the '**DATA rEVOLUTION**'
has begun

by Jenny Mangelsdorf

Regardless of where you are or what industry you work in, how you find and use data is dramatically changing. A data-driven economy is emerging, and organizations across the world are just beginning to understand it and explore the possibilities.

"Your own data is not enough," say the authors of CSC's new report, Data rEvolution, which explores the changing reality of IT where data is increasingly derived from both internal and external sources.

The 65-page report examines the world's rapidly escalating appetite for data, how organizations are creating new ways to manage and see that data and what discoveries and opportunities will arise from it.

"We began this report because we were seeing some pretty interesting and unconventional movements around data," says Paul Gustafson, CSC's Leading Edge Forum[1] Technology Programs director, who led the research. "Data growth was starting to drive what we've dubbed as our own sort of 'bitonomics,' where the data side of business is beginning to hold a new secret to success. It isn't just the factory or the assembly line or business processes — it's the data."

Data: yours, mine and ours

Challenges like evaluating climate change, analyzing property risk, managing energy and maintaining stock exchanges all have data as a primary component. The ability to analyze and act upon data, particularly external data, will be key for both solving challenges and gaining competitive advantage, says the report, which was also authored by Sidney Shek, a CSC solution architect.

Today, for example, climatologists aren't the only ones interested in weather data. Insurance companies, environmental organizations and energy suppliers are just three groups interested in gaining access to climate data, and related analysis, generated by federal agencies such as the U.S. National Oceanic and Atmospheric Administration (NOAA).
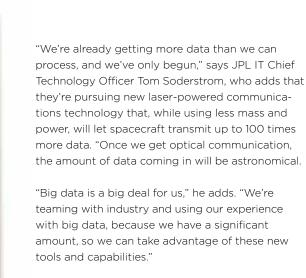
"Climate data is not going to live in a stovepipe any longer," says Gustafson. "Other businesses are creating value from a deeper understanding of climate trends and the symbiosis associated with it."

Flexing muscles to manage big data

It takes technology to unlock, manage and gain value from the volume of data the world is generating. Forget megabytes and gigabytes: Get ready to handle petabytes, exabytes, zettabytes and yottabytes. CSC's own High Performance Computing Center of Excellence[2] manages more than 110 petabytes, or 110 million gigabytes, of data for clients. The Jet Propulsion Laboratory (JPL), which conducts robotic space missions for NASA, is looking at missions that generate many terabytes a day. (As a comparison, its first mission generated a gigabyte of information over its lifetime.)

In the report, the authors also look at the emerging approaches and tools that organizations are using to manage and generate new, unique views on data. These include:

- 'Shared nothing' architectures, where tasks and data are distributed, potentially through a cloud, to separate systems, and worked on independently
- 'Unstructured' data, which is more chaotic and messy than traditional structured data, but potentially rich with contextual insight, and 'jagged' data, which is data in a variety of formats that was not originally intended to be combined
- Google's distributed processing framework, MapReduce, and related developments, such as Hadoop, the open source technology stack that is used by companies such as LinkedIn, Facebook, RackSpace and Yahoo, the largest contributor to Hadoop
- A 'new breed of databases,' made popular by the 'NoSQL' (Not Only SQL) movement, which is being adopted by various organizations to store and process data that has attributes that rarely 'fit' into traditional, highly structured, relational database models

"We're already getting more data than we can process, and we've only begun," says JPL IT Chief Technology Officer Tom Soderstrom, who adds that they're pursuing new laser-powered communications technology that, while using less mass and power, will let spacecraft transmit up to 100 times more data. "Once we get optical communication, the amount of data coming in will be astronomical.

"Big data is a big deal for us," he adds. "We're teaming with industry and using our experience with big data, because we have a significant amount, so we can take advantage of these new tools and capabilities."

Connecting the dots

The Data rEvolution report says new technology solutions will provide the muscle to manage and work data; however, to make real gains in this area, organizations will need to forge new connections. For example, to make energy infrastructure 'smart,' one key will be to have situational and predictive capabilities that can draw on weather data.

"Energy isn't going to get smarter just doing the same thing of rendering power — energy will get smarter when it recognizes consumption, and one of the correlations with consumption is weather," says Gustafson. "That's a new dot we've never connected before because of the systems and their maturity."

Aware of the potential competitive advantages data offers, in 2009, the U.S. launched data.gov to give the American public access to its wealth of data. Since then, the UK has established data.gov.uk, and the European Union is discussing beginning a data.gov.eu initiative.

One of data.gov's communities, health.data.gov, has helped spawn a new company, Asthmapolis.com, which aims to help patients and doctors improve asthma management by giving people global positioning system-equipped inhalers. When used, the inhalers generate a new source of data that physicians, scientists and public health officials can use to improve asthma management and identify asthma-related triggers in the environment.

"Those dots [health and environment] have never fully been connected in this broader system approach," says Gustafson. "Initiatives like data.gov are bringing the data forward so others can use it, and this is where the magic will be.

"For health, it's a step toward being more proactive and predictive, and data is taking the lead in that. A lot of people believe there are golden gems inside an aggregate of health data that will lead to better understanding, drugs and treatment."

Predicting the future

In any industry, the ability to solve or anticipate a problem can be enhanced by asking the right questions. The report cites Endeca's MDEX Engine, a hybrid search-analytical database developed to help users find and understand information.

Toyota used Endeca's solution to help resolve its gas-pedal crisis. The technology let the automaker's engineers sift through years of product and quality data from numerous systems in ways that were not previously possible, and identify patterns they would not have known to look for in the past.

Gustafson says Toyota also had to draw on external data, such as from transportation agencies and other manufacturers, to determine what was happening with its gas pedals. This is another example of an organization's internal data not being enough to solve today's challenges.

Mining the social space

Organizations are also beginning to experience the opportunities available in mining and understanding social media data and the potential pitfalls of not considering that domain.

The authors cite a major consumer products company and reports that its stock price fell after a group of angry parents began saying on Facebook that one of the company's products caused problems. The company, known for protecting its brand, quickly went on the offensive, both traditionally and, notably, online. This included constant social media monitoring.

"The social network has become the feedback loop — you can ignore it or leverage it," says Gustafson. "It's significant that companies are using the same types of tools we would consider counterterrorism technologies to manage their brands. Seeing that was an 'aha' moment for me: If you care about your brand, product or anything you produce, you need to tune in to what people are saying online. You need to monitor the new data channels and learn how to respond."

Much as those working at the frontier of space exploration have generated new products and developments, those working on the web's frontier have created a wealth of innovation and opportunity, too.

"The social media and search companies have done the real pioneering of the technologies we are now seeing commercialized because they were ingesting a lot of data and needed to put it into context," says Gustafson. "They needed a better method than a supercomputer — they needed massive scale and massive storage. Yahoo and Google really are the pioneers of a lot of the under-the-cover tools many commercial organizations are now monetizing on."

When Google launched, few would have predicted the medical community would be able to use its search capabilities to more accurately predict the outbreak of disease. Yet in 2008, Google Flu Trends was able to detect regional outbreaks of the flu in the U.S. almost two weeks faster than public health agencies. Today, people use Google a billion times each day to search for information.

A picture's worth a billion petabytes or more …

With billions of searches and petabytes of data circling the globe, being able to visualize data has become increasingly important, and therefore so have the tools and techniques that enable increasingly sophisticated and analytical visualizations. The LEF report looks at emerging ways to visually represent information using a "multitude of dimensions across time and space," calling it a new form of 21st century digital cubism.

"At DuPont, an increasing number of our employees are mining data and using visualization tools to drive more utility from available information," says Daphne Neveras-Lupfer, DuPont director — Data, DIBM, Six Sigma & Human Resources.

In this new era of information visualization, visual analytics use analytical reasoning supported by highly interactive visual interfaces. Solving the big problems, however, like unlocking the mystery of the human genome, or making truly great discoveries, such as super-fast, super-safe space travel, will take multiple disciplines, people and data. In turn, the potential gains to be realized in solving the big problems are noted by some organizations.

Besides federal initiatives, such as the U.S. National Visualization and Analytics Center, which provides strategic leadership and coordination for visual analytics tools, companies like IBM and GE are sponsoring sites to encourage outside participation in visualization experiments and contests. Visualization conferences, like IEEE's VisWeek, are challenging participants and members to solve large-scale data analysis and visualization problems.

Recent hardware and software advances are also driving augmented realities, such as the ability to view on a smartphone a specific neighborhood, with home prices overlaying it, and immersive realities used to train for industrial mining and race-car driving.

"Visualizing today's data is less about viewing flat passive displays and more about participating with and in the data," says the LEF report. "Data is being shared across boundaries and organizations, be it Toyota (quality and product data), CERN (physics data) or NOAA (ocean data), so that more people can explore and understand the data. Organizations need to prepare for the data-centered economy." ■

[1] CSC's Leading Edge Forum (LEF) provides clients with access to a knowledge base and global network of innovative thought leaders. LEF members work to spot key emerging business and technology trends, and identify specific practices for exploiting those trends for business advantage.

[2] www.csc.com/hpc

JENNY MANGELSDORF is a writer for CSC's corporate office.