Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

Semantic Data Analysis

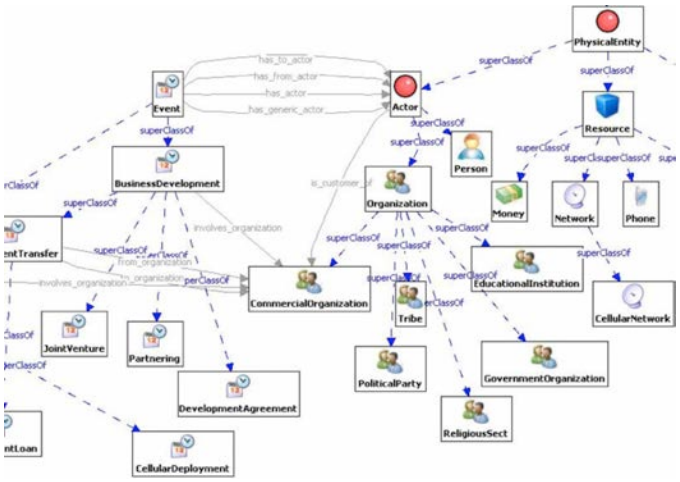
CHALLENGE

Applying high-performance computing technology and advanced mathematical algorithms to big data graph problems requires integration of multiple architectures and software. The resulting hybrid systems—cloud and high-performance computing, tables and graphs, structured and unstructured—motivate new complexity-hiding middleware interfaces for graph query and analytics. Because much of the data of interest are organized as a collection of records from mixed data sources, handling them at large scales adds challenges, including fusion, memory footprint, time-to-solution, and ease-of-use. Moreover, most real-world data sources are extracted from large volumes of noisy data, requiring construction of an enriched information network or knowledge graph from the raw data graph. With large data that typically are continually generated, a system that can periodically ingest more (new) data efficiently is important.

CURRENT PRACTICE

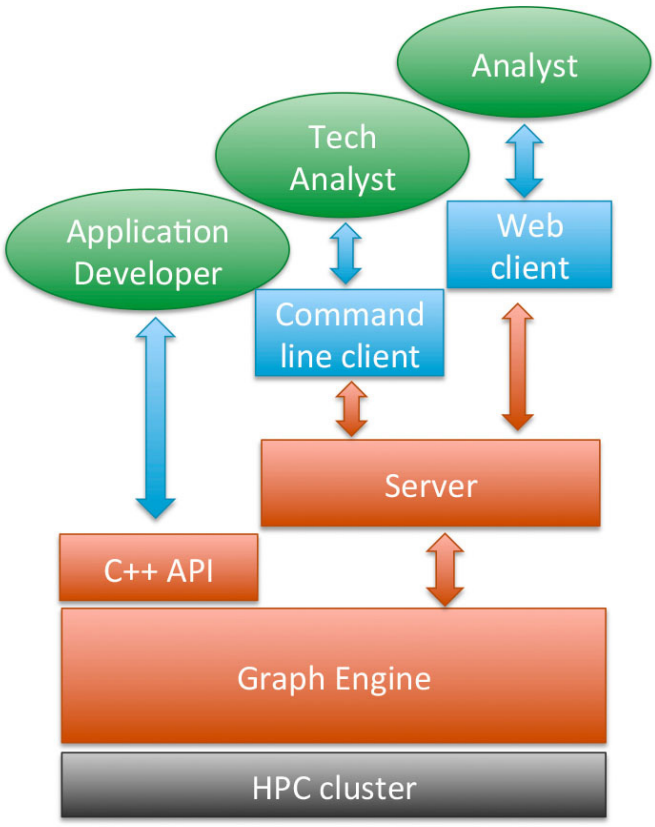
Big data systems are limited by their assumptions and architectures, and there are few attempts to provide common interfaces and capabilities across different big data platforms and high-performance computing systems. An extension of theoretical work from the 1970s and 1980s allows relational queries and complex graph analytics to be parallelized, naturally accommodating Semantic graphs with complex edge and vertex labels and conventional relations. This approach affords connecting different languages for manipulating graphs and relations

Creating high-performance platforms for computing over large sizes of both graph and tabular (SQL-like) data.



Property graphs provide random access to irregularly structured data.

with different backend systems. This approach has not been explored because of implicit assumptions that graph data processing is incompatible with relational data processing and that high-performance computing architectures are incompatible with cloud-oriented computing architectures.



GEMS architecture showing three profiles of users with a diversity of requirements.

TECHNICAL APPROACH

Our primary focus is on the advancement of the Graph Engine for Multithreaded Systems (GEMS), an in-memory property graph database platform, enabled to exploit massive distributed memory by a multi-threading architecture that runs on all commodity x86/Linux architectures, from desktops to clouds. The property graph data model supports ingesting comma-separated values (CSV) files, which means most data are either immediately readable or easily converted. This ingest process supports adding more data without computing new index data structures from scratch.

The flexibility of GEMS software supports a range of user profiles, from an application-domain analyst who knows very little about computing, algorithms, or high-performance computing systems to a more technically oriented analyst using a command-line interface to perform data engineering that best supports the analysts or an application developer writing applications in C++ aimed at computing custom algorithms and/or achieving extreme performance optimization. Analysts typically approach problems in an interactive question-answer manner. Answers to previous questions give rise to new questions. This back and forth requires a high-level query language to allow the analysts to quickly formulate questions. But for those questions that prove to be extremely useful, GEMS supports a C++ API-level programming interface for extreme performance optimization.

IMPACT

The technologies being developed will accelerate massive scaling of graph data query and analysis (on the order of 1T attributed edges or 10T edge attributes) and flexible hybrid computing patterns involving tabular data (SQL) and relational (graphs with relationship edges).

Obtaining insight about a data set (especially one that evolves over time) requires a complex, high-performance, flexible system that can be run on many different scales of computing platforms. GEMS uniquely accommodates all of these features.

Contacts

Vito Giovanni Castellana
Principal Investigator
(509) 375-4421
VitoGiovanni.Castellana@pnnl.gov

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle Since 1965