Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** *Since 1965*

Streaming Hypothesis Reasoning (SHyRe)

Pacific Northwest National Laboratory (PNNL) researchers employ Streaming Hypothesis Reasoning (SHyRe) to better understand the tradeoffs between completeness and performance so that reasoning systems can be more readily optimized in-line with the requirements of each use case.

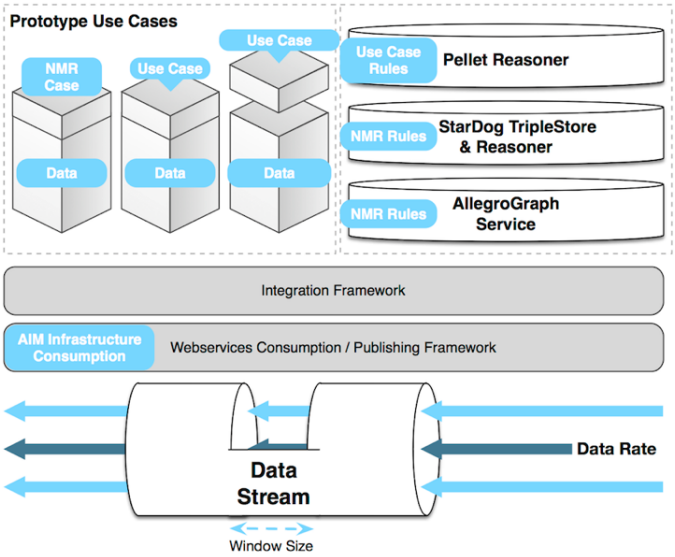
Challenge

Technological advancements in the collection, movement, and storage of data have changed society's ability to measure and analyze information. Yet, data streams provide unique challenges to the analysis process. Often streams of data can be noisy, with extraneous data infused into the stream due to measurement error or interference from an outside source.

Oftentimes a semantic reasoning system is employed to maximize completeness, valued for their rich expressiveness, these data structures quickly grow complex as the knowledge they represent increases. In the streaming context, performance must be taken into consideration at the expense of large graph theory. Options for optimizing this tradeoff include data sampling, data-aware cache eviction techniques, high performance reasoners, theoretical advances in reasoning algorithms, and combining results with inductive methods. The goal of SHyRe is to explore these techniques as well as understand tradeoffs.

Approach

The approaches SHyRe employs primarily consist of dynamic graph creation with semantic triples, description logic queries uniquely composed to return results even when a graph is incomplete, and cache eviction strategies for managing graph size. Because the streaming data within the graphs must periodically be pruned to ensure query completion, an entirely new problem was discovered as graphs increased in size and streaming data was included into the system unchecked.

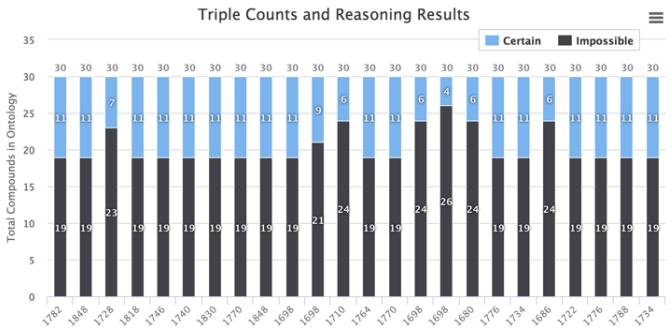


Classic cache-based reasoning systems evict the cache using a time window such that the oldest data is evicted – First In, First Out. Yet it is possible that the oldest data is the most relevant for reasoning. In order to rank the data in order of importance, we introduced the notion of Semantic Importance (SI), which is an indicator that measures a piece of data's contribution to the reasoning and/or query results.

By utilizing SI, researchers explored more sophisticated cache eviction techniques including First-Expired-First-Out, Least-Frequently-Used, and Least-Recently-Used. The SI measure is customized such that less important data, in accordance with the technique, is given a lower SI. Accordingly, data with the lowest SI is evicted. In conjunction with cache management algorithms, we also investigated how performance varies given a fixed cache size and a minimum data eviction requirement.

Impact

SHyRe activities provide the foundation of knowledge upon which to more readily implement deductive reasoning systems on streaming data. The introduction of an architecture that is capable of making operations that bridge streaming data and Semantic Web technologies by providing connectivity between the data stream, graph storage components, and Description Logic reasoner. SHyRe not only includes a means to perform stream ingestion, it also provides a framework for Resource Description Framework (RDF) annotation and querying graph storage systems with in-memory RDF models or SPARQL. By adapting and extending known software development patterns, PNNL researchers designed a



prototype with a reusable architecture that is capable of continuous streaming data collection, RDF annotation, graph storage, and semantic querying.

About the Analysis in Motion Initiative

SHyRe is supported by the Analysis in Motion Initiative. Technology advances have enabled the continuous capture and streaming of data from an increasingly diverse set of sources, from scientific instruments, to the web, to consumer electronics. Timely discovery, evaluation, and response to key insights from these streams require new computing capabilities and a rebalancing of human and computer effort. The Analysis in Motion Initiative is developing analysis and interaction technologies to enable humans to respond more quickly and accurately to data streams reflecting changing events.

About PNNL

Interdisciplinary teams at PNNL address many of America's most pressing issues in energy, the environment and national security through advances in basic and applied science. Founded in 1965, PNNL employs 4,400 staff and has an annual budget of nearly $1 billion. It is managed by Battelle for the U.S. Department of Energy's Office of Science.

For more information, please contact:

Wil Smith
Engineer
(206) 528-3356
william.smith@pnnl.gov

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

U.S. DEPARTMENT OF ENERGY