LabX to Triple Store POC

Semantic Data Models on an Enterprise Scale to Inform Deployment Strategies for the Allotrope Framework

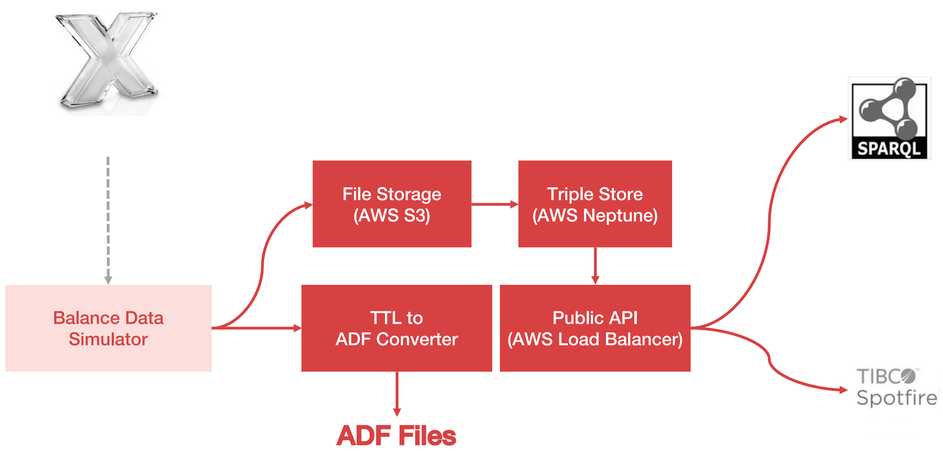
Vincent Chan[1], Milos Grbic[1], Kostadin Alargov[1], Matthew Kramer[2], Benjamin Woolford-Lim[2], Joy Ginocchio[2], James Roberts[2]

[1]TetraScience
[2]GlaxoSmithKline

The objective of this project was to test the use of a gravimetric data model in a real-world production environment. In order to assess this, we measured the performance of SPARQL queries on an enterprise-scale gravimetric dataset within an AWS Neptune triple store. In addition we took a broad assessment over the effectiveness and feasibility of implementing and adopting a gravimetric data model from start-to-finish.

Methodology

In order to model a real-world environment, very large quantities of simulated LabX RDF data was generated and saved in the form of TTL files before being imported into Neptune (see below for system architecture). To assess performance of the system, every step of this process was timed to produce a good picture of performance metrics.



Results

Results were tested using three SPARQL queries that reflect real-world use:

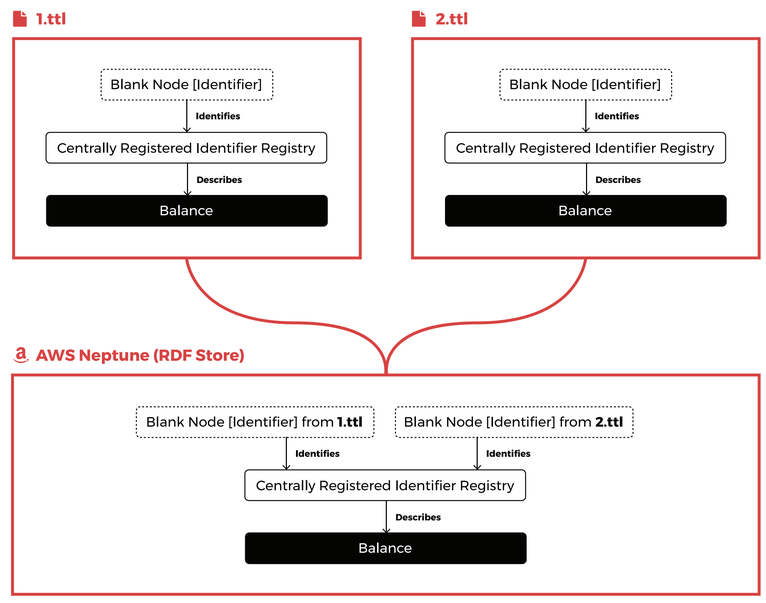
1) Query balances by calibration data (for Metrology)
2) Query weighing events by date-range (for a Quality Investigation)
3) Query by Container ID (To support downstream software)

These queries were selected because their complexity involves multiple many-to-many relationships to solve. Although graph technology is technically better suited for these types of queries, comparison with real-world performance of equivalent relational queries was not measured.

Results revealed some query performance issues, but these were promptly solved through query optimization. There are broader issues found that limit feasibility and practicality. TetraScience has targeted two issues *see right* for further exploration.

CHALLENGE 1: DUPLICATION OF BLANK NODES

Triple store technologies (like Neptune) typically have some form of bulk-loading of RDF graphs. When blank nodes exist, they get duplicated for every graph loaded into the triple store due to their anonymous nature. The illustration below demonstrates how this duplication occurs.

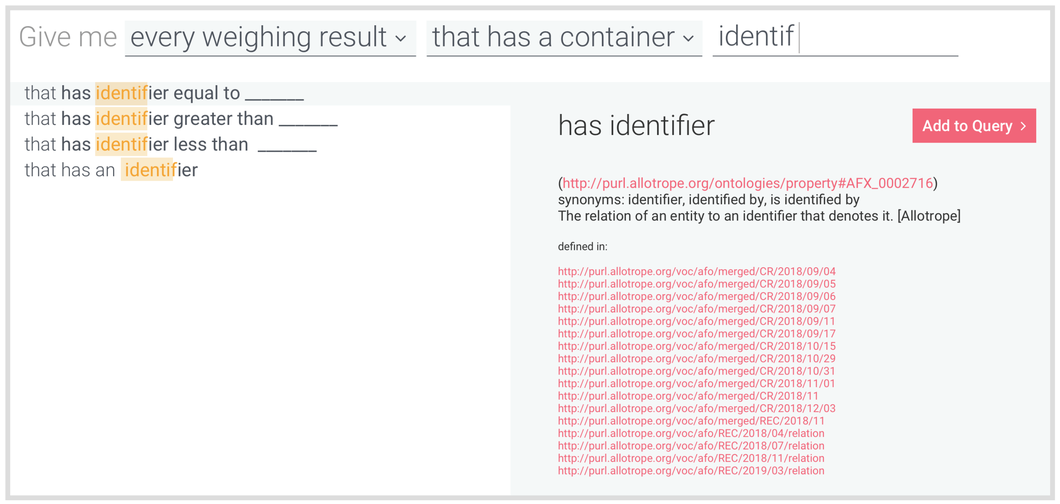


Academic discussion exists around how to address Blank Nodes in graph data, and the W3C has its own guidance[1A]. An accepted long-term solution is to assign transient identifiers (Skolem IRIs) to blank node instances across graphs. This process, also called *Skolemization,* allows the RDF import mechanism to appropriately merge nodes when loading data. The skolemization challenge will likely be encountered in any production-sized dataset, which warrants making a solution.

CHALLENGE 2: USABILITY AND TRIPLE STORE ADOPTION

To justify investment into Semantic Web and the Allotrope Framework, it is important to assess ways it can be applied to produce science value. To this end, triple store technology can be leveraged to more directly expose Allotrope data to scientists.

Putting RDFs into a triple store is not enough to produce scientific value, as the learning curve required to perform SPARQL is a huge barrier to adoption. Efforts can be made to explore new (or existing) user interfaces, such as the semantic query builder mocked up below, which can help make this more accessible to end-users.



[1A] **https://www.w3.org/TR/rdf11-concepts/#section-skolemization**

TETRASCIENCE

To interact with gravimetric data in our AWS Neptune triple store, access

https://tetrascience.com/querytest