CP7019 MANAGING BIG DATA

UNIT I UNDERSTANDING BIG DATA

What is big data – why big data – convergence of key trends – unstructured data – industry examples of big data – web analytics – big data and marketing – fraud and big data – risk and big data – credit risk management – big data and algorithmic trading – big data and healthcare – big data in medicine – advertising and big data – big data technologies – introduction to Hadoop – open source technologies – cloud and big data – mobile business intelligence – Crowd sourcing analytics – inter and trans firewall analytics

UNIT II NOSQL DATA MANAGEMENT

Introduction to NoSQL – aggregate data models – aggregates – key-value and document data models – relationships – graph databases – schemaless databases – materialized views – distribution models – sharding – master-slave replication – peer-peer replication – sharding and replication – consistency – relaxing consistency – version stamps – map-reduce – partitioning and combining – composing map-reduce calculations

UNIT III BASICS OF HADOOP

Data format – analyzing data with Hadoop – scaling out – Hadoop streaming – Hadoop pipes – design of Hadoop distributed file system (HDFS) – HDFS concepts – Java interface – data flow – Hadoop I/O – data integrity – compression – serialization – Avro – file-based data structures

UNIT IV MAPREDUCE APPLICATIONS

MapReduce workflows – unit tests with MRUnit – test data and local tests – anatomy of MapReduce job run – classic Map-reduce – YARN – failures in classic Map-reduce and YARN – job scheduling – shuffle and sort – task execution – MapReduce types – input formats – output formats

UNIT V HADOOP RELATED TOOLS

Hbase – data model and implementations – Hbase clients – Hbase examples – praxis.Cassandra – cassandra data model – cassandra examples – cassandra clients – Hadoop integration. Pig – Grunt – pig data model – Pig Latin – developing and testing Pig Latin scripts. Hive – data types and file formats – HiveQL data definition – HiveQL data manipulation – HiveQL queries.

REFERENCES:

1. Michael Minelli, Michelle Chambers, and Ambiga Dhiraj, "Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses", Wiley, 2013.

2. P. J. Sadalage and M. Fowler, "NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", Addison-Wesley Professional, 2012.

3. Tom White, "Hadoop: The Definitive Guide", Third Edition, O'Reilley, 2012.

4. Eric Sammer, "Hadoop Operations", O'Reilley, 2012.

5. E. Capriolo, D. Wampler, and J. Rutherglen, "Programming Hive", O'Reilley, 2012.

6. Lars George, "HBase: The Definitive Guide", O'Reilley, 2011.

7. Eben Hewitt, "Cassandra: The Definitive Guide", O'Reilley, 2010.

8. Alan Gates, "Programming Pig", O'Reilley, 2011.

Big Data

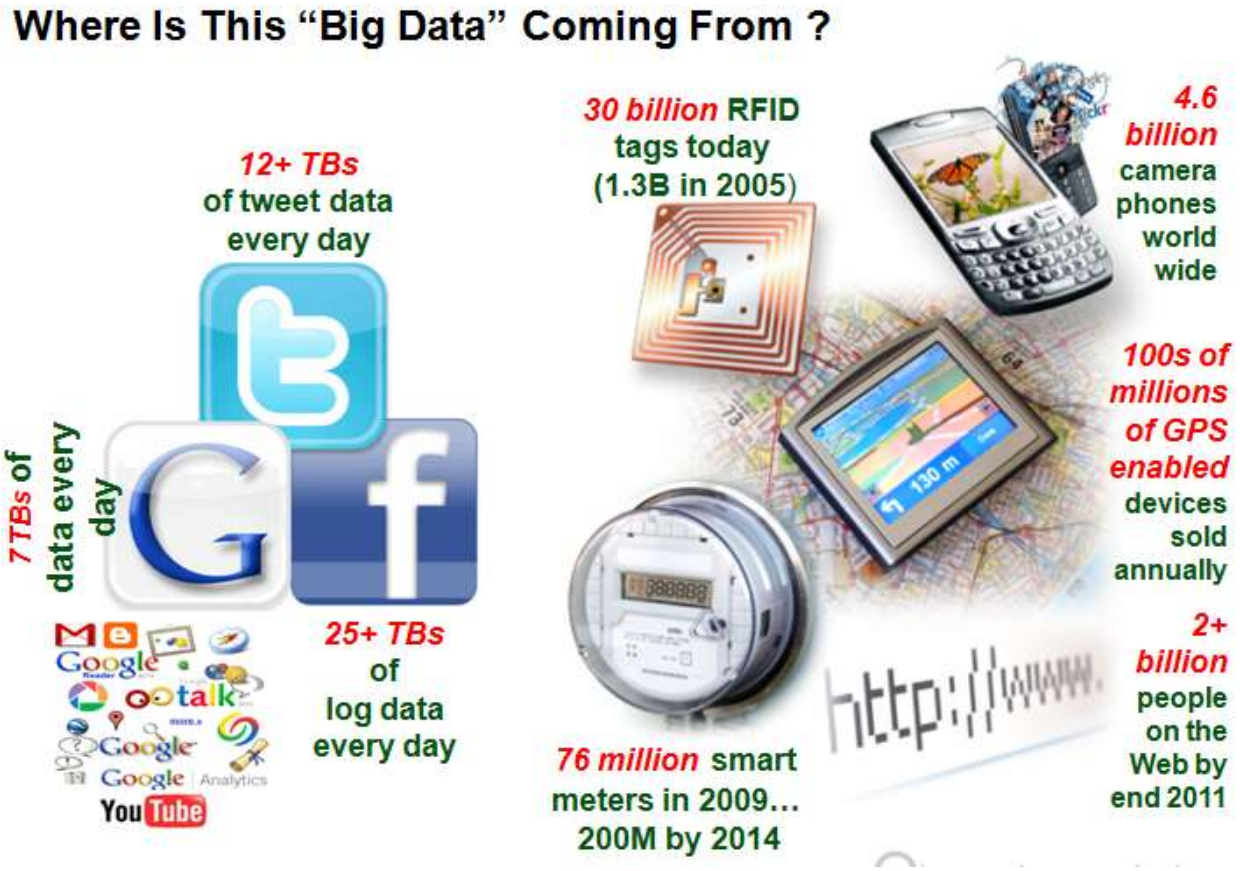
Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of traditional database architectures. In other words, **Big data** is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications. To gain value from this data, you must choose an alternative way to process it. Big Data is the next generation of data warehousing and business analytics and is poised to deliver top line revenues cost efficiently for enterprises. Big data is a popular term used to describe the exponential growth and availability of data, both structured and unstructured.

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. This data is **big data**.

Definition

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, create, manage, and process the data within a tolerable elapsed time

Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making.



Big data is often boiled down to a few varieties including social data, machine data, and transactional data. Social media data is providing remarkable insights to companies on consumer behavior and sentiment that can be integrated with CRM data for analysis, with 230 million tweets posted on Twitter per day, 2.7 billion Likes and comments added to Facebook every day, and 60 hours of video uploaded to YouTube every minute (this is what we mean by velocity of data). Machine data consists of information generated from industrial equipment, real-time data from sensors that track parts and monitor machinery (often also called the Internet of Things), and even web logs that track user behavior online. At arcplan client CERN, the largest particle physics research center in the world, the Large Hadron Collider (LHC) generates 40 terabytes of data every second during experiments. Regarding transactional data, large retailers and even B2B companies can generate multitudes of data on a regular basis considering that their transactions consist of one or many items, product IDs, prices, payment information, manufacturer and distributor data, and much more. Major retailers like Amazon.com, which posted $10B in sales in Q3 2011, and restaurants like US pizza chain Domino's, which serves over 1 million customers per day, are generating petabytes of transactional big data. The thing to note is that big data can resemble traditional structured data or unstructured, high frequency information.

Big Data Analytics

Big (and small) Data analytics is the process of examining data—typically of a variety of sources, types, volumes and / or complexities—to uncover hidden patterns, unknown correlations, and other useful information. The intent is to find business insights that were not previously possible or were missed, so that better decisions can be made.

Big Data analytics uses a wide variety of advanced analytics to provide

1. **Deeper insights.** Rather than looking at segments, classifications, regions, groups, or other summary levels you 'll have insights into all the individuals, all the products, all the parts, all the events, all the transactions, etc.
2. **Broader insights.** The world is complex. Operating a business in a global, connected economy is very complex given constantly evolving and changing conditions. As humans, we simplify conditions so we can process events and understand what is happening. But our best-laid plans often go astray because of the estimating or approximating. Big Data analytics takes into account all the data, including new data sources, to understand the complex, evolving, and interrelated conditions to produce more accurate insights.
3. **Frictionless actions**. Increased reliability and accuracy that will allow the deeper and broader insights to be automated into systematic actions.

Advanced Big data analytics

SQL Analytics	Descriptive Analytics	Data Mining	Predictive Analytics	Simulation	Optimization
• Count • Mean • OLAP	• Univariate distribution • Central tendency • Dispersion	• Association rules • Clustering • Feature extraction	• Classification • Regression • Forecasting • Spatial • Machine learning • Text analytics	• Monte Carlo • Agent-based modeling • Discrete event modeling	• Linear optimization • Non-linear optimization

Business Intelligence — Advanced Analytics

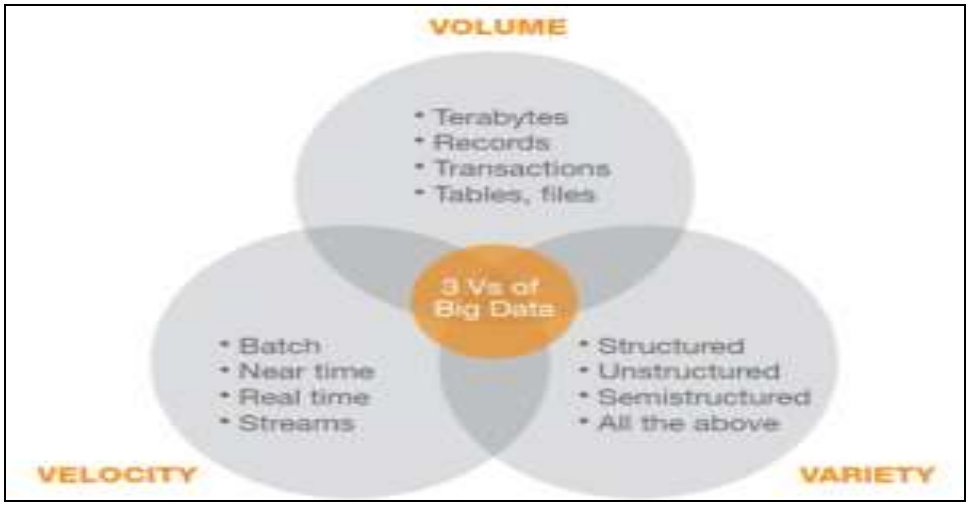
Big data analytic applications

Big Data Analytic Applications

	Improve Operational Efficiencies	Increase Revenues	Achieve Competitive Differentiation
Mature Analytic Applications	■ Supply chain optimization ■ Marketing campaign optimization	■ Algorithmic trading	■ In-house custom analytic applications
Maturing Analytic Applications	■ Portfolio optimization ■ Risk management ■ Next best offer	■ Ad targeting optimization ■ Yield optimization	■ In-house custom analytic applications
Emerging Analytic Applications	■ Chronic disease prediction and prevention	■ Customer churn prevention	■ Product design optimization ■ Design of experiments optimization ■ Brand ■ Product Market Targeting

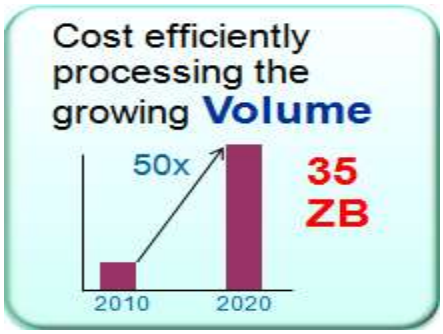
3 dimensions / characteristics of Big data

3Vs (volume, variety and velocity) are three defining properties or dimensions of big data. Volume refers to the amount of data, variety refers to the number of types of data and velocity refers to the speed of data processing.



Volume:

The size of available data has been growing at an increasing rate.



The volume of data is growing. Experts predict that the volume of data in the world will grow to 25 Zettabytes in 2020. That same phenomenon affects every business – their data is growing at the same exponential rate too.

This applies to companies and to individuals. A text file is a few kilo bytes, a sound file is a few mega bytes while a full length movie is a few giga bytes. More sources of data are added on continuous basis. For companies, in the old days, all data was generated internally by employees. Currently, the data is generated by employees, partners and customers. For a group of companies, the data is also generated by machines. For example, Hundreds of millions of smart phones send a variety of information to the network infrastructure. This data did not exist five years ago.

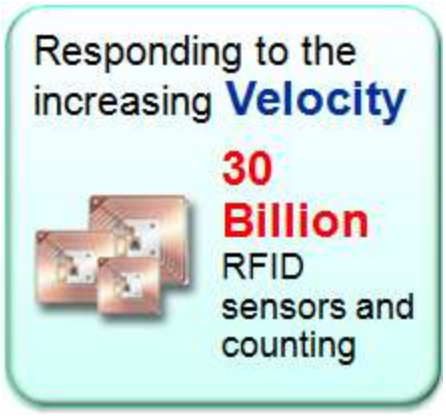
More sources of data with a larger size of data combine to increase the volume of data that has to be analyzed. This is a major issue for those looking to put that data to use instead of letting it just disappear.

Peta byte data sets are common these days and Exa byte is not far away.

Velocity:

Data is increasingly accelerating the velocity at which it is created and at which it is integrated. We have moved from batch to a real-time business.

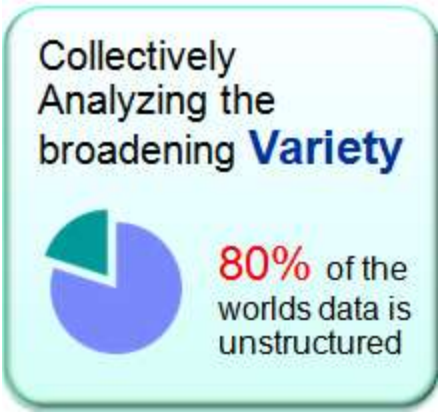
Initially, companies analyzed data using a batch process. One takes a chunk of data, submits a job to the server and waits for delivery of the result. That scheme works when the incoming data rate is slower than the batch-processing rate and when the result is useful despite the delay. With the new sources of data such as social and mobile applications, the batch process breaks down. The data is now streaming into the server in real time, in a continuous fashion and the result is only useful if the delay is very short.



Data comes at you at a record or a byte level, not always in bulk. And the demands of the business have increased as well – from an answer next week to an answer in a minute. In addition, the world is becoming more instrumented and interconnected. The volume of data streaming off those instruments is exponentially larger than it was even 2 years ago.

Variety:

Variety presents an equally difficult challenge. The growth in data sources has fuelled the growth in data types. In fact, 80% of the world's data is unstructured. Yet most traditional methods apply analytics only to structured information.



From excel tables and databases, data structure has changed to loose its structure and to add hundreds of formats. Pure text, photo, audio, video, web, GPS data, sensor data, relational data bases, documents, SMS, pdf, flash, etc. One no longer has control over

the input data format. Structure can no longer be imposed like in the past in order to keep control over the analysis. As new applications are introduced new data formats come to life.

The variety of data sources continues to increase. It includes

■ Internet data (i.e., click stream, social media, social networking links)
■ Primary research (i.e., surveys, experiments, observations)
■ Secondary research (i.e., competitive and marketplace data, industry reports, consumer data, business data)
■ Location data (i.e., mobile device data, geospatial data)
■ Image data (i.e., video, satellite image, surveillance)
■ Supply chain data (i.e., EDI, vendor catalogs and pricing, quality information)
■ Device data (i.e., sensors, PLCs, RF devices, LIMs, telemetry)

Why Big data?

1. Understanding and Targeting Customers

This is one of the biggest and most publicized areas of big data use today. Here, big data is used to better understand customers and their behaviors and preferences. Companies are keen to expand their traditional data sets with social media data, browser logs as well as text analytics and sensor data to get a more complete picture of their customers. The big objective, in many cases, is to create predictive models. You might remember the example of U.S. retailer Target, who is now able to very accurately predict when one of their customers will expect a baby. Using big data, Telecom companies can now better predict customer churn; Wal-Mart can predict what products will sell, and car insurance companies understand how well their customers actually drive. Even government election campaigns can be optimized using big data analytics.

2. Understanding and Optimizing Business Processes

Big data is also increasingly used to optimize business processes. Retailers are able to optimize their stock based on predictions generated from social media data, web search trends and weather forecasts. One particular business process that is seeing a lot of big data analytics is supply chain or delivery route optimization. Here, geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic data, etc. HR business processes are also being improved using big data analytics. This includes the optimization of talent acquisition – Moneyball style, as well as the measurement of company culture and staff engagement using big data tools

3. Personal Quantification and Performance Optimization

Big data is not just for companies and governments but also for all of us individually. We can now benefit from the data generated from wearable devices such as smart watches or smart bracelets. Take the Up band from Jawbone as an example: the armband collects data on our calorie consumption, activity levels, and our sleep

patterns. While it gives individuals rich insights, the real value is in analyzing the collective data. In Jawbone's case, the company now collects 60 years worth of sleep data every night. Analyzing such volumes of data will bring entirely new insights that it can feed back to individual users. The other area where we benefit from big data analytics is finding love - online this is. Most online dating sites apply big data tools and algorithms to find us the most appropriate matches.

4. Improving Healthcare and Public Health

The computing power of big data analytics enables us to decode entire DNA strings in minutes and will allow us to find new cures and better understand and predict disease patterns. Just think of what happens when all the individual data from smart watches and wearable devices can be used to apply it to millions of people and their various diseases. The clinical trials of the future won't be limited by small sample sizes but could potentially include everyone! Big data techniques are already being used to monitor babies in a specialist premature and sick baby unit. By recording and analyzing every heart beat and breathing pattern of every baby, the unit was able to develop algorithms that can now predict infections 24 hours before any physical symptoms appear. That way, the team can intervene early and save fragile babies in an environment where every hour counts. What's more, big data analytics allow us to monitor and predict the developments of epidemics and disease outbreaks. Integrating data from medical records with social media analytics enables us to monitor flu outbreaks in real-time, simply by listening to what people are saying, i.e. "Feeling rubbish today - in bed with a cold".

5. Improving Sports Performance

Most elite sports have now embraced big data analytics. We have the IBM SlamTracker tool for tennis tournaments; we use video analytics that track the performance of every player in a football or baseball game, and sensor technology in sports equipment such as basket balls or golf clubs allows us to get feedback (via smart phones and cloud servers) on our game and how to improve it. Many elite sports teams also track athletes outside of the sporting environment – using smart technology to track nutrition and sleep, as well as social media conversations to monitor emotional wellbeing.

6. Improving Science and Research

Science and research is currently being transformed by the new possibilities big data brings. Take, for example, CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator. Experiments to unlock the secrets of our universe – how it started and works - generate huge amounts of data. The CERN data center has 65,000 processors to analyze its 30 petabytes of data. However, it uses the computing powers of thousands of computers distributed across 150 data centers worldwide to analyze the data. Such computing powers can be leveraged to transform so many other areas of science and research.

7. Optimizing Machine and Device Performance

Big data analytics help machines and devices become smarter and more autonomous. For example, big data tools are used to operate Google's self-driving car. The Toyota Prius is fitted with cameras, GPS as well as powerful computers and sensors to safely drive on the road without the intervention of human beings. Big data tools are also used to optimize energy grids using data from smart meters. We can even use big data tools to optimize the performance of computers and data warehouses.

8. Improving Security and Law Enforcement.

Big data is applied heavily in improving security and enabling law enforcement. I am sure you are aware of the revelations that the National Security Agency (NSA) in the U.S. uses big data analytics to foil terrorist plots (and maybe spy on us). Others use big data techniques to detect and prevent cyber attacks. Police forces use big data tools to catch criminals and even predict criminal activity and credit card companies use big data use it to detect fraudulent transactions.

9. Improving and Optimizing Cities and Countries

Big data is used to improve many aspects of our cities and countries. For example, it allows cities to optimize traffic flows based on real time traffic information as well as social media and weather data. A number of cities are currently piloting big data analytics with the aim of turning themselves into Smart Cities, where the transport infrastructure and utility processes are all joined up. Where a bus would wait for a delayed train and where traffic signals predict traffic volumes and operate to minimize jams.

10. Financial Trading

My final category of big data application comes from financial trading. High-Frequency Trading (HFT) is an area where big data finds a lot of use today. Here, big data algorithms are used to make trading decisions. Today, the majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make, buy and sell decisions in split seconds.

Unstructured data

Unstructured data is information that either does not have a predefined data model and/or does not fit well into a relational database. Unstructured information is typically text heavy, but may contain data such as dates, numbers, and facts as well. The term semi-structured data is used to describe structured data that does not fit into a formal structure of data models. However, semi-structured data does contain tags that separate semantic elements, which includes the capability to enforce hierarchies within the data. The amount of data (all data, everywhere) is doubling every two years. Most new data is unstructured. Specifically, unstructured data represents almost 80 percent of new data, while structured data represents only 20 percent. Unstructured data tends to grow

exponentially, unlike structured data, which tends to grow in a more linear fashion. Unstructured data is vastly underutilized.

Mining Unstructured Data

Many organizations believe that their unstructured data stores include information that could help them make better business decisions. Unfortunately, it's often very difficult to analyze unstructured data. To help with the problem, organizations have turned to a number of different software solutions designed to search unstructured data and extract important information. The primary benefit of these tools is the ability to glean actionable information that can help a business succeed in a competitive environment.

Because the volume of unstructured data is growing so rapidly, many enterprises also turn to technological solutions to help them better manage and store their unstructured data. These can include hardware or software solutions that enable them to make the most efficient use of their available storage space.

Unstructured Data and Big Data

As mentioned above, unstructured data is the opposite of structured data. Structured data generally resides in a relational database, and as a result, it is sometimes called "relational data." This type of data can be easily mapped into pre-designed fields. For example, a database designer may set up fields for phone numbers, zip codes and credit card numbers that accept a certain number of digits. Structured data has been or can be placed in fields like these. By contrast, unstructured data is not relational and doesn't fit into these sorts of pre-defined data models.

In addition to structured and unstructured data, there's also a third category: semi-structured data. Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze. Examples of semi-structured data might include XML documents and NoSQL databases.

The term "big data" is closely associated with unstructured data. Big data refers to extremely large datasets that are difficult to analyze with traditional tools. Big data can include both structured and unstructured data, but IDC estimates that 90 percent of big data is unstructured data. Many of the tools designed to analyze big data can handle unstructured data.

Implementing Unstructured Data Management

Organizations use of variety of different software tools to help them organize and manage unstructured data. These can include the following:

- **Big data tools**: Software like Hadoop can process stores of both unstructured and structured data that are extremely large, very complex and changing rapidly.
- **Business intelligence software:** Also known as BI, this is a broad category of analytics, data mining, dashboards and reporting tools that help companies make sense of their structured and unstructured data for the purpose of making better business decisions.

- **Data integration tools:** These tools combine data from disparate sources so that they can be viewed or analyzed from a single application. They sometimes include the capability to unify structured and unstructured data.
- **Document management systems:** Also called "enterprise content management systems," a DMS can track, store and share unstructured data that is saved in the form of document files.
- **Information management solutions:** This type of software tracks structured and unstructured enterprise data throughout its lifecycle.
- **Search and indexing tools:** These tools retrieve information from unstructured data files such as documents, Web pages and photos.

Industry Examples of Big Data

http://www.safaribooksonline.com/library/view/big-data-big/9781118239155/xhtml/Chapter02.html

Web Analytics

Web analytics is the measurement, collection, analysis and reporting of web data for purposes of understanding and optimizing web usage. Web analytics is not just a tool for measuring web traffic but can be used as a tool for business and market research, and to assess and improve the effectiveness of a web site. The following are the some of the web analytic metrics: Hit, Page view, Visit / Session, First Visit / First Session, Repeat Visitor, New Visitor, Bounce Rate, Exit Rate, Page Time Viewed / Page Visibility Time / Page View Duration, Session Duration / Visit Duration. Average Page View Duration, and Click path etc.

Most people in the online publishing industry know how complex and onerous it could be to build an infrastructure to access and manage all the Internet data within their own IT department. Back in the day, IT departments would opt for a four-year project and millions of dollars to go that route. However, today this sector has built up an ecosystem of companies that spread the burden and allow others to benefit.

Avinash Kaushik believes there is one interesting paradigm shift that the Web mandates, that corporate information officers (CIOs) are and will continue to lose massive amounts of control over data and create large bureaucratic organizations whose only purpose is to support, collect, create, mash data, and be in the business of data “puking.” He believes such CIOs are “losing control in spades”:

> One of the interesting things that I had to grapple with as I embraced the Web and moved to the Web is that the primary way in which data gets collected, processed and stored, and accessed is actually at a third party. I did not have servers any more. I did not actually have implementations. I actually had to massively reduce the site of my implementation and data massaging and data serving and data banking team and rather massively expand the team that analyzes the data. This is the psychologically hard thing for me to do. When I was the BI person that's basically where most of the money of the company went. A little bit then went on analysts.

Kaushik's "elevator pitch" to businesses is that Big Data on the Web will completely transform a company's ability to understand the effectiveness of its marketing and hold its people accountable for the millions of dollars that they spend. It will also transform a company's ability to understand how its competitors are behaving. Kaushik believes that if you create a democracy in your organization where, rather than a few people making big decisions, the "organization is making hundreds and thousands of smart decisions every day and having the kind of impact on your company that would be impossible in the offline world. Not that the offline world is bad or anything, it's just that the way the data gets produced, assessed, and used on the Web is dramatically different."

Why use big data tools to analyse web analytics data?

Web event data is incredibly valuable

- It tells you *how* your customers actually behave (in lots of detail), and how that varies
    - Between different customers
    - For the same customers over time. (Seasonality, progress in customer journey)
    - How behaviour drives value
- It tells you *how* customers engage with you via your website / webapp
    - How that varies by different versions of your product
    - How improvements to your product drive increased customer satisfaction and lifetime value
- It tells you *how* customers and prospective customers engage with your different marketing campaigns and how that drives subsequent behaviour

Deriving value from web analytics data often involves very bespoke analytics

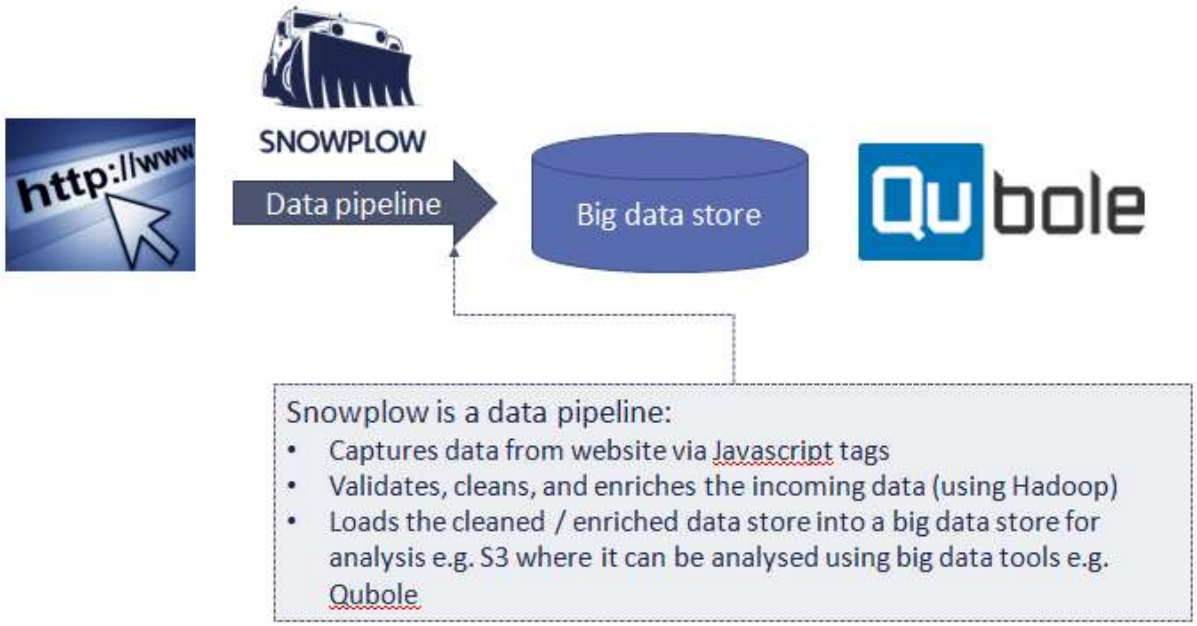
- The web is a rich and varied space! E.g.
    - Bank
    - Newspaper
    - Social network
    - Analytics application
    - Government organisation (e.g. tax office)
    - Retailer
    - Marketplace
- For each type of business you'd expect different :
    - Types of events, with different types of associated data

- Ecosystem of customers / partners with different types of relationships
- Product development cycle (and approach to product development)
- Types of business questions / priorities to inform how the data is analysed

Web analytics tools are good at delivering the standard reports that are common across different business types…

- Where does your traffic come from e.g.
  - Sessions by marketing campaign / referrer
  - Sessions by landing page
- Understanding events common across business types (page views, transactions, 'goals') e.g.
  - Page views per session
  - Page views per web page
  - Conversion rate by traffic source
  - Transaction value by traffic source
- Capturing contextual data common people browsing the web
  - Timestamps
  - Referer data
  - Web page data (e.g. page title, URL)
  - Browser data (e.g. type, plugins, language)
  - Operating system (e.g. type, timezone)
  - Hardware (e.g. mobile / tablet / desktop, screen resolution, colour depth)
- What is the impact of different ad campaigns and creative on the way users behave, subsequently? What is the return on that ad spend?
- How do visitors use social channels (Facebook / Twitter) to interact around video content? How can we predict which content will "go viral"?
- How do updates to our product change the "stickiness" of our service? ARPU? Does that vary by customer segment?

We built Snowplow to address those limitations and enable high value, bespoke analytics on web event data



Big Data and Marketing

Dan Springer, CEO of Responsys, defines the new school of marketing: "Today's consumers have changed. They've put down the newspaper, they fast forward through TV commercials, and they junk unsolicited email. Why? They have new options that better fit their digital lifestyle. They can choose which marketing messages they receive, when, where, and from whom. They prefer marketers who talk with them, not at them. New School marketers deliver what today's consumers want: relevant interactive communication across the digital power channels: email, mobile, social, display and the web."

Big Data and the New School of Marketing

Dan Springer, CEO of Responsys, defines the new school of marketing: "Today's consumers have changed. They've put down the newspaper, they fast forward through TV commercials, and they junk unsolicited email. Why? They have new options that better fit their digital lifestyle. They can choose which marketing messages they receive, when, where, and from whom. They prefer marketers who talk with them, not at them. New School marketers deliver what today's consumers want: relevant interactive communication across the digital power channels: email, mobile, social, display and the web."

Consumers Have Changed. So Must Marketers.

While using a lifecycle model is still the best way to approach marketing, today's new cross-channel customer is online, offline, captivated, distracted, satisfied, annoyed, vocal, or quiet at any given moment. The linear concept of a traditional funnel, or even

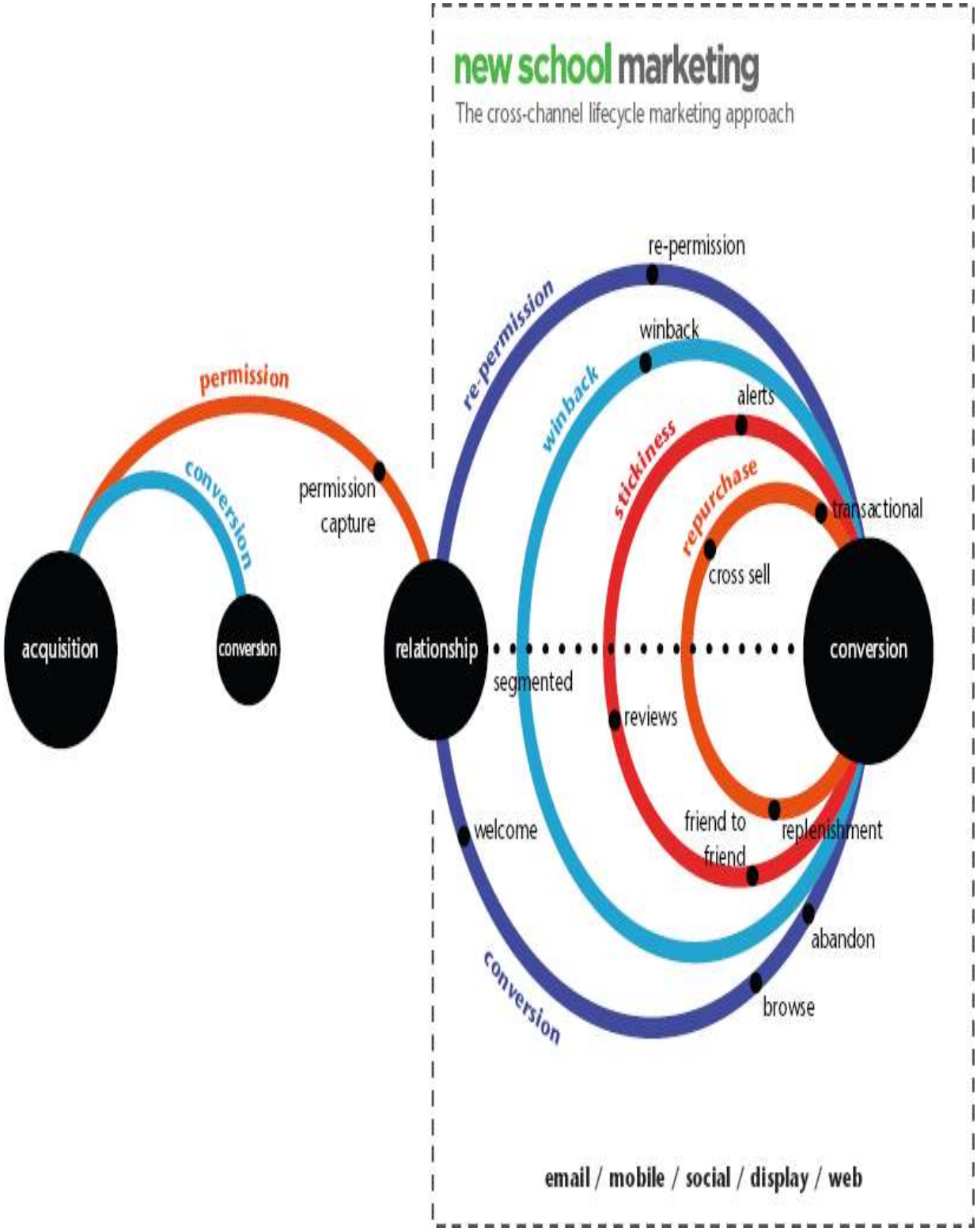
a succession of lifecycle "stages," is no longer a useful framework for planning marketing campaigns and programs.

Today's cross-channel consumer is more dynamic, informed, and unpredictable than ever. Marketers must be ready with relevant marketing at a moment's notice. Marketing to today's cross-channel consumer demands a more nimble, holistic approach, one in which customer behavior and preference data determine the content and timing—and delivery channel—of marketing messages. Marketing campaigns should be cohesive: content should be versioned and distributable across multiple channels. Marketers should collect holistic data profiles on consumers, including channel response and preference data, social footprint/area of influence, and more. Segmentation strategies should now take into account channel preferences.

Marketers can still drive conversions and revenue, based on their own needs, with targeted campaigns sent manually, but more of their marketing should be driven by—and sent via preferred channels in response to—individual customer behaviors and events. How can marketers plan for that? Permission, integration, and automation are the keys, along with a more practical lifecycle model designed to make every acquisition marketing investment result in conversion, after conversion, after conversion.

The Right Approach: Cross-Channel Lifecycle Marketing

Cross-Channel Lifecycle Marketing really starts with the capture of customer permission, contact information, and preferences for multiple channels. It also requires marketers to have the right integrated marketing and customer information systems, so that (1) they can have complete understanding of customers through stated preferences and observed behavior at any given time; and (2) they can automate and optimize their programs and processes throughout the customer lifecycle. Once marketers have that, they need a practical framework for planning marketing activities. Let's take a look at the various loops that guide marketing strategies and tactics in the Cross-Channel Lifecycle Marketing approach: conversion, repurchase, stickiness, win-back, and re-permission (see figure) New School of Marketing



Social and Affiliate Marketing

The Avon Lady has been doing it for over a century. Tupperware parties made buying plastics acceptable back in the 1940s. Word-of-mouth marketing has been the most powerful form or marketing since before the Internet was an idea in Tim Berners-Lee's mind and well before Mark Zuckerberg ever entered that now-famous Harvard dorm room.

It's really just a VERY big Tupperware party.
– Greg Doran, Founder and CEO of TipSpring

What Berners-Lee's and Zuckerberg's ground-breaking concepts and inventions do for word-of-mouth marketers is provide a backbone to bring proven marketing concepts outside of the living room to a scale never before seen.

The concept of affiliate marketing, or pay for performance marketing on the Internet is often credited to William J. Tobin, the founder of PC Flowers & Gifts. In the early 1990s Tobin was granted patents around the concept of an online business rewarding another site (an affiliate site) for each referred transaction or purchase. Amazon.com launched its own affiliate program in 1996 and middleman affiliate networks like Linkshare and Commission Junction emerged preceding the 1990s Internet boom, providing the tools and technology to allow any brand to put affiliate marketing practices to use. Today, one would be hard pressed to find a major brand that does not have a thriving affiliate program. Today, industry analysts estimate affiliate marketing to be a $3 billion industry. It's an industry that largely goes anonymous. Unlike email and banner advertising, affiliate marketing is a behind the scenes channel most consumers are unaware of.

In 2012, the emergence of the social web brings these concepts together. What only professional affiliate marketers could do prior to Facebook, Twitter, and Tumblr, now any consumer with a mouse can do. Couponmountain.com and other well know affiliate sites generate multimillion dollar yearly revenues for driving transactions for the merchants they promote. The expertise required to build, host, and run a business like Couponmountain.com is no longer needed when a consumer with zero technical or business background can now publish the same content simply by clicking "Update Status" or "Tweet." The barriers to enter the affiliate marketing industry as an affiliate no longer exist.

Above and beyond the removal of barriers the social web brings to affiliate marketing, it also brings into the mix the same concepts behind the Avon Lady and Tupperware party—product recommendations from a friend network. As many detailed studies have shown, most people trust a recommendation from the people they know. While professional affiliate marketing sites provide the aggregation of many merchant offers on one centralized site, they completely lack the concept of trusted source recommendations.

Using the backbone and publication tools created by companies like Facebook and Twitter, brands will soon find that rewarding their own consumers for their advocacy is a required piece of their overall digital marketing mix. What's old is new again. While not every company in the past had the resources or knowhow to build an army of Avon Ladies, today there is no excuse. The tools are available to them all and the scale is exponentially larger than ever before. Anyone can recommend a product through the click of a mouse. No more parties needed.

We also spoke with Niv Singer, Chief Technology Officer at Tracx, a social media intelligence software provider. Niv had quite a bit to say about the big data challenges faced in the social media realm and how it's impacting the way business is done today—and in the future.

As a result of the growing popularity and use of social media around the world and across nearly every demographic, the amount of user-generated content—or "big data"—created is immense, and continues growing exponentially. Millions of status updates, blog posts, photographs, and videos are shared every second. Successful organizations will not only need to identify the information relevant to their company and products—but also be able to dissect it, make sense of it, and respond to it—in real time and on a continuous basis, drawing business intelligence—or insights—that help predict likely future customer behavior. And if that sounds like a tall and complex order, that's because it is. Singer explains how this can be a challenging:

> It can sometimes be a real challenge to unify social profiles for a single user who may be using different names or handles on each of their social networks, so we've built an algorithm that combs through key factors including content of posts, and location, among others, to provide a very robust identity unification.

This brings us to the topic of influence and the age old debate of "who is an influencer?" To some brands, influence is measured purely by reach and to others, true influence is more of a function of quality and thoughtfulness of posts showing a real understanding of a given topic, and yet others gauge influence via social engagement or conversations. Because influence is so subjective, Singer believes the client should have the flexibility to sort influencers by any of these characteristics:

> Very intelligent software is required to parse all that social data to define things like the sentiment of a post. We believe using a system that's also able to learn over time what that sentiment means to a specific client or brand and then represent that data with increased levels of accuracy provides clients a way to "train" a social platform to measure sentiment more closely to the way they would be doing it manually themselves. We also know it's important for brands to be able to understand the demographic information of the individual driving social discussions around their brand such as gender, age, and geography so they can better understand their customers and better target campaigns and programs based on that knowledge.

In terms of geography, Singer explained that they are combining social check-in data from Facebook, Foursquare, and similar social sites and applications over maps to show brands at the country, state/region, state, and down to the street level where conversations are happening about their brand, products, or competitors. This capability enables marketers with better service or push coupons in real time, right when someone states a need, offering value, within steps from where they already are, which has immense potential to drive sales and brand loyalty.

These challenges are in the forefront of technology, but also require very creative people and solutions. Every component in the system must be able to be distributed across multiple servers that don't rely on each other. No single point of failure is allowed—the

data must therefore be replicated and stored on different machines, but should still be consistent. The data is later accessed in unpredictable ways. Singer likes to use an analogy to a book in a library:

> Finding a book by title or ISBN number is easy, even in a very big library. Finding, or counting, all the books written by specific authors is also relatively easy. It gets a little more complicated when we try to locate all the books written in a certain year, since we usually keep the books on shelves and sort them according to the author. If we need to count the number of books that contain the word "data" in their title written every year, it gets even more complicated. . . and when we need to locate all the books that contain the phrase "big data" in them, well, you can imagine.

Fundamentally, Singer doesn't view social data as a silo and, instead, believes that the real power comes in mining social data for business intelligence, not only for marketing, but also for customer support and sales. As a result, they've created a system from the ground up that was architected to be open. It's designed to be a data management system that just happens to be focused on managing unstructured social data, but we can easily integrate with other kinds of data sets. It was built with the expectation that social data would not live in an island, but would be pushed out to other applications to provide added business value and insights and that they would be pulling external data in.

This open approach like Singer is suggesting is extremely important because it enables businesses to take action with the data! Examples include integration with CRM systems like Salesforce.com and Microsoft Dynamics to enable companies to get a more holistic view of what's going with their clients by supplementing existing data sets that can be more static in nature with the social data set, which is more dynamic and real-time. Another example is integration with popular analytics platforms like Google Analytics and Omniture, so marketers can see a direct correlation and payoff of social campaigns through improved social sentiment or an increase in social conversations around their brand or product.

Where does Singer think this is all headed next? To the next big holy grail: an ability to take all this unstructured data and identify a customer's intent to buy:

> Customer intent is the big data challenge we're focused on solving. By applying intelligent algorithms and complex logic with very deep, real-time text analysis, we're able to group customers in to buckets such as awareness, opinion, consideration, preference and purchase. That ability let's marketers create unique messages and offers for people along each phase of the purchase process and lets sales more quickly identify qualified sales prospects.

One of Tracx customers is Attention, a heavily data-driven social media marketing agency also based in NYC. The Attention team uses the platform as the backbone of their social market research. Attention's CEO and Founder, Curtis Hougland, had this to say about Big Data's impact on marketing:

> Social media is the world's largest and purest focus group. Marketers now have the opportunity to mine social conversations for purchase intent and brand lift through Big Data. So, marketers can communicate with consumers when they are

> emotionally engaged, regardless of the channel. Since this data is captured in real-time, Big Data is coercing marketing organizations into moving more quickly to optimize media mix and message as a result of these insights. Since this data sheds light on all aspects of consumer behavior, companies are eliminating silos within the organization to align data to insight to prescription across channels, across media, and across the path to purchase. The days of Don Draper are over, replaced by a union of creative and quant.

The capability to understand a customer's intent that Hougland and Singer are referring to is not only good for the corporations; it's also a helpful capability for the client too. Think about it, most people communicate socially because they are looking to share, complain, or find something they need. Wouldn't it be great if someone was listening and knew your intent so that they can provide customer assistance or get you what you need for the best price?

Fraud and Big Data

Fraud is intentional deception made for personal gain or to damage another individual. One of the most common forms of fraudulent activity is credit card fraud. The credit card fraud rate in United States and other countries is increasing and increased 87 percent in 2011 culminating in an aggregate fraud loss of $6 billion.
However, despite the significant increase in incidence, total cost of credit card fraud increased only 20 percent. The comparatively small rise in total cost can be attributed to an increasing sophistication of fraud detection mechanisms. According to the Capgemini Financial Services Team:

Even though fraud detection is improving, the rate of incidents is rising. This means banks need more proactive approaches to prevent fraud. While issuers' investments in detection and resolution has resulted in an influx of customer-facing tools and falling average detection times among credit card fraud victims, the rising incidence rate indicates that credit card issuers should prioritize preventing fraud.

Social media and mobile phones are forming the new frontiers for fraud. Despite warnings that social networks are a great resource for fraudsters, consumers are still sharing a significant amount of personal information frequently used to authenticate a consumer's identity. Those with public profiles (those visible to everyone) were more likely to expose this personal information.

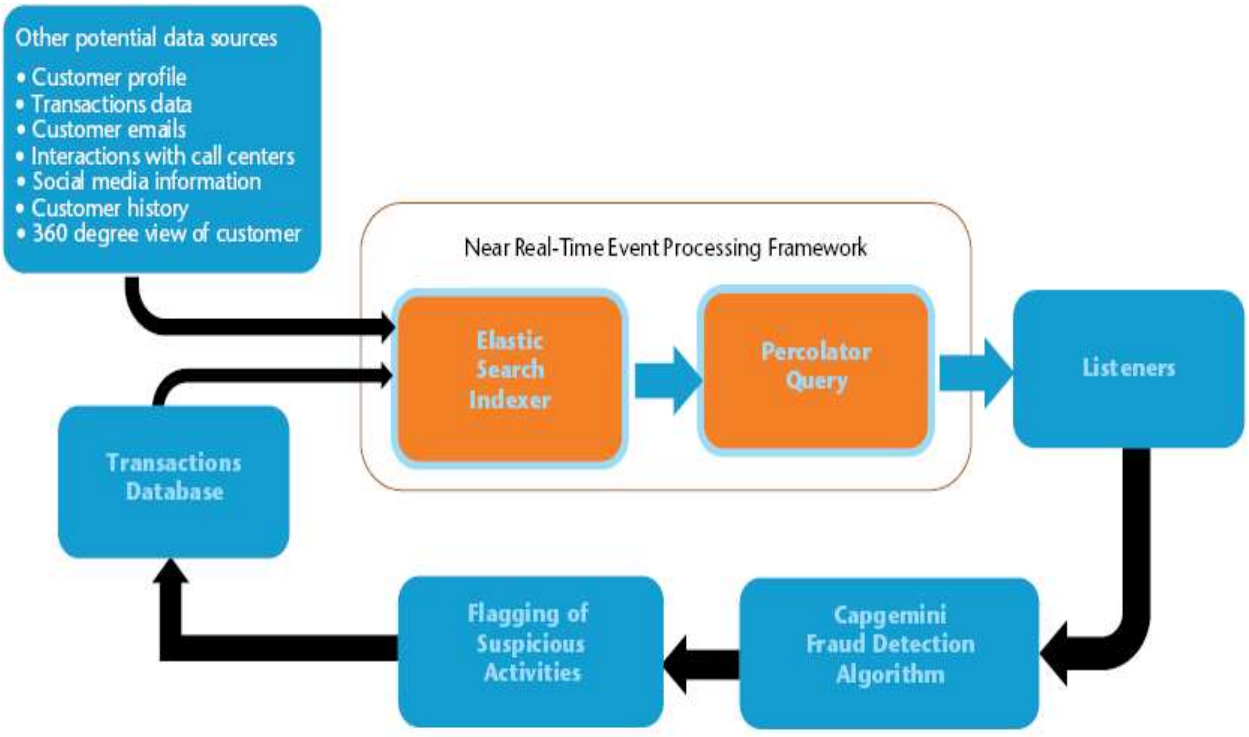
In order to prevent the fraud, credit card transactions are monitored and checked in near real time. If the checks identify pattern inconsistencies and suspicious activity, the transaction is identified for review and escalation.

The Capgemini Financial Services team believes that due to the nature of data streams and processing required, Big Data technologies provide an optimal technology solution based on the following three Vs:

1. High volume. Years of customer records and transactions (150 billion+ records per year)
2. High velocity. Dynamic transactions and social media information
3. High variety. Social media plus other unstructured data such as customer emails, call center conversations, as well as transactional structured data

Capgemini's new fraud Big Data initiative focuses on flagging the suspicious credit card transactions to prevent fraud in near real-time via multi-attribute monitoring. Real-time inputs involving transaction data and customers records are monitored via validity checks and detection rules. Pattern recognition is performed against the data to score and weight individual transactions across each of the rules and scoring dimensions. A cumulative score is then calculated for each transaction record and compared against thresholds to decide if the transaction is potentially suspicious or not.

The Capgemini team pointed out that they use an open-source weapon named Elastic Search, which is a distributed, free/open-source search server based on Apache Lucene shown in following figure. It can be used to search all kind of documents at near real-time. They use the tool to index new transactions, which are sourced in real-time, which allows analytics to run in a distributed fashion utilizing the data specific to the index. Using this tool, large historical data sets can be used in conjunction with real-time data to identify deviations from typical payment patterns. This Big Data component allows overall historical patterns to be compared and contrasted, and allows the number of attributes and characteristics about consumer behavior to be very wide, with little impact on overall performance.



Innovative near-real time event processing framework can be used in multiple applications, which improves processing speed

Capgemini

Once the transaction data has been processed, the percolator query then performs the functioning of identifying new transactions that have raised profiles. *Percolator* is a system for incrementally processing updates to large data sets. Percolator is the technology that Google used in building the index—that links keywords and URLs—used to answer searches on the Google page.

Percolator query can handle both structured and unstructured data. This provides scalability to the event processing framework, and allows specific suspicious transactions to be enriched with additional unstructured information—phone location/geospatial records, customer travel schedules, and so on. This ability to enrich the transaction further can reduce false-positives and increase the experience of the customer while redirecting fraud efforts to actual instances of suspicious activity.

Another approach to solving fraud with Big Data is social network analysis (SNA). SNA is the precise analysis of social networks. Social network analysis views social relationships and makes assumptions. SNA could reveal all individuals involved in fraudulent activity, from perpetrators to their associates, and understand their relationships and behaviors to identify a bust out fraud case.

According to a recent article in bankersonline.com posted by Experian, "bust out" is a hybrid credit and fraud problem and the scheme is typically defined by the following behavior:

- The account in question is delinquent or charged-off.
- The balance is close to or over the limit.
- One or more payments have been returned.
- The customer cannot be located.
- The above conditions exist with more than one account and/or financial institution.

There are some Big Data solutions in the market like SAS's SNA solution, which helps institutions and goes beyond individual and account views to analyze all related activities and relationships at a network dimension. The network dimension allows you to visualize social networks and see previously hidden connections and relationships, which potentially could be a group of fraudsters. Obviously there are huge reams of data involved behind the scene, but the key to SNA solutions like SAS's is the visualization techniques for users to easily engage and take action.

Risk and Big Data

Many of the world's top analytics professionals work in risk management. It would be an understatement to say that risk management is data-driven—without advanced data analytics, modern risk management would simply not exist. The two most common types of risk management are credit risk management and market risk management. A third type of risk, operational risk management, isn't as common as credit and market risk.

The tactics for risk professionals typically include avoiding risk, reducing the negative effect or probability of risk, or accepting some or all of the potential consequences in exchange for a potential upside gain.

Credit risk analytics focus on past credit behaviors to predict the likelihood that a borrower will default on any type of debt by failing to make payments which they obligated to do. For example, "Is this person likely to default on their $300,000 mortgage?"

Market risk analytics focus on understanding the likelihood that the value of a portfolio will decrease due to the change in stock prices, interest rates, foreign exchange rates, and commodity prices. For example, "Should we sell this holding if the price drops another 10 percent?"

Credit Risk Management

Credit risk management is a critical function that spans a diversity of businesses across a wide range of industries. Ori Peled is the American Product Leader for MasterCard Advisors Risk & Marketing Solutions. He brings several years of information services experience in his current role with MasterCard and having served in various product development capacities at Dun & Bradstreet. Peled shares his insight with us on credit risk:

> Whether you're a small B2B regional plastics manufacturer or a large global consumer financial institution, the underlying credit risk principles are essentially the same: driving the business using the optimal balance of risk and reward.

Traditionally, credit risk management was rooted in the philosophy of minimizing losses. However, over time, credit risk professionals and business leaders came to understand that there are acceptable levels of risk that can boost profitability beyond what would normally have been achieved by simply focusing on avoiding write-offs. The shift to the more profitable credit risk management approach has been aided in large part to an ever-expanding availability of data, tools, and advanced analytics.

Credit risk professionals are stakeholders in key decisions that address all aspects of a business, from finding new and profitable customers to maintaining and growing relationships with existing customers. Maximizing the risk and reward opportunities requires that risk managers understand their customer portfolio, allowing them to leverage a consistent credit approach while acknowledging that you can't treat every customer the same.

As businesses grow, what starts out as a manual and judgmental process of making credit decisions gives way to a more structured and increasingly automated process in which data-driven decisions becomes the core. Decisions that impact not only revenue but also operational costs like staffing levels of customer support representatives or collections agents.

The vast amount of both qualitative and quantitative information available to credit risk professionals can be overwhelming to digest and can slow down a process with potential sales at risk. With advanced analytical tools, these abundant and complex data sources can be distilled into simple solutions that provide actionable insights and are relatively easy to implement. As an example, credit scoring solutions allow risk managers to apply a credit policy more efficiently and consistently across the business. Scoring solutions can take various data sources and produce a score that computes the odds of a customer behaving in a specific way. Traditional scoring methods focus on predicting the likelihood of delinquency or bankruptcy but additional scoring solutions can also help companies identify the profitability potential of customers or from a marketing perspective, the propensity to spend. Additionally, companies are leveraging and combining multiple analytical solutions at the same time—this could be a combination of proprietary scoring solutions and third party scoring like those provided by specialized analytics providers and the consumer and commercial bureaus (e.g., Experian, Equifax, D&B, etc.).

As you look across the credit risk management lifecycle, rich data sources and advanced analytics are instrumental throughout. From a customer acquisition perspective, credit risk managers decide whether to extend credit and how much. Lacking any previous experience with the prospect, they depend heavily on third-party credit reports and scores and may assist marketing organizations in employing customized look-alike models to help identify prospective best customers.

From an existing customer standpoint, the focus shifts to ongoing account management and retaining profitable accounts. This requires periodic customer risk assessments that influence key decisions on credit line increases and decreases. Again, advanced analytical solutions come into play, especially in larger organizations where the volume of accounts dictate a need for automated decisioning solutions that leverage behavior scores and other data sources. Continuous monitoring of an existing portfolio can also help credit risk managers forecast expected losses and better manage their collections efforts. Advanced analytics in the collections phase can help identify customers most likely to pay or even respond to different collection strategies and approaches.

The future of credit risk management will continue to change as we leverage new data sources emanating from a highly digital and mobile world. As an example, social media and cell phone usage data are opening up new opportunities to uncover customer behavior insights that can be used for credit decisioning. This is especially relevant in the parts of the world where a majority of the population is unbanked and traditional bureau data is unavailable.

As the following figure illustrates, there are four critical parts of the typical credit risk framework: planning, customer acquisition, account management, and collections. All four parts are handled in unique ways through the use of Big Data.

Credit Risk Framework

Big Data and Algorithmic Trading

Partha Sen is the CEO of Fuzzy Logix, a company that specializes in high-performance, cross platform in database, and GPU (graphics processing unit) analytics. Sen spent over 15 years as a quantitative analyst in the financial services industry. Over the course of his career, he developed over 700 highly parallelized algorithms in C/C. He, along with a team of very talented quantitative professionals, now leverages his formidable expertise to help customers across a number of industries.

Sen has seen a significant shift in the use of data in the financial services industry over the past decade. "Financial institutions," he says, "particularly investment banks, have been at the forefront of applying analytics for risk management, proprietary trading, and portfolio management."

As most of you know, many investment banks use algorithmic trading, a highly sophisticated set of processes in which "insights" are made "actionable" via automated "decisions." Algorithmic trading relies on sophisticated mathematics to determine buy and sell orders for equities, commodities, interest rate and foreign exchange rates, derivatives, and fixed income instruments at blinding speed. A key component of algorithmic trading is determining return and the risk of each potential trade, and then making a decision to buy or sell. Quantitative risk analysts help banks develop trading rules and implement these rules using modern technology. Algorithmic trading involves a huge number of transactions with complex interdependent data, and every millisecond matters.

It's fair to say that these days banks focus more closely on market risk today than ever before. Market risk is basically the risk due to a fluctuation in the value of assets in the

marketplace. For a given portfolio, you are trying to determine the probability that the value of the portfolio will fall within a certain threshold within five days, within seven days, within one month. With asset volatilities as high as they have been observed in the last few years, a lot of stress is being put on market risk. Sophisticated methods for managing market risk depend very heavily of modern technology.

Apart from investment banks, corporate and retail banks also rely very heavily on quantitative techniques. Two areas that readily come to mind are marketing, where they solicit households for financial products like credit cards, and credit risk management, where banks try to understand the probability that borrowers will default on loan. The models used in these areas for future outcomes are created with huge number of variables. For example, a model of the default risk for credit cards could be influenced by demographic factors, whether people have a job or not, what is the growth in the economy, and interest rates. There can be hundreds of factors or variables for each credit card. A typical retail bank will evaluate somewhere north of 5,000 factors for one given model to establish or calculate the probability of each of the borrowers defaulting. The number of calculations just for the risk factor can easily climb into billions of calculations being performed to calculate risk for a portfolio.

Crunching Through Complex Interrelated Data

In a frictionless economy, time is the critical driver to gain and sustain a competitive advantage. Every second, or more correctly, every millisecond counts today. Banks have graduated from daily evaluation of risk to intra-day risk evaluations.

"Risk management on a daily basis is a thing of the past because there are higher volatilities," says Sen of the marketplace today. "You can see the volatilities unfold in the marketplace when there is an event or an announcement about macro events—unemployment rate or interest rate going up or down or important geo-political events. News often causes uncertainty in the minds of investors and thus volatilities in financial markets increase. When volatility goes up during the course of a day or trading session, it has instantaneous effect on the value of financial instruments."

For market risk, the data explodes very quickly. Today, the portfolios being evaluated are quite large and include multiple financial instruments. For example, an investment bank will have a portfolio of equities, along with a portfolio of options—both calls and puts on equities. In addition, there will be foreign exchange trades, a portfolio of interest rate instruments, and interest rate derivatives. Some banks may have more complex products in their portfolios like exotic options—Bermudans, Asian options, digital options, and such.

An often used convention, according to Sen, is to calculate the mark-to-market value of the underlying financial instruments and thus calculate the risks. To show how this works, he gave us this example:

> [L]et's say that you have an investment bank that has an equity derivative in its portfolio and the value of this derivative will change in the future. That change is

going to be influenced by the spot price of the underlying stock, the volatility of that stock, interest rate, and time to maturity.

The convention is that every day you take the value of that derivative and you perform scenario analysis over a time horizon—the next 30 days—to determine what will be the value. Will it be $3.00 instead of $10.00 that the bank has on the books? Or will it be $13.00? In this type of scenario analysis, you create multiple scenarios and price the derivative against all the scenarios. Even for a single instrument it is possible to have hundreds of thousands of scenarios. Naturally, when you have hundreds of thousands of equity derivatives in your portfolio different equities, different maturities, and different strike prices, the problem of scenario analysis becomes very complex.

Intraday Risk Analytics, a Constant Flow of Big Data

To maintain competitive advantage, banks need to continuously evaluate their models, including the performance of the production models, and also continuously try to build new models to incorporate new variables with new and evolving macroeconomic conditions in a faster way. Banks have also moved from daily risk management to intraday risk management. Intraday risk management involves pricing the entire portfolio and calculating the risk limits of each of the counter-parties within the bank's portfolio. The problem gets very complex and computationally intensive.

Let's take an example of intraday risk evaluation of equities. The potential changes within a day include the spot price, the volatility of the underlying equity, and the risk free rate. If we do some basic scenario analysis—say 100 risk-free rate scenarios that could manifest themselves during the course of the day—that means calculating 100 scenarios for the spot price of the equity during the course of the day, 100 scenarios for volatility during the course of the day, and 100 scenarios for risk-free rate during the course of the day. For the bank to do their basic scenario analysis, it takes a million calculations for determining the value at risk for just that one instrument. And it must happen fast enough so that risk limits on the entire portfolio can be evaluated several times during the course of the day

"The only option that I can currently see," says Sen, "is to be solving these problems using a very large amount of parallelized computations and that is only possibly doable with GPUs. Using this type of high performance compute technology, we can determine value at risk for 100 million scenarios in less than ten milliseconds using just one of these GPU cards. The real power comes into play when you use multiple cards and parallelize the entire workload. That's when you can do scenario analysis across your entire portfolio in about 15 minutes."

Calculating Risk in Marketing

While risk analytics is used for risk management, banks are using risk predictive analytics for marketing as well. For example, when a bank scores its customers and prospects for credit card solicitations, it will use some risk management tools as well. In addition to determining who has a high likelihood of responding to promotional offers, the bank will want to consider the underlying risk for each of the prospects to whom

the solicitations are being sent. Without taking into account risk profiles of individuals, bank promotion responses can result in customers with a higher risk profile.

"One of the challenges for retail banks," according to Sen, "is to score such large numbers of people for its marketing initiatives"

Given people's exact situation, you have to determine what are the right products to promote. Maybe somebody has a home loan but doesn't have a credit card or debit card. In addition, you also have to score your existing customers to determine the borrowers whose probabilities of not paying on their credit card, or on the mortgage, is rising. Once you know who these potential defaulters could be, you can see what you can do to mitigate risk of default. The sheer volume of the population that you have to score compounds the problem. You have to score it quickly because you have to take action promptly be it promotion or risk mitigation.

Other Industries Benefit from Financial Services' Risk Experience

Outside of financial services there are other industries that can benefit from this work, such as retail, media, and telecommunications. They are following suit to include evaluation of adverse select in their promotional offers.

While the adoption of analytics has been slower in other industries, momentum is starting to build around Big Data analytics. Marketing is an area that is clearly more mature in terms of adopting analytics in the areas of for marketing campaign management, targeted micromarketing (sending of different offers to different types of people depending on their likelihood to buy), and market basket analysis, which indicates what people buy together and more.

For example, in retail, forecasting is a key area where analytics is being applied. Customer churn analysis has been used by banks to determine who is likely to cancel their credit card or account. This is the same technique that is being used by telecommunication companies and retailers today to determine customer defection. Churn is also a factor used in determining customer lifetime value. Customer lifetime value indicates how much money a firm can make over the customer's lifetime, that is the period of association of the customer with the firm. Companies typically use the customer lifetime value to segment their customers and determine which are the customers to focus on.

The insurance industry today uses actuarial models for estimating losses. However, the emerging trend is to use Monte-Carlo simulations for estimating potential losses in insurance portfolios. These computationally complex models require a large footprint of hardware in order to handle the massive calculations. The cost of acquiring and maintaining such hardware sometimes becomes the impediment to adoption of analytics in enterprises. "With the advent of GPU technology, however, that will change," says Sen.

Another use for Big Data analytics in banks is identifying manipulative behavior or fraudulent activities in real-time so that you can mitigate or penalize the behavior

immediately. For this, you have to dig through the voluminous transactions and find the patterns quickly.

"It's always good to catch the culprit but by that time—five years or five days later—a lot of honest players have been disadvantaged." And what can you do? "Well, not much. says Sen. However, Sen observes, "if you can catch it [the fraud] while it's happening, then you can focus on immediate enforcement so the manipulators can't twist the prices in the market to negatively impact the retail investor or even the institutional investor, who is a fair player. By quickly catching and correcting this market manipulative behavior you're creating a fair trading platform."

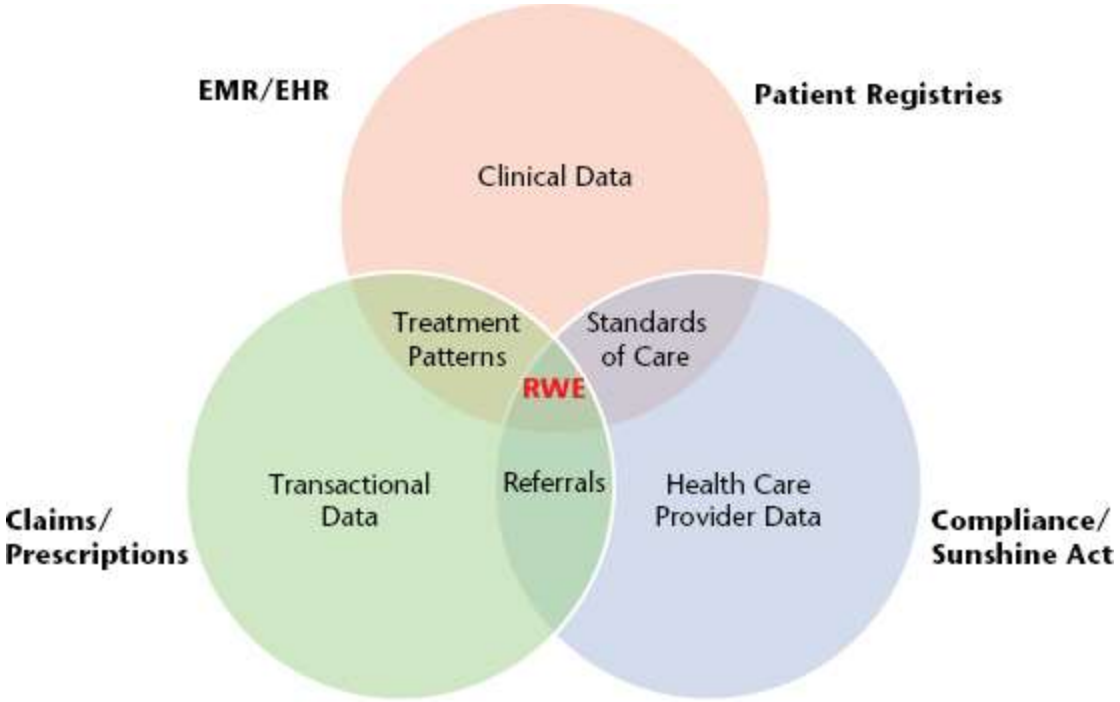
Big Data and Advances in Health Care

So far, most of our conversation around Big Data has focused on activities such as marketing, offer optimization, budget planning, business process management, supply chain management, anti-money laundering, fraud monitoring, and risk management.

Let's be honest. The average person can't relate their personal life to all of those topics. So here's a topic that has an impact on everyone's life: health care.

Big Data promises an enormous revolution in health care, with important advancements in everything from the management of chronic disease to the delivery of personalized medicine. In addition to saving and improving lives, Big Data has the potential to transform the entire health care system by replacing guesswork and intuition with objective, data-driven science see the following figure

Data in the World of Health Care



The health care industry is now awash in data: from biological data such as gene expression, Special Needs Plans (SNPs), proteomics, metabolomics to, more recently, next-generation gene sequence data.

This exponential growth in data is further fueled by the digitization of patient-level data: stored in Electronic Health Records (EHRs) and Health Information Exchanges (HIEs), enhanced with data from imaging and test results, medical and prescription claims, and personal health devices.

The U.S. health care system is increasingly challenged by issues of cost and access to quality care. Payers, producers, and providers are each attempting to realize improved treatment outcomes and effective benefits for patients within a disconnected health care framework. Historically, these health care ecosystem stakeholders tend to work at cross purposes with other members of the health care value chain. High levels of variability and ambiguity across these individual approaches increase costs, reduce overall effectiveness, and impede the performance of the health care system as a whole.

Recent approaches to health care reform attempt to improve access to health care by increasing government subsidies and reducing the ranks of the uninsured. One outcome of the recently passed Accountable Care Act is a revitalized focus on cost containment and the creation of quantitative proofs of economic benefit by payers, producers, and providers. A more interesting unintended consequence is an opportunity for these health care stakeholders to set aside historical differences and create a combined counterbalance to potential regulatory burdens established, without the input of the actual industry the government is setting out to regulate. This "the enemy of my enemy is my friend" mentality has created an urgent motivation for payers, producers, and, to a lesser extent, providers, to create a new health care information value chain derived from a common health care analytics approach.

The health care system is facing severe economic, effectiveness, and quality challenges. These external factors are forcing a transformation of the pharmaceutical business model.

Health care challenges are forcing the pharmaceutical business model to undergo rapid change. Our industry is moving from a traditional model built on regulatory approval and settling of claims, to one of medical evidence and proving economic effectiveness through improved analytics derived insights.

The success of this new business model will be dependent on having access to data created across the entire health care ecosystem. We believe there is an opportunity to drive competitive advantage for our LS clients by creating a robust analytics capability and harnessing integrated real-world patient level data.

"Disruptive Analytics"

The changing health care landscape is an excellent example of where data science and disruptive analytics can have an immediate beneficial impact. We believe transformation of the health care system will come through Big Data-driven decisions and improved insights. Over time, evidence of value measured in patient outcomes tied to costs derived from multiple health care Big Data assets will become the common currency across all health care sectors. Let's introduce one of the health care analytics experts we interviewed, James Golden.

James Golden is a Partner at Accenture in the pharmaceutical R&D practice. His work has included the development of systems and software for the integration and analyses of structured and unstructured health care and life science data. Jim's most recent work focuses on *evidence-based medicine* (EBM). Although the idea of EBM has been around for a while, the arrival of Big Data analytics makes it possible to transform the vision into reality, creating a transparent approach to pharmaceutical decision making based on the aggregation and analysis of health care data such as electronic medical records and insurance claims data.

Prior to joining Accenture, Golden was the chief technology officer of SAIC's Commercial Life Sciences Office, where he focused on search and intelligence analysis, including unstructured text mining, competitive intelligence, and social networks. He is a major in the U.S. Air Force Reserve and spent several years on the staff of the Air Force Test Pilot School.

According to Golden's insight, health care Big Data analytics presents an opportunity to unify the health care value chain in a way not achieved to date, a virtual form of unification with significant benefits for all stakeholders. Creating a health care analytics framework has significant value for individual stakeholders:

- For providers (physicians), there is an opportunity to build analytics systems for EBM—sifting through clinical and health outcomes data to determine the best clinical protocols that provide the best health outcome for patients and create defined standards of care.
- For producers (pharmaceutical and medical device companies), there is an opportunity to build analytics systems to enable *translational medicine* (TM)—integrating externally generated postmarketing safety, epidemiology, and health outcomes data with internally generated clinical and discovery data (sequencing, expression, biomarkers) to enable improved strategic R&D decision making across the pharmaceutical value chain.
- For payers (i.e., insurance companies), there is an opportunity to create analytics systems to enable *comparative effectiveness research* (CER) that will be used to drive reimbursement—mining large collections of claims, health care records (EMR/EHR), economic and geographic, demographic data sets to determine what treatments and therapies work best for which patients in which context and with what overall economic and outcomes benefits.

- A Holistic Value Proposition
- James Golden explains the theory of a holistic value proposition:
- If we believe that data is indeed a platform, then we should begin to manage it like one. It is the ability to collect, integrate, analyze and manage this data that make health care data such as EHR/EMRs valuable. Longitudinal patient data is one source of the raw material on which evidence based insight approaches can be built to enable health care reform.
- To date, there has been little attempt to "see where the data takes us" and create a holistic health care value proposition built on quantifiable evidence that clarifies business value for all stakeholders.
- Because of our client relationships across the health care ecosystem, we are facilitating unique partnerships across payer, provider, pharma, and federal

agencies to work on problems of health care data analytics together and create value for all health care stakeholders.

- At Accenture, we are working across our life science and health care practices to identify the breadth of health care data sources that exist in order to better understand how our client's pharmaceutical and health care delivery products perform in the real world. Working with our clients, we have taken a "big data" approach to the analysis of health care data—by that we mean creating methods and platforms for the analysis of large volumes of disparate kinds of data—clinical, EMR, claims, labs, etc.—to better answer questions of outcomes, epidemiology, safety, effectiveness, and pharmacoeconomic benefit. We are leveraging big data technologies and platforms such as Hadoop, R, openhealthdata, and others to help our clients create real-world evidence-based approaches to realizing solutions for competitive effectiveness research, improve outcomes in complex populations, and to improve patient cohort segmentation and formulary decision making.
- By "big data," we also mean that health care data sets are big enough to obscure underlying meaning; that traditional methods of storing, accessing, and analyzing those data are breaking down; the data itself is both structured and unstructured; and large-scale analytics are needed for critical decision making, specifically in the face of cost containment, regulatory burden, and requirements of health care reform.
- Over the last decade companies such as Google, LinkedIn, eBay, and Facebook have created enormously valuable businesses that rely on the skills of new data scientists, who are linking together large data sets from disparate sources, visualizing and mining data in new ways to create novel modeling techniques, and developing new platforms for predictive analytics and decision support that impact the lives of millions of people on a daily basis. Statisticians and experts in machine learning and artificial intelligence, once relegated to academia, are becoming the new rock stars of Silicon Valley and bring multidisciplinary mathematical approaches to e-commerce and social media.
- As a result, engineers and data scientists at these companies are creating the technology and expertise required to turn raw data into information assets with tangible business value. The challenge is to discover how to leverage these rapidly evolving and nonstandardized skills into an enterprise analytics group usable by health care stakeholders. IT resources within the enterprise are often older and reluctant to embrace a "hacker ethos" needed to create patient data mash-ups, new data products, and on-the-fly data mining techniques. There is very little knowledge transfer between tomorrow's Big Data scientist and today's health care CIOs.

Pioneering New Frontiers in Medicine

The new era of data-driven health care is already producing new kinds of heroes. For example, Murali Ramanathan is co-director, Data Intensive Discovery Institute, State University of New York at Buffalo. Ramanathan uses Big Data analytics to identify the genetic variations that predispose individuals to multiple sclerosis (MS). Here's a brief description of what he does, in his own words:

> I am not a computer scientist, I'm a pharmaceutical scientist and work on the role of environmental factors, interactions between environmental factors in multiple sclerosis, which is a neurological disease. What we've got are data sets that contain thousands of genes. Typically, our data sets contain somewhere between 100,000 to 500,000 genetic variations.
>
> Our algorithms identify the interactions (between environmental factors and diseases) and they also have rapid search techniques built into them. We also want to be able to do statistical analysis. In our case, we are doing permutation analysis, which can be very, very time consuming if not properly done. With today's technology, we can get about 500,000 genetic variations in a single patient sample.

Nate Burns, a colleague of Ramanathan's at SUNY Buffalo, paints a vivid description of the challenge facing pioneers of data-intensive quantitative pharmacology:

> The data set is very large—a 1,000 by 2,000 matrix. What makes it interesting is when you try to do an interaction analysis for first and second order interactions, basically each of those 500,000 genetic locations is compared to each of all the rest of the 500,000 genetic locations for the first order; then you have to do that twice, so you've cut 500,000 to a third, for second order interactions and so on. It becomes exceedingly challenging as you move into more interactions. Basically, a second order interaction would be 500,000 squared, a third order would be 500,000 cubed, and so on.

The good news is that results from the MS study can potentially help researchers understand other autoimmune diseases (such as rheumatoid arthritis, diabetes, and lupus) and neurodegenerative diseases such as Parkinson's and Alzheimer's. "MS actually occupies a space between these two categories of diseases," says Ramanathan. "Our goal is finding the similarities and differences by looking at the data sets."

Advertising and Big Data: From Papyrus to Seeing Somebody

Let's take one of the oldest business practices, advertising, which dates back to the days when ancient Egyptians used Papyrus to make sales banners to promote their businesses. Back then it was a simple matter of promoting your business through the use of visual promotional advertising.

Now let's fast forward to an incredible scene on the streets of Madison Avenue, New York City, during the 1960s. Every present-day businessperson smirks in either jealousy or disbelief when they see the work-life of the advertising executive character from AMC's drama *Mad Men*, Donald Draper. Draper's character is partially based on Draper Daniels, the creative head of the Leo Burnett advertising agency in 1950s who created the Marlboro Man campaign. As Don said, "Just think about it deeply, then forget it . . . then an idea will jump up in your face." A bunch of executives ideating about the next big idea over Manhattan cocktails and Lucky Strike cigarettes wasn't that far from reality in those days.

With a history dating back to 1873, Foote, Cone & Belding Worldwide is one of the oldest providers of advertising and marketing services. Fairfax Cone inherited the agency from Albert Lasker, who can justifiably be called the founder of the modern advertising industry. Together with his colleagues, Emerson Foote, and Don Belding, Cone led the agency for over 30 years.

Advertising is what you do when you can't (afford to) go see somebody. That's all it is.
—Fairfax Cone, principal of Foote, Cone & Belding, 1963

Cone's quote was stated around the same time when companies were using rather primitive sales and marketing tactics to "go see someone." Salesmen were lugging Electrolux Vacuum cleaners from house to house pitching their high-end equipment and competing against manufacturers like Hoover and Oreck. The salesmen had a tough job because the only customer information they had was picking a neighborhood where they felt people could afford their product. The truth is they were not welcome at any home and were greeted with a scowl or totally ignored by people pretending to not be home. There was one key question each salesman would ask that increased their chance of being invited in, "Do you own an Electrolux?" If the answer was yes, the salesman would offer to service their equipment with a tune-up. This would result in an infrequent upsell to a new vacuum, but most of the time they were lucky if they sold a pack of new bags!

While in Great Britain, the firm launched an advertising campaign with the slogan "Nothing sucks like an Electrolux." This clever slogan was accompanied by a visual of the Leaning Tower of Pisa next to the latest series of the Electrolux vacuum. Since Electrolux is a Scandinavian-based company, most people thought this double entendre was a blunder. However, the slogan was deliberate and designed to have "stopping power," and it certainly succeeded in grabbing the audience's attention. The Electrolux vacuum brand sold very well in Great Britain for some time and people still remember the slogan. Although the campaign was rejected in the United States in the late 1960s, some think that this campaign humor would work in America today, especially during the Super Bowl.

The early advertising executives had a powerful means to reach their audiences, which was billboard, newspaper, radio, and eventually television. However, their clients were focused on the big idea because they were desperate to get their messages through these channels. As the industry matured, they demanded to learn more about their audiences, which created demand for firms such as Nielsen Media Research, which would statistically measure which television programs are watched by different segments of the population. This would help the advertisers pick the best place to place their ads (media spend). After years of refinement, clever media planning, and the inclusion of more and more data, marketers got pretty savvy at targeting their ads.

Big Data Feeds the Modern-Day Donald Draper

To get a feel for how Big Data is impacting the advertising market, we sat down with Randall Beard, who is currently the global head of Nielsen's Global Head of Advertiser Solutions. Essentially what Beard's team does is connect what people watch and what people buy to help their clients optimize advertising and media return on investment.

The Nielsen experience is great, but the best part of interviewing Beard is that before Nielsen he actually worked on the client side for 25 years at companies such as the big advertising spender P&G. Needless to say, he knows his stuff.

> *What's driving all of this is not just that they're spending a lot of money but CEOs/CFOs are looking at the chief marketing officers and saying look, if we're going to spend $100 million, $500 million, a billion in advertising and media, show me the money. Show me this stuff works or we're not going to spend the money on it. There was this huge demand for accountability that's driving the need for the marketing heads to answer these questions.*
>
> —Randall Beard

Reach, Resonance, and Reaction

Beard explained that big data is now changing the way advertisers address three related needs:

1. **How much do I need to spend**? "It's basic and fundamental but it's amazing to me that we're sitting here in 2012 and advertisers still have a really hard time answering the question, How much do I need to spend next year? I know what my revenue goal is next year, and I know how much profit I need to deliver. What do I need to spend to deliver that?"

2. **How do I allocate that spend across all the marketing communication touch points**? "Answering the question, "how do I allocate my marketing communications spending across paid, owned and earned media is increasingly difficult. If you think about that question, it's getting harder and harder to answer because of the fragmentation of media, the rise of digital, and the increasing importance of social media. If I'm going to spend money in digital, how much do I allocate to banner versus rich media versus online video versus search? How about mobile, how about apps? It's further complicated by the fact that all these things work together in a complementary manner. You can't even think about them as independent things."

3. **How do I optimize my advertising effectiveness against my brand equity and ROI in real-time**. "The old paradigm was I go out and run a campaign. Maybe after the fact, I measure it . . . maybe . . . try to determine some ROI then plan for the next cycle of advertising. Basically, advertisers are saying that's not good enough. I want to know within days, or at least weeks, how my advertising is performing in the market and what levers to pull, what knobs to turn, so that I get a higher ROI."

Given these needs, advertisers need to be able to measure their advertising end to end. What does this mean?

To start with, they need to identify the people who are most volumetrically responsive to their advertising. And then answer questions such as: What do those people watch? How do I reach them? "With more and more data, and the ability to measure what people watch and buy at the household level, there is the capability to identify those people who were most volumetrically responsive to your advertising. Then you can figure out: What TV programs do those people watch? What do they do online? How do I develop my media plan against that intended audience? That's the first part of reach," explained Beard.

Now the second part of the "reach" equation is to understand if you are actually reaching your desired audience. If you think about the online world, it's a world where you can deliver 100 million impressions but you never really know for sure who your campaign was actually delivered to. If your intended audience is women aged 18 to 35, of your 100 million impressions, what percentage of impressions were actually delivered to the intended audience? What was the reach, what was the frequency, what was the delivery against the intended audience? For all the great measurement that people can do online, that hasn't been well measured historically. This is the other part of reach—delivering your ads to the right audience.

Let's now talk about resonance. If you know whom you want to reach and you're reaching them efficiently with your media spend, the next question is, are your ads breaking through? Do people know they're from your brand? Are they changing attitudes? Are they making consumers more likely to want to buy your brand? This is what I call "resonance."

Lastly, you want to measure the actual behavioral impact. If you've identified the highest potential audience, reached them efficiently with your media plan, delivered ads that broke through the clutter and increased their interest in buying your brand—did it actually result in a purchase? Did people actually buy your product or service based on exposure to your advertising? At the end of the day, advertising must drive a behavioral "reaction" or it isn't really working.

Beard explained the three guiding principles to measurement:

1. End to end measurement—reach, resonance and reaction
2. Across platforms (TV, digital, print, mobile, etc.)
3. Measured in real-time (when possible)

The Need to Act Quickly (Real-Time When Possible)

When you start executing a campaign, how do you know on a daily basis whether your advertising campaign is actually being delivered to your intended audience the way it's supposed to?

For example, in digital, ad performance will differ across websites. Certain websites are really good; certain websites are really bad. How do you optimize across sites "on the fly?" By moving money out of weak performing sites and into better performing sites.

Beard describes how real time optimization works:

> I'm one week into my new ad campaign. There's good news and bad news. The good news is that my ad is breaking thru and is highly memorable. The bad news is that consumers think my ad is for my key competitor. I work with my agency over the weekend to edit the spot, and it goes back on air. Presto! Branding scores increase.

A week later, I see that of my three ads on air, two have high breakthrough but one is weak. I quickly take the weak performing ad off air and rotate the media spend to the higher performing ads. Breakthru scores go up!

My campaign soon moves from running only: 30's to a mix of: 15's and: 30s, a fairly typical plan. Real time data shows me that my 15s work as well as my 30s. Why spend money on 30s? I move all the weight to 15-second ads—and see scores continue to grow.

In digital, I see that brand recall increases with exposure frequency up to two exposures, and then levels off. My agency caps exposure frequency at two. I use the savings from reduced frequency to buy more sites and extend reach.

You have real-time optimization that's going on, which is data driven instead of just gut driven! The measurement tools and capabilities are enabling this and so there's a catch-up happening both in terms of advertising systems and processes, but also just the industry infrastructure to be able to actually enable all of this real-time optimization."

Measurement Can Be Tricky

Beard gave an example of the complexity of measurement. There are tools that allow you to tag digital advertising and, typically, through a panel of some kind, you can read those people who were exposed to the advertising and those who were not and measure their actual offline purchase behavior.

In doing this for a large beer client, we could see that this campaign generated (after the fact) a 20 percent sales increase among consumers exposed versus not exposed to the advertising. You (the average person) would look at that and say, wow, looks pretty good—my advertising is working.

But the sales results aren't everything. Beard elaborates on the first part of the end-to-end measurement, involving the reach:

When we looked at reach for this particular client, their intended audience was males, aged 21–29. Of their 100 million delivered impressions, only about 40 million were actually delivered to males aged 21–29. Sixty million went to someone other than their intended audience; some went to kids (not good for a beer brand); some went to people 65+. You start out by saying wow, how much better could I have done, if instead of 40% of my impressions hitting my intended audience, I had 70 or 80% of the impressions hitting them.

When you look at the 40 percent of impressions that hit the intended audience, the reach and frequency of those was something like a 10 percent reach and a 65 frequency. In other words, they only hit about 10 percent of their intended audience, but each of these people was bombarded with, on average, 65 ads! That's not quite the optimization one would hope for. There's a lot of science in advertising that shows that by maximizing reach and minimizing frequency, you get your best response. If they had been measuring all of this in real time, they could have quickly adjusted the plan to increase delivery to the intended audience, increase reach, and reduce frequency.

Content Delivery Matters Too

Let's now look at ad performance by website. The ads were on twelve websites: four were terrible; the breakthrough was terrible, branding was terrible—the ads didn't perform well in those sites. The other ones were really good. If they had measured that in flight, they could have moved spending out of the bad performing sites, into good performing sites, and further improved results.

Beard explains the importance of end-to-end measurement:

> When I think about it, it's almost like the reach times resonance equals reaction. Of course, this isn't arithmetically true, but it illustrates that while measuring the sales impact alone is great, it's not enough. You could have great sales impact and still be completely non-optimized on the reach and resonance factors that caused the reaction."

Optimization and Marketing Mixed Modeling

Marketing mixed modeling (MMM) is a tool that helps advertisers understand the impact of their advertising and other marketing activities on sales results. MMM can generally provide a solid understanding of the relative performance of advertising by medium (e.g., TV, digital, print, etc.), and in some cases can even measure sales performance by creative unit, program genre, website, and so on.

Now, we can also measure the impact on sales in social media and we do that through market mixed modeling. Market mixed modeling is a way that we can take all the different variables in the marketing mix—including paid, owned, and earned media—and use them as independent variables that we regress against sales data and trying to understand the single variable impact of all these different things.

Since these methods are quite advanced, organizations use high-end internal analytic talent and advanced analytics platforms such as SAS or point solutions such as Unica and Omniture. Alternatively, there are several boutique and large analytics providers like Mu Sigma that supply it as a software-as-a-service (SaaS).

MMM is only as good as the marketing data that is used as inputs. As the world becomes more digital, the quantity and quality of marketing data is improving, which is leading to more granular and insightful MMM analyses.

Beard's Take on the Three Big Data Vs in Advertising

Beard shared his perspective on how the three Vs (volume, velocity, and variety) have impacted advertising:

Volume

> In the old days, this is not that old, not even *Mad Men* days, maybe 20 to 25 years ago, you would copy test your advertising. The agency would build a media plan demographically targeted and you'd go execute it. That was pretty much it. Maybe

6 to 12 months down the road, you'd try to use scanner data or whatever sales data you had to try to understand if there was any impact.

In today's world, there is hugely more advertising effectiveness data. On TV advertising, we can measure every ad in every TV show every day, across about 70 percent of the viewing audience in the U.S. We measure clients digital ad performance hourly—by ad, by site, by exposure, and by audience. On a daily or weekly basis, an advertiser can look at their advertising performance. The volume of information and data that is available to the advertiser has gone up exponentially versus what it was 20 years ago.

Velocity

There are already companies that will automate and optimize your advertising on the web without any human intervention at all based on click-thru. It's now beginning to happen on metrics like breakthrough, branding, purchase intent, and things like that. This is sometimes called programmatic buying. Literally, you'll have systems in place that will be measuring the impact of the advertising across websites or different placements within websites, figuring out where the advertising is performing best. It will be automated optimization and reallocation happening in real-time. The volume and the velocity of data, the pace at which you can get the data, make decisions and do things about it is dramatically increased.

Variety

Before, you really didn't have a lot of data about how your advertising was performing in market. You have a lot more data and it's a lot more granular. You can look at your brand's overall advertising performance in the market. But you can also decompose it to how much of a performance is due to the creative quality, due to the media weight, how much is due to the program that the ads sit in. How much is due to placement: time of day, time of year, pod position, how much is due to cross-platform exposure, how much is due to competitive activity. Then you have the ability to optimize on most of those things—in real time. And now you can also measure earned (social) and owned media. Those are all things that weren't even being measured before."

Using Consumer Products as a Doorway

As an experienced business executive, what would you say if you were asked by your boss to consider entering into the mobile phone or PC/tablet business, which the company has never tried to do before? Chances are your answer would be no way! First of all, the hardware business has a lot of manufacturing overhead, which means margins are razor thin. Second, the space is overcrowded with mature players and other low-cost providers. Last, there are dozens of consumer hardware business case nightmares such as Research in Motion (RIM), the maker of Blackberry. "Over the last year, RIM's share price has plunged 75 percent. The company once commanded more than half of the American smartphone market. Today it has 10 percent."[5] Hopefully, RIM will have the fortune to turn things around for their employees and shareholders, but we can't help but to remind ourselves of powerhouses like Gateway computing that

disappeared in 2007. And companies with deep pockets and resources such as Hewlett Packard (HP) that failed to enter the tablet market, while Apple is selling iPads in its sleep.

It made a lot of sense that Apple entered into the mobile and tablet market because, after all, it is a software and hardware player that made the iPod, which crushed giants like Sony in the MP3 market. That one was a hard pill to swallow for Sony when their Walkman was all the rage through the cassette and CD years. For Apple, the market was not just about selling hardware or music on iTunes. It gave them a chance to get as close to a consumer as anyone can possibly get. This close interaction also generated a lot of data that help them expand and capture new customers. Again it's all about the data, analytics, and putting it into action.

Google gives away product that other companies, such as Microsoft, license for the same the reason. It also began playing in the mobile hardware space through the development of the Android platform and the acquisition of Motorola. It's all about gathering consumer data and monetizing the data. Do you use Google? Check out your Google Dashboard. You can see every search you did, e-mails you sent, IM messages, web-based phone calls, documents you viewed, and so on. How powerful is that for marketers? We'd say that would be similar to meeting somebody!

Who would have thought that an online retailer, Amazon, would create hardware with their Kindle Fire and that Barnes and Noble would release the Nook? Imagine that both companies know every move you make, what you download, what you search for, and now they can study your behaviors to present new products that they believe will appeal to you. It all comes down to the race for the connection with consumers and more importantly taking action on the derived data to win the marathon.

Big Data Technology

Technology is radically changing the way data is produced, processed, analyzed, and consumed. On one hand, technology helps evolve new and more effective data sources. On the other, as more and more data gets captured, technology steps in to help process this data quickly, efficiently, and visualize it to drive informed decisions. Now, more than any other time in the short history of analytics, technology plays an increasingly pivotal role in the entire process of how we gather and use data.

The Elephant in the Room: Hadoop's Parallel World

There are many Big Data technologies that have been making an impact on the new technology stacks for handling Big Data, but Apache Hadoop is one technology that has been the darling of Big Data talk. Hadoop is an open-source platform for storage and processing of diverse data types that enables data-driven enterprises to rapidly derive the complete value from all their data.

We spoke with Amr Awadallah, the cofounder and chief technology officer (CTO) of Cloudera, a leading provider of Apache Hadoop-based software and services, since it was formed in October 2008. He explained the history and overview of Hadoop to us:

> The original creators of Hadoop are Doug Cutting (used to be at Yahoo! now at Cloudera) and Mike Cafarella (now teaching at the University of Michigan in Ann Arbor). Doug and Mike were building a project called "Nutch" with the goal of creating a large Web index. They saw the MapReduce and GFS papers from Google, which were obviously super relevant to the problem Nutch was trying to solve. They integrated the concepts from MapReduce and GFS into Nutch; then later these two components were pulled out to form the genesis of the Hadoop project.
>
> The name "Hadoop" itself comes from Doug's son, he just made the word up for a yellow plush elephant toy that he has. Yahoo! hired Doug and invested significant resources into growing the Hadoop project, initially to store and index the Web for the purpose of Yahoo! Search. That said, the technology quickly mushroomed throughout the whole company as it proved to be a big hammer that can solve many problems.

In 2008, recognizing the huge potential of Hadoop to transform data management across multiple industries, Amr left Yahoo! to co-found Cloudera with Mike Olson and Jeff Hammerbacher. Doug Cutting followed in 2009.

Moving beyond rigid legacy frameworks, Hadoop gives organizations the flexibility to ask questions across their structured and unstructured data that were previously impossible to ask or solve:

- The scale and variety of data have permanently overwhelmed the ability to cost-effectively extract value using traditional platforms.
- The scalability and elasticity of free, open-source Hadoop running on standard hardware allow organizations to hold onto more data than ever before, at a transformationally lower TCO than proprietary solutions and thereby take advantage of *all* their data to increase operational efficiency and gain a competitive edge. At one-tenth the cost of traditional solutions, Hadoop excels at supporting complex analyses—including detailed, special-purpose computation—across large collections of data.
- Hadoop handles a variety of workloads, including search, log processing, recommendation systems, data warehousing, and video/image analysis. Today's explosion of data types and volumes means that Big Data equals big opportunities and Apache Hadoop empowers organizations to work on the most modern scale-out architectures using a clean-sheet design data framework, without vendor lock-in.
- Apache Hadoop is an open-source project administered by the Apache Software Foundation. The software was originally developed by the world's largest Internet companies to capture and analyze the data that they generate. Unlike traditional, structured platforms, Hadoop is able to store any kind of data in its native format and to perform a wide variety of analyses and transformations on that data. Hadoop stores terabytes, and even petabytes, of data inexpensively. It is robust and reliable and handles hardware and system failures automatically, without losing data or interrupting data analyses.

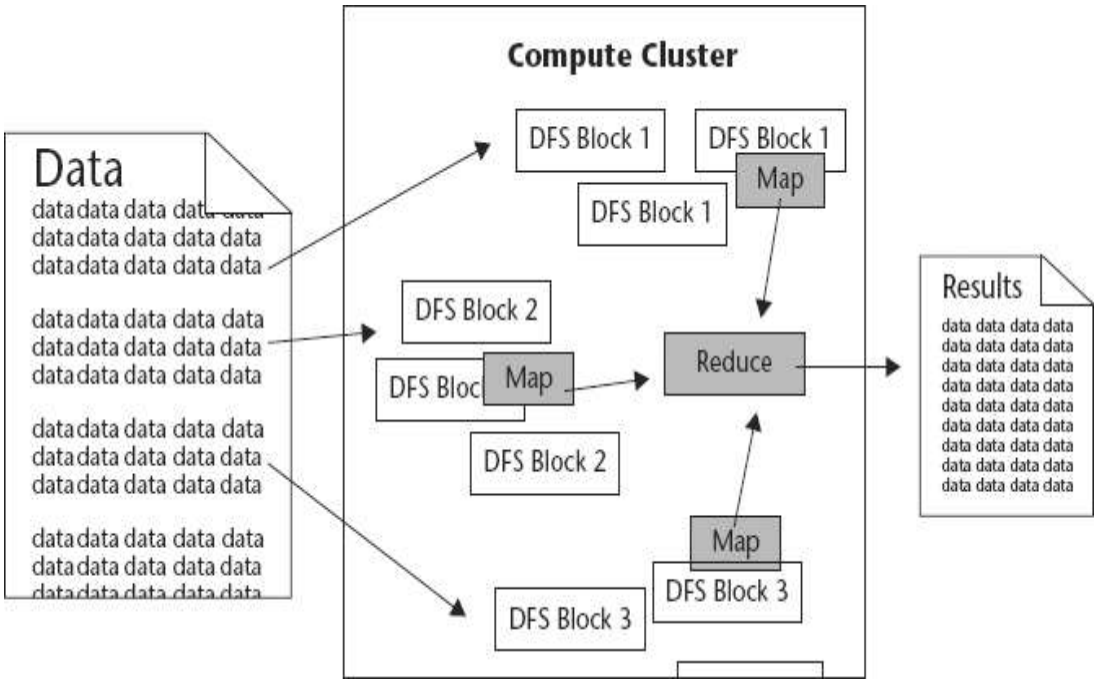
- Hadoop runs on clusters of commodity servers and each of those servers has local CPUs and disk storage that can be leveraged by the system.

The two critical components of Hadoop are:

1. **The Hadoop Distributed File System (HDFS)**. HDFS is the storage system for a Hadoop cluster. When data lands in the cluster, HDFS breaks it into pieces and distributes those pieces among the different servers participating in the cluster. Each server stores just a small fragment of the complete data set, and each piece of data is replicated on more than one server.

2. **MapReduce**. Because Hadoop stores the entire dataset in small pieces across a collection of servers, analytical jobs can be distributed, in parallel, to each of the servers storing part of the data. Each server evaluates the question against its local fragment simultaneously and reports its results back for collation into a comprehensive answer. MapReduce is the agent that distributes the work and collects the results.

Both HDFS and MapReduce are designed to continue to work in the face of system failures. HDFS continually monitors the data stored on the cluster. If a server becomes unavailable, a disk drive fails, or data is damaged, whether due to hardware or software problems, HDFS automatically restores the data from one of the known good replicas stored elsewhere on the cluster. Likewise, when an analysis job is running, MapReduce monitors progress of each of the servers participating in the job. If one of them is slow in returning an answer or fails before completing its work, MapReduce automatically starts another instance of that task on another server that has a copy of the data. Because of the way that HDFS and MapReduce work, Hadoop provides scalable, reliable, and fault-tolerant services for data storage and analysis at very low cost.



Old vs. New Approaches

We interviewed data guru Abhishek Mehta to get his perceptions of the differences between the "old" and "new" types of big data analytics. Mehta is a former Bank of America executive and MIT Media Lab executive-in-residence. He recently launched Tresata, a company that is developing the first Hadoop-powered Big Data analytics platform focused on financial industry data. Here is a summary of what Mehta told us:

> The old way is a data and analytics technology stack with different layers "cross-communicating data" and working on "scale-up" expensive hardware. The new way is a data and analytics platform that does all the data processing and analytics in one "layer," without moving data back and forth on cheap but scalable ("scale out") commodity hardware. This is a mega shift and a complete game changer!
>
> The new approach is based on two foundational concepts. Number one, data needs to be stored in a system in which the hardware is infinitely scalable. In other words, you cannot allow hardware (storage and network) to become the bottleneck. Number two, data must be processed, and converted into usable business intelligence where it sits. Put simply, you must move the code to the data and not the other way around. That is a fundamental departure and the primary difference between the old way and the new way. In the old ways, you had the multiple tiers of the stack and in the new way we have what is essentially a horizontal platform for data. The data sits in one place, you never move it around. That's the "secret" to big data analytics.
>
> And here's another important point to remember: The technology stack has changed. New proprietary technologies and open-source inventions enable different approaches that make it easier and more affordable to store, manage, and analyze data. So it's not a coincidence that all of this change is occurring right now.
>
> Hardware and storage are more affordable than ever before, and continuing to get cheaper [thanks to Dr. Moore], which allows for increasingly larger and more ambitious massively parallel architectures. As the sheer quantity and complexity of data increases, our ability to handle complex and unstructured data is also rising.
>
> Today we can run the algorithm, look at the results, extract the results, and feed the business process—automatically and at massive scale, using all of the data available.

We continue our conversation with Mehta later in the book. For the moment, let's boil his observations down to three main points:

1. The technology stack has changed. New proprietary technologies and open-source inventions enable different approaches that make it easier and more affordable to store, manage, and analyze data.

2. Hardware and storage is affordable and continuing to get cheaper to enable massive parallel processing.

3. The variety of data is on the rise and the ability to handle unstructured data is on the rise.

Data Discovery: Work the Way People's Minds Work

There is a lot of buzz in the industry about *data discovery*, the term used to describe the new wave of business intelligence that enables users to explore data, make discoveries, and uncover insights in a dynamic and intuitive way versus predefined queries and preconfigured drill-down dashboards. This approach has resonated with many business users who are looking for the freedom and flexibility to view Big Data. In fact, there are two software companies that stand out in the crowd by growing their businesses at unprecedented rates in this space: Tableau Software and QlikTech International.

Both companies' approach to the market is much different than the traditional BI software vendor. They grew through a sales model that many refer to as "land and expand." It basically works by getting intuitive software in the hands of some business users to get in the door and grow upward. In the past, BI players typically went for the big IT sale to be the preferred tool for IT to build reports for the business users to then come and use.

In order to succeed at the BI game of the "land and expand model," you need a product that is easy to use with lots of sexy output. One of the most interesting facts about Tableau Software is that the company's chief scientist and cofounder, Pat Hanrahan, is not a BI software veteran—he's actually an Academy Award–winning professor and founding member of Pixar! He invented the technology that helped change the world of animated film. Harahan's invention made it possible to bring some of the world's most beloved characters to the big screen, such as Buzz Lightyear and Woody the cowboy. Imagine the new creative lens that Pat brought to the BI software market!

When you have a product that is "easy to use," it also means that you have what Harahan and his colleagues call the "self-service approach," versus the traditional approach with heavy reliance on IT. Pat, co-founder Chris Stolte, and colleague Dan Jewett stated in a recent whitepaper:

> Analytics and reporting are produced by the people using the results. IT provides the infrastructure, but business people create their own reports and dashboards.
>
> The most important characteristic of rapid-fire BI is that business users, not specialized developers, drive the applications. The result is that everyone wins. The IT team can stop the backlog of change requests and instead spend time on strategic IT issues. Users can serve themselves data and reports when needed.
>
> The traditional practice of trying to anticipate the analytic needs of each employee is impossible—can an IT department really read the minds of business users? Business users are more productive when answering questions with their own tools.[1]

There is a simple example of powerful visualization that the Tableau team is referring to. A company uses an interactive dashboard to track the critical metrics driving their business. Every day, the CEO and other executives are plugged in real-time to see how their markets are performing in terms of sales and profit, what the service quality scores look like against advertising investments, and how products are performing in terms of revenue and profit. Interactivity is key: a click on any filter lets the executive look into

specific markets or products. She can click on any data point in any one view to show the related data in the other views. Hovering over a data point lets her winnow into any unusual pattern or outlier by showing details on demand. Or she can click through the underlying information in a split-second.

We also spoke with Qliktech's CTO, Anthony Deighton, to get his view on the world of data discovery. Deighton is an ex-Seibel executive who has been with Qliktech since 2005. He is responsible for guiding product strategy and leads all aspects of the company's R&D efforts for its product suite, named QlikView. Deighton started off the interview with a very simple message: "Business intelligence needs to work the way people's minds work. Users need to navigate and interact with data any way they want to—asking and answering questions on their own and in big groups or teams."

One capability that we have all become accustomed to is search, what many people refer to as "Googling." This is a prime example of the way people's minds work. Qliktech has designed a way for users to leverage direct—and indirect—search. With QlikView search, users type relevant words or phrases in any order and get instant, associative results. With a global search bar, users can search across the entire data set. With search boxes on individual list boxes, users can confine the search to just that field. Users can conduct both direct and indirect searches. For example, if a user wanted to identify a sales rep but couldn't remember the sales rep's name—just details about the person, such as that he sells fish to customers in the Nordic region—the user could search on the sales rep list box for "Nordic" and "fish" to narrow the search results to just the people who meet those criteria.

Open-Source Technology for Big Data Analytics

Open-source software is computer software that is available in source code form under an open-source license that permits users to study, change, and improve and at times also to distribute the software. The open-source name came out of a 1998 meeting in Palo Alto in reaction to Netscape's announcement of a source code release for Navigator (as Mozilla).

Although the source code is released, there are still governing bodies and agreements in place. The most prominent and popular example is the GNU General Public License (GPL), which "allows free distribution under the condition that further developments and applications are put under the same license." This ensures that the products keep improving over time for the greater population of users.

Some other open-source projects are managed and supported by commercial companies, such as Cloudera, that provide extra capabilities, training, and professional services that support open-source projects such as Hadoop. This is similar to what Red Hat has done for the open-source project Linux.

"One of the key attributes of the open-source analytics stack is that it's not constrained by someone else's predetermined ideas or vision," says David Champagne, chief technology officer at Revolution Analytics, a provider of advanced analytics. "The open-source stack doesn't put you into a straitjacket. You can make it into what you want and what you need. If you come up with an idea, you can put it to work

immediately. That's the advantage of the open-source stack—flexibility, extensibility, and lower cost."

"One of the great benefits of open source lies in the flexibility of the adoption model: you download and deploy it when you need it," said Yves de Montcheiul, vice president of marketing at Talend, a provider of open-source data integration solutions. "You don't need to prove to a vendor that you have a million dollars in your budget. With open source, you can try it and adopt it at your own pace."

David Smith of Revolution Analytics has written many blogs and papers about the new open-source analytics stack. Smith is vice president of marketing at Revolution Analytics in Palo Alto. He observes that the pace of software development has accelerated dramatically because of open-source software. He follows this observation by describing how this phenomenon is setting the stage for a new "golden age" of software development:

> In the past, the pace of software development was moderated by a relatively small set of proprietary software vendors. But there are clear signs that the old software development model is crumbling, and that a new model is replacing it.
>
> The old model's end state was a monolithic stack of proprietary tools and systems that could not be swapped out, modified, or upgraded without the original vendor's support. This model was largely unchallenged for decades. The status quo rested on several assumptions, including:
>
> 1. The amounts of data generated would be manageable
> 2. Programming resources would remain scarce
> 3. Faster data processing would require bigger, more expensive hardware

Many of those underlying assumptions have now disappeared, David writes:

> The sudden increase in demand for software capable of handling significantly larger data sets, coupled with the existence of a worldwide community of open-source programmers, has upended the status quo.

The traditional analytics stack is among the first "victims" of this revolution. David explains how it has changed the game of enterprise software:

> The old model was top-down, slow, inflexible and expensive. The new software development model is bottom-up, fast, flexible, and considerably less costly.
>
> A traditional proprietary stack is defined and controlled by a single vendor, or by a small group of vendors. It reflects the old command-and-control mentality of the traditional corporate world and the old economic order.

David then makes the case for an open-source analytics stack. For David, who is a leading proponent of open-source analytics, it's a logical leap:

> An open-source stack is defined by its community of users and contributors. No one "controls" an open-source stack, and no one can predict exactly how it will

> evolve. The open-source stack reflects the new realities of the networked global economy, which is increasingly dependent on big data.

It's certainly fair to argue whether the new analytics stack should be open, proprietary, or a blend of the two. From our perspective, it seems unlikely that large companies will abandon their investments in existing proprietary technologies overnight.

Our hunch is that open-source and proprietary solutions will coexist for a long time, and for many good reasons. In fact, most proprietary vendors have been designing their solutions to plug and play with technology such as Hadoop. For example, Teradata Aster designed SQL-H, which is a seamless way to execute SQL and SQL-MapReduce on Apache Hadoop data.

Tasso Argyros is copresident of Teradata Aster, leading the Aster Center of Innovation. In a recent blog, Argyros explained the significance of his firm's integration with open-source Hadoop:

> This is a significant step forward from what was state-of-the-art until yesterday. This means that [in the past] getting data from Hadoop to a database required a Hadoop expert in the middle to do the data cleansing and the data type translation. If the data was not 100% clean (which is the case in most circumstances) a developer was needed to get it to a consistent, proper form. Besides wasting the valuable time of that expert, this process meant that business analysts couldn't directly access and analyze data in Hadoop clusters. SQL-H, an industry-first, solves all those problems.[2]

The Cloud and Big Data

It is important to remember that for all kinds of reasons—technical, political, social, regulatory, and cultural—cloud computing has not been a successful business model that has been widely adopted for enterprises to store their Big Data assets. However, there are many who believe that some obvious industry verticals will soon realize that there is a huge ROI opportunity if they do embrace the cloud.

> *There will be Big Data platforms that companies will build, especially for the core operational systems of the world. Where we continue to have an explosive amount of data come in and because the data is so proprietary that building out an infrastructure in-house seems logical. I actually think it's going to go to the cloud, it's just a matter of time! It's not value add enough to collect, process and store data.*
>
> —Avinash Kaushik, Google's digital marketing evangelist

Abhishek Mehta is one of those individuals who believes that cloud models are inevitable for every industry and it's just a matter of when an industry will shift to the cloud model. He explains that his clients are saying, "I don't have unlimited capital to invest in infrastructure. My data is exploding—both structured and unstructured. The models that I use to price products or manage risks are broken. I'm under immense pressure to streamline my operations and reduce headcount. How am I going to solve these problems?"

Market economics are demanding that capital-intensive infrastructure costs disappear and business challenges are forcing clients to consider newer models. At the crossroads of high capital costs and rapidly changing business needs is a sea change that is driving the need for a new, compelling value proposition that is being manifested in a cloud-deployment model.

With a cloud model, you pay on a subscription basis with no upfront capital expense. You don't incur the typical 30 percent maintenance fees—and all the updates on the platform are automatically available. The traditional cost of value chains is being completely disintermediated by platforms—massively scalable platforms where the marginal cost to deliver an incremental product or service is zero.

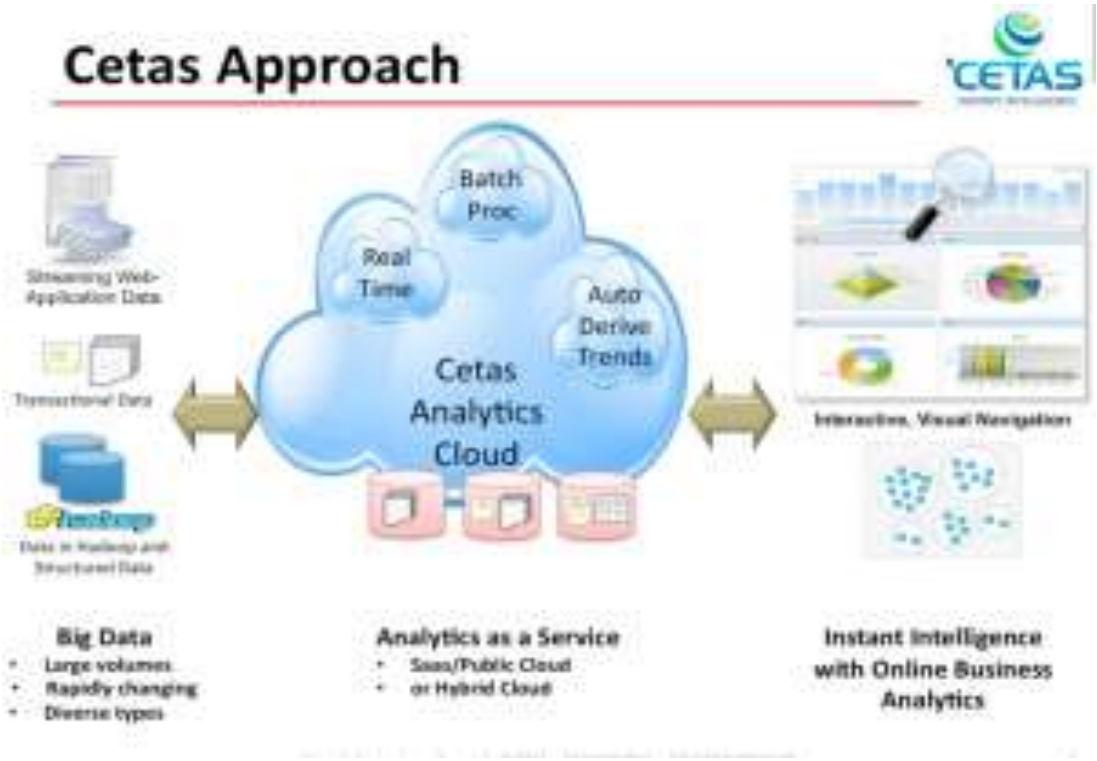
The ability to build massively scalable platforms—platforms where you have the option to keep adding new products and services for zero additional cost—is giving rise to business models that weren't possible before. Mehta calls it "the next industrial revolution, where the raw material is data and data factories replace manufacturing factories." He pointed out a few guiding principles that his firm stands by:

1. **Stop saying "cloud."** It's not about the fact that it is virtual, but the true value lies in delivering software, data, and/or analytics in an "as a service" model. Whether that is in a private hosted model or a publicly shared one does not matter. The delivery, pricing, and consumption model matters.

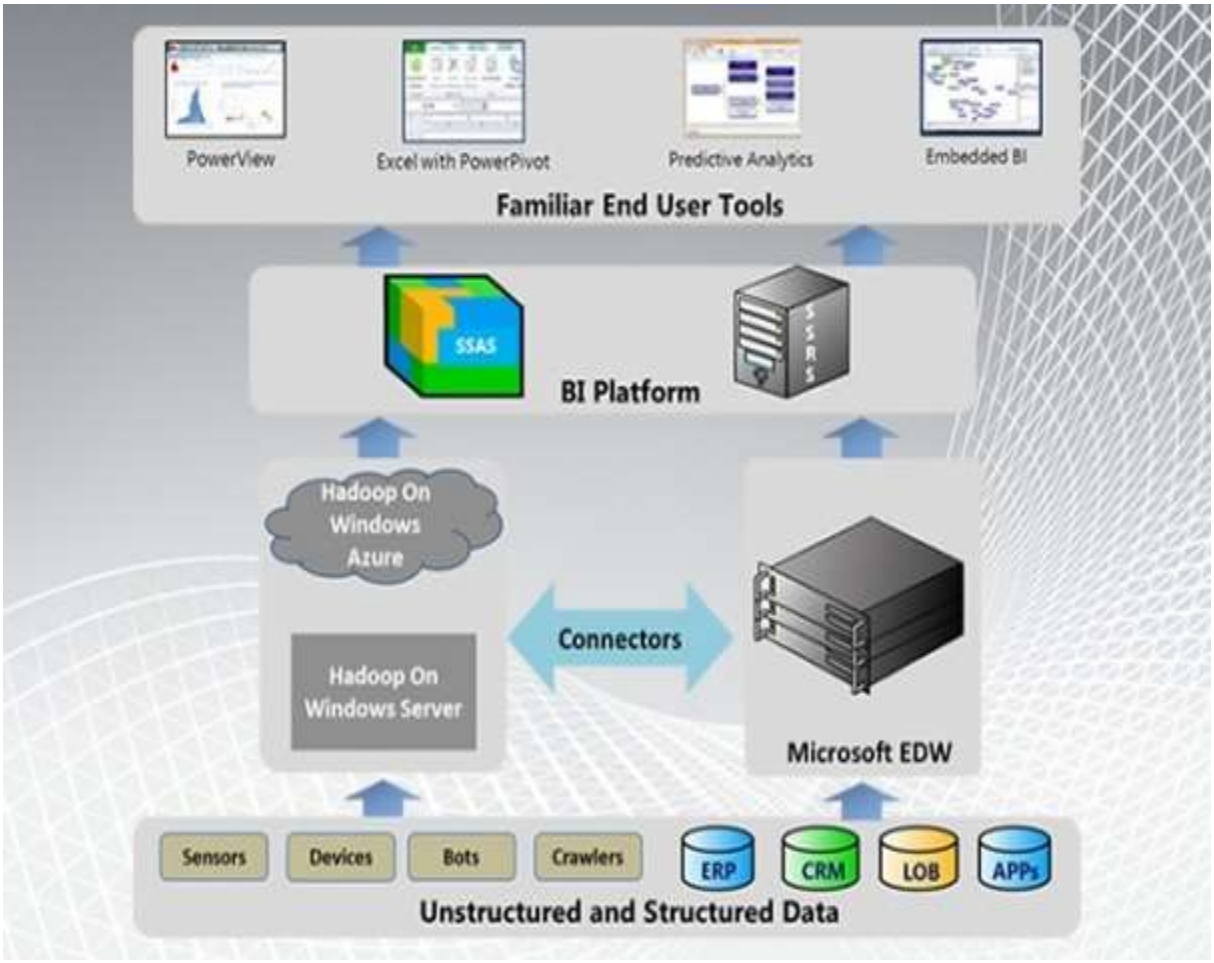
2. **Acknowledge the business issues**. There is no point to make light of matters around information privacy, security, access, and delivery. These issues are real, more often than not heavily regulated by multiple government agencies, and unless dealt with in a solution, will kill any platform sell.

3. **Fix some core technical gaps**. Everything from the ability to run analytics at scale in a virtual environment to ensuring information processing and analytics authenticity are issues that need solutions and have to be fixed.

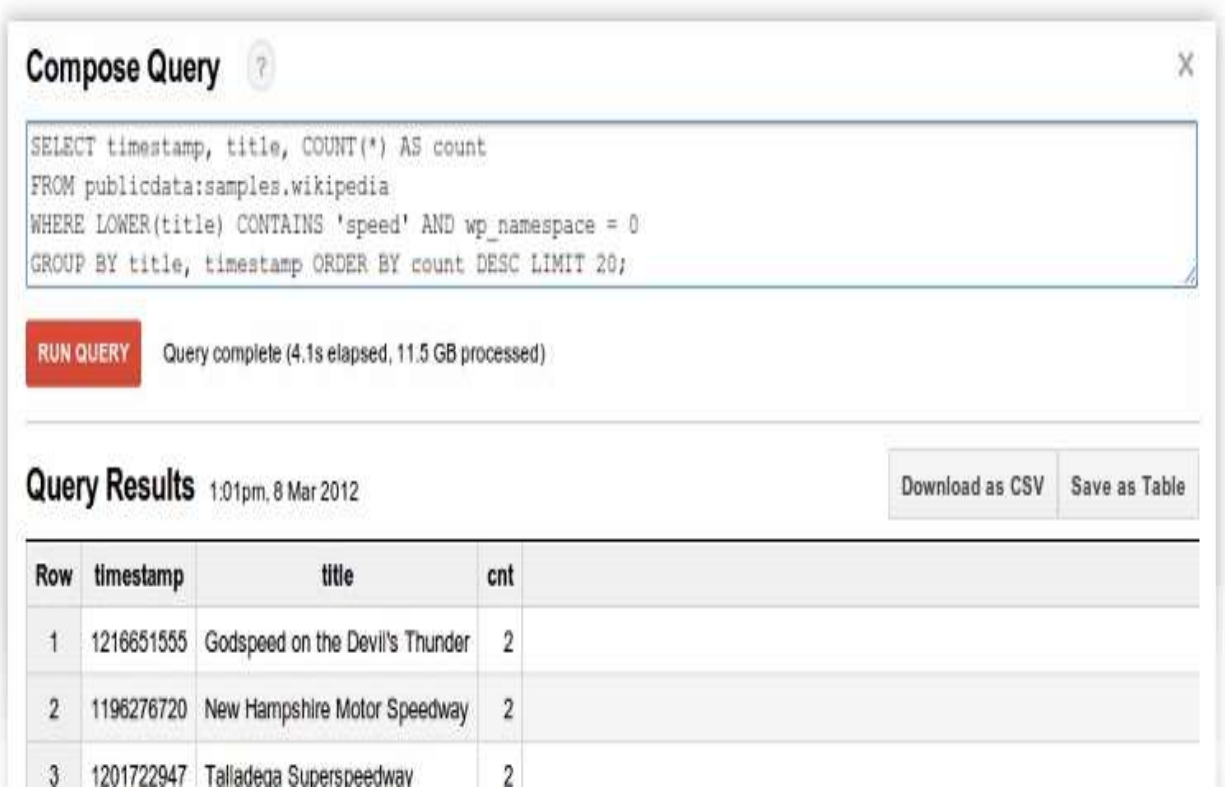
Cetas: Cetas (see disclosure) is a stealth-mode startup focused on providing an entire analytics stack in the cloud (or on-premise, if a customer prefers). The driving theory is to let companies running web applications get the types of user analytics that Facebook and Google are able to get, only without the teams of expensive engineers and data scientists, Cetas VP of Products Karthik Kannan told me. While most of that functionality is prepackaged now into core capabilities, Kannan said Cetas plans to let power-users build their own custom models and tie Cetas into existing analytic platforms.



Microsoft: Microsoft was late to the Hadoop game, but has been making up for lost time since October. I recently described its progress on the Hadoop on Windows Azure offering: "The company opened a preview of the service to 400 developers in December and on [March 6], … opened it up to 2,000 developers. According to Doug Leland, GM of product management for SQL Server, … Microsoft is trying 'to provide a service that is very easy to consume for customers of any size,' which means an intuitive interface and methods for analyzing data. Already, he said, Webtrends and the University of Dundee are among the early testers of Hadoop on Windows Azure, with the latter using it for genome analysis."



Google: Google has a multi-pronged strategy on cloud-based big data services, but the two services that stand out most are BigQuery and the Google Prediction API. Google describes BigQuery, available in limited preview now, as a service that "allows you to run SQL-like queries against very large datasets, with potentially billions of rows. … BigQuery works best for interactive analysis of very large datasets, typically using a small number of very large, append-only tables." The Google Prediction API is just what it sounds like, a service that puts machine learning and pattern-detection capabilities in developers' hands so they can analyze application data for things such as sentiment analysis, system analytics and recommendation engines.



Compose Query

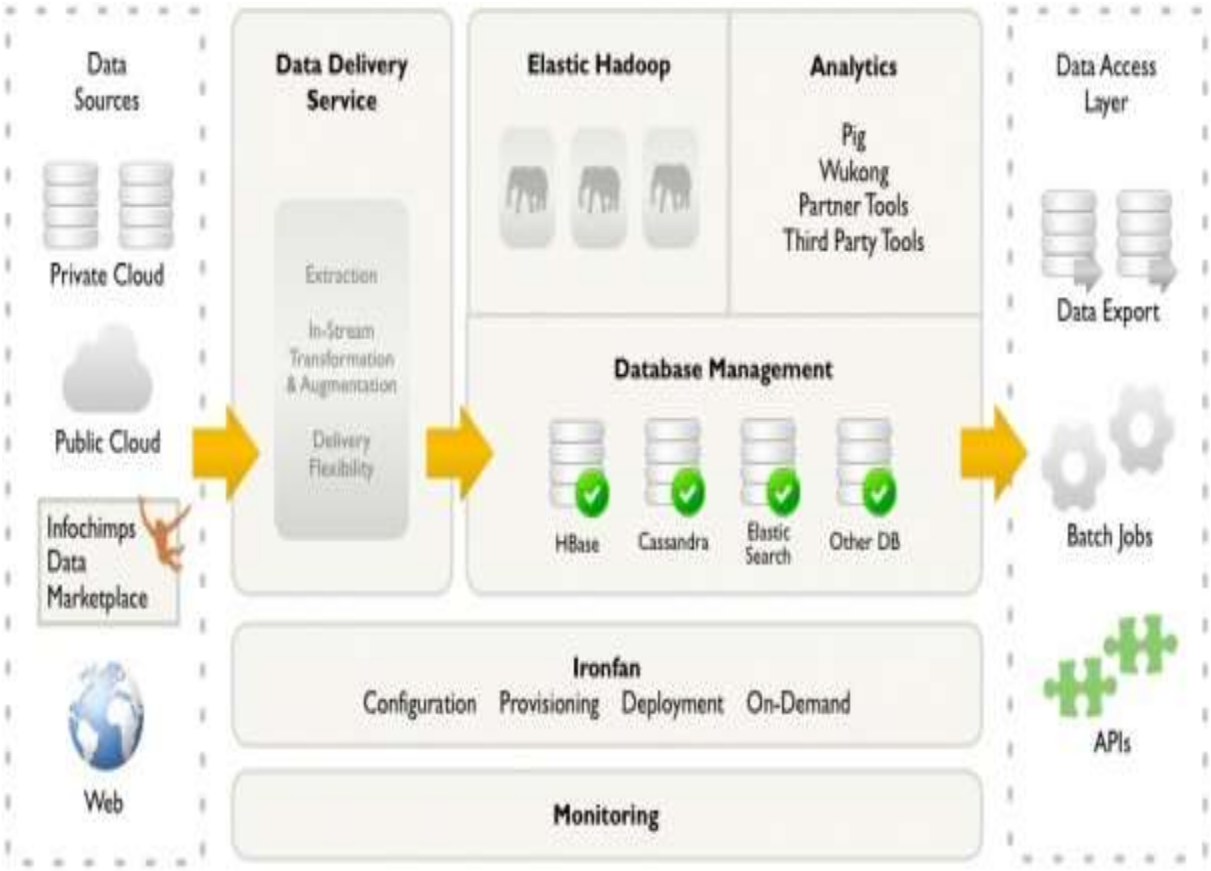
```
SELECT timestamp, title, COUNT(*) AS count
FROM publicdata:samples.wikipedia
WHERE LOWER(title) CONTAINS 'speed' AND wp_namespace = 0
GROUP BY title, timestamp ORDER BY count DESC LIMIT 20;
```

RUN QUERY Query complete (4.1s elapsed, 11.5 GB processed)

Query Results 1:01pm, 8 Mar 2012 Download as CSV Save as Table

Row	timestamp	title	cnt
1	1216651555	Godspeed on the Devil's Thunder	2
2	1196276720	New Hampshire Motor Speedway	2
3	1201722947	Talladega Superspeedway	2

Infochimps: Once a startup focused on its data marketplace, Infochimps has morphed into a provider of big data infrastructure as a service that provides its marketplace data as a value-added feature. Describing the new Infochimps Platform in February, I wrote, "The platform is hosted in the AWS cloud and supports Hadoop, various analytical tools on top of that — including Apache Pig and Infochimps' own Wukong (a Ruby framework for Hadoop) — and a variety of relational and NoSQL databases." But the key is the platform's automated nature, which CEO Joe Kelly hopes "will help answer the question of 'what does a Heroku for big data look like?'"



Predictive Analytics Moves into the Limelight

To master analytics, enterprises will move from being in reactive positions (business intelligence) to forward leaning positions (predictive analytics). Using all the data available—traditional internal data sources combined with new rich external data sources—will make the predictions more accurate and meaningful.

Because the analytics are contextual, enterprises can build confidence in the analytics and the trust will result in using analytic insights to trigger business events. By automatically triggering events, the friction in business will be greatly reduced. Algorithmic trading and supply chain optimization are just two typical examples where predictive analytics have greatly reduced the friction in business. Look for predictive analytics to proliferate in every facet of our lives, both personal and business. Here are some leading trends that are making their way to the forefront of businesses today:

- Recommendation engines similar to those used in Netflix and Amazon that use past purchases and buying behavior to recommend new purchases.
- Risk engines for a wide variety of business areas, including market and credit risk, catastrophic risk, and portfolio risk.
- Innovation engines for new product innovation, drug discovery, and consumer and fashion trends to predict potential new product formulations and discoveries.
- Customer insight engines that integrate a wide variety of customer-related info, including sentiment, behavior, and even emotions. Customer insight engines will be the backbone in online and set-top box advertisement targeting, customer loyalty programs to maximize customer lifetime value, optimizing marketing

campaigns for revenue lift, and targeting individuals or companies at the right time to maximize their spend.

- Optimization engines that optimize complex interrelated operations and decisions that are too overwhelming for people to systematically handle at scales, such as when, where, and how to seek natural resources to maximize output while reducing operational costs—or what potential competitive strategies should be used in a global business that takes into account the various political, economic, and competitive pressures along with both internal and external operational capabilities.

Today we are at the tip of the iceberg in terms of applying predictive analytics to real-world problems. With predictive analytics you can realize the uncontested market space [competitive free] that Kim and Mauborgne described in *Blue Ocean Strategy*

Software as a Service BI

The software industry has seen some successful companies excel in the game of *software as a service*(SaaS) industry, such as salesforce.com. The basic principal is to make it easy for companies to gain access to solutions without the headache of building and maintaining their own onsite implementation. When you add up the costs of the people and technology, SaaS is far less expensive too. The solutions are typically sold by vendors on a subscription or pay-as-you-go basis instead of the more traditional software licensing model with annual maintenance fees.

According to a recent article in *TechTarget*, "SaaS BI can be a good choice when there's little or no budget money available for buying BI software and related hardware. Because there aren't any upfront purchase costs or additional staffing requirements needed to manage the BI system, total cost of ownership (TCO) may be lower than it is with on-premise software—although overall SaaS BI costs will depend on the amount of usage the tools get."

Another common buying factor for SaaS is the immediate access to talent, especially in the world of information management, business intelligence (BI), and predictive analytics. More than a decade ago, analytics legend John Brocklebank of SAS (not to be confused with SaaS) created a thriving analytics on-demand center that allows companies to gain access to Ph.D.-level statisticians who deliver sophisticated output within a simple BI portal. This is now one of SAS's fastest growing lines of business, which is logical given the shortage in predictive analytics talent.

In the world of web analytics, there was another significant SaaS BI invention named Omniture (now owned by Adobe). Omniture's success was fueled by their ability to handle Big Data in the form of weblog data. We spoke with Josh James, the creator of Omniture and now the founder and CEO of Domo, a SaaS BI provider. Our first question for James was why his business was so successful:

> In addition to the Omniture people, several other reasons stand out to me. They include:

- **Scaling the SaaS delivery model**. We built Omniture from the ground up to be SaaS and we understood the math better than the competition. We invented a concept called the Magic Number. The Magic Number helps you look at your SaaS business and helps you understand the value you are creating when standard GAAP accounting numbers would lead you to believe the opposite.
- **Killer sales organization**. Once we had a few well-known customers like HP, eBay, and Gannett, we stepped on the pedal from a competitive standpoint and really went to battle against the other sales organizations and we won. We focused the whole company on sales.
- **A focus on customer success**. We had 98 percent retention rate. Customer happiness and success were always first because in a SaaS business, unlike traditional enterprise software, it's too easy for customers to leave if they are not happy. James explained the three market reasons why he started Domo, knowing we had to fix three problems in traditional BI. Here is a summary in his own words:

1. **Relieving the IT choke point**. Removing the friction for BI to become useful and enabling IT to be more strategic by enabling self-service BI.
2. **Transforming BI from cost center to a revenue generator**. Addresses a very common frustration that I've experienced as a CEO and that other CEOs have shared with me . . . now that we've invested in capturing all this data—how do we benefit from it?
3. **The user experience**. Is where we are putting all our marbles. Today's BI is not designed for the end user. It's not intuitive, it's not accessible, it's not real time, and it doesn't meet the expectations of today's consumers of technology, who expect a much more connected experience than enterprise software delivers. We'll deliver an experience with BI that redefines BI and is unlike anything seen to date.

Although this model makes a lot of sense, we can't help but remind ourselves that there are two sides to every decision. Tech Target pointed out: "With SaaS BI the analysis tools may not have all the features that on-premise software products do—which may make them less complex and difficult to use, but also less functional. Sending corporate data beyond the firewall also raises red flags for some IT managers. To try to assuage those concerns, some vendors have created private analytic clouds that run inside a customer's firewall."

Mobile Business Intelligence Is Going Mainstream

Analytics on mobile devices is what some refer to as putting BI in your pocket. Mobile drives straight to the heart of simplicity and ease of use that has been a major barrier to BI adoption since day one. Mobile devices are a great leveling field where making complicated actions easy is the name of the game. For example, a young child can use an iPad but not a laptop. As a result, this will drive broad-based adoption as much for the ease of use as for the mobility these devices offer. This will have an immense impact on the business intelligence sector.

We interviewed Dan Kerzner, SVP Mobile at MicroStrategy, a leading provider of business intelligence software. He has been in the BI space for quite a while. People have been talking about mobile BI for quite some time, especially since the 1999 release

of the good-old BlackBerry. However, it seems as though we have finally hit an inflection point. Kerzner explains his view on this topic:

> We have been working on Mobile BI for a while but the iPad was the inflection point where I think it started to become mainstream. I have seen customers over the past decade who focused on the mobile space generally and mobile applications in particular. One client in particular told me that he felt like he was pushing a boulder up a hill until he introduced mobility to enhance productivity. Once the new smart phones and tablets arrived, his phone was ringing off the hook and he was trying to figure out which project to say yes to, because he couldn't say yes to everyone who suddenly wanted mobile analytics in the enterprise.

That experience of folks who have been trying to use mobility for a long time to drive productivity and having really only pockets of success and then suddenly flipping over and becoming very pervasive is starting to be well understood now. In terms of why that's the case, Dan's perspective on that is that with the advent of touch-driven devices, you get a set of phones that are really much more about software than they are about being a phone:

> You turn off the iPhone and it's kind of a brick, nothing to it. It doesn't look like a phone. But you turn it on and the animating experience of it is the screen and the software that flows through that screen and the usability you get from having that touch-driven device. What's happened is suddenly you get a world where you actually have a form factor which lends itself to the power and flexibility, creativity, and innovation that comes with software development. That hadn't really been the case before. You sort of had it with some of the Palm organizer devices that were out there and you started to have it in a light-touch way with the early Blackberries. But it was always still your phone first, your messaging, you weren't fundamentally software driven. I think the combination of multi-touch and having a software oriented device is what has unlocked the potential of these devices to really bring mobile analytics and intelligence to a much wider audience in a productive way.

Ease of Mobile Application Deployment

Another inflection point for the industry is the development and deployment of mobile applications. In the past, that was controlled by the relationship with the carrier. It used to be that if you wanted to push out a mobile application, the only way you could get that application on the phone for the most part was to go through the carriers. That meant there were development environments that were sometimes very proprietary or you had to develop one set of applications for one carrier and another set of applications for a different one, maybe a third for the RIM BlackBerry environment. It didn't lend itself to very fast detonation because there was a real channel control now. Kerzner elaborated:

> One of the things that's happened recently is that with the advent of these app stores and the maturing of the browsers on the devices into something much more powerful, now as a software provider, you can go directly to the end user. I can go to a corporation and say I'm going to roll out a powerful global reporting

application that's also going to do deal approvals and it's going to totally change a whole business process. I think something that was previously locked in a desk will now give people insights into the purchasing patterns, as well as the ability to take that action. I can roll out that whole application—I never have to talk to anybody but that customer because the devices that everybody's lugging around are really little computers and of course you can put any software you want on a little computer and that really wasn't the case historically in the mobile space.

Three elements that have impacted the viability of mobile BI:

1. Location—the GPS component and location . . . know where you are in time as well as the movement.

2. It's not just about pushing data; you can transact with your smart phone based on information you get.

3. Multimedia functionality allows the visualization pieces to really come into play.

Three challenges with mobile BI include:

1. Managing standards for rolling out these devices.

2. Managing security (always a big challenge).

3. Managing "bring your own device," where you have devices both owned by the company and devices owned by the individual, both contributing to productivity.

Crowdsourcing Analytics

In October 2006, Netflix, an online DVD rental business, announced a contest to create a new predictive model for recommending movies based on past user ratings. The grand prize was $1,000,000! While this may seem like a PR gimmick, it wasn't. Netflix already had an algorithm to solve the problem but thought there was an opportunity to realize additional model "lift," which would translate to huge top-line revenue. Netflix was an innovator in a space now being termed *crowdsourcing*. Crowdsourcing is a recognition that you can't possibly always have the best and brightest internal people to solve all your big problems. By creating an open, competitive environment with clear rules and goals, Netflix realized their objective and, yes, they did create a lot of buzz about their organization in the process.

Crowdsourcing is a great way to capitalize on the resources that can build algorithms and predictive models. Let's face it, you can't "grow" a Ph.D. (or big brain) overnight. It takes years of learning and experience to get the knowledge to create algorithms and predictive models. So crowd sourcing is a way to capitalize on the limited resources that are available in the marketplace.

It's often been said that competition brings out the best in us. We are all attracted to contests; our passion for competing seems hardwired into our souls. Apparently, even predictive modelers find the siren song of competition irresistible.

That's what a small Australian firm, Kaggle, has discovered—when given the chance, data scientists love to duke it out, just like everyone else. Kaggle describes itself as "an innovative solution for statistical/analytics outsourcing." That's a very formal way of saying that Kaggle manages competitions among the world's best data scientists.

Here's how it works: Corporations, governments, and research laboratories are confronted with complex statistical challenges. They describe the problems to Kaggle and provide data sets. Kaggle converts the problems and the data into contests that are posted on its web site. The contests feature cash prizes ranging in value from $100 to $3 million. Kaggle's clients range in size from tiny start-ups to multinational corporations such as Ford Motor Company and government agencies such as NASA.

According to Anthony Goldbloom, Kaggle's founder and CEO, "The idea is that someone comes to us with a problem, we put it up on our website, and then people from all over the world can compete to see who can produce the best solution."

In essence, Kaggle has developed a remarkably effective global platform for crowdsourcing thorny analytic problems. What's especially attractive about Kaggle's approach is that it is truly a win-win scenario—contestants get access to real-world data (that has been carefully "anonymized" to eliminate privacy concerns) and prize sponsors reap the benefits of the contestants' creativity.

Crowdsourcing is a disruptive business model whose roots are in technology but is extending beyond technology to other areas. There are various types of crowdsourcing, such as crowd voting, crowd purchasing, wisdom of crowds, crowd funding, and contests. Take for example:

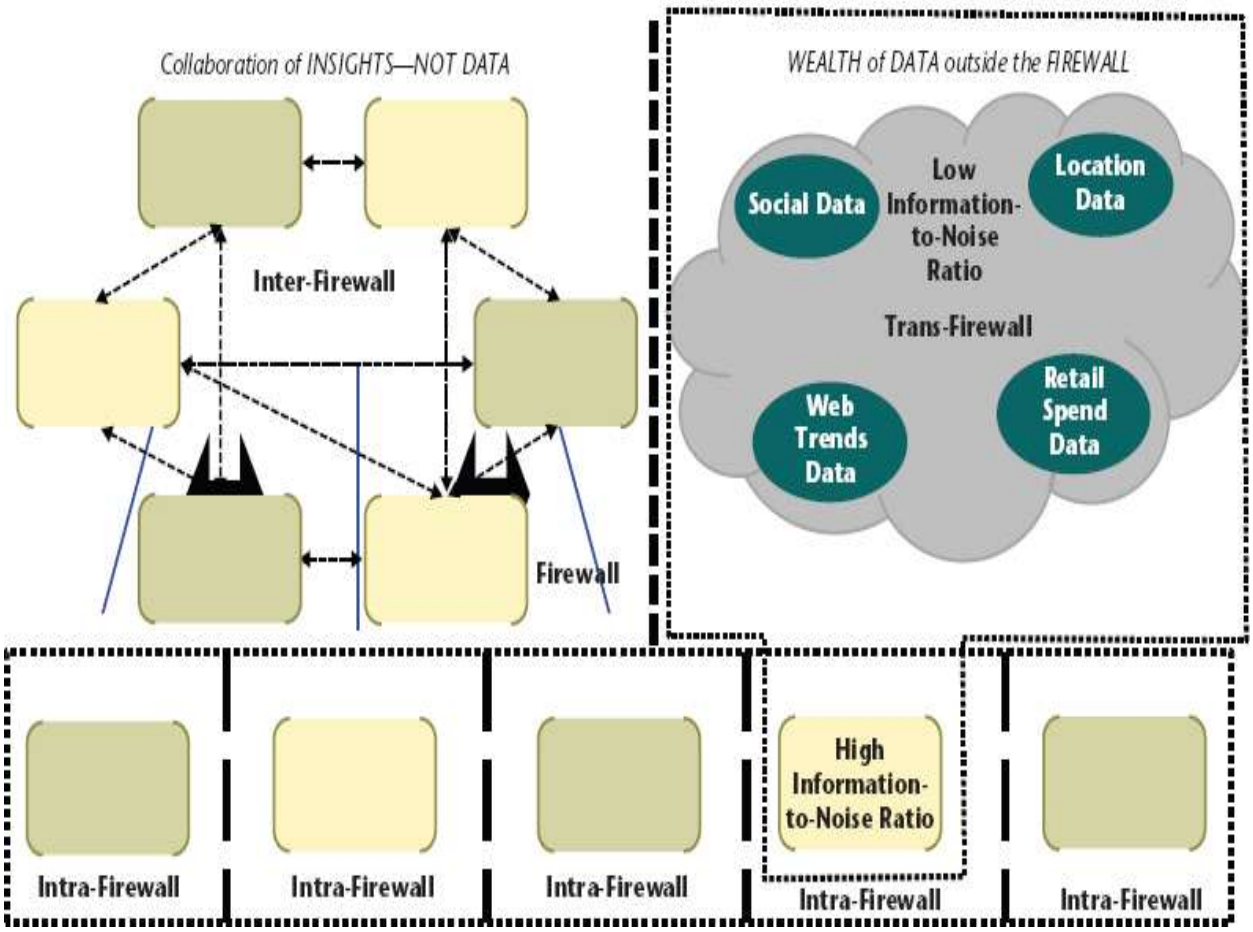
- 99designs.com/, which does crowdsourcing of graphic design
- agentanything.com/, which posts "missions" where agents vie for to run errands
- 33needs.com/, which allows people to contribute to charitable programs that make a social impact

Inter- and Trans-Firewall Analytics

Over the last 100 years, supply chains have evolved to connect multiple companies and enable them to collaborate to create enormous value to the end consumer via concepts such as CPFR, VMI, and so on. Decision science is witnessing a similar trend as enterprises are beginning to collaborate on insights across the value chain. For instance, in the health care industry, rich consumer insights can be generated by collaborating on data and insights from the health insurance provider, pharmacy delivering the drugs, and the drug manufacturer. In-fact, this is not necessarily limited to companies within the traditional demand-supply value chain. For example, there are instances where a retailer and a social media company can come together to share insights on consumer behavior that will benefit both players. Some of the more progressive companies are taking this a step further and working on leveraging the large volumes of data outside the firewall such as social data, location data, and so forth. In other words, it will be not very long before internal data and insights from within the firewall is no longer a differentiator. We see this trend as the move from intra- to inter- and trans-firewall

analytics. Yesterday companies were doing functional silo-based analytics. Today they are doing intra-firewall analytics with data within the firewall. Tomorrow they will be collaborating on insights with other companies to do inter-firewall analytics as well as leveraging the public domain spaces to do trans-firewall analytics (see following figure)

Organizations will need to complement just intra-firewall insights with inter- and trans-firewall analytics

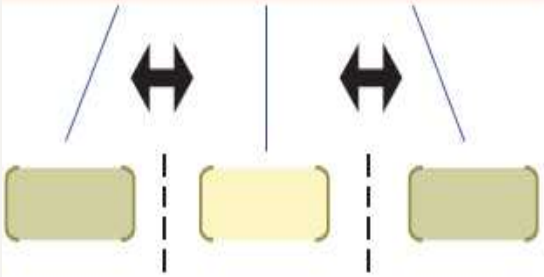


As Figure depicts, setting up inter-firewall and trans-firewall analytics can add significant value. However it does present some challenges. First, as one moves outside the firewall, the information-to-noise ratio increases, putting additional requirements on analytical methods and technology requirements. Further, organizations are often limited by a fear of collaboration and an overreliance on proprietary information. The fear of collaboration is mostly driven by competitive fears, data privacy concerns, and proprietary orientations that limit opportunities for cross-organizational learning and innovation. While it is clear that the transition to an inter- and trans-firewall paradigm is not easy, we feel it will continue to grow and at some point it will become a key weapon, available for decisions scientists to drive disruptive value and efficiencies.

Value Chain for Inter-Firewall and Trans-Firewall Analytics

Disruptive value and efficiencies can be extracted by cooperating and exploring outside the boundaries of the firewall

Inter-Firewall



Enables significant breakthroughs based on synergies in insights

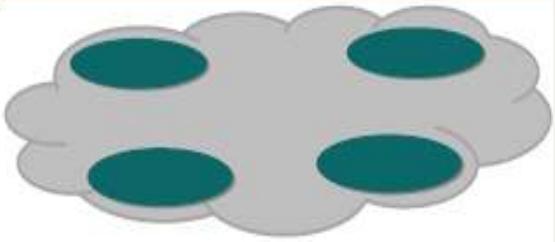
Value Chain

- Health Insurance + Pharmacy + Drug Maker
  - Customer health care insights – How does the consumer value his options?

Outside the Value Chain

- Search Engine + Retailer
  - Behavioral insights and outcome – How did the customer choose what they finally bought?

Trans-Firewall



Internal data is no longer a strong differentiator/game changer
Large volumes of data outside the firewall

New data explains previously unsolvable problems

- Consumer Social Interaction
  - Social feed data (outside firewall) + clickstream data (within firewall)
- Customer Price Elasticity
  - Price tests data (within firewall) + competitive prices (outside data)
  - What is the sensitivity to price changes in the presence of competitor pricing?