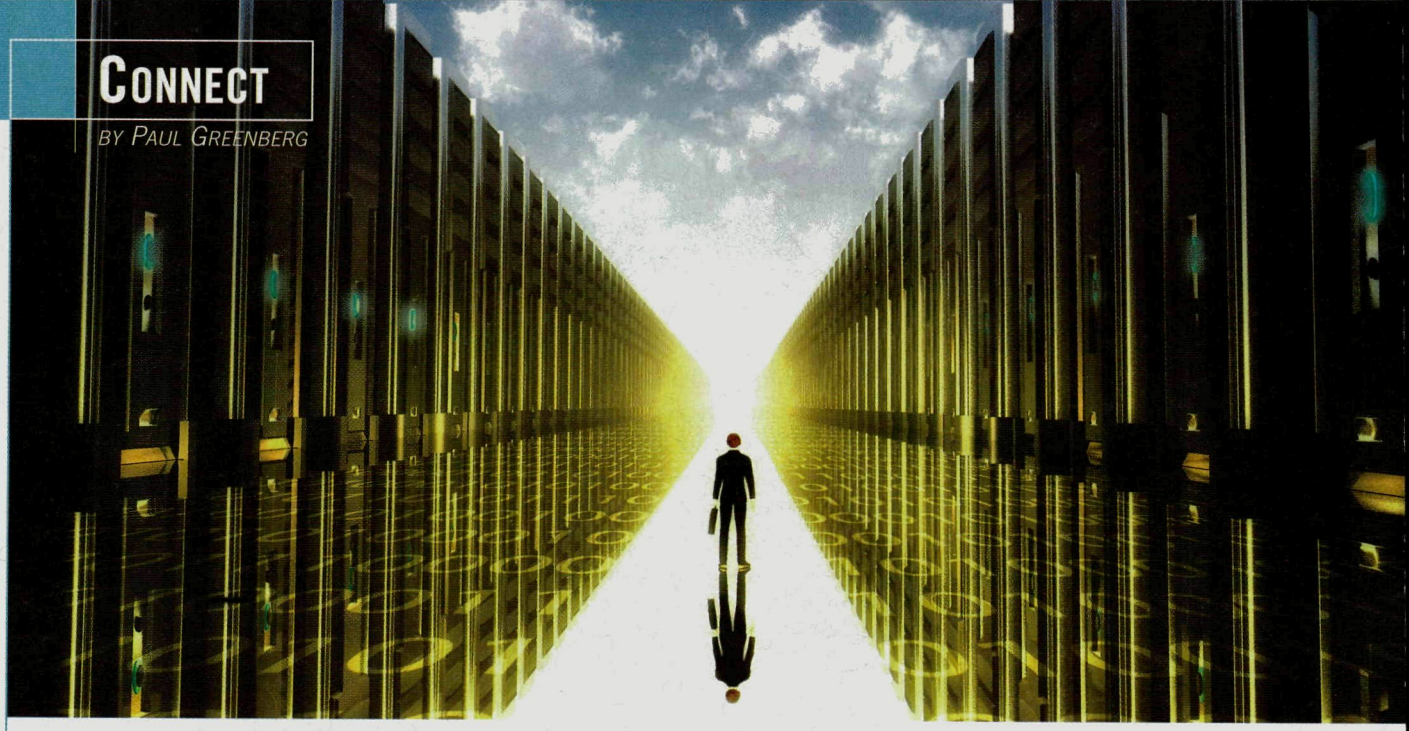
BY PAUL GREENBERG



Big Data, Big Deal

What you don't know can hurt your business

THE TITLE says it all, doesn't it? Or does it? We are in what has been called, often enough, the era of big data. Yet you read the title of this article, four words, extremely little data, and even with those few words to analyze, you can't figure out if I mean Big Deal = who cares or Big Deal = really important, can you?

This lovely irony just highlights one of the biggest problems that so-called big data presents. What can you make of it?

Let's make this a really basic primer for a businessperson on big data, given the level of conversation about it.

What is big data, anyhow?

For those of you who haven't been paying much attention, big data is the name given to the prodigious amounts of fast-moving data that typically can't be handled by existing data tools. Andrew Brust, a fellow ZDNet blogger, gives it a simple definition:

"We can safely say that big data is about the technologies and practice of handling data sets so large that conventional database management systems cannot handle them efficiently, and sometimes cannot handle them at all."

The data in question can be structured or unstructured, though the latter is where it really comes from. We are talking the proliferation of zettabytes of data daily, weekly, monthly. The amounts are staggering. For example, in 2011, the ubiquitous "we" created 1.8 zettabytes of data. A zettabyte is 10 to the 21st power. By 2015, this number is expected to be 7.9 zettabytes (thank you, CenturyLink, for these numbers).

We are also not limited here to text data; big data can include video, audio, and images too.

Here are some of the driving forces behind big data. In a month, a teen will text, on average, 4,732 times. In a day, two billion YouTube videos are watched and four billion pieces of content are shared on Facebook—about half of those pieces by a couple of friends of mine alone. According to a neat infographic by Get Satisfaction, on average, a securities firm with fewer than 1,000 employees will have 3.8 petabytes (defined as 10 to the 15th power) of data stored—and that number only grows.

While this is dramatic, amazing, and cool in some geeky way, we have a problem. What do we have to do to make sense of it all?

UNDERSTANDING BIG DATA

In a Gartner Twitter chat (#GartnerChat) earlier this year, Seth Grimes, a mega thought leader in the text and sentiment analysis space, lamented that "questionable data is the rule rather than the exception in my specialization areas."

This statement is downright scary.

Because, in order to actually meet the escalated requirements of contemporary customers, we have to understand their behaviors, try to interpret what they are thinking, and—since we need to do that on a larger scale than ever before, even if we have a small, growing business—the data we're attempting to interpret needs to be accurate and must be as easily interpretable as possible.

In the world of big data, especially the unstructured variety, the data quality issue escalates big data to levels that are likely to be unprecedented. For example, here are several issues that make it obvious why this can be a problem:

1. Misspelled words in an unstructured text.
2. Abbreviated terms—(e.g., LMTC: left message to call—an oldie, and the obvious: LOL) or social media terminology (e.g., a Twitter message, such as r u ok 2 go 2nite?).
3. Industry-specific terminology or organizations—e.g., CRM Playaz.
4. Categories that need to be interpreted—e.g., demographic data, financial data, geographic/geospatial data, property characteristics, and personal identifiable information.
5. Contextually appropriate data—e.g., "Overheard in Stoughton, MA—Why do red sox fans wear white sox?" Could the lowercase red sox be a reference to the Boston team? Or something related to some people who love red socks? It's from Massachusetts. You figure it out.
6. Emotionally interpreted semantic data—e.g., the title of this article. What's the context? Could it be sarcastic or a recognition of the magnitude? Do you know the answer to that yet?

In other words, the interpretation of the data doesn't only depend on its accuracy but on its clarity as well.

WHAT CAN BE DONE?

Even with clarity, how do we interpret this mass of data flying at us at the speed of light? After all, data is useless unless it becomes information and then is translated into actionable knowledge—aka insight—especially in the CRM world, which relies on accurate interpretation of information about customers to keep the customers interested enough in the company to minimally continue purchasing. Some is quantifiable information, some emotional and behavioral information. To make it even more difficult, we only have a small period before the data is dated and no longer valuable. Michael Wu, chief scientist for Lithium, calls this the "predictive window."

We do have tools that are now available, notably the Apache Foundation's free and open source Hadoop framework that can at least theoretically handle these large datasets by distributing up to petabytes of data files across multiple clusters of computers. This gives us the speed and flexibility to handle a high volume data of unprecedented scope.

What would be an example of how it works? Think about this one. Amazon has more than 100 million registered members. Most of them actually shop there. When you get on the site and look at a product, you almost always see "Customers Who Bought This Item Also Bought.... " Think about the incredible number of data points that are being examined almost instantaneously to give you results. They are looking at individual and general purchasing patterns and identifying products that fit profiles for "people like me"—a person who has similar transactional behavior and profiles to you. This effort is called "collaborative filtering." It's also called "personalizing the buying experience." It was big data that was sifted in almost real time to come up with something that you routinely expect from Amazon. So much so that you barely glance at it there—but you do look.

So tools like Hadoop can likely handle high-volume unstructured data at the level that we're calling big data. That's good.

WHAT'S THE BENEFIT?

There is no question that if we can master the scope of the data and the scale of some of the issues, there will be obvious benefits.

Let me clue you into something that I read in the *New York Times* about the U.N. Global Pulse initiative. Here's the way the paper describes it:

"The group will conduct so-called sentiment analysis of messages in social networks and text messages—using natural-language deciphering software—to help predict job losses, spending reductions, or disease outbreaks in a given region. The goal is to use digital early-warning signals to guide assistance programs in advance to, for example, prevent a region from slipping back into poverty."

DATA IS USELESS UNLESS IT BECOMES INFORMATION AND THEN IS TRANSLATED INTO ACTIONABLE KNOWLEDGE.

Partnered with SAS, what the initiative is preparing to do is use the behaviors of constituents to determine likely trends that they can act on. The data gives them markers that help the Global Pulse initiative forecast issues and prevent likely problems. For example, they were able to use social channel–focused conversations on increasing use of public transportation, cutting back on grocery expenditures, and downgrading cars, among other markers, to identify an unemployment spike.

That's amazing stuff—and only the beginning of what the availability of all this information portends. So pay attention. Big data is really a big deal.

Paul Greenberg (@greenbe on Twitter) is president of consultancy The 56 Group (the56group.typepad.com) and cofounder of training company BPT Partners. He is also the conference chair of CRM Evolution (www.destinationCRM.com/conference). The fourth edition of his book, CRM at the Speed of Light, *is available in bookstores and online.*