SGI DataRaptor with MarkLogic Database

Eat Big Data for Lunch

IDC estimates that in the next five years, the amount of information created and replicated will grow by a factor of nine to more than 1.8 zettabytes (1.8 trillion gigabytes), while according to The Economist, only 5% of information created is structured - exacerbating the problem of how to analyze and derive quality business insights from the remaining 95% that is unstructured or semi-structured data.

Unstructured or ‘Big Data’ is data that comes in different shapes and sizes, is frequently changing, can be very short-lived, generally arrives in huge volumes and originates from a broad variety of sources such as experimentation, web stream, point-of-sale, RFID tags, sensor arrays, etc. Unstructured data is also very difficult to analyze and manage using traditional means such as relational databases. Many market segments can benefit from optimal solutions for managing and analyzing Big Data such as:

- **Biosciences**—for pharmacological trials
- **Federal & defense**—for fraud detection, predictive demographics, signal analysis, trend analysis and security analysis
- **Financial services**—for automated and algorithmic trading, risk analysis and detecting fraud in credit transactions/insurance claims
- **Retail**—for analysis of customer buying behaviors and inventory management
- **Science and research**—for large-scale experiments (e.g. the Large Hadron Collider), continental-scale experiments and environmental monitoring, instruments and sensors (e.g. the Large Synoptic Survey Telescope)
- **Social media**—for click stream analytics, user search patterns and behavioral analysis
- **Telecom**—for customer trend analysis, network usage patterns and fraud detection

Analysis in Real Time

Big Data can bring terrific insights to enterprises and government agencies. Many popular methods for analysis, though, work in a batch mode, whereas often the insights are more valuable when they are revealed in real time. Some of the most critical Big Data applications in operation today use the MarkLogic enterprise NoSQL database and SGI and MarkLogic have created a combined solution to expand the world’s access to Big Data capabilities. The combined solution is specially tuned by SGI and MarkLogic engineers working together and arrives at your site, ready to be plugged in and providing answers right away.

Many Big Data needs can be addressed in a batch mode, but the solution becomes better when the failing manufacturing process is improved as soon as possible, or the fraud is stopped before the loss occurs. Also, using current methods, many organizations need months to develop analysis tools to harness the Big Data in their midst. The following are examples of applications where Big Data analysis in real time has big pay-offs:

- Manufacturers need to optimize process yields based on sensor data.
- The intelligence community needs to coordinate diverse sources of intelligence data to predict events.
- In genomics, sequence data needs to be mapped to reference genomes.
- Media firms want to bring the right content to users at the right time.
- Financial services firms need to conduct market analysis, as well as fraud analysis, risk analysis and compliance processing.
- Retailers need to perform customer buying analysis and inventory optimization.

Traditional RDBMS vs. the MarkLogic Database

While one maintains a traditional RDBMS for processing and analyzing structured data, dealing with the volume, velocity and variety of unstructured and semi-structured information becomes a challenge. The rigidity of the relational model, in which all data is held in tables with a fixed structure of rows and columns, has increasingly been seen as a limitation when handling information that is richer or more varied in structure.

One solution is NoSQL technology comprised of XML databases, for example. An XML database is a data persistence software system—this means that changes to the database result in a new structure, versus replacement of the old structure. For example, the previous version of a document stored in the database is preserved even if it is modified, resulting in at least two versions of the document. XML databases are usually associated with document-oriented databases designed for storing, retrieving, editing and managing documents, semi-structured data and information. This data can then be queried, exported and serialized into the desired format. The MarkLogic database is one such XML database. Unlike many other NoSQL databases, the MarkLogic database is an ACID-compliant, W3C-standardized XML database. It is especially useful for loading, searching, unifying diverse content formats for real time access with high efficiency. The MarkLogic database also combines a database with a search engine and an applications platform in one package, versus other products that include only the database.

Many corporations and institutions have a SQL-based RDBMS in their 'back-office.' And they have access to lots of Big Data through their normal operations: documents, customer click-through data on their website, sensor data, video assets, etc. As they see possible value in utilizing the Big Data in their midst, they are at a crossroads—should they continue to invest in their current structured data approach, or join the "Wild West" of Big Data methods? The proprietary "stack" vendors would like them to use methods such as Hadoop to reduce the data so that they can apply standard ETL routines to make it 'structured.' Then they'll offer analysis tools to look at the data in the structured database.

Conversely using the MarkLogic database, a customer can develop applications that allow them to access the data, in real time and with ACID transactions. ETL processes that convert the data into a structured format are unnecessary. For example, a media company might decide to make its video assets available for sale over the internet. This would represent a new income opportunity for the company. Searching, retrieving and selling media assets is an activity that would require ACID transactions, making it an ideal application for the MarkLogic database.

SGI DataRaptor: SGI and MarkLogic Together

Even though NoSQL technology is now becoming popular for addressing such Big Data and information needs, challenges like sizing, configuring and optimizing the NoSQL databases to meet end-user requirements is difficult. Next, applications need to be written to scale with this content processing framework. The other challenge lies in applying the knowledge of search and analytics in the structured space to the unstructured domain. Lastly, users need to ingest, search and analyze high velocity content in real time. Thus, an integrated solution for speed and scale is required to address the different phases like ingestion, search, retrieval and insight for a seamless information flow across an enterprise.

To address these needs and more, SGI and MarkLogic have partnered to deliver solutions that reduce time to insight and remove the risk of harnessing Big Data. SGI brings its heritage in high-performance computing, producing hardware and software optimized for managing large data ingestion, split-second computation and real-time reaction to a myriad of inputs. The SGI and MarkLogic solution is based upon the Intel® Xeon® E5-2600 processor and includes:

- Hardware and software in a complete, factory-integrated package.
- A choice of either performance-optimized or capacity-optimized solutions.
- Quarter-rack, half-rack, or full-rack options for each solution.
- Support offered by SGI for the complete solution including MarkLogic software, worldwide.

SGI DataRaptor

The core to the SGI DataRaptor is the Rackable™ ISS3112 or ISS3124 storage server with similar capabilities but different disk sub-chassis for different size disks.

- "RP2" two-socket motherboard with the latest Intel® Xeon® E5-2600 microprocessor
- 16 cores of computing power per server
- 384 GB of main memory, up to 1600 MHz in speed per server
- Six PCIe gen 3 x 8 slots, full-height, or two PCIe gen 3 x 8 slots, full-height and two PCIe gen3 x 16 slots, double-width
- Up to twelve 3.5 inch disks or twenty-four 2.5 inch disks, SAS, SATA or SSD

Database Nodes (with high performance drives): SGI ISS3124-RP2 servers with Intel® Xeon® Processor E5 Family

Database Nodes (High Capacity drives): SGI ISS3112-RP2 servers with Intel® Xeon® Processor E5 Family

SGI DataRaptor Speeds and Feeds

Full Rack Features	Performance Configuration	Capacity Configuration
Uncompressed Disk Bandwidth	Up to 47 GB/s	Up to 26 GB/s
Uncompressed Flash Data Bandwidth	Up to 20 GB/s	Up to 10 GB/s
Database Disk IOPS	Up to 144,000 8K reads	Up to 32,500 8K reads
Database Flash IOPS	Up to 2,100,000 8K reads	Up to 1,500,000 8K reads
Formatted Disk Capacity (after RAID 6)	300 TB	504 TB
Uncompressed Usable Capacity	~135 TB	~ 224 TB
Rack	42U/standard rack (24 in. width)	42U/standard rack (24 in. width)
Nodes	21	21
CPU	42 Intel E5	42 Intel E5
Cores	336	336
Memory	2688 GB	2688 GB
Networking	2x 10 GigE Data, 1x GigE for Management	2x 10 GigE Data, 1x GigE for Management

Summary

SGI DataRaptor with MarkLogic provides:

- All you need for Big Data built into a single system, factory-integrated and with plug-and-play capability that allows you to be up and analyzing Big Data in hours, not days or weeks.
- Rock solid Big Data capability, containing the MarkLogic database.
- Real ROI for your business or enterprise, with rapid prototyping to develop applications quickly so that you can deploy new applications in record time, monetize your Big Data and crowd source your decisions quickly.
- Big Data. Ready. Set. Go. Starting at 100s of TBs, and expanding quickly, at ease and under load, the SGI DataRaptor offers one price and one support line to call.

To find out more, please visit www.sgi.com/products/bigdata

Global Sales and Support: sgi.com/global