Q&A

BILL WONG | Embedded/Systems/Software Editor
bill.wong@penton.com

# LSI's Kimberly Leyenaar Discusses Big Data

"Big Data" takes lots of storage, and LSI is at the forefront in providing platforms that can deliver the massive amounts of storage that these applications require. Kimberly Leyenaar, principal Big Data engineer and solution technologist for LSI's Accelerated Storage Division, focuses on discovering innovative ways to solve the challenges surrounding Big Data applications and architectures.

**ED: What are the challenges in dealing with Big Data?**

**LEYENAAR:** Most organizations have heard of Big Data, and certainly data mining and business intelligence are no strangers to most businesses. However, the challenges are great. When IT managers begin implementing Big Data solutions, the first question is always, "What data do I keep?" While certainly not a barrier to entry, this has been cited as the number one challenge.

The cost of storing data has gone down significantly, leading many businesses to the "save everything" conclusion for fear that latent information not previously known could be mined from these sources ex post facto. It's well accepted that "more data beats better algorithms." However, this also brings other challenges such as how to handle transfer of data for eventual hardware end-of-life cycles as well as potential legal issues that come with storing large amounts of (potentially private) information. Another challenge is finding the right talent to help implement, manage, and apply the proper analytics in order to ultimately extract the business value within the data sources.

Enter the newest job title, the data scientist, a professional with a special mix of business acumen, statistical and mathematical expertise, software engineering, high-performance computing, patterns, and a little bit of magical pixie dust. Today, dozens of colleges and universities offer data scientist degrees, but the supply is far below the demand as the skill sets needed to effectively implement and fully exploit the value of a Big Data project cannot be found in the general IT population. Because Big Data ecosystems are fairly premature, we are still discovering challenges every day.

**ED: What application tools are being used in this arena?**

**LEYENAAR:** While Hadoop is the primary software, the landscape is fertile and I counted over 200 different software applications that, in some way, help IT professionals and data scientists in their quest to derive the value from the Big Data infrastructures. We can divide the application tools into three distinct groups: applications that help build/analyze data queries, applications that help manage server clusters, and finally applications that help manage data sources, structures, and flows.

Hadoop is written in Java, while many business intelligence personnel are accustomed to SQL-based data interaction, and there exists an abundant amount of excellent software meant to alleviate the burden of learning new languages such as Hive, Pig, and Ruby, as well as tools meant to integrate workflows (Oozie) or apply well known query algorithms (Mahout).

Because Hadoop can interface with diverse data sources, there are numerous tools meant to extract data and/or provide connectors to and from other resources (e.g., databases, online data stores) such as Sqoop, as well as software that allows easy data ingest to Hadoop data nodes such as Flume. There is a popular trend to apply organization on top of Hadoop data stores such as noSQL, key/value, or columnar or modular structured data layouts using tools like Cassandra or Hbase.

Finally, because clusters require continuous care and feeding, software such as Zookeeper, Ambari, and Zettaset helps monitor cluster health and maintenance. With so many software choices, tool integration can become daunting. Thus, we will see more Hadoop distributions integrate their own complete toolsets, removing this burden from the end users. We saw this same phenomenon with Linux over the past 20 years, and I anticipate we will continue to see this trend as Hadoop matures.

**ED: What types of underlying hardware and software does LSI offer**

**designers that support these applications?**

**LEYENAAR:** LSI is committed to helping the community solve the problems associated with Big Data. The ability to run a query against a whole dataset and get results quickly is transformative, and it changes the way we all think about data. By improving a Hadoop cluster's efficiency, businesses can run more queries, more batch processes, and that means unlocking the value of their data so they can start innovating their marketing and sales strategies and processes in ways they never knew sooner.

Hadoop clusters are a major investment, and smart IT managers are looking for ways to significantly improve the TCO (true cost to own) and ROI (return on investment) while maintaining cluster availability. The LSI Nytro WarpDrive product line provides the lowest-latency, highest-performing managed PCI Express flash, with capacities up to 3.2 Tbytes, enabling the data to be even closer to the processing and the business closer to their answers. The LSI Nytro MegaRAID provides the best solution with enterprise RAID (redundant array of independent disks) capabilities to add up to 128 devices and integrated on-board flash-based cache with capacities up to 800 Gbytes *(see the figure)*.

Nytro MegaRAID sets the standard for OEM hardened storage, and with eight lanes of 6-Gbyte/s SAS (serial-attached SCSI), it can push up to 4200 Mbytes/s and well over 400k I/Os per second with enterprise features such as RAID, automatic data repair, and SMART (self-analysis, monitoring, and reporting technology) support. The intelligent, integrated cache automatically detects frequently accessed "hot" data and moves it to the high-performing flash.

In our lab, we observed the map-reduce scratchpad operations being moved to the flash (spillover files, intermediate result files, etc.), while HDFS (Hadoop distributed file system) data is read or written sequentially on LSI's high-performing SAS-based architecture. The result is a perfect marriage of high throughput and fast computations that allows most Hadoop jobs to complete in a third less time.


LSI's Nytro MegaRAID supports up to eight lanes of 6-Gbyte/s SAS and transfers up to 4200 Mbytes/s.

Although Hadoop is designed for inexpensive commodity hardware, Nytro MegaRAID can actually lower the total infrastructure cost while providing highly reliable storage. After all, Big Data is about their data, and since an organization's success will be dictated by its ability to extract value and derive innovation from the data available to it, why would an organization trust poor-quality storage hardware?

**ED: How does the OEM partnership with Intel work, and what areas will you be addressing together?**

**LEYENAAR:** LSI has entered into an expanded OEM relationship with Intel, whereby LSI Nytro MegaRAID technology will be available as part of the Intel RAID product family for Intel server boards and systems. Specifically, Intel will offer Nytro MegaRAID technology within its Intel RAID SSD (solid-state disk) cache controllers RCS25ZB040 and RCS25ZB040LX, which include embedded flash of 256 Gbytes and 1 Tbyte, respectively.

The OEM relationship with Intel will broaden LSI's global presence and reach for Nytro MegaRAID technology through Intel's extensive channel network. LSI will also continue to offer Nytro MegaRAID adapters through the LSI worldwide network of distributors, integrators, and VARs (value-added resellers).

LSI Nytro MegaRAID adapters help enterprise customers cut latency and cost-effectively boost performance for applications such as online transaction processing, Web serving, Hadoop/NoSQL, VDI (virtual desktop infrastructure), and other data-intensive workloads. Benchmark testing using Nytro MegaRAID cards have achieved up to a 33% improvement in the time it takes to complete Hadoop jobs and delivered support for up to twice as many VDI sessions compared to a non-caching storage implementation. Nytro MegaRAID cards also enable HDD (hard-disk drive) array rebuilds to complete up to 10 times faster.

**ED: How will this partnership improve the handling of Big Data, and what can you do together that would not work as well individually?**

**LEYENAAR:** The partnership will benefit both companies as organizations look to accelerate key business applications while preserving existing investments in HDD and DAS (direct attached storage) infrastructure. The OEM relationship will also benefit both companies as more and more organizations adopt Hadoop implementations and flash-enhanced big data analytics. ed

---

**KIMBERLY LEYENAAR** is a principal Big Data engineer and solution technologist for LSI's Accelerated Storage Division. She is an electrical engineering graduate of the University of Central Florida and has been a storage performance engineer and architect for more than 14 years.