SOS SPOTLIGHT ON GOVERNMENT SOLUTIONS

High-Performance Cluster Computing

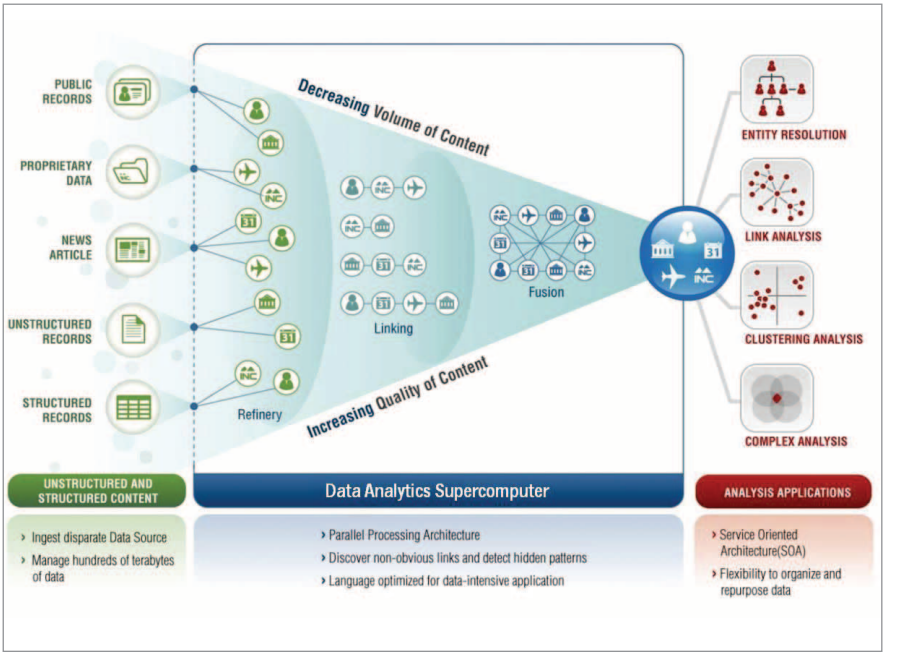
Making Sense of Data

The LexisNexis® Data Analytics Supercomputer (DAS) Delivers Results

The rapid growth of the Internet has led to vast amounts of information available online. In addition, business and government organizations routinely create large amounts of both structured and unstructured information which needs to be processed, analyzed, and linked. An IDC white paper estimates the amount of information currently stored in digital form totals 281 exabytes, and the overall compound growth rate of that information is 57%, with information in organizations growing at even a faster rate. Federal agencies that deal with massive volumes of data are often faced with the dilemma of not only looking for the right data, but also linking disparate information together to produce actionable intelligence.

LexisNexis, an industry leader in data content, data aggregation, and information services, faced this information explosion challenge early on. The company independently developed a massively parallel processing (MPP) solution for data-intensive computing called the Data Analytics Supercomputer (DAS). For more than a decade, LexisNexis has been using the DAS internally to power its multi-billion dollar LexisNexis Accurint® public records business. LexisNexis Accurint is the fusion of over 24,000 structured and unstructured data sources that are cleansed, linked, and disambiguated on the DAS.

Shortly after the 9/11 attacks, various government agencies approached LexisNexis to learn how they could successfully integrate and analyze many disparate data sources, and the result was the creation of LexisNexis Special Services Inc. (LNSSI). Created in 2004, LNSSI has since been offering the DAS and LexisNexis data directly to its government customers for deployment behind their own firewalls as a High Performance Computing Cluster (HPCC) appliance for fusion of their internal and external data in order to solve their most complex and challenging data management problems. The DAS offers the scale, power, and flexibility to take on just about any large-data problem and turn it into actionable intelligence.



Managing Petabytes of Data

The DAS is a platform designed to refine, link, and fuse large amounts of data from disparate sources for complex analysis and queries. Leveraging the speed from its parallel processing architecture, the DAS extracts and normalizes key elements of any data set, clusters all known information around an entity (linking), and discovers the relationships and patterns between these entities (fusion). DAS can manage hundreds of terabytes and even petabytes of content coming from a wide variety of sources. The massively parallel supercomputer leverages service-oriented architecture (SOA) interfaces to integrate with existing networks, and can work with both structured data and unstructured data – such as email messages and network log files – to unearth non-obvious relationships between various entities and uncover hidden patterns.

Architecturally, the DAS is a HPCC based on commodity server hardware, which can be scaled up to thousands of processors to handle any amount of

High-Performance Cluster Computing

Drill Down: Cyber Security

Security professionals often need to establish usage patterns of IT systems so that anomalies to these patterns that could indicate suspicious behavior are discovered. In order to do so, sources such as network flow data, deep-packet inspection data, alerts, and firewall logs need to be considered. But finding patterns and relationships among terabytes worth of data in varying formats from disparate sources is a daunting task.

With the DAS, agencies refine hundreds of terabytes of data, rapidly perform complex queries as well as massive joins and merges, and return actionable intelligence – all at a performance level that is one to two orders of magnitude faster than the competition. A LexisNexis customer reported that one complex query that takes an hour to run on its legacy system took just seconds on the DAS. These capabilities enable agencies to analyze larger time spans of network traffic for suspicious behavior and obtain a comprehensive view of network activity. As a result, agencies can identify the "low and slow" threats presented by the most sophisticated attackers that could only be found by noticing patterns in many months worth of data - much longer than most other cyber security products can manage.

data, and runs on the Linux operating system. System software and middleware components were developed and layered on to provide the execution environment, distributed file system, and SOA interfaces required to support data-intensive computing in an enterprise environment. Additionally, to abstract the inherent complexities of coding against a large massively parallel processing system, LexisNexis created Enterprise Control Language (ECL). ECL is a revolutionary high-level language for parallel data processing that allows the programmer to focus on the data and desired results, rather than on the multitude of sequencing, messaging, and management tasks associated with running a massively parallel processing engine. The power, flexibility, advanced capabilities, and ease of use of the ECL programming language are the primary distinguishing factors between the LexisNexis HPCC and other data-intensive computing solutions.

The Advantages of Enterprise Control Language (ECL)

ECL is a declarative, non-procedural programming language and is optimized for large-scale data management and query processing, automatically managing workload distribution and all other management tasks associated with spreading the processing power across the multiple servers (nodes) of the platform. ECL allows data relationships to be defined by the programmer and achieves flexibility by assigning reusable attributes to data, as opposed to putting information into distinct tables as is required in relational database models. ECL was designed from the ground up by LexisNexis specifically to make analyzing large data sets easier, more accurate, and far more efficient in terms of development resources. ECL includes a mature and robust library of out-of-the-box functions such as joins, merges, and sorts that are easy to execute across massive data sets and performs those functions with greater speed and efficiency than other data manipulation languages.

"We own a lot of expensive technology…but the DAS distinguishes itself for its ability to do the serious 'heavy-lifting' that can't be effectively handled on competing parallel RDBMS technologies, such as solving extremely complex, data-intensive problems that require the correlation, linking, and fusion of dirty data," said one LexisNexis government customer.

LexisNexis has noticed that MapReduce users appreciate the DAS for its flexibility and performance, but most importantly for the data-flow programming model of ECL which allows complex algorithms to be wholly expressed. This data-flow programming approach allows for the processing to be expressed in terms of data flows and transformations, abstracts the underlying complexities that exist in conventional supercomputing and distributed computing systems, and results in the ability for many independent tasks to access data in parallel. This approach achieves what one of the preeminent national labs described as "data parallelism – where tasks are triggered by the availability of data."

SOS SPOTLIGHT ON GOVERNMENT SOLUTIONS

High-Performance Cluster Computing

Case Study: Sandia National Laboratories

Finding Relationships Among Unstructured Data

Turning data into actionable intelligence is instrumental to fulfilling Sandia National Laboratories' mission. Based in Albuquerque, N.M., the lab is owned by the U.S. Department of Energy's National Nuclear Security Administration and tasked with developing science-based technologies to support national security.

In order to perform research and development in its four key areas of concentration – nuclear weapons; energy, resources, and nonproliferation; defense systems and assessments, and homeland security and defense - Sandia must deal with the exponential growth of data sets and next-generation informatics applications. While mining vast amounts of data is essential to Sandia's enterprise, it also makes finding relationships among pieces of information and drawing conclusions particularly challenging.

The LexisNexis Data Analytics Supercomputer (DAS) has been proven to help solve large, complex data challenges such as national security issues. Sandia acquired the high-performance computing cluster to prove it could be integrated with the lab's system of large-scale scientific computing systems.

"Traditional supercomputing technology allows us to run complex physics applications and visualize detailed simulations," said Dr. Richard Murphy, a member of the technical staff at Sandia. "However, these systems are not ideal for the informatics challenge of sorting through petabytes of data to find correlations and generate hypotheses. Our tests show that the DAS is a strong platform for helping us address these challenges."

The DAS looks for specific patterns and non-obvious relationships in massive volumes of data. When used in concert with traditional supercomputers, scientists at Sandia will be able to better identify possible outcomes from their simulations. By accurately managing the massive data sets generated in these operations, DAS enables traditional systems to more quickly extract relevant data to process scientific calculations in their core memory at high speeds. Sandia is also working with LNSSI to assess the applicability of the DAS for the development of the next generation of high performance systems.

Before bringing the DAS in house, Sandia ran the platform through a battery of performance tests against other large-data analysis systems; in several of these tests, the DAS performed ten times faster than the next best system. Through this process the DAS proved what it was designed to do – process, analyze, and find links and associations in high volumes of complex data significantly faster and more accurately than competing technology systems.

The Dirty Data Problem

Garbage in, garbage out. That describes most data-crunching systems that require 'clean' data to produce accurate results. Thanks to ECL and the DAS' algorithms, the system can cleanse the data as it takes input from thousands of different sources. DAS also uses clustering algorithms so it can understand that "J. Smith" and "John Smith" and "JSmith" (or "Muhammad", "Mohammad", "Muhammed") are the same person by considering other attributes linking those pieces of data to create a comprehensive set of attributes for each person, establishing links between each person based on common attributes, and analyzing the information accordingly.

SOS SPOTLIGHT ON GOVERNMENT SOLUTIONS

High-Performance Cluster Computing

DAS Deployments

The LexisNexis Data Analytics Supercomputer is both flexible and powerful enough to be deployed in a wide variety of data-intense scenarios. Here are just a few examples:

- **Data Fusion for Intelligence Missions** – The DAS excels at finding relationships among various massive data sets where the data in each set is sparse and the linkages may not be readily apparent. By utilizing the power of the DAS, sparse data are grouped together into common data clusters and then linked together to disambiguate entities and find non-obvious relationships;
- **Cyber Security:** The DAS sorts through months of network logs to quickly identify patterns and suspicious behavior while allowing security experts to drill down on the data;
- **Medical Informatics:** The DAS uses algorithms to improve communication, understanding, and management of medical information;
- **Maritime Domain Awareness:** The DAS analyzes data from key entities such as ships, containers, cargo, and crew to identify patterns and threats;
- **International Data Enrichment:** The DAS helps identify and locate persons of interest, using foreign datasets and linking information about entities within them together.

Quick Facts

- Benchmark tests show the LexisNexis Data Analytics Supercomputer (DAS) sorted 1 terabyte of data in one quarter of the time and 10% of the code vs. Hadoop:

2009 Test on Equal Hardware	Hadoop	DAS HPC
Create Data	6 min 45 sec	2 min 35 sec
Sort	25 min 28 sec	6 min 27 sec
Lines of Code	562	10

In late 2009, LexisNexis reran the 1TB sort test on its latest generation of hardware in 1 minute 42 seconds, as compared to 6 minutes 27 seconds in the test above.

- An example of the complex type of query the DAS can answer: Find all male persons aged 35 to 45 who lived in an apartment in Seattle from 2005 to 2007 and owned a car with the letters "O" and "H" in any position the license plate number and may go by John or Jim or Joe or Jose Smith.
- A national credit bureau replaced its mainframe system with DAS to run a complex algorithm that predicted the likelihood of loan defaults, and cut processing time from 26 days to 8 minutes.
- DAS is pre-tested and configured to customer specifications prior to delivery, making its implementation fast and seamless-customers can be up and running in thirty to sixty days.

About LNSSI

As a trusted leader in enabling government agencies to transform data into mission critical decisions, LexisNexis® Special Services Inc. leverages over 30 years of information solutions expertise to provide customers with global sources of data, data fusion technology, and advanced analytics that address their most challenging analytical and decisioning needs. LNSSI is committed to serving the U.S. government by providing advanced information to the U.S. intelligence, homeland security, federal law enforcement and defense communities.

LexisNexis® is a leading global provider of business information solutions to a wide range of professionals in the legal, risk management, corporate, government, law enforcement, accounting and academic markets. LexisNexis® originally pioneered online information with its Lexis® and Nexis® services. A member of Reed Elsevier [NYSE: ENL; NYSE: RUK] LexisNexis® serves customers in more than 100 countries with 13,000 employees worldwide.

LexisNexis®

For more information call 800.291.3670 or visit lexisnexis.com/government