Datanami article – The Bright Future of Semantic Graphs and Big Connected Data

The big data revolution is generating a mess of unruly data that's difficult to parse and understand. This is to be expected–explosions don't generally occur in a nice, orderly fashion, after all. But if the folks at Cloudera and Franz have their way, the world of connected data will become more accessible and useful when viewed through the lens of semantic graph technologies.

Semantic graph technology is shaping up to play a key role in how organizations access the growing stores of public data. This is particularly true in the healthcare space, where organizations are beginning to store their data using so-called triple stores, often defined by the Resource Description Framework (RDF), which is a model for storing metadata created by the World Wide Web Consortium (W3C).

One person who's bullish on the prospects for semantic data lakes is Shawn Dolley, Cloudera's big data expert for the health and life sciences market. Dolley says semantic technology is on the cusp of breaking out and being heavily adopted, particularly among healthcare providers and pharmaceutical companies.

"I have yet to speak with a large pharmaceutical company where there's not a small group of IT folks who are working on the

open Web and are evaluating different technologies to do that," Dolley says. "These are visionaries who are looking five years out, and saying we're entering a world where the only way for us to scale….is to not store it internally. Even with Hadoop, the data sizes are going to be too massive, so we need to learn and think about how to federate queries."

By storing healthcare and pharmaceutical data as semantic triples using graph databases such as Franz's AllegroGraph, it can dramatically lower the hurdles to accessing huge stores of data stored externally. "Usually the primary use case that I see for AllegroGraph is creating a data fabric or a data ecosystem where they don't have to pull the data internally," Dolley tells *Datanami*. "They can do seamless queries out to data and curate it as it sits, and that's quite appealing."

Cloudera and Franz today announced that AllegroGraph has been certified to run atop Cloudera's Distribution of Hadoop (CDH). The companies see the software becoming a reference architecture for creating semantic data lakes from big connected data. Healthcare providers and pharmaceutical companies are expected to be the biggest adopters, but the partners also see the approach being adopted in the financial services, intelligence/national security, and publishing sectors.

Succeeding at cognitive computing demands that we tackle this big data integration problem, according to Franz CEO Jans Aasman. "We need to combine unstructured data with structured data to fuel real-time analysis, predictive analytics, and deep learning," Aasman says in a press release.

"But the ease of data integration largely depends on the type of database. With the semantic flexibility of AllegroGraph, integrating databases is a virtually effortless, since the data can remain in its original databases and database designers do not have to create a schema up front. This capability is particularly valuable if organizations want to

tap into the growing number of public datasets to enrich their analytics," he continues.

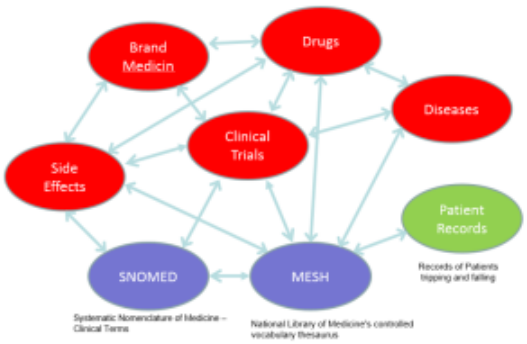
This is leading-edge stuff, and there are few mission-critical deployments of semantic graph technologies being used in the real world. However, there are a few of them, and the one that keeps popping up is the one at Montefiore Health System in New York City.

Montefiore is turning heads in the healthcare IT space because it was the first hospital to construct a "longitudinally integrated, semantically enriched" big data analytic infrastructure in support of "next-generation learning healthcare systems and precision medicine," according to Franz, which supplied the graph database at the heart of the health data lake. Cloudera's free version of Hadoop provided the distributed architecture for Montefiore's semantic data lake (SDL), while other components and services were provided by tech big wigs Intel (NASDAQ: INTC) and Cisco Systems (NASDAQ: CSCO).

This approach to building an SDL will bring about big improvements in healthcare, says Dr. Parsa Mirhaji MD. PhD., the director of clinical research informatics at Einstein College of Medicine and Montefiore Health System.

"Our ability to conduct real-time analysis over new combinations of data, to compare results across multiple analyses, and to engage patients, practitioners and researchers as equal partners in big-data analytics and decision support will fuel discoveries, significantly improve efficiencies, personalize care, and ultimately save lives," Dr. Mirhaji says in a press release.

Of course, AllegroGraph is not the only graph database on the market, and it's not even the only one that runs on Hadoop. But to Cloudera's way of thinking, it was a natural to team up with Franz because of its presence in the marketplace.



How a semantic data lake may connect different data sets

“The beauty of Franz AllegroGraph is they’ve been in the marketplace for so long an

d
ha
ve
be
en
do
in
g
th
is
fo
r
so
lo
ng
th
at
th
e
un
de
rl
yi
ng
fo
un
da
ti
on
of
th
e
te
ch
no
lo
gy

is very solid and it’s very simple,” Cloudera’s Dolley says. “Franz has been an early

y
in
no
va
to
r.
”

Similarly, the folks at Franz are happy to team up with Cloudera, which is a leader in the emerging Hadoop software ecosystem. According to Dolley, Franz’s customers benefit from having a secure, scalable, and flexible platform upon which to build SDLs that are composed of data pulled from internal as well as external sources.

The flexibility of Hadoop is key when tapping external data sets as part of a virtual SDL, Dolley says. “From the data architect’s perspective, using Franz on Hadoop means I get an environment that’s lower cost, more salable, more secure, and fits with the schema on read proclivities of data that did not originate from my own mainframes,” he says.