YarcData

Urika™:

Enabling Real-Time Discovery in Big Data

Discovery is the process of gaining valuable insights into the world around us by recognizing previously unknown relationships between occurrences, objects and facts. Traditionally a manual process, big data has accelerated discovery by enabling the formulation and validation of theories electronically, through a collaboration between man and machine in areas as diverse as fraud and risk analysis, drug discovery, personalized healthcare, micro-targeted marketing campaigns, cybersecurity, law enforcement and counter-terrorism operations. Data discovery enables organizations to identify connections between disparate pieces of data: to find not just a single "needle in a needle stack," but hundreds—and the invisible threads that link them together.

YarcData's Urika is an appliance purpose built for the challenges of the discovery process, transforming massive amounts of seemingly unrelated data into relevant insights. With the world's most scalable shared memory architecture, Urika can surface hidden relationships and non-obvious patterns in big data with unmatched speed and simplicity, facilitating breakthroughs that can give your enterprise a measurable competitive advantage. Why is such an appliance necessary? Because traditional data analysis tools don't have the capability to effectively perform real-time data discovery:

"YarcData's Urika appliance offers organizations an extremely fast, effective means of implementing discovery analytics through a graph solution that includes a graph database.

("Discovering Big Data's Value with Graph Analytics," Enterprise Strategy Group)

Discovery cannot know all the data relationships in advance.
The essence of discovery is to identify and create relationships dynamically as new data sources are added. Traditional analytics tools fail because they are schema based, forcing time consuming and error prone schema extension for each new data source.

Urika addresses this challenge with a schema-free graph database. New sources of structured, semi-structured and unstructured data can be incrementally fused without upfront modeling. Furthermore, Urika's graph database also provides a range of powerful capabilities for querying and reasoning about relationships that are unavailable in traditional tools.

Discovery cannot know all the questions to ask in advance.
Data discovery is an iterative, real-time process, where the answers to the first set of questions determine the next questions to ask. Traditional analytics tools fail because their performance depends upon optimizing the data model for specific questions.

Urika addresses this challenge with a purpose built hardware accelerator leveraging massive multi-threading technology, capable of returning real-time results to the most complex, ad-hoc questions as dataset sizes grow. Furthermore, this technology supports the breadth of analytic and visualization approaches required for rapid exploration of big data.

Discovery doesn't access data in a predictable pattern.
Data access during discovery follows no predictable pattern, accessing data virtually randomly from anywhere in the massive trove of data. Traditional analytics tools fail because they depend upon effective partitioning of the data across a computing cluster, which requires advance knowledge of data access patterns.

YarcData

"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)

Urika addresses this challenge with a data model held entirely in a large memory of up to 512 Terabytes, shared by up to 8,192 multi-threaded processors with 128 hardware threads apiece. The large shared memory avoids the need for data partitioning required by distributed memory systems. Furthermore, hardware multi-threading enables the processors to tolerate memory access latency, delivering excellent scaling and real-time performance as dataset sizes and processor counts increase.

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

Urika can fulfill the true promise of big data in your organization. What will you discover?

The Power of Graphs for Discovery Analytics

Graphs are the ideal data model for discovery analytics, and Urika's unique hardware architecture enables graphs to scale to handle big data.

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 a relationship. Graphs provide a holistic view of the relationships in which an entity participates.

Graphs represent the relationships between entities directly, allowing data from multiple structured, semi-structured and unstructured sources to be fused without complex, time consuming and error prone modeling or schema definition. Furthermore, graph databases support a variety of sophisticated, ad-hoc analytic techniques, specifically designed to surface previously unknown patterns of relationships.

Achieving the real-time performance required for collaborative data discovery imposes certain architectural requirements:

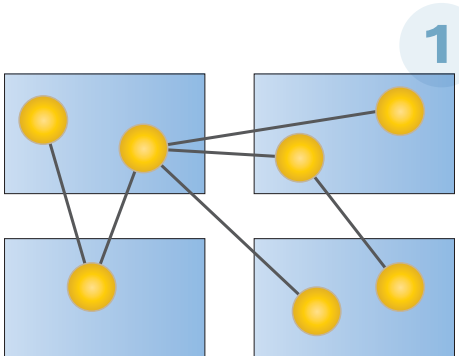

Figure 1. High cost to follow relationships that span cluster nodes

Graphs must not be partitioned.
Discovery analytics involve following the edges (the relationships) in the graph. It is an in-memory problem: the iterative nature of discovery dictates that the edges to be traversed will not be known in advance, so strategies based on pre-fetching from disk will not work. Similarly, partitioning the graph across a distributed memory compute cluster will result a large number of edges crossing cluster node boundaries, requiring time-consuming network transfers. Compared to local memory, even a fast commodity network such as 10 gigabit Ethernet is at least 100 times slower at transferring data. Users gain a significant processing advantage on discovery problems if the entire graph is held in a large shared memory.

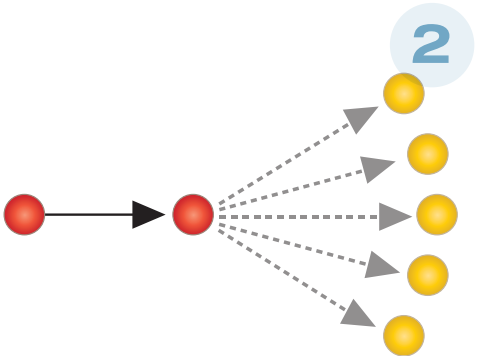

Figure 2. High cost to follow multiple competing alternatives which cannot be pre-fetched/cached

Graphs are not predictable.

Analyzing 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, with the result that 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 through the concurrent investigation of multiple, changing hypotheses.

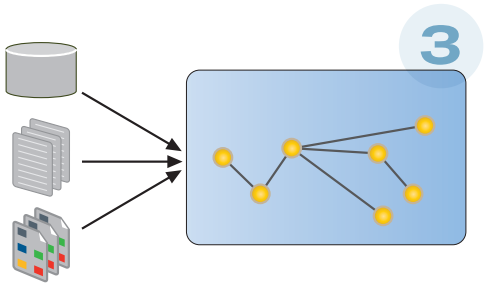

Figure 3. High cost to load multiple, constantly changing datasets into in-memory graph models

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.

Urika: Purpose Built

Urika is a purpose-built appliance for data discovery, integrating hardware, software and storage. Urika hardware delivers three key innovations to analyze the largest datasets in real time:

1. Large, global shared memory, whose architecture can scale up to 512 terabytes, enables uniform, low latency access to all of the data in the graph, with no need to consider partitioning, layout in memory or memory access patterns, eliminating the delays associated with 100 times slower network access.

2. Threadstorm™ massively multithreaded hardware accelerator 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), which enables global access of multiple, random dynamic memory references simultaneously without pre-fetching or caching, and eliminates the waits caused by memory speed lagging the processor.

3. Highly scalable I/O gets data into and out of Urika with transfer rates of up to 350TB/hr, allowing diverse data sets to be dynamically added and alleviating the problem associated with storage I/O being 1,000 times slower than memory I/O.

4

The Urika software stack consists of the Graph Analytic Database and the Graph Analytic Tools

The Graph Analytic Database is a high performance, in-memory, W3C standards compliant implementation of an RDF (Resource Description Framework) quad store. It can be queried using SPARQL 1.1 (SPARQL Protocol and RDF Query Language), providing 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 expressivity of SPARQL has been further extended with support for a range of whole graph algorithms, opening the door to new discovery approaches. For example, consider how "shortest path" analysis aids in understanding how two entities are related, while "community detection" and "between-ness centrality" enable identification of cliques and influencers for highly targeted marketing applications.

The Graph Analytic Tools 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. The Graph Analytics Manager (GAM) performs database management and is designed to provide a familiar environment to database administrators.

Urika: Easy to Deploy

Urika hardware consists of a standard blade configuration, low power profile and air-cooled cabinet making it easy to deploy in enterprise datacenters. It connects easily to standard networking and storage solutions. Installation is quick and users can load data and perform graph analytics immediately.

The Graph Analytic Tools provide a familiar environment to systems and database administrators. Existing skillsets are sufficient to provision, manage, secure and monitor Urika.

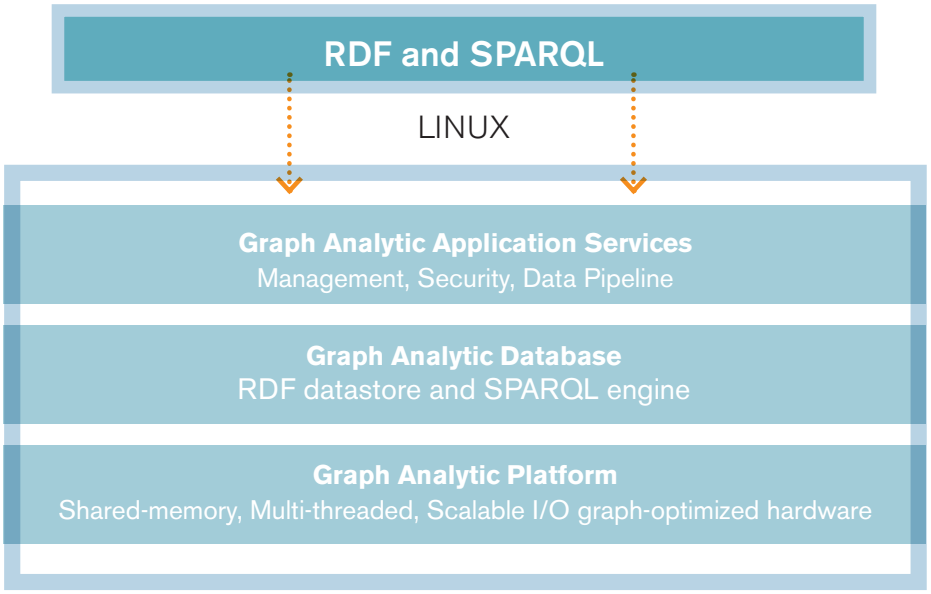

Figure 4.

5 Urika: Easy to Integrate

Urika augments existing analytic environments and easily integrates with data warehouses, Hadoop, other big data appliances and visualization solutions for a rapid return on investment (Figure 5). A connector model extracts data and returns results, enabling enterprises to offload graph workloads to an appliance specifically designed for the task. Urika also provides a graph database endpoint accessible through an open, industry standard interface, making integration easy for in-house application developers.

Urika's support for industry standards, including Linux, Java/ J2EE, JDBC, RDF and SPARQL, enables enterprises to leverage existing IT skill sets and expertise to solve big data discovery problems, while avoiding vendor lock-in.

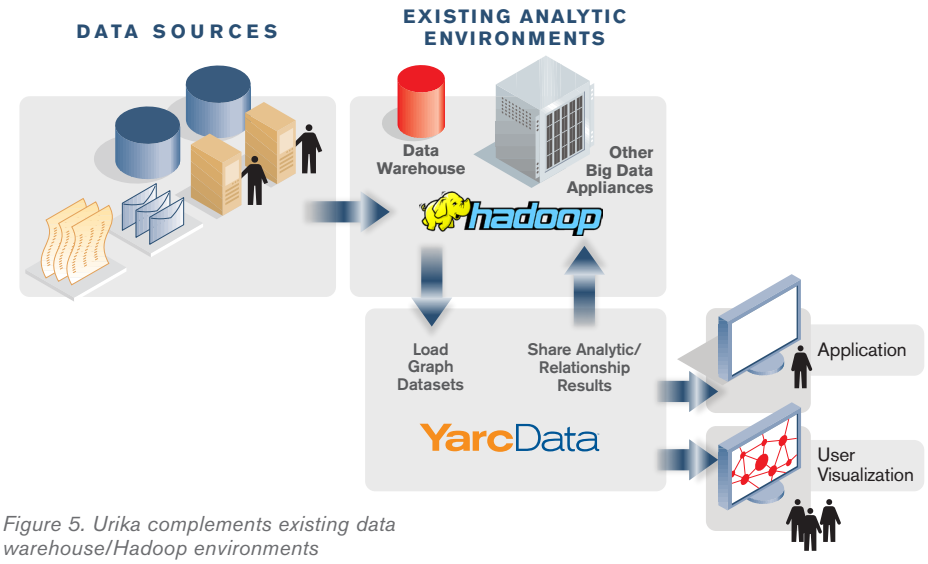

Figure 5. Urika complements existing data warehouse/Hadoop environments

Surface unknown linkages or non-obvious patterns without advance knowledge of relationships in the data

Shared memory model enables uniform, low-latency access to all the data regardless of data partitioning, layout, or access pattern

Investigate multiple, changing hypotheses in real-time simultaneously

Hardware Accelerator enables global access of multiple, random, dynamic memory references in parallel without pre-fetching/caching

Gain new insights by easily fusing diverse data sets without upfront modeling and independent of linkage

In-memory Graph Analytical Database enables merging of structured/semi-structured/ unstructured data without schemas/layouts and querying data without pre-specifying the connections between the data

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

About YarcData YarcData, a Cray company, delivers a big data appliance for real-time data discovery, enabling enterprises to gain game-changing business insights by surfacing unknown relationships and non-obvious patterns. Adopters include the Swiss National Supercomputing Centre (CSCS), the Mayo Clinic, Noblis, Oak Ridge National Laboratory, QinetiQ, Pittsburgh Supercomputing Center and Sandia National Laboratories, as well as leading government and intelligence organizations, financial services firms, life sciences companies, and telecommunications providers. YarcData is based in the San Francisco Bay Area and more information is at www.yarcdata.com.

P/N: xxxxxxxx