YarcData
Getting to **Eureka!** faster™

Solution Brief

How a government organization is “connecting the dots” to eliminate terrorist threats and ensure public safety

The Business Problem

A government organization was challenged to quickly identify “persons of interest” in its massive data stores comprised of multiple data types from multiple sources. What makes the task daunting is that information is inaccurate, incomplete and likely contains deliberately falsified data from individuals actively trying to cover their tracks.

Suspicious individual activities are usually easy to spot, but individual activities by themselves do not necessarily represent a genuine threat. The real challenge was to find abnormal patterns of activities, especially those representing new forms of undiscovered threats.

The Technical Challenge

Representing seemingly disconnected information in a graph enables interactive exploration of relationships between people, places, things, organizations, communications, etc., thus making it easier to detect patterns.

The best performance on graph queries is achieved when the data is all held in memory, as opposed to residing on disk. On a mass-market system, when the data size of the graph exceeds the maximum memory size of a single node, which is a few terabytes, the graph has to be partitioned among multiple nodes. Unfortunately, partitioning graphs efficiently onto multiple nodes is a problem proven not to have efficient solutions.

For the persons-of-interest use case, this might be illustrated by data showing that communication between a person in Germany and a person in Malaysia is essential to a plot whose target is in the Philippines. While almost all of the communication of each person may be to people outside the plot, ignoring these rare but essential communications may overlook the existence of the plot. The Urika™ system’s innovative solution to this problem is an immense memory system, up to 512 terabytes, shared by all processors, which avoids the partitioning problem.

Further, even on a single node, these “long distance” communications, among graph nodes scattered in memory, violate one of the core performance principles of mass-market processors, which is that needed data is close to recently used data and hence its use can be predicted and the data pre-fetched. Mass-market systems have a limited ability to cope with this unpredictability, and thus graph problems execute notoriously slowly on mass-market systems. The Urika system is designed for this unpredictability, supporting numerous (128) outstanding requests from each processor compared to just a few (4-8) for mass-market processors.

The Urika™ Solution

To discover hidden and unknown relationships in big data, a graph analytics data warehouse has to be constructed, integrating data from multiple, disparate data sources and data formats into a consistent set of relationships, represented as a graph.

Further, the Urika appliance augments existing relationships through inference/deduction, such as noting that a new compound is classified as an explosive and so any access to that compound by a person of interest is noteworthy, so the resulting warehouse contains the most complete set of relationships available. It can be continuously updated either through the addition of new sources of data or through updates from existing sources with additional inference/deduction taking place each time to reveal new relationships.

Due to the tangle of relationships in highly interconnected data, activities can appear innocuous in isolation but can be suspicious when coupled with other activities by the same or related individuals. For example, buying fertilizer or renting a truck becomes suspicious when the activities are connected with a geographic location not normally associated with agriculture. A search pattern is used in Urika to identify such plots.

This makes it easy for the data analyst to iteratively define and find suspicious activity even if the exact threat is not well defined. With the differentiated performance delivered by the graph-focused Urika architecture, this work can be done interactively even on very large data, leading to a pace of discovery unattainable with mass-market systems.

Because Urika's software stack is enterprise ready and complies with WS02 standards, the government organization is able to deploy quickly. Standard interfaces like Java Python, and SuSE Linux reduce the learning curve and allow reuse of existing skillsets in a familiar environment. The RDF data representation and SPARQL querying language, on which the Urika appliance depends, are emerging standards being used and implemented by a variety of organizations that are originating data and building tools and systems to process such data and queries.

Eventually, the government organization will be able to replace a large number of commodity systems with a better performing, more reliable, easier-to-program and -manage system that lets them run far more queries far faster. By accelerating investigations, they are able to take effective action sooner.
Urika is enabling the organization to proactively identify terrorists, criminals, and countless nefarious plots by processing – interactively and in real time – a highly dynamic graph database that can scale to hundreds of terabytes in size.

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.