Graph Analytics
Two Goals…

Supercomputing Market Leadership

Grow Profitably

…Three Business Units

HPC Systems
- Breakthrough supercomputing performance and scalability

Storage & Data Management
- Scale-out, manageable high performance storage systems

YarcData
- Solutions for Big Data analytics
Created a new division within Cray

100% focused on Big Data solutions

Rapidly-growing, multi-billion market

Hired experienced management team with deep Enterprise experience
Search vs. Discovery

“Needle in a Haystack”

Paradigm: SEARCH

Needle in a Needlestack!

Paradigm: DISCOVERY

Discovery is about finding QUESTIONS as well as finding the answers to questions
The hype about big data is mostly on Hadoop or data warehouses, but big data involves a much wider and varied set of needs, practices and technologies. We offer recommendations for IT organizations seeking a solution to "graph" problems, including use of the uRiKA graph appliance.

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

Source: Gartner
Graphs are Everywhere

Introducing the Knowledge Graph: things, not strings
May 16, 2012

DebiRank: Too Central to Fail? Financial Networks, the FED and Systemic Risk

DebtRank decreases by moving outwards and leftwards along the spiral. The diagram allows at the same time to visualise the structure of the network and to compare the importance of any two given nodes. The size and the color of the node reflects the DebtRank value (larger and red nodes have higher DebtRank). The color of a link reflects the DebtRank of the node from which it originates (red links originate from node with high DebtRank and make high impact to the destination nodes).

Figure 4
Debt Rank, asset size and fragility.

Scatter plot of DebtRank versus asset size, measured as a fraction (in %) of the total of the asset size in the network. For sake of simplicity, in the experiment, asset size was assumed constant during the time span of the data. Notice that institutions such as UBS, or CITIGROUP alone account for almost 10% of the total assets. The size of each bubble is proportional to the outstanding debt of the institution while the color reflects its fragility, defined as the ratio of debt over market capitalization in the given period, as in the previous section.

(a) Period one. Since the outstanding debt was very low or zero, most nodes appear small and have levels of DebtRank below 0.3, but comparable among each other.

(b) Period four. Many institutions have a Debt Rank larger than 0.5, i.e. each can impact, alone, the majority of the economic value in the network. The outstanding debt in this period is close to the peak for all the institutions, as reflected by the size of the bubbles. Notice, also a higher fragility, most bubbles are red, although with some heterogeneity.
Many Analytics problems are about Relationships and Discovery

- Social Networking
  ![Social Networking](image)
- Life Sciences/Biology
  ![Life Sciences/Biology](image)
- Finance
  ![Finance](image)
- Intelligence/Security
  ![Intelligence/Security](image)
- Telecom/Mobile
  ![Telecom/Mobile](image)
- Healthcare/Medicine
  ![Healthcare/Medicine](image)
- Internet/WWW
  ![Internet/WWW](image)
- Supply Chain
  ![Supply Chain](image)

Company Confidential – Do Not Distribute
Common Theme:
DISCOVERING RELATIONSHIPs spanning large, fragmented data sets, and doing so at an increasing speed

Domain
- Risk
- Compliance
- Customer Insight
- Cyber Security

Current
- Departmental Computations & aggregations
- Recognition of known behaviors
- Profitability, retention and marketing
- Siloed examination of network incidents

Emerging
- Risk as a function of the connections between counterparties
- Proactively discover new patterns of suspicious behavior
- Follow connections across customer social networks to drive revenue
- Find hidden threats through a connected view of network activity
Current Analytical processing depends on partitionable data structures

- RDBMS
- Column Oriented
- No SQL
- Relational Extensions
- No ACID
- In Memory
- Key Value
- Document Stores
Current Graph approaches are poorly suited to relationship and discovery “Graph” analytics

**Graphs are hard to Partition**
High cost to follow relationships that span Cluster Nodes

**Graphs are not Predictable**
High cost to follow multiple competing paths which cannot be pre-fetched/cached

**Graphs are highly Dynamic**
High cost to load multiple, constantly changing datasets into in-memory graph models

uRiKA: Purpose-built Graph Appliance

Business Challenge

- Graphs are hard to Partition
- Graphs are not Predictable
- Graphs are highly Dynamic

URiKA Advantages

- Large Shared Memory
  Up to 512 TB
- Massively Multi-threaded
  128 Hardware Threads/processor
  Up to 8,192 Processors
- Highly Scalable I/O
  Up to 350 TB/hr

Real-time, Interactive Analytics on Big Data Graph Problems
uRiKA: … designed for Enterprise Adoption

- **Data Center** sees another Linux server
- **Applications** see industry standard interface
  - RDF, SPARQL, Java, Gadgets...
- **Reuse Existing Skillsets**
  - Java, OSGI, App Server, SOA, ESB, Web toolkit...
- **No Lock-in**
  - All applications and artifacts built on uRiKA can be run on other platforms
- **Subscription Pricing Model**
Subscription Pricing Model

● Install in your Data Center like any other purchased equipment
  • ... But pay a subscription fee for use of the box instead of purchasing the box up front

● Incrementally add processors and memory as use-cases expand

● License for as little as a one-year term

● Enables firms to prove out the value incrementally
Graph Analytics Augments Your Analytics Infrastructure

Data Sources
- External
  - Data Gateways
  - Market Data
  - News Feeds
  - Social Media
  - IDS Logs
  - Call Data Records
- Internal
  - Core Systems
  - Ops Data
  - Emails
  - Netflow Data
  - Web Logs
  - Incident Reports

Structured
- Departmental DWs

BI/DW Infrastructure
- EDW
- Customer Mart
- Risk Mart
- Finance Mart
- Compliance Mart

Unstructured
- Graph Analytics

Analytical Outputs
- Regulatory Capital
- Average Balance
- Relationship Depth
- Regulatory Capital
- Product Profitability
- Paths to Fraud
- Developing plots
- Developing cyber attacks
- Impactful market events
- Ultimate counterparty exposure

YarcData
A CRAY COMPANY
“In the amount of time it takes to validate one hypothesis, we can now validate 1,000 hypotheses – massively improving our success rate and systematizing serendipity.” (YarcData Govt Customer)
A Graph Database

- Stores information in the following format – an RDF Triple

<table>
<thead>
<tr>
<th>subject</th>
<th>predicate</th>
<th>object</th>
</tr>
</thead>
<tbody>
<tr>
<td>David Jones</td>
<td>WorksFor</td>
<td>Bank of Nebraska LLC</td>
</tr>
</tbody>
</table>

- Each RDF triple is a pair of vertices and an edge in a directed graph:
Use Cases for Scalable Graph Analytics
Government Organization: Discovering new plots

- **Goal**: Proactively identify patterns of activity and threat candidates by aggregating intelligence and analysis

- **Data sets**: Reference data, People, Places, Things, Organizations, Communications…

- **Technical Challenges**: Volume, Variety and Velocity of data; Inaccurate, incomplete and falsified data

- **Users**: Intelligence Analysts

- **Usage model**: Search for patterns of activity and graphically explore relationships between entities for candidate behavior and activities

- **Augmenting**: Existing Hadoop cluster and multiple data appliances
uRiKA Customer Use Case: Healthcare Provider

Discover “similar” Patients to optimize treatment

The Challenge

- Longitudinal, historical data spanning all events, symptoms, diagnoses, diseases, treatments, prescriptions, etc of 10M patients including genetics and family history
- Ad-hoc, constantly changing definition of “similarity” based on thousands of parameters
- Interactive, real-time response during consultation

uRiKA Solution

- uRiKA holds entire relationship graph in memory
- Identify “similar” patients based on ad-hoc physician specified patterns
- Interactive, real-time access by entire physician community

Business Value

- Consistent selection of the most effective treatment for each patient by each doctor every time
uRiKA Financial Services Use Case: Trade Surveillance

Quickly adapt surveillance to new rules and patterns

The Challenge
- Continuous stream of possible compliance events
- New compliance rules require constant updates of signals and patterns
- Current solutions require offline preparation and are based on rigid rules

uRiKA Solution
- Entire relationship graph in memory
- New Patterns/templates can be identified and added in real-time
- Graphical interactive exploration of relationships between people, places, things, organizations, communications, etc.

Business Value
- A responsive, flexible event detection platform that adapts to new knowledge

Figure Source: *Graph-based technologies for intelligence analysis*, T. Coffman, S. Greenblatt, S. Marcus, Commun. ACM, 47(3):45-47, 2004
uRiKA Financial Services Use Case: Operational Risk

Monitor, analyze complex operational processes to reduce risk

The Challenge

- In large institutions, operations (e.g., Trade Lifecycle) can be complex, inter-connected processes involving people, technology, and process steps
- Multiple data collection points exist, but hard to combine to get a holistic picture
- Hard to monitor if critical controls are effective

uRiKA Solution

- Ability to hold entire ‘operational model’ in memory
- Can seamlessly add additional incremental data sources relevant to analysis
- Interactive, real-time querying capability allows ‘constant’ monitoring of operations

Business Value

- Better control of operational risk by combining disparate, disconnected instrumentation data for a complex operation
uRiKA Financial Services Use Case: Systemic Risk

Understand financial risk emerging from connections in a financial system

The Challenge

- The financial crisis illustrated that counterparty risk is a function of the network of counterparties, not just each in isolation
- Eg. Ford, its suppliers, employees are all counterparties with interconnected exposures to a lender like BofA

uRiKA Solution

- Ability to hold entire dependency graph of positions in memory
- Foundation to run complex graph analytic algorithms without losing interactivity

Business Value

- Ability to gain additional perspective on complex counterparty risk in addition to existing procedures for risk calculations
**Healthcare Provider: Discovering new treatments**

- **Goal:** Proactively identify optimal treatments for patients based on treatment results for “similar” patients
- **Data sets:** Longitudinal patient data, Family history, Genetics, Reference data,
- **Technical Challenges:** Ad-hoc constantly changing definition of “similarity” across thousands of constantly changing attributes
- **Users:** Doctors
- **Usage model:** Compare results of candidate treatment options for “similar” patients based on ad-hoc physician specified patterns
- **Augmenting:** Existing data warehouse
Government Organization: Discovering new cyber threats

- **Goal**: Proactively identify unknown cyber threats by examining all relationships
- **Data sets**: IP, MAC, BGP, Firewall, DNS, Netflow, Whois, NVD, CIDR...
- **Technical Challenges**: Volume and Velocity of data; Temporal dependencies; Real-time response
- **Users**: Cyber Analysts
- **Usage model**: Iterative analysis of all patterns across all traffic to explore deviations in frequency of occurrence, derivative patterns of known threats and linking patterns through relationships in offline data
- **Augmenting**: Existing data appliances
Healthcare Payer: Discovering new fraud patterns

- **Goal:** Proactively identify new patterns of healthcare fraud (perpetrator/provider/patient nexus) by examining all healthcare relationships
- **Data sets:** Provider, Beneficiary, Policy, Claims, Billing, Services, Outcomes, Social Network, Organization…
- **Technical Challenges:** Volume and Velocity of data; Entity Resolution; Complex Inter-relationships; Temporal dependencies
- **Users:** Fraud Inspectors/Analysts
- **Usage model:** Analyze outcome and cost for various disease trajectories and identify outliers for treatment optimization and fraud investigation; Separate fraud patterns from benign treatment or legitimate errors
- **Augmenting:** Existing data warehouse, Predictive Analytics
The Key to Leveraging Graph Analytics

- When you can interactively ask complex, graph-oriented questions of large, diverse data sets, you can discover patterns and relationships that are beyond the capability of other systems.

- You can solve analytical challenges you thought were intractable.
Q & A