Big Data Advances that Make a Difference

Big data is big business for government. It's clear in the funding provided for big data-related projects, and it's clear in the response from agencies at all levels of government. In 2012, the Obama Administration announced the Big Data Research Development Initiative, providing more than $200 million to launch new big data-related projects across federal government. And state and local governments are just as keen on harnessing the power of big data to boost security, prevent fraud, enhance service delivery and improve emergency response.

Agencies have been doing what they can to gain control of and manage the variety, volume and velocity of structured and unstructured data. That's increasingly important, as the pace of big data continues unabated. Deltek estimates that the volume of data stored by federal agencies alone will increase from 1.6 to 2.6 petabytes within two years.

There are advances in the field that can help manage these growing stores of structured and unstructured data, including:

Better analytics: This is an area of rapid and progressive growth. Advanced capabilities now allow for in-memory analysis, complex event processing and visualization along with real-time stream processing and stream analysis capabilities.

Better search: Moving beyond simple word-based search, many big data tools are incorporating semantic search, which examines the intent of the question and presents answers that more closely matches what the searchers seek.

Advancements in NoSQL: NoSQL is an important technology in making big data come to life. This horizontally scalable database technology is very flexible: It can ingest and combine a vast variety of data sources to produce more value. Advanced NoSQL products now offer semantic search, support for a variety of file systems and the ability to work directly with Hadoop. Like Hadoop, NoSQL is a distributed computing model that uses commodity hardware.

Advancements in Hadoop: Hadoop has matured from a batch-processing model to a much more versatile tool appropriate for real-time big data applications. The most maturation has occurred in HDFS (Hadoop Distributed File System), which manages data storage across nodes, and MapReduce, a distributed processing paradigm.

SQL VS. NoSQL: YOU BE THE JUDGE

For decades, relational, or SQL-based databases, have been the database schema of choice to store and manage data. But in the era of big data, the traditional database model often can't keep up. It has difficulty dealing with the multitude of unstructured data types, as well as the massive amounts of data that must be stored, managed and manipulated.

To address these issues, many vendors have built next-generation databases with NoSQL ("Not only SQL"), a database structure optimized for scalability, retrieval and large amounts of data that combines traditional SQL with innovative ways of querying and access. With NoSQL, it's easier to manage projects with large, varied data stores, such as medical health records processing, geospatial mapping and fraud detection. Check out this chart for some differences between the two.

	SQL	NoSQL
Types of databases	One (SQL)	Many, including key value stores, document databases, wide column stores and graph databases
Strong support for unstructured documents	No	Yes
Ease of integrating information from other systems	Difficult, and sometimes expensive	Simple
Scalability	Vertically scalable. Requires additional resources to truly scale	Horizontally scalable. Automatically spreads data onto multiple servers; servers can be added or removed from the data layer without application downtime
Schema	Fixed structure and data types make it difficult to add information about new data items	Data can be inserted in a NoSQL database without first defining a database schema. The format of the data being inserted can be changed at any time, without application disruption.
Hadoop integrated	No	Yes

Big Data: The Geospatial Opportunity

If there is one area of data growth that can rival social media, it's geospatial data. In the past decade, technical advances in the areas of global navigation satellite systems, satellite remote sensing, sensor networks and radar and LiDAR, combined with the exponential growth in location-aware mobile devices, have created the ultimate big data problem—and opportunity. By harnessing all of this valuable geospatial data and combining it with content such as video, imagery and text, agencies have the data they need to see the bigger picture. That information, in turn, can be very useful in determining where to allocate resources, create new models for delivering services, and developing public policy.

According to the United Nations Global Geospatial Information Management (UN-GGIM), there are 2.5 quintillion bytes of data generated every day, and a significant portion of that data is location-aware. That is both the challenge and the opportunity. The opportunity is significant. For example:

- **Crisis management:** With access to information combined from location-aware devices including roadside sensors and security cameras, along with social media feeds, first responders can respond more quickly and efficiently to natural disasters and wounded citizens.
- **Natural resource monitoring and management:** Location-aware data can help agencies better track and predict where oil will go following a spill, model earthquake recurrence rates, detect probable locations of mercury to assess mercury risk, and understand and manage the effects of global change, to name a few.
- **Optimize traffic routes and building locations:** With access to automotive telematics, mobile phone-based location services, smart traffic cameras and other devices, cities and states can optimize traffic routes and improve urban planning.

The challenge lies in harnessing these rich sources of geographic data in systems

that can handle the processing of large amounts of unstructured location data. Traditional databases and algorithms aren't powerful or fast enough to process that amount or variety of data. Newer processing engines, based on NoSQL, Hadoop and other big data-friendly technologies, allow for fast filtering, processing, visualization and analysis of data.

FINDING MEANING IN WORDS LEFT UNSAID

If you haven't heard much about semantic technology, it's only a matter of time. Semantic technology — which takes text analytics to a new level by extracting meaning from both structured and unstructured data — can make a real difference in interpreting data and increasing the accuracy of projections and trends, especially when it comes to big data. Here's what you need to know.

WHAT IS SEMANTIC TECHNOLOGY? Language is ambiguous — one word or phrase can have multiple meanings. In the past, that has made accurate analysis of text difficult. Semantic technology extracts the meaning from data using machine learning, natural language processing, text mining and advanced statistics. The technology also can tag, categorize and classify data in relationship to each other and other sources.

WHY IS IT IMPORTANT IN THE CONTEXT OF BIG DATA? Big data is all about the volume, velocity, variety and variability of data. That means massive amounts of unstructured data — data from video, social media, voice and images, among others. Semantic technology greatly improves the processing and understanding of rhetorical references, comparisons, slang and acronyms, as well as distinguishing whether an expression is intentional or emotional. This is especially helpful with social media and voice data.

WHY DOES MY ORGANIZATION NEED SEMANTIC TECHNOLOGY? Semantic technology supports better decision-making. Here are some examples:

- Speeds up reaction times, helpful in critical decision-making areas such as anti-terrorism, intelligence, defense, fraud detection and detection, crisis and emergency management
- Increases accuracy of trends and projections, useful in security, decision support and risk management
- Improves the knowledge base. By linking to a CRM system, semantic technology can help agencies understand what types of questions and answers are most useful to citizens. This, in turn, can be used to improve self-help web sites.

NoSQL expands the art of the possible with agency data

Thinking innovatively: Scenario 1: It's a scene all levels of government can relate to — volumes of data scattered in databases and file systems in different formats, unable to be accessed, searched or analyzed easily. To more quickly respond to property history requests, Fairfax County, Va., developed a repository for land use data residing in a MarkLogic NoSQL database where it can be viewed, correlated and analyzed. Tasks that once took days or weeks now take minutes or even seconds.

Thinking innovatively: Scenario 2: The Defense Department created the Distributed Common Ground System (DCGS) to enable all branches of the DOD, including the military, to share information via the cloud. But with different agencies and military services using different data collection and storage systems based on relational database technology, searching and sharing is complex and laborious. An independent test consisting of 20 million records using MarkLogic Server showed that the same amount of work could be done 59 times faster, with average response time 32 times faster than the previous system.

Thinking innovatively: Scenario 3: The Federal Aviation Administration needed a way to make faster, more accurate decisions during crises. Built on the MarkLogic platform, the FAA's Emergency Operations Network Dashboard combines all data sources including valuable semi- and unstructured data sources such as PDFs, email, social media posts, weather reports and news stories. The system includes powerful search, discovery and collaboration, creating a common operating picture that allows personnel to properly manage emergencies on the ground and in the air.

TAPPING THE UNREALIZED VALUE IN DATABASES

While these agencies might have achieved similar results with existing technology, it would have taken much more time and money, and involved multiple databases. Faced with growing volumes of both traditional structured and newer unstructured data — social media, text, geographic data and much more — along with mandates to increase efficiency, share information and improve service to the citizen, agencies are beginning to face the fact that in many cases, traditional approaches and tools just aren't measuring up.

But it's even more than just becoming more efficient; it's about tapping the unrealized value in government databases by combining structured data in SQL-based databases with new sources of unstructured data that make better use of technologies more suited to handle them, like NoSQL and Hadoop.

"Agencies no longer have to limit their hypotheses to traditional data sources, but can expand their world to look

at other data sources," said Kevin Shelly, Vice President of MarkLogic Public Sector. "By finding patterns that you haven't been able to find before or took too long to find, agencies can take the quick, decisive action they need to take to save lives, find and mitigate risk, improve efficiency, save money and improve service to the citizen."

MarkLogic is a database platform with a difference. As the only Enterprise NoSQL database, it combines billions of data points and documents from many highly disparate data sources into a secure, searchable repository. Not only does this create a single view of data, but it provides rich search, query and real-time alerting capabilities. And it's lightning-fast: it can search and analyze petabyte scale data with sub-second response times.

It's also extremely cost-effective. With systems built on MarkLogic, agencies can use less expensive, distributed commodity hardware. That cost savings, combined with information ingestion that is 120 times faster than relational databases, improves performance on unstructured information by as much as 100 times (versus relational databases). It also reduces the amount of software code by 70 percent, decreases development time for new software applications by 50 percent, and reduces overall development costs by 90 percent.

By combining out-of-the-box thinking with next-generation technology, agencies can achieve and surpass their goals and know that they are prepared for whatever the mission requires. •

MarkLogic®

For more information please visit
www.MarkLogic.com