# Datanami article – Hadoop, Triple Stores, and the Semantic Data Lake

Hadoop-based data lakes are springing up all over the place as organizations seek low-cost repositories for storing huge mounds of semi-structured data. But when it comes to analyzing that data, some organizations are finding the going tougher than expected. One solution to the dilemma may be found in Hadoop-resident graph databases and the notion of the semantic data lake.

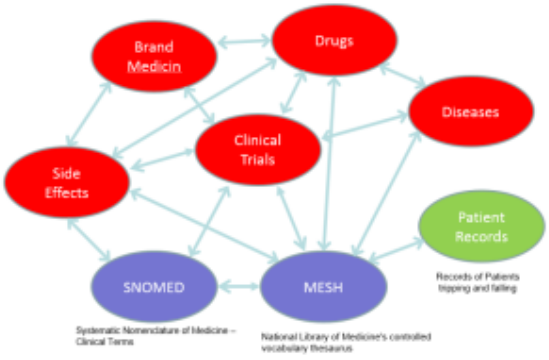
Despite their growing popularity, data lakes have taken a bit of heat lately as analyst firms like Gartner call into question their long-term viability. Without a way to organize the schemaless data that people are shunting into Hadoop en masse, the data lakes risk becomes convoluted quagmires, where data goes in and nothing useful comes out.

Franz Dr. Jans Aasman, president and CEO of graph database developer Franz, knows a thing or two about organizing and querying schemaless data. “Gartner wrote a piece about it, saying data lakes are great but you have to make sure they’re semantically consistent,” Aasman tells *Datanami*. “So we took it a step further and created what we call the sematic data lake.”

Semantic data lakes, as Aasman envisions them, build on the structure that a graph database can provide to unstructured data. Franz’s graph database, called AllegroGraph, stores data

as “quad-tuple” that adhere to Resource Description Framework (RDF) standards and can be queried in SPARQL, a SQL-like language that the W3C recognizes as the standard language for semantic graph database.

“What we’re investing most of our time in now is the semantic data lake, where we store data in a key value store in Hadoop [Hbase], but then index it with our graph database so that we can do these SPARQL queries,” Aasman says. “In general, Hadoop is the ultimate in scalability but no one has yet succeed in putting really complex data into Hadoop. It’s usually simple data, and everybody is working hard to make it easier to put complex data into Hadoop.”




***A semantic data lake opens Hadoop analytics to external linked data***

While banks and intelligence agencies have been early adopters of Franz’s graph database, Aasman sees healthcare companies being the first adopters of his vision for the semantic data lake. Most healthcare firms today rely on data marts when they want to analyze data, but this approach is limited due to the time it takes to set up the data mart, the cost, and the resulting isolation of the data and the analytics that can run there.

But with a semantic data lake running atop Hadoop and powered with an RDF graph database, healthcare companies can begin analyzing and finding useful connections hidden amid huge amounts of data, Aasman says. Not only can healthcare companies analyze electronic medical records (EMRs), financial data, and unstructured data such as doctor’s notes, but thanks to the standardization of terminologies within the healthcare industry, they can easily add so-called Linked Open Data

sources into the mix.

The hospital world uses six different vocabulary systems, according to Aasman. That means, within a given hospital, the words and terminologies they use is consistent, but it's a crap shoot as to whther a neighboring hospital will speak the same language. "It's hell because everybody uses different standards," he says.

Franz did the hard work of linking all those terminologies as triples, thereby providing real-time translation services, if you will, among the six different standards. That opens up a whole new world of health analytic capabilities, which Franz recently demonstrated at the HIMSS 2015 conference with its partners, Cisco and Intel.

"Now we can take any hospital standard, and as long as we use one of the six standards, we can link it in," he says. "So when we talk about a particular failure of the left ventricle of the heart….independent of what hospital your data comes from, we talk always about the same thing, because we unify the terminology."

The capability to link common data sets is critical to the idea of the semantic data lake, because often the data that will make the biggest impact on your analyses is external to your organization. One of Franz's customers, the pharmaceutical giant Pfizer, used AllegroGraph to create a "life science cloud" (or a semantic data lake, if you will) that enabled the company to link internal lab data to external data sets, including EMRs sourced from UK's National Health Service via MESH and SNOMED.

During a demo, Aasman showed *Datanami* how a researcher is able to explore linked data residing in multiple internal and external data sets, and to do so in real time. For example, a researcher seeking to gain insight into the likelihood of patients suffering nasty falls while taking an anti-depressant

medicine, for example, can easy execute that query across many linked data sources using Franz's fat client interface.


***Bloor Research puts AllegroGraph in the "champion" sector for graph databases***

That sort of query would be quite difficult to pull off using traditional relational databases, ETL jobs, and data marts. But with an RDF store like AllegroGraph finding structures hidden within disparate data sets, it's quite easy.

AllegroGraph isn't open source and it's not free, but it may be one of the most under-appreciated parts of the Hadoop stack. While there are other graph databases available on Hadoop, such as Titan and Giraph and Apache Spark's GraphX, none of them are semantic graphs, and therefore none of them deliver on the promise of a semantic data lake, according to Aasman.

The main benefit of a semantic graph databases is that it's designed to share data, Aasman says. "So if multiple people make their own databases, as long as they make sure that they use the same names for the same things, then all these databases can be automatically linked together in a very straightforward way," he says.

The other main type of graph database is a property graph database. According to Aasman, semantic graph databases maintain certain advantages over their property graph database

cousins due to the several factors, including their to the RDF framework and use of the SPARQL language.

***Franz CEO and president Dr. Jans Aasman***

“If you want to use property graph databases, you have to re-invent the wheel every time,” Aasman says. “If you wanted to use Titan, you have a huge problem because Titan has a fixed data model. This world is so complex—you can’t write a fix data model for this kind of data. It’s very complicated…Dumb graphs don’t give you knowledge. You need a little bit more.”

As more open data goes online, ensuring that the data is readily accessible and “clickable” will make that data much more useful for wider audiences. “We take having a semantic graph very seriously because we have to work with all these ontologies and all these external data sets that are available as triples,” Aasman. “We’re set up to make that easy.”

Franz seems well-positioned to take the semantic data lake idea further. While the 31-year-old Oakland, California, firm is no startup (see its website for a colorful history of life in the artificial intelligence business in the 1970s and 1980s), Franz appears to have some of the tools that people—especially those in healthcare—are clamoring for to get in front of the big data challenge.