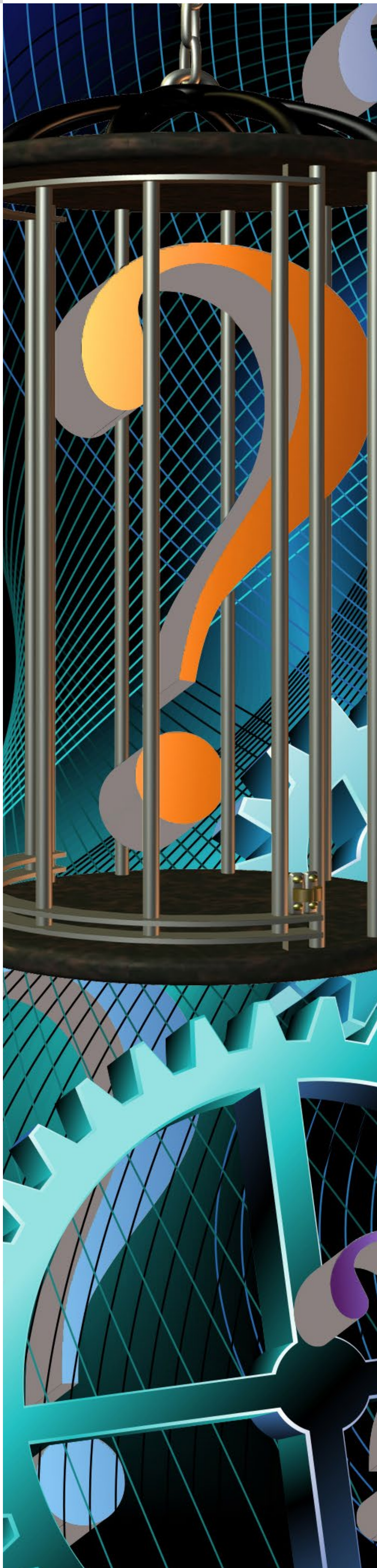
YarcData
Getting to **Eureka!** faster™

Solution Brief

Forbidden No More: Removing the restrictions on ad hoc analytic queries



The Business Problem

IT groups frequently struggle to find a balance between the conventional analytics that underpin operational needs and demands for complex, ad-hoc analytics from business units. Massive data warehouses and analytic environments, operating under change control, have been built in response to the former. Often, however, the ad-hoc analytics involve queries for which the data warehouse was not optimized, resulting in unacceptably long query execution and impact to existing workloads.

The demand for ad-hoc analytics is driven by growing awareness of the business value of discovering hidden and unknown relationships in big datasets: for example, uncovering new trading strategies, determining counter-party risk, identifying insider trading or exposing new forms of cyber-attack. Discovery typically entails iteratively posing complex queries to the analytics systems, further increasing workload.

This leaves IT executives with very difficult choices:

- Buy additional analytics hardware? But the expenditure required will typically be millions of unbudgeted dollars.
- Optimize the data warehouse for each new query? But IT resources are already stretched very thin, and optimizing warehouse data schema is a complex, error-prone and time-consuming process, not to be lightly undertaken.
- Forbid these complex ad hoc queries – or limit them to off-hours and weekends only? But there are legitimate business drivers for these analytics, and senior business leaders would be displeased with IT responsiveness.

It's time for a new approach—one that recognizes the reality that massive and growing big data sources are here to stay; that discovering new relationships in big data is the new norm, requiring complex ad hoc queries; and that timescales for responding to queries are shrinking to real time to support iterative processes. What's required is an approach which augments and optimizes existing analytics investments.

The Technical Challenge

Executive demands for complex, ad hoc, real time analysis is causing IT tremendous stress because of three very significant challenges.

Running complex, ad-hoc queries against un-optimized datasets cripples analytic system performance.   Data warehouses depend upon schema optimized for the queries to be run to achieve good performance.   However, optimization for complex, ad-hoc queries is not possible because, by definition, the queries to be run are not known in advance.   They also frequently involve exploring relationships between data items, necessitating table joins.   The net result is that these queries frequently result in computationally expensive, deeply nested and self-referential table joins, which consume enormous analytics resources and may never complete when confronted with big data. So despite their business value, IT groups are forced to ban many such queries.

Preparing the dataset needed to run the analysis is complex, error prone and labor-intensive. The difficulty arises because the required data often comes from different data sources, each with their own schema, but existing analytic environments require a pre-defined, common schema accommodating all the required data. Schema extensions can tie up several people for days or weeks – not a good investment for a query which will only be run once.

Achieving real time response is problematic with existing analytic environments designed for batch execution.   Discovery is an iterative process, involving collaboration between man and machine. Every step involves the exploration of a hypothesis, and the results guide the next step.   Real time response is a pre-requisite to analysts reaching their conclusions in a useful timeframe.

The Urika™ Solution

A major financial services organization augmented their existing analytic environment with YarcData's Urika appliance to enable complex ad hoc analytics in real time. Data from all relevant sources were loaded into a very large graph, and new data relationships and data sources could be dynamically defined as needed to run a query, providing transparent scalability.   Urika does not require schema to be defined in advance, removing a major barrier to executing ad hoc queries. Finally, the Urika appliance is designed to provide deterministic, real time performance on the most complex queries.

Urika owes its performance to its large shared memory, scalable I/O system and purpose-built Threadstorm graph processor. Its huge, global shared memory architecture of up to 512TB can hold the entire graph of relationships, and 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.

The massively multi-threaded architecture of the Threadstorm™ processor (128 independent threads) 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 delivers several orders of magnitude better performance on graph problems than commodity hardware.

Query results are provided in human time – seconds rather than overnight or multi-day runs – and presented graphically so that it's easier to understand relationships at a glance. The overall speed allows refinements to be made multiple times if necessary within the time-sensitive window of real time transactions.

So rather than forbidding complex, ad hoc queries from the building, Urika wins them a welcome invitation to the party.

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.