 1 is a graphical depiction of a single customer's journey through the online and offline landscape in search of the right product. Analyzing many such journeys provides answers to key questions such as what channels are most effective in driving interest, when and where do potential customers most often lose interest, where is investment needed to fix problems or encourage purchase.

It is often the case today that a number of different businesses cooperate to deliver the full clicks and mortar service. This adds further complexity to data collection and preparation. A good example is in the auto retailing industry, where web log data from a search and comparison website must be combined with inventory data about cars available on dealers' lots, call logs, social media data, and actual sales, as well as a range of demographic data to allow the website provider to track trends, measure satisfaction and improve service to its customers. Customers benefit from more accurate and timely information, dealers gain through more focused inquiries, and the search and comparison provider gets more accurate facts about today's market and an improved ability to predict tomorrow's.

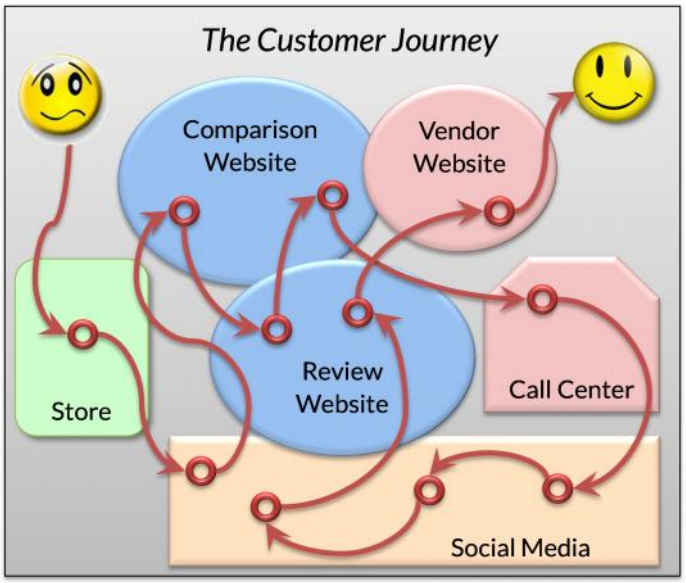

Figure 1: The customer journey

Entertainment—the times they are a-changin'

Online gaming, from smartphone apps to MMOGs (massively multiplayer online games) are big business, but keeping players engaged and spending money can be a challenge. The more sophisticated games are heavily instrumented with software routines to track user behavior and actions: how long is it taking to overcome a challenge, are players dropping off in specific places, what interactions precede a purchase, and so on. These probes within the game communicate with the central server allowing developers to change aspects of the game on an ongoing basis. The probes are also changed regularly, as well as their message payloads. For larger game providers, messages run to tens of terabytes per day and new messages collected every millisecond.

Rapid, in-depth analysis of the information coming in such messages can be difficult in a traditional BI environment because of their changing form and content. Predefined database models populated through pre-built ETL processes lack the agility to easily react to changes. IT can quickly fall behind with a backlog of database and ETL changes. Business users are not amused. The solution is to store the messages in their original format as they arrive and allow analysts to dynamically define and map their contents at the last moment before analysis.

Rapidly evolving markets require agile business analytics, enabled by users who can control much of the data definition and preparation.

Similar thinking applies even in cases where message formats change more slowly, but multiple generations must be supported. Set-top boxes in cable TV also send an increasing number and variety of messages back to the providers of service to allow analysis of service quality, viewer activity and more. Newer generations of hardware produce a growing variety of messages, as can firmware upgrades. All these generations of messages must be supported for analysis, although privacy is a growing concern as increasingly personal information can be captured and transmitted[4].

This late-binding of meaning is a powerful and productive approach. (Note, however, that business-critical messages still require a higher level of checking and preparation, as found in traditional data warehousing.) Business users are empowered to take control of their own destiny and IT is relieved of the burden of trying to model and manage every incoming message. The business becomes sufficiently agile to identify and react to emerging trends in customer behavior.

IT, PHONE HOME

Take some thermostats intelligent enough to infer when to turn heating on and off. Add a home security system, maybe a refrigerator that can monitor food spoilage and consumption, connect them all via the Internet to an app on your smartphone, and you are joining the Internet of Things. Home management, supported by information technology, is emerging as a key market, if Google's recent purchase of Nest and hype at the 2014 Consumer Electronics Show is anything to go by. The foundation of intelligent home management is ongoing analysis of the events occurring in the home, an activity that requires messages from diverse and unrelated devices to be collected together, understood semantically, and combined in the right temporal sequence. Given the highly personal nature of the events involved, a thorough understanding of privacy and security issues is mandatory.

The solution required here combines aspects of the two previous use cases. First, we have the messaging data, the structure and content of which may be unknown in advance. As we've seen, an environment where data is stored as received, with context and usage determined at analysis time offers the benefit of agility and flexibility. Second, the event data from the different streams must be segmented into related activity sets, analogous to the sessionization of data familiar in web log analysis.

This final use case raises one further important need. Companies moving from purely physical world to the hyper-connected and IT-rich world of the Internet of Things are unlikely to have the highly skilled data scientists we find in the long-standing financial institutions or in Web pioneers. Simplicity and ease-of-use for business analysts is mandatory. Hadoop MapReduce programming skills are a major barrier to broader use of big data analytics. Business analysts today require a graphical interface to drive real business value at speed of thought.

THE MODERN LANDSCAPE AND HOW TO INHABIT IT

"Information about transactions, at some point in time, will become more important than the transactions themselves." [5]

The previous section described a number of modern analytical business needs. These are but a small sampling. As more and more novel analytical needs and opportunities emerge, it's vital that practitioners fully understand the new and changing data landscape and how to choose between the different IT architectural approaches that are beginning to appear.

THE NEW DATA LANDSCAPE

Given the unfortunate term *big data*, much of the focus is on data volumes. These are, indeed, large. By the end of 2013, according to IDC's Digital Universe study[6], there existed some 4,000 exabytes of data in the world. Furthermore, the volume is doubling every two years. Excluding video and image data, transient and other data of no obvious interest to the typical business analyst, IDC estimate that perhaps 25% of this data might be valuable if analyzed. However, of more fundamental importance is the recognition that the newer types of data have very different characteristics than traditional business data. The new data landscape consists of three distinct types of data/information.

Process-mediated data has long been generated and collected by traditional business processes such as buying an item or cashing a check. Such data is characterized by its well-defined and long-lasting meaning and structure, and its known, internal sourcing. It is the basis for the operational applications

and business intelligence / data warehousing architectures and technologies we have used for decades. Deeper thought about such data shows that it comes from two sources: (i) people entering it on keyboards and (ii) machines, such as ATMs, telephone exchanges, and more, generating it as a byproduct of human actions. As we've moved into the era of big data, both of these once-unseen, ultimate sources of data have become recognized as direct and important. *Human-sourced information*, the current focus for much of predictive analytics, comes directly from people—typing, photographing, speaking, videoing and more. While some is internally generated, much more is external, coming from social media sites and is very loosely defined in both structure and content. *Machine-generated data* is also growing rapidly, produced by computers in web logs, smartphones and a wide variety of devices in the Internet of Things.

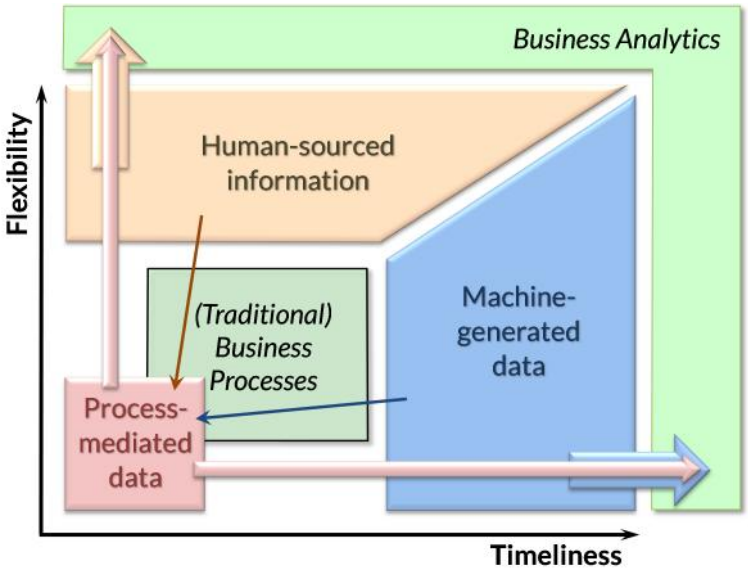

Figure 2: The tri-domain information model

The relationships between these three types of data are shown in Figure 2. This plots these data types against two key characteristics of data, timeliness and flexibility, which determine how we can process and analyze data. The two solid-line arrows show how business processes and applications traditionally create our managed representations of what is happening in and around the business. Human-sourced information and machine-generated data, previously unnoticed or uncollected, are processed into transactional and business intelligence databases with reduced flexibility and timeliness. However, as business accelerates and increases focus on external data, the more flexible and timely human-sourced information and machine-generated data become ever more important and are thus physically stored, often in non-relational stores such as Hadoop. The thicker arrows emphasize that modern business analytics must combine these three data types to ensure that the quality and consistency of process-mediated data is applied to the flexibility and timeliness of the new data sources. Further discussion of these important issues can be found in my book, *"Business unIntelligence—Insight and Intuition Beyond Analytics and Big Data"*[7].

Architectural Options for "Big Data" Analytics

Practitioners of business intelligence (BI) and analytics (BA) have long experience of the need to combine and cleanse data from multiple sources as a prelude to providing business user access. There are two broad project approaches (although these are often combined in a hybrid method today):

- *Enterprise integration:* typically implemented via a data warehouse, this approach emphasizes reconciliation and consolidation of cleansed data from multiple sources in anticipation of user needs
- *Departmentally driven:* usually involving the creation of many data marts, this approach starts from specific business needs and delivers focused data stores first

When used solely for process-mediated data, the technology of choice has long been a relational database or some variation thereof. As human-sourced information and machine-generated data increase in volume and, more importantly, value to the business, the two approaches and their hybrids

above remain valid, but the technology choices become more varied. In particular, the relative roles of relational technology and Hadoop must be evaluated.

The agility and late-binding model of Hadoop make it particularly suitable for human-sourced and machine-generated data, as well as for the speed of deployment usually demanded by departmentally driven business needs. Of course, process-mediated data can also be loaded and processed here. However, the lack of a formal and agreed data model makes proper enterprise integration problematical. Therefore, although we may have an extensive quantity and variety of data in Hadoop—sometimes called a "data lake"—the result is not equivalent to a classical data warehouse. It is, however, a particularly appropriate store for rapid-delivery, multi-sourced departmentally driven big data analytics, a modern and more extensive version of the independent data marts of the past.

Big data analytics poses the same dilemma as BI: how to integrate diverse data sources and deliver early business value.

In effect, what is required are the tools and techniques to create and use "data pools" (to continue the watery metaphor) in Hadoop, populated with all three types of data. Such tools and techniques must mitigate the structural roadblocks and ease of use issues, previously described, associated with the Hadoop environment.

Building an analytic environment for big data

"Knowing is not enough; we must act." [8]

As noted in the earlier list of needed improvements, an important aspect of big data analytics for many business needs is the combination of data from traditional and new sources. Hadoop can natively store all three types of information described in the previous section. External human-sourced information and machine-generated data is already commonly loaded directly here in any of a range of common formats, such as CSV, JSON, web logs and more, in fact in any file format desired. Process-mediated data, however, must first be extracted from its common relational database environments and before loading in a flat-file format. Careful analysis and modeling is needed to ensure that such extracts faithfully represent the actual state of the business. Such skills are often to be found in the ETL (extract-transform-load) teams responsible for traditional business intelligence systems, and should be applied here too.

With data now in a variety of file formats, the agility to process it any manner is assured. However, with little or no formal metadata to describe the content, potential users need to be able to define the meaning of the data before exploring and playing with it, in order to address improvement #2 above. Given analysts' familiarity with tabular data formats, such as spreadsheets and relational tables, a simple modeling and enhancement tool that overlays such a structure on the data is a useful approach. The aim is to separate the user from the underlying programming methods and allow access via visual (often SQL-based) business intelligence tools.

At the level of the physical data access and processing required to return results to the users, one approach is to translate the users' queries into MapReduce programs to run directly against the Hadoop file store. Another approach, which circumvents the batch nature of MapReduce, adds a columnar, compressed, in-memory appliance. This provides iterative, speed-of-thought analytics, in line with improvement #3, by offering an analytic data mart or data pool sourced from the Hadoop data lake.

Creating data pools that are directly related to key business entities, such as customers, transactions, products, and so on, provides users with highly intuitive and relevant data for their work. An important design aspect here is to ensure that the population and updating of data pools is generated and run in a fully automated manner in the background.

With such an analytic appliance in place, user-friendly, visual dashboards where the analyst interacts iteratively with the data can be easily constructed. This is analogous to the BI tool environment, operating on top or a relational database, with which most business analysts would be familiar today. This top layer provides for the fourth required improvement: a business-user-friendly analytical interface.

Conclusions

"Experience is the name everyone gives to his mistakes." [9]

The Web has developed over the past decade into both a pervasive business environment and a burgeoning social network. Now, and in coming years, the added Internet of Things is binding together the online and physical environments in a complex and heavily interdependent network. It is in this world that modern business must operate and strive for success. Such success depends implicitly on the ability of business analysts to examine the past, interpret the present and predict the future of the full scope of the business' activities and engagement with all its external counterparts. Only comprehensive big data analytics can offer this breadth of vision.

A number of new architectural approaches are emerging that offer different solutions to the challenges of big data analytics. This paper focuses on an approach that is driven explicitly by the early delivery of business value based on a single data storage environment and empowered business users.

We envisage an analytic environment where real business value can be delivered quickly and iteratively by business users without recourse to IT skills in programming or data preparation.

The Hadoop environment has emerged as the preferred platform for storing and processing much of the new big data, especially machine-generated and human-sourced. Traditional, process-mediated data can also be included, offering a potential single platform for business analytics. However, in its basic form, Hadoop poses significant challenges to users in terms of ease-of-use and ability to iterate quickly in analytic processes.

The hybrid analytic environment sketched above shows how these issues can be addressed, allowing real business value to be delivered quickly and iteratively without recourse to programming skills or IT data preparation. Analysts simply access the data, create a contextual model for analysis and immediately receive a graphical representation of the results, which they can explore with ease and speed to deliver the insights required by the business.

Dr. Barry Devlin is among the foremost authorities on business insight and one of the founders of data warehousing, having published the first architectural paper on the topic in 1988. With over 30 years of IT experience, including 20 years with IBM as a Distinguished Engineer, he is a widely respected analyst, consultant, lecturer and author of the seminal book, "Data Warehouse—from Architecture to Implementation" and numerous White Papers. His new book, "Business unIntelligence—Insight and Innovation Beyond Analytics and Big Data" was published in October 2013.

Barry is founder and principal of 9sight Consulting. He specializes in the human, organizational and IT implications of deep business insight solutions that combine operational, informational and collaborative environments. A regular contributor to BeyeNETWORK, TDWI and other publications, Barry is based in Cape Town, South Africa and operates worldwide.

---

[1] Marlowe, Christopher, *"Doctor Faustus"*, act 1, scene 1, (c.1592)

[2] Lewis, Michael, *"Moneyball—The Art of Winning an Unfair Game"*, W. W. Norton & Company, (2004)

[3] Brynjolfsson, Erik, Hitt, Lorin M. and Kim, Heekyung, *"Strength in Numbers: How Does Data-Driven Decision-making Affect Firm Performance?"* (April 22, 2011). Available at SSRN: http://ssrn.com/abstract=1819486

[4] Devlin, Barry, *"Intention in decision making"*, (November 2013), Blog at B-Eye-Network.com, http://bit.ly/1c4JXY8

[5] Wriston, Walter, former CEO and chairman of Citicorp, 1967-1984

[6] IDC, *"The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East"*, (2012), http://www.emc.com/leadership/digital-universe/index.htm

[7] Devlin, Barry, *"Business unIntelligence—Insight and Intuition Beyond Analytics and Big Data"*, Technics Publications, (2013), http://bit.ly/BunI-Technics

[8] Goethe, Johann Wolfgang von, *"Kritik der praktischen Vernunft"*, (1788)

[9] Wilde, Oscar, *"Lady Windermere's Fan"*, Act III, (1892)