Big Data Analytics

Disruptive Technologies for Changing the Game

Dr. Arvind Sathi

Big Data Analytics

Dr. Arvind Sathi

MC Press Online, LLC
Boise, ID 83703

Big Data Analytics:
Disruptive Technologies for Changing the Game
Dr. Arvind Sathi

First Edition
First Printing —October 2012

MC Press Online, LLC, 3695 W Quail Heights Court, Boise, ID 83703-3861 USA
Customer Service: Toll Free: (877) 226-5394; cust.srv@mcpressonline.com
Permissions and Special/Bulk Orders: mcbooks@mcpressonline.com

ISBN: 978-1-5S347-380-1

In memory of Professor Herbert Simon, who sparked my curiosity in qualitative reasoning

To Neena, Kinji, Kevin, and Conal for giving me the time, the encouragement, and the support in writing this book

About the Author

Dr. Arvind Sathi is the World Wide Communication Sector architect for the Information Agenda team at IBM®. Dr. Sathi received his Ph.D. in Business Administration from Carnegie Mellon University and worked under Nobel Prize winner Dr. Herbert A. Simon. Dr. Sathi is a seasoned professional with more than 20 years of leadership in Information Management architecture and delivery. His primary focus has been in creating visions and roadmaps for Advanced Analytics at leading IBM clients in telecommunications, media and entertainment, and energy and utilities organizations worldwide. He has conducted a number of workshops on Big Data assessment and roadmap development.

Prior to joining IBM, Dr. Sathi was the pioneer in developing knowledge-based solutions for CRM at Carnegie Group. At BearingPoint, he led the development of Enterprise Integration, MDM, and Operations Support Systems/Business Support Systems (OSS/BSS) solutions for the communications market and also developed horizontal solutions for communications, financial services, and public services. At IBM, Dr. Sathi has led several Information Management programs in MDM, data security, business intelligence, and related areas and has provided architecture oversight to IBM's strategic accounts. He has also delivered a number of workshops and presentations at industry conferences on technical subjects including MDM and data architecture, and he holds two patents in data masking. His first book, *Customer Experience Analytics,* was released by MC Press in October 2011. Dr. Sathi has also been a contributing author in a number of Data Governance books written by Sunil Soares.

Acknowledgements

First and foremost, I would like to acknowledge the hard work from the Information Agenda community in creating a world-class reference material. I have heavily referenced the material here, including the Business Maturity Model, the Solution Architecture framework, and a number of case studies. I would like to acknowledge Bob Keseley, Wayne Jensen, and Mick Fullwood for conceiving the ideas and organizing the reference material. I would like to acknowledge Tim Davis for his encouragement and for providing financial services examples. Jeff Jonas provided me with inspiration for experimenting with the ideas and provided me with much of the backbone for this book. The technical ideas were created with help from Beth Brownhill, Paul Christensen, Elizabeth Dial, Ram Dorairaj, Tommy Eunice, Rich Harken, Eberhard Hechler, Bob Johnston, Noman Mohammed, Peter Harrison, Daryl BC Peh, Steve Rigo, and Barry Rosen. The Dallas Global Solutions Center team—Christian Loza, Tom Slade, Mathews Thomas, and Janki Vora—provided valuable experimentations on the ideas. Mehul Shah, Emeline Tjan, Livio Ventura, Wolfgang Bosch, Steve Trigg, Don Bahash, and Jessica White have provided valuable business value analysis components in this book. I would also like to thank the Communication Sector Industry Consulting team—Ken Kralick, Dirk Michelsen, Tushar Mehta, Richard Lanahan, Rick Flamand, Linda Moss, and David Buck—for providing the opportunities, customers, and contributions to the Big Data Analytics solutions.

Next, I would like to acknowledge the excellent work from the IBM Business Analytics and Optimization consulting team. In particular, Adam Gersting, Joseph Baird, Anu Jain, Bruce Weiss, Aparna Betigeri, and John Held provided the ideas behind the business scenarios and use cases through their consulting activities. I would also like to thank Mark Holste for collaborations and brainstorms on these solutions.

The IBM Software Group product teams provided the much-needed case studies and product examples. I would like to thank Roger Rea, Dan Debrunner, and Vibhor Kumar for their help on the InfoSphere® Streams® product; Arun Manoharan and Patrick Welsh for their support in getting Vivisimo® information; Andrew Colby for help on the Netezza™ Analytics Engine; Shankar Venkataraman, Girish Venkatachaliah, and Karthik Hariharan for Big Insights®; Claudio Zancani for Optim™ Privacy; and Mike Zucker for SPSS®.

I worked closely with the practitioners as I studied Big Data business opportunities. This includes Anthony Behan, Ash Kanagat, Audrey Laird, Bob Weiss, Christine Twiford, Carmen Allen, Dave Dunmire, Doug Humfries, Duane Gabor, Gautam Shah, Girish Varma, Harpinder Singh Madan, Harsch Bhatnagar, Jay Praturi, Jessica Shah, Jim Hicks, Joshua Koran, Judith List, Kedrick Brown, Ken Babb, Lindsey Pardun, Mahesh Dalvi, Maureen Little, Neil Isford, Norbert Herman, Oliver Birch, Perry

McDonald, Philip Smolin, Piyush Sarwal, Ravi Kothari, Randy George, Raquel Katigbak, Richa Pandey, Rob Smith, Robert Segat, Sam King, Sankar Virdhagriswaran, Sara Philpott, Steve Cohen, Steve Teitzel, Sumit Chowdhury, Sumit Singh, Teresa Jacobs, Umadevi Reddy, Vasco Queiros, Vikas Pathuri, Von McConnell and Yoel Arditi. I am grateful for the insightful discussions and implementations in understanding business opportunities as well as current Big Data practices.

I would like to thank Cheryl Daugherty for her review of the book and Sunil Soares for inspiring me to write the book. Gaurav Deshpande did a fair amount of work behind the scenes to help me organize and fund the book. It was also Gaurav's inspiration to introduce the cartoon strip, which was eventually co-authored between the two of us. Susan Visser provided valuable help organizing the publication process. Katie Tipton provided valuable publication and editorial guidance.

Last, but not least, I would like to thank my wife Neena, my daughter Kinji, my son-in-law Kevin, and my son Conal for their inspiration, support, and editorial help.

Foreword

by Bob Keseley

We are seeing an unprecedented interest in Big Data Analytics around the globe. Top performers have declared themselves "Analytics driven" organizations. Savvy business and IT leaders are starting to leverage Big Data Analytics to drive substantial enhancements in their business models, partnerships, and business processes. While almost everyone is talking about Big Data at the tool or product level, successful organizations are focused on Big Data use cases and techniques that drive the greatest business value. They are focused on the "business" of Big Data Analytics. Arvind has taken the same perspective in *Big Data Analytics: Disruptive Technologies for Changing the Game.*

Over the past three years, our Information Agenda team has worked extensively helping organizations shape their Big Data Analytics strategies and solutions. Starting with the business is fundamental to the success of any organization. I am pleased to see a book starting with the business as the primary focus and exploring best practices across sales, marketing, customer service, and risk management, before linking them to the solutions and architectures that make it all possible. We hope you enjoy this book about evolving best practices and their impact on the competitive landscape. May it facilitate the right dialogue between your business and IT leaders.

Bob Keseley
Vice President, WW Information Agenda
IBM Software Group

Foreword

by Jeff Jonas

This book covers a number of Big Data use cases, architecture considerations, and the rise of emerging observation spaces (social, geospatial, etc.) and covers some of the thorny issues around data privacy. An organization's available observation space (data they can get their hands on within law and policy) is growing faster than their ability to make sense of it. As organizations struggle to keep up, they are being forced to reconsider what kind of infrastructure will be required to harness Big Data.

Going forward, organizations must be able to sense and respond to transactions happening now and must be able to deeply reflect over what has been observed—this deep reflection is a necessary activity to discover relevant weak signal and emerging patterns. Following fairly recent experiments involving how humans piece jigsaw puzzles together, I have witnessed the criticality of tightly coupling discovery from deep reflection right back into the real-time sense and respond analytics. In fact, as the feedback loop gets faster and tighter, it significantly enhances the discovery.

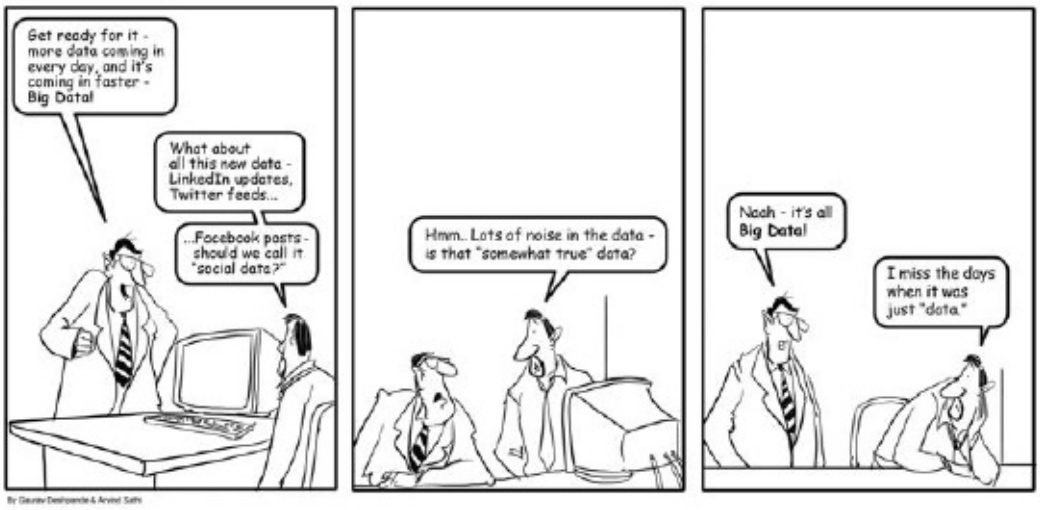
The organizations that figure out how to make sense of what they learn fast enough to do something about it while it is happening will be more competitive.

Jeff Jonas IBM Fellow and Chief Scientist IBM Entity Analytics

Contents

Big Data Analytics:

Disruptive Technologies for Changing the Game


Get ready for it - more data coming in every day, and it's coming in faster - Big Data!
What about all this new data - LinkedIn updates, Twitter feeds...
...Facebook posts - should we call it "social data?"
Hmm.. Lots of noise in the data - is that "somewhat true" data?
Nah - it's all Big Data!
I miss the days when it was just "data."

Chapter 1
Introduction

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.

1.1 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]

1.2 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 (Extract, Transform, Load) or ELT (Extract the data and Load it in the warehouse, then Transform 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 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.

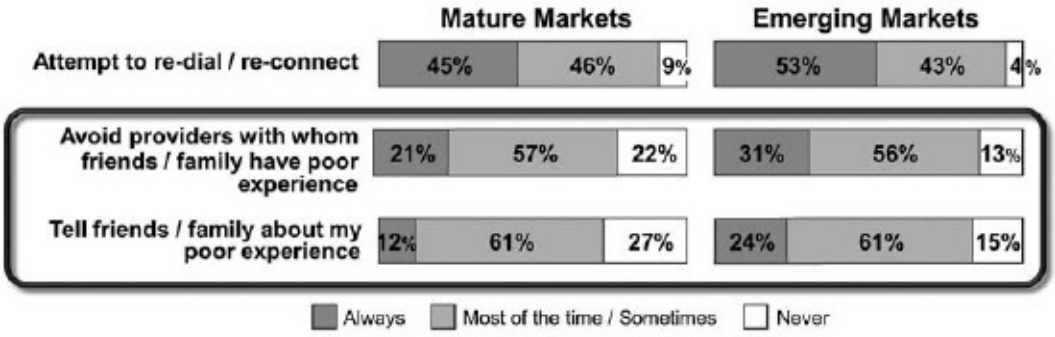


Figure 2.1: Behaviors in response to poor service quality experience

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.