WHITE PAPER

LEXISNEXIS OFFERS ENTERPRISE DATA FUSION TO GOVERNMENT AGENCIES TO MEET NATIONAL SECURITY CHALLENGES:

Data Integration Platform Offers Immediate Solution for Large Scale Disparate Data Challenges

Table of Contents

Introduction

> *"We in the NSA are encountering problems with the flood of information that people (in the outside world) won't see for a generation or two. We've been into the future and we've seen the problems of a 'tidal wave' of data."*
>
> *> Former Head of Research, National Security Agency, 2004*

In a recent report titled, 'Textaware Applications: The Endgame for Unstructured Data Analysis', the 451 Group wrote that, "The future success of companies and organizations will increasingly be based on their ability to unlock hidden intelligence and value from unstructured data, and text in particular." For government professionals working in National Security, this acknowledgement from a commercial technology analyst will simultaneously cause both relief and a concern. National Security Agencies have been aware of the scope of the unstructured data challenge for years. The global data collection and analysis efforts of intelligence agencies span multiple languages, media and communications protocols and clearly invite challenges far beyond those seen by most enterprises. What the commercial sector calls unstructured data, the intelligence community calls 'real world' data. The fact that the commercial sector is beginning to take these real world data challenges seriously means a wave of innovation will follow that could provide relief to National Security Agencies. But the reality is that the commercial sector views unlocking intelligence from real world data as critical to its future success. For national security agencies this is cause for concern because intelligence hidden within real world data is critical to its current success and waiting for the commercial sector to evolve its technology is not an effective option.

LexisNexis, the world's leading information solutions provider, depends upon its ability to harness heterogeneous data sources to serve its customers. Like the national security agencies, LexisNexis is on the frontier of large scale data management and analysis and has made substantial progress in evolving technology to handle the disparate data challenges of the modern enterprise. As traditional enterprise solution providers adopt their technology to manage the challenges of real world data, LexisNexis recognizes that they may not come in time for the most pressing national security needs. As such, LexisNexis is making its core platform technology available as a stand-alone data integration platform to government agencies. Known as the *Data Analytics Supercomputer (DAS)*, this platform represents the state-of-the-art in high processing computer technology for large scale disparate data analysis. The purpose of this white paper is to articulate the changing nature of enterprise data integration and analysis to help government decision-makers determine if they can wait for the evolution of traditional commercial enterprise infrastructure or adopt the LexisNexis *Data Analytics Supercomputer*.

Unstructured Data Enters the Mainstream

Three Trends Trigger Enterprise Adoption

Large enterprises – both government and commercial – are realizing that their success will increasingly depend upon their ability to discover hidden intelligence and value from a wide range of data sources. Within the information technology community, there is a general understanding that substantial quantities of unstructured data within the enterprise present both a strategic opportunity and an infrastructural challenge. While an acknowledgement of the problem represents progress, few enterprises have implemented broad data exploitation and management strategies that harness the breadth of the relevant data available. But that is beginning to change because of three major trends in data management:

- **Dramatic growth of data volume:** The unstructured data available to enterprises is varied and rich and generated from both internal and external sources. The continuum from structured to unstructured data types is broad and the growth along that continuum is substantial. This growth can be defined in terms of volume; according to one study the amount of new information created in a single year in the world today would fill 37,000 new Libraries of Congress. It can be defined in terms of variety as the type of data available grows with the evolution of new communications media. This growth subsequently leads to a growth in the need for data velocity or the speed at which new information is consumed, updated and processed. And finally this growth can be defined in terms of data veracity, or the increasing need to distinguish between accurate and inaccurate data that exists because of the growth in volume.

- **Increased value of new data:** Inside this data growth there is value for enterprises. Leading firms and agencies are finding innovative ways to improve performance by providing their professionals with more sound intelligence through new data sources. For example, online product reviews, call-center notes, and other web content are mined to pinpoint customer sentiment and improve customer services or publicly-available government filings are used to detect fraudulent activity or investigate crimes.

- **Improved underlying technologies:** The technology for converting raw data into information for analysis has simultaneously improved and made new data sources available in data warehouses for storing data, customer relationship management (CRM) or case management systems for recording interactions and natural-language processing (NLP) for helping computers understand written human languages. The advance of such technologies has made the impossible seem possible - unstructured data can now be as useful as structured data in the modern day enterprise and enterprise managers are eager to exploit the new possibilities for data management and analysis.

Adopting Enterprise Infrastructures is Evolutionary

Technology standards struggle to keep pace with data growth

While enterprise managers are hopeful of the possibilities for unstructured data for good reason; this large scale disparate data problem is somewhat ambiguous and can mean so many different things to different people. Modern technology lore maintains that 80% of all data available for use in an enterprise is unstructured, with the remaining 20% as the structured data that has traditionally been managed by enterprise infrastructure in database and enterprise applications. Structured data is the staple of enterprise technology deployments, represented in rows and columns, persisting in a database management system, defined by syntax (i.e. file format, location), and enforced by a known schema. Customer information, accounting records and employee data are primary examples of structured data, and their ongoing management has fostered a thriving industry for structured data management, storage and analysis tools. Unstructured data, by comparison, is less-well defined, usually unmanaged or relegated to a content repository, requires semantic interpretation to understand its meaning and is usually generated with little or no schematic rigor to aid in its manipulation. Word processing documents, blogs and news articles are examples of unstructured data, while spreadsheets and email messages offer some structure, but not much.

The dynamics of this growing tidal wave of data are multifarious and can manifest themselves to enterprise decision-makers either as lost opportunity or new costs for system integration, increased computer processing power, new data storage and data analytic tools, or new personnel. Often, the significance of data growth is defined in terms of the structured data applications that run the operations of most enterprises. For enterprises with growing amounts of data that can be generated, stored and delivered

according to the same data model, this is the appropriate paradigm. But for national security agencies with growing amounts of data that is generated, at best, according to different data models or, more likely, according to no data model at all, the dynamics of the data growth problem offers a particularly daunting challenge: How to collect, identify and link relevant data attributes across a wide set of disparate data sources so they can be leveraged in a business application and to do so without crippling the performance of the underlying system.

Three Approaches Have Emerged

None have evolved to fully handle the big data problems of national security

To exploit new and disparate data sources for specific applications requires the basic physical requirement of moving data from one place to another so that it can be made useful to end users. In the structured data world, there are well-defined options for extracting data from its source of origin, transforming it for effective use, loading it into a system so it can be accessed, and managing that data over time to reflect changes and new business requirements. In the unstructured data world there are many niche applications for searching sets of unstructured information that have been indexed for specific needs and extracting relevant information from the text of documents. Such solutions can be useful for finding a piece of information, but lack the ability that is taken for granted in a structured database to comprehensively link that information to similar data to support a business application. The challenge increases when the application requires both structured and unstructured data. Organizations that don't simply ignore unstructured data, approach the challenge by using some combination of the following three approaches:

- **Relational Databases:** the core infrastructure of any enterprise is the standard for structured data management and has some capabilities to handle unstructured text.
- **Keyword Search:** the popular Internet tool works well for retrieving documents but struggles with analysis.
- **Point Analytic Solutions:** The next evolution of search is point analytics applications that perform tagging and fact extraction for very specific needs.

The Relational Data Base Approach

The process of building a data model for a relational database is relatively straightforward and well understood. But relational databases are built upon the principle that the physical data representation should be entirely separate from the way that the data is viewed by people accessing the database. When unstructured data is required and the content structure varies between each source, it is difficult to fit such data into orderly database tables and create relationships with other data because these new content sources do not natively support the required physical representations. Despite this, the wide availability of skilled database programmers in the marketplace has allowed organizations to architect solutions using RDBMS that ultimately result in substantial operational overhead.

A premise of the RDBMS concept is that the data is generated, stored and delivered according to the same data model. For 'unstructured' data collected from external sources this premise is fundamentally broken. Each data source that is collected will have, at best, been generated according to a different data model. Far more commonly the data has not been collected according to a data model at all. The procedures in place to ensure an RDBMS has integrity simply do not apply to that 80% of unstructured data that is available today. As a result, systems that rely on RDBMS solutions to harness both structured and unstructured data require recoding or extra database tables, extraneous querying, and a multitude of indexes to implement and manage. This adds financial costs and impedes performance. The relational model is effective for server-hosted business applications leveraging structured data stored in an RDBMS. But for assimilating real world data from all directions and in all formats, it is insufficient.

These challenges with the RDBMS model are compounded by the availability of other tools designed to support the acquisition and integration of multiple data sources and facilitate changing data structure as that data is delivered to target relational databases. While ETL tools increasingly include some ability to handle unstructured data, those capabilities remain limited to the connectivity of a few types of unstructured data sources, acceptance of XML structures as an input type; and the parsing of PDF files and Microsoft® Office documents. The aforementioned overhead increases as software tools designed to process structured data are used to perform unnatural acts on unstructured data. For the government decision-maker evaluating the application of a data warehouse

or structured data ETL tool against a large scale disparate data problem, the following criteria should be carefully considered:

- **Data type variability**: Unstructured and semi-structured data will go through an entirely different process than the ETL process of integration, which operates on records of data with keys, fields, and indexes. These things don't exist in unstructured data, which must first be filtered, edited, translated, and linked to other data. For a small set of unstructured data, this is possible. For large scale disparate data sets, this is a daunting challenge.
- **Flexibility to change**: Structured data solutions do not lend themselves to "on the fly" changes that leverage disparate data sets. Once data is stored in a data base, such changes are possible, but if data needs evolve, an RDBMS solution will not have the flexibility to adapt.
- **Analytic freedom**: Business intelligence tools that rely on RDBMS systems restrict users' analysis to the fields and indexes defined in the database. This limits analytic freedom outside the RDBMS paradigm, as it is difficult to create new indexes quickly.

The Keyword Search Solution

Search solutions include software that allows end users to search files in databases, content management systems, email servers and file servers. Google, Yahoo! and MSN have popularized the well known concept of search by searching files on the publicly available Internet. LexisNexis has provided search solutions for decades against specialized data sources such as legal research, public record and news content. The concept of 'enterprise search' has recently entered the market lexicon to connote a specific search solution for discovering information inside an enterprise.

The concept of enterprise search is appealing. End users type in a query and return the information they want. In practice, enterprise search is impractical for harnessing the power of large scale disparate data. In order for search to work, the target data sources must first be analyzed so relevant information can be created and stored in an index database for use in later queries. It is unlikely that a company or agency will be able to index every piece of data within the enterprise to make available for a single search solution. The cost and complexity of the development effort would be prohibitive and the end result would undermine data and interdepartmental security policies. And the

seemingly unstoppable growth in unstructured data will perpetuate those problems over time. Further, the short ambiguous queries that lend themselves to search do not exploit the full power of disparate data discovery required by many users. Regardless, if all of the aforementioned problems could be isolated and resolved, Search still does not provide the solution required by enterprises with large scale, disparate data problems: integrating data from these sources.

Search alone may be useful for retrieving data from specific data sources, but it does not link or integrate results across multiple sources for business application use. Ultimately, national security enterprises require an aggregate view of data, not just a segregated view. For the government decision-maker evaluating the application of keyword search against a large scale, disparate data problem, the following criteria should be carefully considered:

- **Query Complexity:** Search tools were designed for ease of use and generally rely on simple Boolean expressions. Results rely upon combinations of the right keywords which can have different meanings in different contexts, so often return irrelevant responses or omit relevant ones. For agencies with complex analytic needs, search will not be enough.
- **Data Linking**: Keyword search is effective for retrieving individual documents quickly. It will not link relevant meaning within those documents and an end user will still have to read through and extract the meaning. For agencies with substantial amounts of data to be linked, search will be inadequate.
- **Storage of Results**: While keyword search is effective for the instant gratification of results, it is not effective for storing and quantifying results over time. For agencies that track activity over time or the evolution of ideas, search is not an effect solution.

The Point Analytics Solution

There are a number of niche solutions offerings that take raw text, often in different languages, perform linguistic analysis on it of some sort, and transform it into something useful for an application or a user. Such offerings include content classification; categorization and clustering; fact and entity extraction; taxonomy creation and management; and information presentation such as visual analytics or guided navigation. These techniques are highly useful to making unstructured text valuable, but no clear

standard has emerged in this space for enterprises to depend upon. Typically these solutions are offered as pure play packaged middleware products that require enterprise users to build something before there is a direct application, and still that application will lack the benefit of structured data within the enterprise. These tools can provide value by resolving documents to a more granular level. For specific uses, certain unstructured data analysis tools will be very useful, but for managing large scale disparate data, they are limited. For the government decision-maker evaluating the application of point analytic solutions against a large scale, disparate data problem, the following criteria should be carefully considered:

- **Analytic specificity**: the immaturity of this technology group is such that each point solution may be good at one thing – such as name matching – but not good at another thing – such as event extraction. If agencies analytic needs are specific, there is most likely a point data analytic solution that will work well. But if the agencies needs are broad, or even unknown, no one solution will suffice, and integrating multiple applications will be expensive and time consuming.
- **Custom integration**: In order to run analytics against enterprise documents, a point analytic solution will require access to information stored in systems for document management, content management, database management, metadata management, and business intelligence. Such integration can be time consuming, costly and can affect system performance. Agencies with decentralized information should first consider the custom integration required to use point analytic solutions.
- **Integrated scalability**: Pursuant to the custom integration criteria, it is important to consider the impact to scalability of implementing point solutions in the enterprise. Many of these solutions are unable to handle data volumes in the 50-100TB range.
- **Toolkits require time and resources**: Point analytic solutions typically provide customers with a toolkit that requires that business rules are re-written in the platform. This requires training and time and often requires agencies to invest in skill sets and algorithms that do not have broader application to wider data analysis needs across the enterprise.
- **User types**: Many point analytic solutions require end users to have specialized linguistic background or training.

National Security Agencies Cannot Wait

Commercial Technologies Evolve Against a Different Timeline

Business level users quickly grasp the potential of augmenting data management systems with new data sources, but find it difficult to assess a viable strategy for enterprise wide implementation or an obvious return on investment – hence the conclusion of the 451 Group that the future success of organizations depends upon new data sources. The reason for this, as demonstrated above, is that presentable options do not offer an elegant solution, compounding the ambiguity of the underlying data problem. National security agencies with large investments in many of the technologies mentioned are presented with a difficult question: do they wait for commercial enterprise technologies to catch up with their data needs or do they develop their own systems? Clearly pressing national security needs do not lend themselves to waiting for industry to develop new technologies, but custom solutions present their own problems. For agencies with very large disparate data problems, LexisNexis offers another option, and that is the use of its data fusion platform. Like national security agencies, LexisNexis considered these same questions after evaluating the options for integrating disparate data sources at scale and concluded that it needed to develop its own system. That system – the *Data Analytics Supercomputer (DAS)* – has evolved through multiple generations and is now currently in use by several national security agencies. Specifically, this system is useful for agencies with the following needs:

- **Complex data linking**: Agencies that wish to show linkages between known and unknown entities – person, place, thing – that exist within unstructured content.
- **Disparate unstructured & semi-structured data sources**: Agencies that collect data that is generated by different data models or no data models at all.
- **Ongoing analysis dependent on large data quantities**: Agencies that have to maintain data in its native state and perform ongoing analysis.
- **Evolving analysis**: Agencies that may not always know what they are looking for but know they have to look.

Disparate Data Analysis at Scale

Defining a Solution to an Esoteric Problem

Of course, the objective of harnessing large scale disparate data is to better support applications and business processes by providing readily available access to more types of data — structured and unstructured. To accomplish this, mechanisms are needed for the data to be pulled together and integrated in a useful way to support emerging business requirements. Tools for data integration are still limited in their capabilities to deal with both structured and unstructured data, yet for many business applications a combination of the solutions mentioned above will provide an effective solution. Business processes and applications that have to scale the mix and match approach will prove to be inefficient and disappointing. It is important to note that the challenge of scale is not simply defined by volume of data or variety of data type. There are high quality solutions available for processing multiple terabytes of structured data and there are integration platforms that are suitable for integrating the most common forms of data found inside the enterprise.

For LexisNexis, scale is defined by *the computational complexity of the data analysis* required by a given business application. Data analysis, of course, is another ambiguous term that many rightfully claim to perform, although very few can agree on its meaning. For some, data analysis occurs if data is simply presented on screen for an analyst to review. Others would say analysis has occurred if the data is aggregated or summarized. Others may consider data analysis to be data represented in some alternative format. While these may be fully viable forms of data analysis, LexisNexis defines data analysis as *the process of concentrating information from a very large disparate data stream with low information content to a small data stream with high information content.* Clearly this definition introduces its own ambiguity – large, low, small, and high -- but suffice it to say, this is an apt starting point for describing the types of large scale disparate data problems LexisNexis is concerned with solving and also one of the most technologically daunting and opportunity rich data problems facing the modern enterprise. The solution proposed below is not meant to serve as a substitute for the RDBMS, Search, & data analytics solutions already described. Those solutions are the best for resolving the types of problems they were engineered to address. But they were not engineered to transform information from a very large disparate data stream with low information content to a

small data stream with high information content, and to do so without disrupting the integrity of the underlying data.

A Data Centric Solution for Large Scale Disparate Data Analysis

LexisNexis Develops the Data Analytics Supercomputer

In 1999, a team of world leaders in the field of utilizing RDBMS technology conducted an evaluation of existing commercially available technology and concluded that the RDBMS and other available solutions were unsuitable for large scale, disparate data integration. In response to this conclusion, they began with blank sheet of paper and constructed a model to handle huge amounts of real world data. The result was a system that currently processes 40+ billion records and maps each to 200+ million distinct entities for use by over 600,000+ customers of LexisNexis's Accurint product line. This system is also used as a stand alone data analysis solution by several U.S. government agencies.

From its conception, the LexisNexis enterprise solution was engineered to specifically address the large scale disparate data problem. A fundamental underlying premise is that the logical and physical data layers did not need to be separated for programmatic reasons, as doing so would inhibit the required performance levels at scale. A key factor in the integration of disparate datasets is the disparity of the datasets. Attempting to fix those disparities in different layers with different skill sets would be counterproductive to the end goal. For large scale disparate data analysis, skills needed to be focused around the data; not specifically around the processes that are used to manipulate the data. The layers in a data process would represent the degree of data integration, not the particular representation used by the database. The advantages of this approach include:

- **Data analysis over data integration**: resources typically allocated to managing and manipulating data can be allocated to understanding the relevance of data.
- **Speed and breadth of analysis**: with less data manipulation logic, analysis is not constrained by process and able to move quickly through large amounts of content.

Another premise of the enterprise solution was that it would exploit advances in parallel computing. From the outset, this system was designed to process massive volumes of disparate data at levels of performance unattainable on traditional systems. It would be

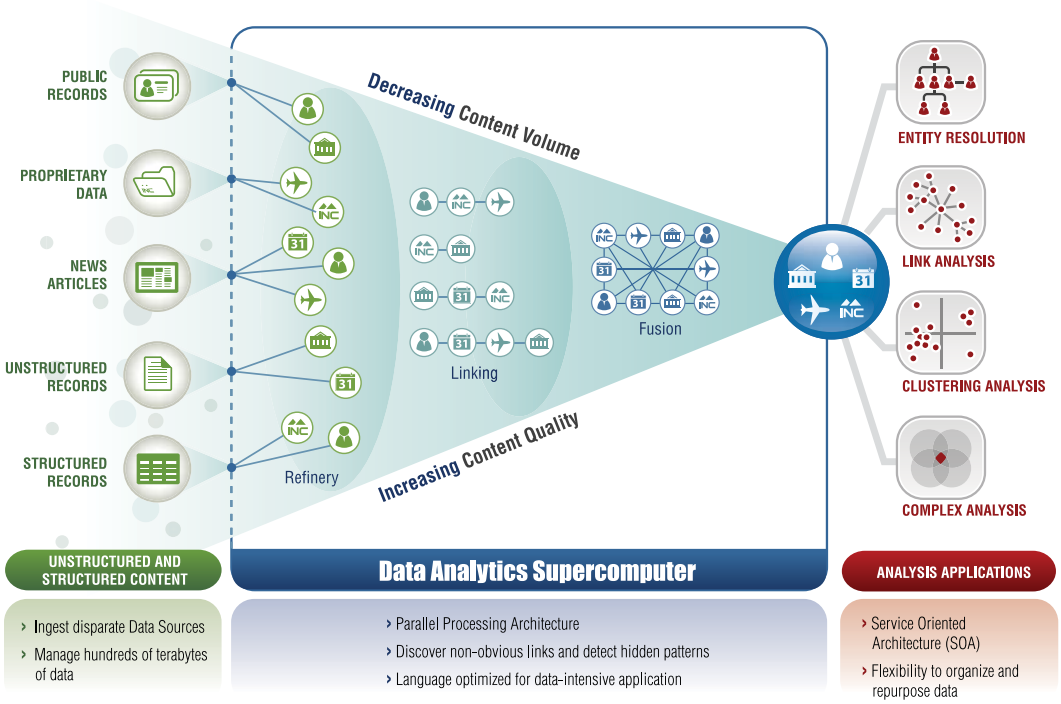
critical to the success of such a system that performance is matched to the analytical demands placed on it; otherwise congestion would lead to dramatic slowdown in performance undermining the benefits of using a high performance computing system. The field of parallel execution and sequencing had progressed to the point where a well specified problem was more likely to be correctly optimized automatically than by hand. Therefore a commitment was made to invest the required resources in a parallel processing architecture to ensure that performance could be tuned by an expert system, leaving the data programmer with the sole responsibility to correctly specify the correct manipulation of the data. The advantages of this approach include:

- **N^2 algorithmic processing**: Linking and clustering algorithms can be run on multiple dimensions and at scales not previously possible.
- **Rapid index building**: New data models can be built and more rapidly populated with a corpus of data. Large amounts of data can be indexed and made available for rapid querying.
- **Enhanced analytic point solution capability**: Specialized extraction and transliteration tools can exploit parallel processing to achieve enterprise scale.

The final premise was the development of a new programming language designed to have all of the data processing capabilities required by the most advanced SQL or ETL systems but also to have the code encapsulation mechanisms demanded by systems programmers. There were very few viable candidates for a language that could easily manipulate data and concurrently produce complex logic or efficient performance. ECL, or Enterprise Control Language, is the query and control language developed to manage all aspects of the data within the enterprise solution: from the creation or ingestion of data; through its refinement; and ultimately to querying, analysis and output. ECL is now a 5th generation declarative language optimized for the manipulation of data. It is modular, allowing ECL programmers to create attributes which can be stored and re-used for multiple purposes. The ability to modularize and save queries alone can permit standardization and produce efficiencies where previously there were none. ECL was designed to make it simple to the new user and ultimately flexible to the advanced user. The advantages of this new language include:

- **Analytic speed**: ECL code is compiled and executed in the fastest way rather than in sequence

- **Data integrity for disparate data formats**: data remains in its original form or on disk; transformations and filters are computed as the data is used at query time. Data sets can be queried or merged without actually modifying the original data.
- **Subfield level control**: ECL enables subfield level queries to be made (i.e. all the names with 'smith' in them that do not start with 'smith')
- **Complex matching capabilities**: ECL enables multi-criteria matching capabilities for the sophisticated merging of disparate data sets. Complex data relationships can be treated as re-usable attributes.



The Benefits of Enterprise Data Fusion

Supercomputing power previously not available to the enterprise

The benefits of such a system may not be immediately obvious, but for large scale disparate data analysis they are significant. Because the data sources are stored unmodified there is never any 'loss' of information or difficulty in re-mapping the incoming files to the target formats. Toward the objective of processing large streams of data with low information content to small streams of data with high information content, it is important to preserve what original information is available. Further, data analyst teams can be segmented by data type rather than language skill, allowing for every file type to be handled by individuals skilled in that field. And for increased processing power, the parallel architecture allows for components to be added to meet rising data integration requirements. LexisNexis has invested several years of development effort and several generations of network topology configuration to come up with its current configuration. All aspects of the network topology have been carefully configured to ensure that there are no bottlenecks as data analysis requirements scale.

The resulting solution allows for huge scale data ingest and integration. It allows for scaling of hardware and personnel and encourages deep data comprehension and code reuse which enhances development productivity and improves the quality of results not previously possible. It is a solution designed to concentrate information from a large disparate data stream with low information content to a small data stream with high information content.

This is a type of data problem that some enterprises are beginning to see, particularly in national security, and that should continue to emerge as the tidal wave of data grows. Consider the following example: Multiple data sets are available for ingestion into the enterprise from many different sources. One data set may contain the name and address data available from an Electronic Directory Assistance (EDA) source. Another data set may contain property listings associated with individuals at various addresses. Other data sets may contain professional licenses, motor vehicle registrations, business affiliations, and watercraft licenses. Each is generated, stored and delivered according to a different data model and in aggregate is a very large data stream with little inherent

information value to an enterprise. Indexing these records to enable a user to find a known query – such as Jon Smith's Driver's License Number -- is clearly possible with a search solution. Integrating these records into a known RDBMS schema to enhance a business process that relies upon specific and known application logic – such as integrating all known addresses into a contact record -- is also possible with an RDBMS solution. But to generate a small data stream with high information content -- such as linking all the entities within those record sets, defining non-obvious relationships and providing a 'fused' view of the entity information -- and doing so while maintaining the integrity of the underlying data to enable rapid updates and new discovery is not within the engineering design of generally available commercial technologies.

When LexisNexis originally attempted to implement a data analysis solution to discover non-obvious relationships such as relatives and associates within such a data set, it encountered the expected problems. To query across the physical, logical and application layers, thousands and sometimes tens of thousands of fetches per query were required. The result, even on high performance hardware, was that many relationship trees would fail the system's 45 second timeout requirement. The query logic was then moved into an ECL process on the enterprise solution that simultaneously computes the associates and relatives for every person in the US and the result was impressive. The process at one point evaluated 750 billion intermediate computations stored in a 16 terabyte data file. The result is presented through web services and delivered up with a sub-second response time. Further, these relationships now exist as enriched data 'facts' which can themselves be utilized in other non-obvious relationship computations. A large data stream with low information content was effectively concentrated into a small data stream with very high information value.

When applied to national security challenges, the *Data Analytics Supercomputer (DAS)* has demonstrated impressive benefits to agency missions. Such benefits include:

- Discovery of non-obvious relationships between entities
- Identification of relationships between structured and unstructured data
- Best in class entity resolution and disambiguation
- Flexibility to organize and repurpose data for new and different problem sets
- Simultaneously allow aggregate and segregate views of data
- The speed to obtain data analysis results when needed

The tidal wave of data is real and growing and will require new approaches to exploiting the inherent data analysis opportunities. LexisNexis has designed a platform architecture that allows enterprise teams to integrate large volumes of data quickly and extremely efficiently for complex analysis. It is a solution designed to solve a very specific, high value and complex type of data integration problem that can significantly benefit national security agencies.