The Business Problem

Organizations typically address the mining of rich unstructured text resources by indexing a set of documents using a variety of techniques and displaying ranked results similar to a search engine like Google.

The Technical Challenge

CFL Discover provides a solution for the handling of text within a resource description framework (RDF). All textual data is stored in an n-triple store, independent of text origin. It is more than just an index of keywords or categories; relevant material is identified by matching the whole complex contexts in which terminology appears. This is not an indexing solution, but rather a descriptive system that can be searched flexibly.

In this dynamic system, the structures and sequences inherent to individual documents are all that is needed to encode them. New material is easily added to existing stores and is immediately available for use by the search queries. Furthermore, searching new documents in their entirety is as simple as searching terms, because each incoming document has a built-in relationship with all other documents that contain the exact word pairings. This potential for gathering subsets of related documents means that incoming material can be scanned against sets known to be of interest, which are useful not only in intelligence applications, but also in such applications as e-discovery where relevance of terminology is a primary issue.

The fact that CFL Software Ltd. stores this rich discovery method in RDF calls for a platform that can handle path processing at scale. CEO David Woolls of CFL Discover comments that “on a standard graph platform, the processing of all the interactions between the individual segments to find relationships rapidly causes the heap size limit to be reached and therefore discovery becomes curtailed.”
This is where “Eureka!” moments occur. Following paths and relationships at scale is achievable with the Cray® Urika-GD™ graph analytics solution. The system’s Threadstorm™ processor and its large shared memory with uniform latency gives CFL Discover far greater capability in mining rich text and uncovering hidden relationships in big data.

A Wiki Research Exercise
To demonstrate the combined power of the Urika-GD appliance and CFL Discover, Wikipedia was preprocessed into a micro-level format as used by the text analytics. Articles were randomly selected and SPARQL queries juxtaposed them against the 1 million extracted using Urika-GD, so as to prove similar representation — all through just the title of the article. The program’s parameters define similarity between segments and how many segments need to be similar to count.

The Cray® Urika-GD™ Platform Solution
Organizations manage large repositories of unstructured text documents, and it is very difficult to know who else, past or present, might have worked on your particular problem. The Urika-GD system’s ability to contain a relevant set of documents can save many hours of wasted effort in re-solving a problem, quite apart from the time it would take to find relevant documents.

In the world of mergers and acquisitions, the same principles apply concerning the need for conducting due diligence on large and complex datasets. Neither party will know exactly what it should be looking for, but indications of related material can give a clear and rapid overview of where to direct attention.

The power to introduce additional data from external sources with ease will add new dimensions to your research. The provenance of data inherent in a graph structure will be beneficial when tracking how decisions and conclusions are made, ensuring a richer foundation of knowledge upon which to base the final outcome.

The Urika-GD appliance owes its performance to its large shared memory, scalable I/O system and purpose-built Threadstorm graph processor. The system’s huge, globally shared memory architecture of up to 512 TB can hold the entire graph of relationships in memory. The scalable I/O subsystem, which can scale up to 350 TB of I/O per hour, enables continuous updates to the graph as new data streams in from on-demand analysis computations.

The Threadstorm’s massively multithreaded architecture (128 independent threads per processor) is specially designed for analyzing graphs and allows threads to continue executing even if some are waiting for data to be returned from memory. This architecture facilitates relationship exploration with performance that can be several orders of magnitude better on graph problems than commodity hardware. Life isn’t structured. Don’t worry if your data isn’t either.

About Urika-GD
The Urika-GD big data appliance for graph analytics helps enterprises gain key insights by discovering relationships in big data. Its highly scalable, real-time graph analytics warehouse supports ad hoc queries, pattern-based searches, inferencing and deduction. The Urika-GD appliance complements an existing data warehouse or Hadoop® cluster by offloading graph workloads and interoperating within the existing analytics workflow.

About Cray
Global supercomputing leader Cray Inc. provides innovative systems and solutions enabling scientists and engineers in industry, academia and government to meet existing and future simulation and analytics challenges. Leveraging more than 40 years of experience in developing and servicing the world’s most advanced supercomputers, Cray offers a comprehensive portfolio of supercomputers and big data storage and analytics solutions delivering unrivaled performance, efficiency and scalability. Go to www.cray.com for more information.

©2014 Cray Inc. All rights reserved. Specifications subject to change without notice. Cray is a registered trademark and Urika-GD is a trademark of Cray Inc. All other trademarks mentioned herein are the properties of their respective owners.

20140915