YarcData
Getting to **Eureka!** faster™

Solution Brief

Don't let your data devolve from unstructured to anarchic

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's. These capabilities allow us to search for information that is already known, but what about research and discovery of the unknown? How do we uncover those hidden patterns and relationships in heterogeneous rich text data that are the fabric of knowledge in our world?

The Technical Challenge

CFL Discover provides a solution for the handling of text within an RDF framework. 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 eDiscovery where relevance of terminology is a primary issue.

The fact that CFL Discover 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 Urika moments occur. Following paths and relationships at scale is achievable with Urika's supercomputing know-how and its ease-of-use paradigm. The Threadstorm™ processor and the large shared memory with uniform latency allows 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 Urika and CFL Discover, Wikipedia was pre-processed 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, 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.

One such search started with "anarchism." The returned list of related material included "libertarianism," "types of socialism," "individualism," "capitalism" and "collectivism." It is probably clear on review that these articles are relevant to the overall concept, although not all would necessarily have occurred to a researcher at the outset.

What is also interesting is that the number of identifiably related articles is generally a very small proportion of the overall number. Most importantly, Urika returns the results of the enormous number of comparisons within seconds. This is possible thanks to the data structure, the queries, the Threadstorm™ multithreaded architecture, and the ability to hold all the data in memory.

The Urika™ 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. Urika'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.

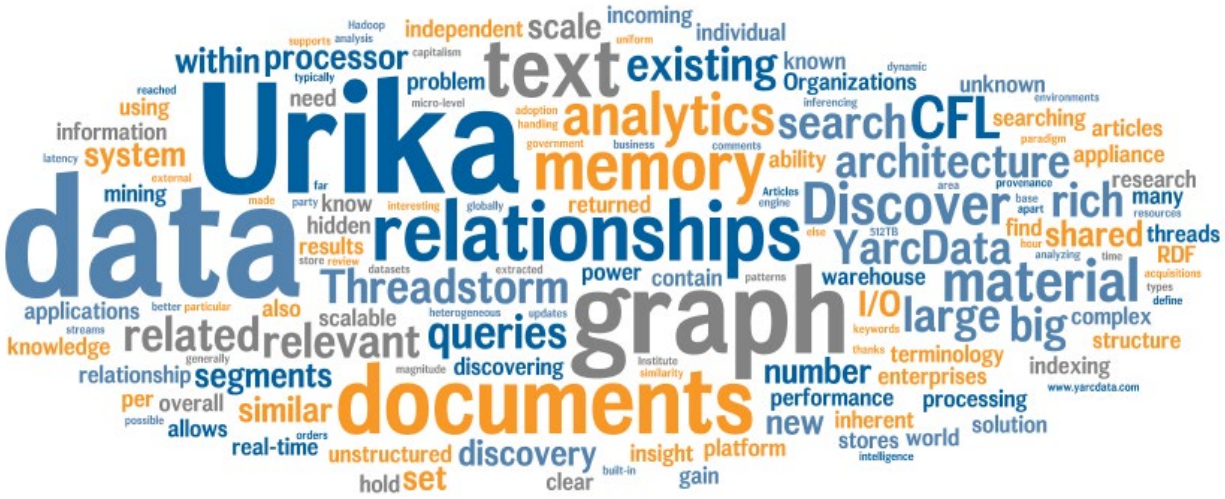
In the world of mergers and acquisitions, the same principles apply concerning the need for conducting due diligence on large and complex datasets. Neither party will know exactly what it should be looking for, but indications of related material can give a clear and rapid overview of where to direct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

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

The massively multi-threaded architecture of the Threadstorm™ (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 at several orders of magnitude better performance on graph problems than commodity hardware.

Life isn't structured. Don't worry if your data isn't either.



About Urika

YarcData's Urika is a big data appliance for graph analytics. Urika helps enterprises gain business insight by discovering relationships in big data. It's highly-scalable, real-time graph analytics warehouse supports ad hoc queries, pattern-based searches, inferencing and deduction. Urika complements an existing data warehouse or Hadoop cluster by offloading graph workloads and interoperating within the existing analytics workflow. Subscription pricing or on-premise deployment of the appliance eases Urika adoption into existing IT environments.

About YarcData

YarcData, a Cray company, delivers business-focused real-time graph analytics for enterprises to gain insight by discovering unknown relationships in big data. Adopters include the Institute of Systems Biology, the Mayo Clinic, Noblis, Sandia National Labs, as well as multiple deployments in the US government. YarcData is based in the San Francisco bay area and more information is at www.yarcdata.com.