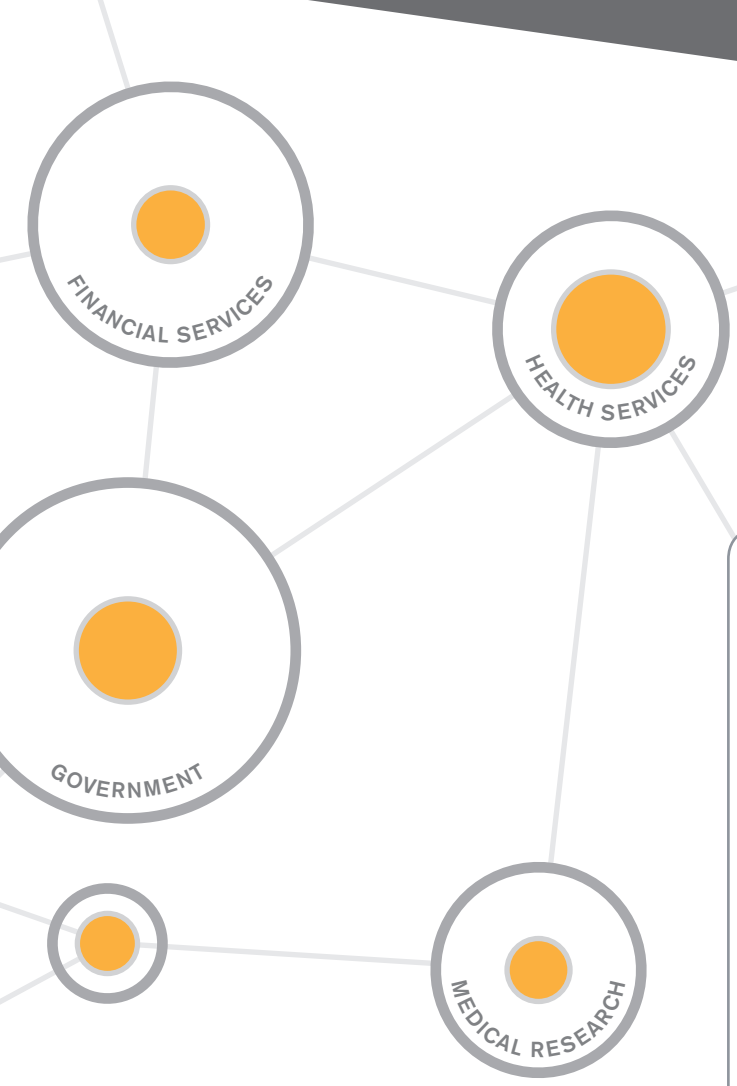
Urika

YarcData™

Enabling Real-Time Discovery in Big Data



Discovery is crucial to advancing knowledge. This essential process of uncovering hidden relationships and unknown patterns is difficult to automate and has long been viewed as unique to the human intellect. But with the advent of Big Data the sheer volume of data has made automation mandatory.

An iterative process, discovery involves formulating, testing and refining hypotheses. Thus, systems supporting discovery must be able to respond in real-time to complex, ad hoc queries and quickly add new relationships and data sources to the knowledge base

The Urika graph analytics appliance from YarcData is purpose-built to meet these challenging requirements, transforming massive amounts of seemingly unrelated data into relevant insights. With the world's most scalable shared memory architecture, Urika can discover hidden relationships and unknown patterns in Big Data, do it with an unmatched level of speed and simplicity, and facilitate the kinds of breakthroughs that can give your enterprise a measurable competitive advantage.

Urika complements existing data warehouses and Hadoop clusters by offloading graph analytics, while still interoperating with the existing analytics workflow. Subscription pricing for on-premise deployment eases Urika adoption into existing IT environments.

What is a graph?

A graph is a data structure capable of representing any kind of data in an accessible way. Fundamentally, a graph is an abstract representation of a set of objects where some pairs of the objects are connected by links. A graph consists of "nodes" and "edges" and is typically depicted in diagrammatic form as a set of dots for the nodes, joined by lines or curves for the edges. A node represents an entity (a person or thing) and an edge represents a relationship. Graphs provide a holistic view of the relationships in which an entity participates.

"IT organizations faced with previously infeasible graph-style discovery problems may succeed using a focused solution like Urika."

YarcData's Urika Shows Big Data is More than Hadoop and Data Warehouses (Carl Claunch, Sept 11, 2012)

The Advantages of Graphs for Discovery

Discovery generally involves analysis of the relationships between entities. Graphs represent these relationships directly, enabling relationship analytics. Adding new types of relationships to a graph is straightforward, as is the incremental addition of new data sources. Graph databases support ad hoc queries, a key requirement for discovery.

Collaboration between man and machine for discovery in Big Data requires real-time performance. Common Big Data implementations partition the data structures across a commodity cluster for performance. This approach is effective for storing, searching and retrieving individual entities, but not for analyzing relationships within the data. Despite this limitation, most efforts to implement graph analytics on clusters rely on partitioning.

Partitioning approaches result in low performance on Big Data graphs for three reasons:

1

Graphs are hard to partition.

Analyzing graph relationships requires following the edges in the graph. Regardless of the scheme used, partitioning the graph across a cluster will result in edges spanning cluster nodes. In most cases, the number of edges crossing cluster nodes is so large it requires a time-consuming network transfer each time those edges are crossed. Compared to local memory, even a fast commodity network such as 10 gigabit Ethernet is at least 100 times slower at transferring data. Given the highly interconnected nature of graphs, users gain a significant processing advantage if the entire graph is held in sufficiently large shared memory.

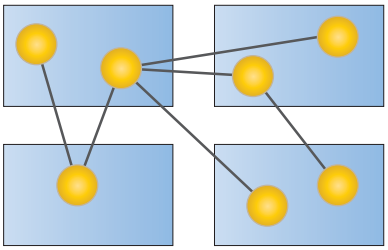


Figure 1. High cost to follow relationships that span cluster nodes

2

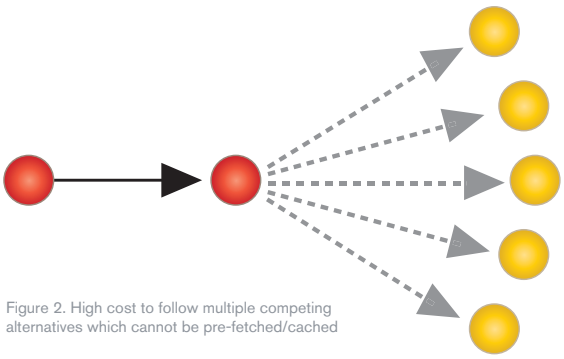


Figure 2. High cost to follow multiple competing alternatives which cannot be pre-fetched/cached

Graphs are not predictable.

Analyzing and graphing relationships in large datasets requires the examination of multiple, competing alternatives. These memory accesses are very data dependent and eliminate the ability to apply traditional performance improvement techniques such as pre-fetching and caching. Given that even the fastest memory is 100 times slower than processors and that most graph analytics computational work consists of exploring alternatives, the processor sits idle most of the time waiting for delivery of data. Using multithreading technology can help alleviate this problem. Threads can explore different alternatives and each thread can have its own memory access. As long as the processor supports a sufficient number of hardware threads, it can be kept busy. Given the highly non-deterministic nature of graphs, a massively multithreaded architecture enables a tremendous performance advantage.

3

Graphs are highly dynamic.

Graph analytics for discovery involves examining the relationships and correlations between multiple datasets and, consequently, requires loading many large, constantly changing datasets into memory. The sluggish speed of I/O systems – 1,000 times slower compared to the CPU – translates into graph load and modification times that can stretch into hours or days — far longer than the time required for running analytics. In a dynamic, real-time enterprise with constantly changing data, a scalable processing infrastructure enables a tremendous performance advantage for discovery.

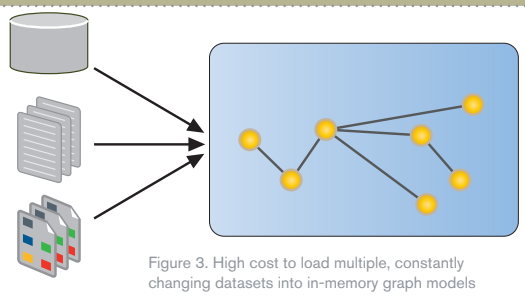


Figure 3. High cost to load multiple, constantly changing datasets into in-memory graph models

Urika: Purpose Built

Purpose built for graph analytics, the Urika hardware delivers three key innovations. These innovations mean Urika can handle the largest graph problems and enterprises can analyze real world Big Data graph problems in real time.

- Large, global shared memory whose architecture can scale up to 512 terabytes, eliminating the delays associated with 100 times slower network access.
- Threadstorm™ massively multithreaded graph processor supports 128 hardware threads in a single processor (65,000 threads in a 512 processor system and over 1 million threads at the maximum system size of 8,192 processors). This technology eliminates the waits caused by memory speed lagging the processor.
- Highly scalable I/O gets data into and out of Urika with transfer rates of up to 350TB/hr, alleviating the problem associated with storage I/O being 1,000 times slower than memory I/O.

Urika: Enterprise-Ready Appliance

The Urika appliance consists of the Graph Analytic Platform, the Graph Analytic Database and the Graph Analytic Application Services (Figure 4) each carefully designed for easy enterprise adoption.

The Graph Analytic Platform consists of a standard blade configuration, low power profile and air-cooled cabinet making it easy to deploy in enterprise datacenters. Installation is quick and users can load data and perform graph analytics immediately.

The Graph Analytic Database is a high performance, in-memory implementation of an RDF (Resource Description Framework) triple store. It can be queried using SPARQL 1.1 (SPARQL Protocol and RDF Query Language) which provides sophisticated pattern matching and dynamic data update capabilities. The database has been carefully tuned to the hardware and delivers orders of magnitude better performance than alternatives.

The Graph Analytics Application Services provide a comprehensive, simple and familiar set of management tools for the appliance and database, security and the data pipeline. Appliance management is provided through a comprehensive set of Linux-based tools, giving Urika the appearance of a Linux server. Database management is performed using the Graph Analytic Manager (GAM), which provides a familiar web-based management environment to database administrators.

With support for industry standards (including Linux, Java/J2EE, RDF, and SPARQL), Urika enables enterprises to leverage existing IT skill sets and expertise to solve Big Data graph problems, while avoiding vendor lock-in.

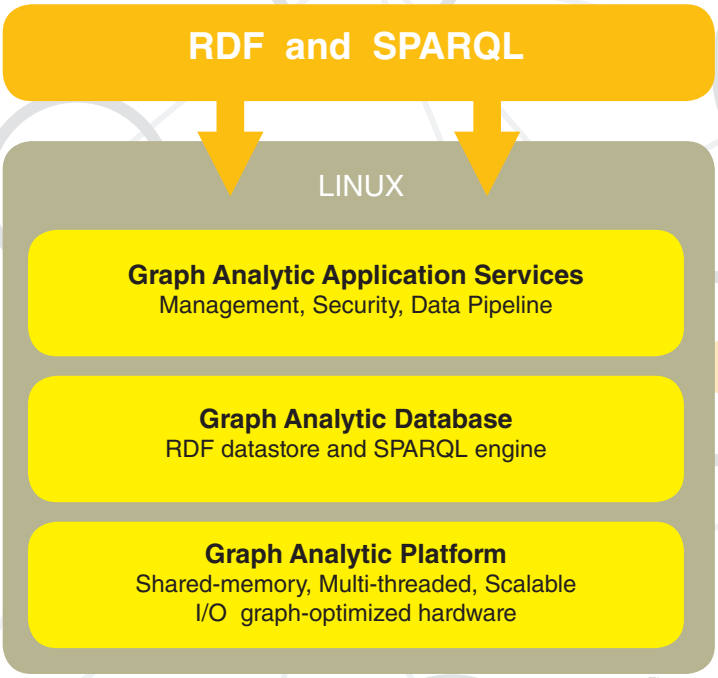


Figure 4

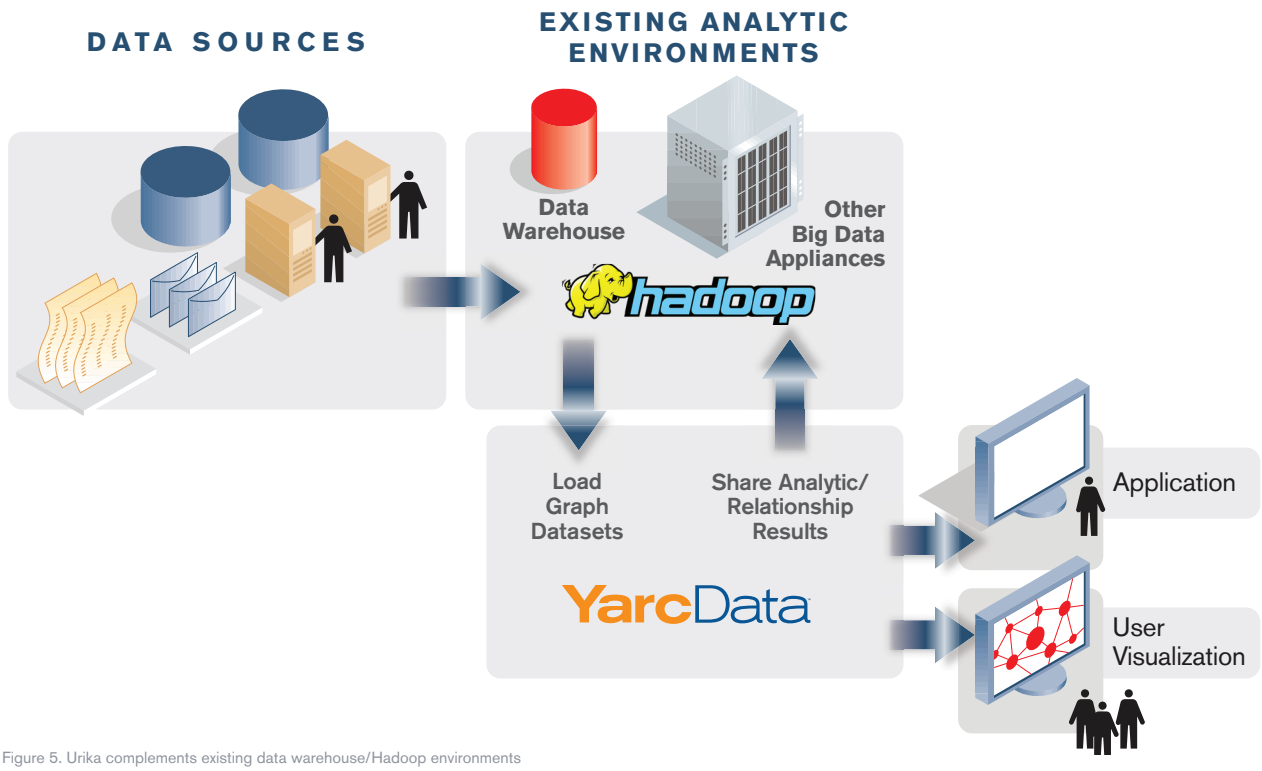


Figure 5. Urika complements existing data warehouse/Hadoop environments

Urika Integration in Existing Analytic Environments

The Urika appliance is designed to integrate with existing analytic environments and interoperate with data warehouses, Hadoop and other Big Data appliances (Figure 5). A connector model extracts data and returns results, enabling enterprises to offload graph workloads to an appliance specifically designed for the task. The approach also lets Urika interoperate within an existing enterprise analytics workflow/pipeline.

Gain insight by discovering unknown relationships in Big Data
Graph analytics warehouse supports ad hoc queries, pattern-based searches, inferencing and deduction on dynamic datasets

Achieve competitive advantage with scalable real-time graph analytics
Purpose built to solve Big Data graph problems with large shared memory and massive multithreading

Ease adoption with subscription pricing and industry standards support
Datacenter-ready appliance with open interfaces enables re-use of in-house skill sets, no lock-in and simplified integration

About YarcData Inc.

YarcData Inc., a Cray Company, delivers business-focused real-time graph analytics for enterprises to gain business insight by discovering unknown relationships in Big Data. Adopters include the Institute of Systems Biology, the Mayo Clinic, Noblis, Oak Ridge National Labs, Sandia National Labs and the Canadian government, as well as multiple deployments in the US government. YarcData is based in the San Francisco Bay Area. More information is available at www.yarcdata.com.

YarcData
*Getting to **Eureka!** faster*