A New Analytics Paradigm in the Age of Big Data

How Behavioral Analytics Will Help You Understand Your Customers and Grow Your Business Regardless of Data Sizes

Table of Contents

Abstract

The explosion of business data, especially customer-related data, carries very high business expectations and contains huge amounts of value. But an "analytics action chasm" has emerged separating the vast amounts of data from a company's ability to effectively understand and use that data. The chasm is widened by the inability of current business intelligence and web analytics solutions to effectively bridge the chasm and unleash the data's potential to the business managers who need it most.

This white paper explains Quantivo's entirely new and patented technological approach to behavioral analytics and details Quantivo's ability to uncover valuable online and offline customer behavior patterns quickly. This paper also demonstrates the applicability of Quantivo's solution to the analysis of customer behavior, and explains how it exposes underlying business value.

The Growing "Analytics Action Chasm"

Mountains of Data, But Very Little Knowledge

Companies are capturing more data at an ever-increasing rate, but the value derived from that data is not increasing at the same pace. While companies capture every click, call, purchase, page views, and event, data warehouses become burdened and relational databases strain under the volumes. And while a few billion records used to be referred to as "big data," the pervasiveness of data collection systems has enabled even mid-sized companies to collect immense volumes of information. What used to be considered big data is now a daily reality for most companies, stressing their IT departments and overloading their business intelligence tools.

IDC recently estimated that 45GB of data currently exists for each person on the planet: that's a mind-blowing 281 billion gigabytes in total (or 281 "exabytes"; see sidebar). Furthermore, IDC forecasts this volume of data will grow at a staggering 60 percent per year! Wal-Mart alone handles more than one million customer transactions every hour, feeding databases estimated at more than 2.5 petabytes, according *The Economist* magazine.

Compounding the challenge are the executives who ask, frustrated after spending millions of dollars on IT hardware and advanced analytics solutions and hiring dedicated experts to run proprietary queries, "Why can't you answer my questions? The data is being collected, so why can't I see the value?" The bottom line: today's enterprises are overwhelmed by their data yet can't understand what it's telling them. This is the "analytics action chasm." (See Figure 1.)

What's an "Exabyte"?

1000 Kilobytes	=	*1 Megabyte*
1000 Megabytes	=	*1 Gigabyte*
1000 Gigabytes	=	*1 Terabyte*
1000 Terabytes	=	*1 Petabyte*
1000 Petabytes	=	*1 Exabyte*

Or...1 Exabyte = 1 billion Gigabytes or 10^{18} bytes of data

The challenge is no longer how to gather and store information, but how to turn very large amounts of data into useful knowledge. As relational databases and related technologies struggle to adapt, the common approach is to just throw more hardware at the problem. While this may work as a short-term solution, it increases costs dramatically while increasing performance only marginally.

Over the past few decades, companies have poured money into the same relational database technologies, fractured data collection systems, and monolithic analytics and business intelligence (BI) tools with little progress on the original and ultimate goal: getting answers into the hands of the business decision makers. This analytics action chasm – the distance in time, expertise, and cost between the data and the marketers, merchandisers, managers, and business analysts – has continued to grow even as companies have thrown billions of dollars at it.

Three things continue to widen this chasm:

1. **The tired and expensive fundamental approach to BI.** Since the first databases were developed, BI and analytics tools have been an expensive proposition for companies in terms of software, training, and IT resources. And even with decades of attempts at advancement, business users still remain frustrated at the lack of answers. As more data is collected on the web, in stores, via

mobile channels and CRM systems, and across social networking venues, the ability for standard relational technologies to digest huge data processing loads – let alone provide insights and understanding – becomes questionable.

Figure 1: *The "analytics action chasm" is the gap between data and the business users who need it. It is created by processes that require too much time to provide insights, tools that require layers of highly-trained resources to extract insights, and the fundamental Business Intelligence model which relies on decades-old solutions that cost millions of dollars to deploy and maintain.*

2. **The expertise required to ask even simple questions.** With traditional relational database technology, knowledge discovery relies on labor-intensive data mining techniques that typically allow only limited amounts of knowledge to be uncovered. Analysts must make educated guesses about the kinds of knowledge they believe exists in their data sets or the kinds of questions they will be expected to answer, and then formulate computationally-intensive SQL queries to actually "discover" that knowledge. While new user interfaces have been developed, BI and advanced analytics tools remain the domain of PhDs, statisticians, and hard-core analysts. Even a simple question, such as when a marketer asks, "What content did these set of customers view online?", lacks an easy answer, and drilling deeper, such as asking, "What *other content* did these same customers view afterwards?", forces a multi-step process with various analytics intermediaries and a time-consuming formalized request process. One cause of the problem is simple data accessibility. A recent Forrester Research survey found that nearly 60 percent of respondents claimed that they were unable to access the data needed for their typical analyses.

3. **The time required to get actionable answers.** These first two issues create this third issue: it takes too long to get an answer. Data warehouse users spend hours, days, or weeks digging into their data trying to find meaningful patterns and information clusters to answer a single question. In today's fast-paced, web-driven world, it's no longer viable to wait a week for an answer. Trends and opportunities spike and diminish in days, and the frontline decision-makers need access to answers, and the inevitable follow-up questions, in minutes. The difference between success and failure is the difference between answers today versus next week.

To be successful today, businesses need a complementary solution that takes the pressure off of their existing BI infrastructure, allows lightly-trained business users to get answers to both simple and complex questions, and provides insights even when customer data grows into the billions of records.

The Breakthrough Technology Underlying Quantivo's Behavioral Analytics

Quantivo has developed an advanced, patented analytics solution that represents a fundamentally new approach to uncovering value from the mountains of data that companies are collecting. Our database is a fully-indexed, intelligent database that recognizes and stores the patterns and affinities within customer behaviors as well as other important statistics – such as distinct counts, averages, and correlations – all while maintaining the original source data. By focusing on the hierarchical patterns among data elements, Quantivo can quickly identify associations and opportunities across all interactions, regardless of the volume of data, the segment being analyzed, the target being compared, or the data sources being combined.

But how is this really different, and why is that better?

Counts Instead of Additional Records

Let's start with an example. Imagine reading every book available on Amazon.com. After reading the first book, you will have seen thousands of words and word combinations. The second book will introduce more words and combinations, as will the third, the fourth, and so on. After maybe a dozen books you will have encountered most of the words and word combinations that appear throughout the entire Amazon library. Your "database" will contain a large majority of the words and phrases from the entire library after reading a small fraction of the books.

This same concept of quickly accumulating most of the data points that will ever occur also applies to business data. However, in corporate databases and especially in customer behaviors, information is even more redundant than the written word, so this process occurs even faster.

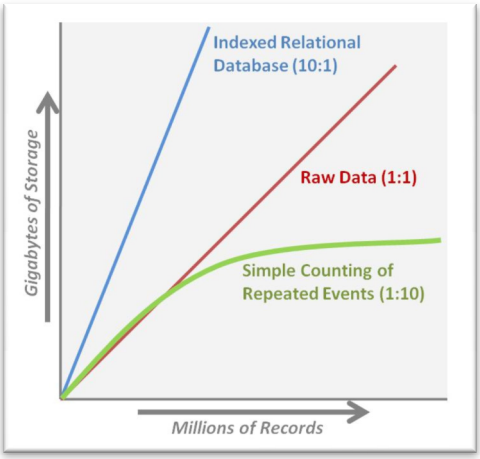

Figure 2: *Indexed relational databases significantly increase storage requirements – by 10X as conservatively shown here – while Quantivo increases storage requirements by only 10 percent.*

Business- and customer-related information is comprised of repeated events. Natural patterns form and segmentations quickly emerge from those patterns. A typical grocery store offers thousands of items for sale, creating billions of possible item combinations. But a simple market basket analysis would uncover that only a relatively small number of combinations actually occur, and that there is a huge number of product combinations that have never been purchased together in the same basket. It's the difference between "possible" combinations and "actual" or observed combinations that enables Quantivo's massive compression of data. (See Figure 2.)

The first key difference between Quantivo and relational databases is that we exploit this redundant nature of data, tracking specific data points (names, SKUs, colors, URLs, clicks, events, etc.) and combinations of data points that have actually occurred, and then simply count repeated values within the data set. This is in stark contrast to the unscalable model used by conventional relational databases that store every data point and combination as unique.

To highlight this difference, let's return to our book example. If the word "Caribbean" appeared 50 times in the first 100 books you read, Quantivo would store that as a single record with a count of 50, whereas relational databases would create 50 distinct records.

One Important Result – Lower Storage Requirements

All analytics applications need additional space for the "analyzed database" that is created from the original data. But because of how Quantivo stores information, that need diminishes as the data is transformed. The analyzed database is a fraction of the size of the original data, and as more data is added to the database the rate at which the analyzed database grows continually slows. This is because nearly all possible data points and data combinations have already been captured, and any new data is just an increment in an already-noted pattern. As such, Quantivo makes much more efficient use of available storage (and subsequently the processing power and time required for analysis), generating analyzed databases that are on the order of 10 percent of the size of the original data. So, 1GB of original data requires a total of 1.1GB of storage.

Conventional, indexed relational databases create additional records for all of the original data as described earlier, plus add indexes and other incremental information to aid in searching and analysis. The result is an analyzed database that, according to the Transaction Processing Performance Council (www.ptc.org), can be, *on average*, 48 times larger than the original data. So that 1GB of original data now requires a total of 49GB of storage. This not only increases storage costs, it also means that searches and analysis must struggle through massively larger amounts of data than Quantivo, slowing the process considerably.

Diving Deeper – Associations Among Data Points

Just counting how many times an event occurs is not enough. A further distinction in how the Quantivo database works is rooted in the concept of associations. Association is a quantitative measure of the correlation or relevance between two data points, derived from past data that can be used as a predictor of future behavior.

Quantivo enables the sorting of patterns by relevance, or more precisely, by an association metric. Association, represented as a multiple, is used when considering a specific behavior of a specific segment of the population as compared to the overall population. For example, if something is six and a half times more likely to occur within a particular segment than in the population as a whole, then the association is 6.5.

As shown in Figure 3, association can be positive or negative:

- A positive association indicates that the behavior being analyzed is *more likely* to occur in the target segment than the overall population.
- A negative association indicates that the specific behavior under analysis is *less likely* to occur in the target segment than the overall population.
- An association of zero indicates that the behavior is uncorrelated with the target population and that it is no more likely to occur within the target population as it is across the entire population.

Associations are calculated within Quantivo based upon a proprietary calculation that is essentially the correlation. Let's look at a simple example to further explain Quantivo associations.

An electronics retailer may want to understand how to increase overall sales by increasing the lift that HDTVs have on other products. With current BI tools, they can find simple metrics that show sales by region or week, or the sales revenue generated by various products. What does this tell them and what decisions might the merchandiser make to increase revenue? Alone, these data points mean little, but with Quantivo and associations, much more can be understood.

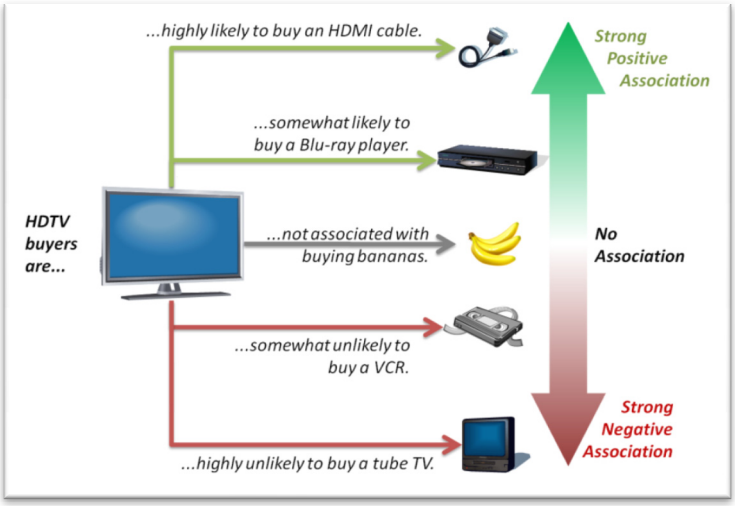

Figure 3: *Association compares behavior in a specific segment, such as everyone who previously purchased an HDTV, to the overall population of all customers. Positive associations suggest a correlation between the behavior being analyzed and the target segment, while negative associations indicate a repulsive correlation. Buyers of HDTVs have a positive association to the purchase of a Blu-ray player, but a negative association to the purchase of a VCR. The association between the purchase of an HDTV and the purchase of bananas is at or near zero, since there is no correlation between these two behaviors.*

For the purpose of this example:

- The overall population is everyone who has ever purchased anything.
- The target segment is everyone who has purchased an HDTV.
- The behavior being analyzed is what else customers purchase in the two weeks after an HDTV purchase.

Looking more closely at the sales of HDTVs, Quantivo calculates an association of -8.7, for example, with VCRs, which tells us that customers who purchased an HDTV are very unlikely to purchase a VCR after the HDTV purchase. On the other hand, an association of +7.6, for example, with Blu-ray players indicates that HDTV buyers are very likely to make a Blu-ray purchase.

By understanding this simple affinity – which is difficult for standard tools to calculate, especially as data volumes grow into the tens or hundreds of millions of transactions – the retailer has quickly identified that there is no need to promote Blu-ray players along with HDTVs, since the strong association indicates that these purchases tend to occur within a week of each other anyway. However, they do want to ensure that the purchase happens at their store, so their next promotion could combine this obvious product lift with a small incentive: save 10 percent when you buy a Blu-ray player *with* your HDTV!

While there is an obvious connection between HDTVs and Blu-ray players, especially to those who already own HDTVs, Quantivo customers have been able to quickly identify other, less obvious but more valuable examples, such as:

- A home improvement retailer discovered that buyers of holiday decorations had a high association to the purchase of extension cords *and* large plastic storage containers, encouraging them to move these products closer together during the holiday buying season.

- An online media content provider found that recommendations based on content genre had low or negative associations with the content being viewed, while the content's release year had positive associations that helped increase both page views and session frequencies.

- A pharmaceutical company uncovered that visitors to a product aimed at teen girls would not click on the "what to ask my parents" link until after their fourth or fifth session, pushing their marketing team to create more-compelling follow-up campaigns to ensure a return visit, while also easing navigation to that "conversion" point.

- A cross-channel home furnishings retailer discovered that customers who purchased a small quantity of bath towels in-store, had a high association with purchasing the remaining matching set via their online channel within a few days, opening an opportunity to salvage sales lost to competitors and also reducing the costs related to product returns.

So Why Are Associations Important?

Most associations are answers to questions yet to be asked. Because queries of large data sets often require significant querying expertise, processor time and IT resources (i.e. they are expensive), with other tools, users often restrict queries to only a fraction of the information known to be available in the data set. Relying on a sample dataset reduces querying times, but eliminates any chance of drilling down into the underlying data or discovering the unexpected.

Furthermore, Quantivo's associations depict only observed events, rather than possible events. Within analytics, this distinction is critically important. By tracking only observed events, or "what has actually happened" instead of "what could possibly happen," Quantivo can accommodate larger and more complex data sets for analysis while still presenting all of the associations in the data, without programmed queries and without guesswork (see sidebar for an example). This makes Quantivo ideally suited to the massive amounts of data captured by various web, support, and in-store channels.

Counting "What actually happened" vs. "What could happen"

Consider all possible three-letter permutations of "A," "R," and "T." There are 3^3, or 27, possible permutations: AAA, AAR, AAT, ARA, ATA, etc. However, in a database of three-letter English words (i.e. what actually happened), only four are observed: ART, RAT, TAT, and TAR. Therefore, the number of observed three-letter permutations is a small subset, only 15%, of possible combinations.

The implications for data analytics is that when asking how many times something has occurred (i.e. the word "art" occurring within a database), Quantivo only needs to search through a small fraction of records that a relational database would need to search, so can more quickly find the result. On the contrary, conventional relational databases must search for all possible three-letter combinations for this particular instance and count along the way, significantly complicating the search and therefore the time needed to complete it.

For most business questions, there are billions of associations that have never occurred. Guessing at what may have happened and asking the question, "Did X and Y ever sell together?" with a relational database requires a query of the entire database. Ironically, due to the enormous volumes of data, associations that previously would be immediately evident are missed by traditional methods and the underlying value is never discovered. So, more data actually translates to less knowledge.

Quantivo needs to query only the observed combinations, which is exponentially smaller, and therefore faster to analyze. In addition, associations that a database analyst may never have considered are still

captured by Quantivo just by the fact that they occurred and have a count associated with them, eliminating the guesswork about what may have happened. As the possible combinations increase, such as with website page or content views, Quantivo's advantages in both cost and time become even more pronounced.

Technology Ideally Suited To Analyzing Big Data Without Losing The Details

Two things are needed to go beyond merely coping with data overload: a new technology for understanding what lies within the data, and a new infrastructure and delivery mechanism that doesn't require major involvement from IT. The Quantivo "pattern-store" database described earlier addresses the first issue of understanding what lies within the data, and its cloud computing infrastructure addresses the second: providing the stability, scalability, and low cost without IT involvement or high hardware and maintenance investments.

Without cloud-based solutions such as Amazon's EC2 focusing entirely on advancing and streamlining data processing and storage, it would almost be physically impossible for a business to expand and upgrade their IT infrastructure at the pace needed to keep up with all the data, not to mention the cost and logistics involved with such hardware changes. The cloud offers cutting-edge processing power and virtually limitless storage at just a few dollars per month – and you only pay for what you actually use, not what you might need to prepare for your most extreme possibilities.

While other "new" analytics technologies resort to data aggregations and summaries to speed analysis (at the cost of detail and granularity), Quantivo utilizes a lossless and fully-reversible transformation of the source data – there is no aggregation – preserving the original context, relationships, and facts. This allows for questions to be asked directly against the data, value to be uncovered, and all original records, such as customer account numbers or email addresses, to be immediately available without any additional effort.

The "Cloud" and Why It's Important

"Cloud computing" means using computer resources – servers, processors, storage, software – that resides on the Internet, not within a company's own facility. The immediate benefits of this are eliminating fixed up-front costs, reduced downtime, improved back-up and security, and the ability to scale as demand increases. And this is the extent to which most applications utilize the cloud.

Quantivo takes it a step further. The Quantivo architecture was designed to take full advantage of the cloud. The idea behind the cloud is that there are virtually unlimited processors out there available to be used. So, as the number of queries increases, Quantivo uses more processors. For more parallel queries, it divides the work between more processors. So, it automatically scales as the work load increases, delivering equally fast results regardless of the workload being tackled.

Converting Associations Into Business Value

Unlike data mining or other analytical methods, which rely primarily on guesswork and brute force querying, Quantivo quickly and easily reveals comprehensive behavioral information and presents it in an interactive and intuitive form. And, being cloud-based, Quantivo eliminates any concern over data storage, scale, or processing power – each of which is there in abundance when you need it.

Quantivo's behavioral analytics takes existing website, contact history, financial, and other online and offline data and turns it into actionable insights that identify the patterns that matter most across transactions, page views, purchases, returns, support calls, and any other customer behavior.

By providing a deep view into customer actions, Quantivo's behavioral analytics enables the discovery of valuable patterns and trends hidden in mountains of data and behind complex, expensive business intelligence, web analytics, and data collection systems, helping decision-makers understand the page views that precede a likely transaction, campaigns that drives conversions, and factors that increase stickiness.

Quantivo offers three distinct advantages over traditional business intelligence and web analytics tools:

1. **Relevant, measurable behavioral analytics.** Simple operational reports only give a sense of what happened, but do not answer why it happened or what to do next. With Quantivo, business managers can uncover the behavioral patterns that show how best to segment visitors, where they spend their time, when they will click on an ad, and which campaigns work with which product and segment combinations. By focusing on patterns and affinities across data and over time, Quantivo identifies the most profitable and monetizable segments, content, ads, and strategies that provide the edge needed to increase visitors, stickiness, conversion rates, and ultimately drive more revenue.

2. **The speed of the cloud.** Quantivo's on-demand, cloud-based infrastructure provides the speed and capacity that gives decision makers the power to dynamically explore virtually limitless data volumes, breaking free from the complexities and time requirements of predefined reports and proprietary query languages. Additionally, Quantivo's distributed architecture was built from the ground up to take advantage of cloud computing. Unlike other so-called cloud solutions which just throw traditional technology into a SaaS environment, Quantivo truly exploits the cost, speed, and scalability advantages of the cloud.

3. **Actionable insights to the decision makers.** Quantivo puts these actionable insights into the hands of frontline decision makers in marketing, sales, support, the website, and other business functions with easy-to-use discovery tools that deliver behavioral insights needed to drive new opportunities. Additionally, Quantivo extends the dynamic nature of analysis, allowing decision makers to follow their train of thought as they dive into a specific hypothesis, or to work faster when new or unforeseen behavior patterns are uncovered.

Summary

Compared to traditional business intelligence and advanced analytics tools, Quantivo offers a revolutionary way for companies to analyze their increasingly large mountains of data. Whether focused on speed, flexibility, ad hoc analysis, cost, or scalability, Quantivo is vastly different from – but complementary to – current BI solutions.

A patented core technology allows Quantivo to leap beyond the tired, decades old column/row approach to data analysis and uncover patterns that are both actionable and measurable. A distributed architecture designed from the ground up allows Quantivo to exploit the advantages of cloud computing for large-scale databases. Offering the solution at a low cost and in just days allows companies of all sizes to close the analytics action chasm and start to identify compelling business opportunities in their customer behavior data.

For more information, please visit www.quantivo.com.

About Quantivo

Quantivo is pioneering a fundamentally new approach to Behavioral Analytics. For the first time, corporate decision-makers can quickly uncover behavioral patterns across any aspect of their customer interactions and act on these timely insights to increase customer acquisition, retention, up-sells, cross-sells and web monetization. By combining large-scale analytics with rapid SaaS delivery, Quantivo is the only company that puts impactful revenue-generating insights directly into the hands of a wide set of executives and analysts for applications such as market basket and loyalty analysis, merchandising and marketing optimization and online analytics.