SPARQL on Hadoop using Apache Hive and Jena SDB

Eric Nagler\textsuperscript{1,*}, Alex Vertlieb\textsuperscript{1}

1. CUBRC

* Contact author: eric.nagler@cubrc.org

Keywords: Semantic Web, Big Data, Apache Hive, RDF, SPARQL-to-SQL Translation

Semantic web technologies are frequently being used to align, store, and query data due to their expressiveness and powerful vocabularies. Our research seeks to mesh the utility of the semantic web and the query capability, availability and active research and development of Hadoop. To accomplish this we are developing an RDF triple store that utilizes the Apache Hadoop storage and processing framework along with the Apache Hive SQL-on-Hadoop processing engine. One goal of this research is to be as flexible as possible by avoiding a coupling to any particular Big Data SQL query execution engine. To accomplish this we are leveraging an open-source SPARQL-to-SQL translation module, Apache Jena SDB, and extending it to be compatible with multiple SQL-on-Hadoop engines, allowing us to issue queries and return results using the universal Hive Query Language. Currently, we have integrated this tool to run against two different SQL-on-Hadoop providers, the batch oriented Apache Hive and the real-time oriented Cloudera Impala. Our research also focuses on leveraging the best Hadoop data processing features to access the semantic web data as efficiently as possible. The two features we are leveraging are Hive’s data partitioning and advanced file storage formats (Sequence Files, Apache Parquet) to efficiently store and access semantic web graphs.

Currently, we are targeting full SPARQL 1.1 capability, expanding past the current simple graph pattern matching prototype. In order to achieve this we are developing our software against the SPARQL 1.1 testing suite\cite{Polleres:2015} and using two industry standard benchmarks, the LeHigh University Benchmark (LUBM)\cite{Guo:2005} and the Berlin SPARQL Benchmark\cite{Bizer:2009}. Internally, we have seen positive initial results with a 9x reduction of space required to store the LUBM semantic graph and query speed improvements using different query execution engines and optimizations to the Jena SDB query translator. This software is in active development with support from the Federal Highway Administration’s (FHWA) Exploratory Advanced Research Program to develop a Transportation Research Informatics Platform (TRIP) providing a data intensive RDF triple store to assist data scientists and researchers with transportation safety research challenges.

References


\cite{Polleres:2015} Polleres, A., S. Harris, J. Broekstra, and L. Feignbaum. SPARQL 1.1: Test case structure.