Big Data: Big Hole or Big Opportunity?

There is no doubt about the potential of this technology revolution but companies should remember that it is imperative to have a clear vision and value definition along with the technology prowess to achieve success

— Apurba Dutta, Founder & CEO, vCloudIT

Let's begin by understanding what is driving this revolution. In most organizations, information travels along familiar routes. Proprietary information is lodged in databases, analyzed in reports and then rises up the management chain. Information also originates externally—gathered from public sources, harvested from the Internet, or purchased from information suppliers.

But the predictable pathways of information are changing: the physical world itself is becoming a type of information system. In what's called the *Internet of Things*, sensors and actuators embedded in physical objects—from roadways to pacemakers—are linked through wired and wireless networks, often using the same Internet Protocol (IP) that connects the Internet. These networks churn out huge volumes of data that flow to computers for analysis. When objects can both sense the environment and communicate, they become tools for understanding complexity and responding to it swiftly. What's revolutionary in all this is that these physical information systems are now beginning to be deployed, and some of them even work largely without human intervention.

At the recently concluded Google's annual I/O conference for developers in San Francisco, The Data Sensing Lab, a project of O'Reilly Media, had deployed over 500 sensor motes at key locations around the Moscone West Centre. Each phone-sized mote is a self-contained computer, some measure temperature, pressure, noise, and humidity and light levels. Others were tracking air quality, motion of crowds and number of mobile phones being used; a network producing over 4,000 streams of data that was uploaded to the Google cloud for processing. Some interesting live visualizations depicted attendees flowing out of seminars and flocking around Google glass stand, the noisiest vs. quietest locations among others.

It is predicted that by 2020, there will be well over a billion machines talking to each other and performing tasks without human intervention. From phones to cars to bridges, embedded technologies are increasingly making the things we use smarter every day (building a *Smarter Planet*). For example, some of the newest cars use cameras mounted in the rear to see if something is in the way when you are backing up. If there is something in the way, the car will apply the brake even if you don't or you are slow to react. Likewise, the concrete in new bridges has embedded chips that can let engineers know when the concrete is cracking, stressed, and in need of repair before the bridge collapses. In addition, sensors on the surface of the road going over the bridge will detect ice and wirelessly communicate the information to your car. If you don't slow down, the car will slow down to a safe speed for you.

Augmented reality (AR) may seem like a futuristic concept, but it will be a reality of our digital lives in 2013. AR is a new technology that blurs the line between what's real and what's computer generated by enhancing what we hear, see and feel. Few companies have made some pretty impressive technological leaps: Vuzix debuted their monocular AR prototype; Innovega showcased their Augmented Reality contact lens, and this year was all about Google's Project Glass. Phones will be AR equipped as OEMs, handset manufactures and carriers are planning to integrate optimized AR, visual display and camera performance in the next generation of smartphones, because it's basically already happening. This year we are also expected to see the next generation of location-based data going beyond the on board GPS, gyroscope and camera to using Near Field Communication (NFC) to create things like 3D-Augmented city maps and hyper local information.

So as you can see, Big Data is a revolution – fundamental, economic transformation. It's not about size,

velocity, variability or any of the other Vs we could throw at it. For more and more businesses, data is not a convenience; it is in fact a factor of production. Essential to the way we deliver our goods and services to our market of choice. Fundamentally, true across industries, size of organizations, public or private sector boundaries, very few of us will escape this economic transformation.

"Big data" is high-volume, -velocity and -variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making. With too many pieces to put together, the Big Data ecosystem can be a bit overwhelming when you first try to put all the pieces together; there are just too many products, vendors and technologies. A simplified view of a Big Data strategy would depict the following stages:

1. **Acquire** and ingest information from federated sources both structured (as relational databases) and unstructured (log files, sensor information, social media streams…)
2. **Organize** this information and store it in a distributed file systems for integration. After that Map Reduce programs that will extract the value out of all that data process it, the results are then moved to a more traditional storage, like a data warehouse.
3. **Analyze** that information in the data warehouse using a BI tool that allows users to formulate questions that leverage all the original data.

Remember, numbers can't speak for themselves, and data sets - no matter their scale - are still objects of human design. The tools of big-data science, such as the Apache Hadoop software framework, do not immunize us from skews, gaps, and faulty assumptions. Those factors are particularly significant when big data tries to reflect the social world we live in, yet we can often be fooled into thinking that the results are somehow more objective than human opinions. Big Data is a powerful tool for inferring correlations, not a magic wand for inferring causality.

Velocity is the most misunderstood data characteristic: it is frequently equated to real-time analytics. Yet, velocity is also about the rate of changes, about linking data sets that are coming with different speeds and about bursts of activities, rather than habitual steady tempo.

Hadoop wasn't designed for enterprise data. Its original purpose was to manage publicly available information such as web links, and it was designed to format large amounts of unstructured data within a distributed computing environment, specifically Google's. It was not written to support hardened security, compliance, encryption, and policy enablement and risk management. □