# Datanami article – Medical Insight Set to Flow from Semantic Data Lakes

The potential for data analytics to disrupt healthcare delivery is large, and getting larger by the day. But in many cases, the need to hammer data into a structured format creates a barrier to productivity. Now a hospital chain in New York City is hoping to change that by adopting a Hadoop-based semantic data lake.

Located in the Bronx, Montefiore Health System is the first hospital to implement a semantic data lake as part of the New York City Clinical Data Research Network (NYC-CDRN), an association of seven hospitals in the NYC area that are sharing data. As the pioneer, Montefiore is working with several technology providers, including Intel and Franz, to test a big data system capable of delivering precision medicine.

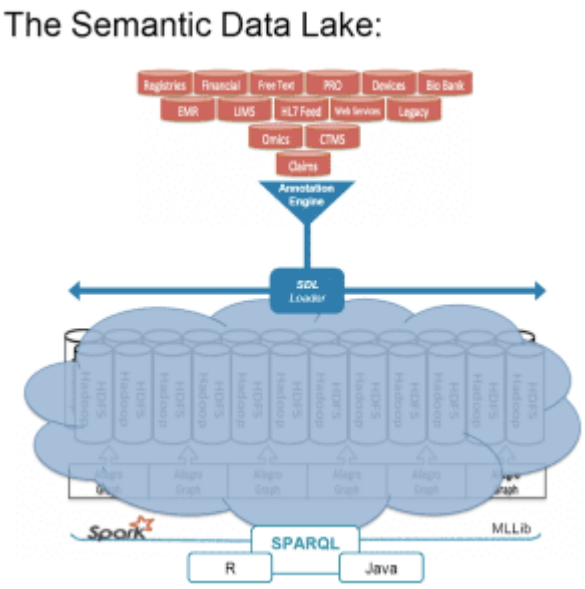
Most *Datanami* readers are familiar with the term “data lake,” but a “semantic data lake” provides an interesting twist on the familiar concept. According to Franz CEO Jans Aasman, a semantic data lake employs a combination of technologies, including Hadoop, graph analytics, a semantic “triple store,” the SPARQL query language, and Spark-based machine learning, to allow doctors to connect the dots between patient conditions and a world of knowledge contained in structured internal systems, as well as unstructured data sources outside

of the organization.

## Medical Analytics—Not Just Semantics

The idea behind the semantic data lake is to leapfrog traditional healthcare analytics, which today is hamstrung by relational technologies, Aasman says.

“For each question they had, they built a data mart,” he says. “The problem is, these data marts are usually silos. It’s kind of hard to combine the data from one data mart to another data mart. The other problem is it can be very expensive.”




Aasman and his colleagues at Franz think the semantic data lake approach provides a more powerful and elegant solution. Instead of hammering the data into a relational schema and storing it in a data mart, they want to pull all the pertinent data—including any “linked open data” such as drug interaction databases, genetic test results, or a demographics database sorted by ZIP code—into Hadoop and HDFS.

The semantic magic starts once the data is in Hadoop and HBase. Montefiore transforms all the data into semantic triples using a special ETL tool called a semantic annotation engine. Then the Franz AllegroGraph graph databases indexes all the triples and powers the SPARQL queries across the joined corpus, in addition to machine learning powered by Apache Spark ML or R.

As a result of this semantic approach, users can run queries that traverse different data sources, which was the stumbling block of traditional relational approaches. Because

MOntefiore has taken the time to build a unified clinical event model that maps all the different medical coding terminologies used—not to mention coding all the possible events that can occur in a hospital (there are about 350 of them, it turns out)—knowledge can be extracted by writing queries in a relatively simple and straightforward manner that follows the Resource Descprition Format's (RDF)'s "subject->predicate->object" approach.

"The hard part is, how do you take data form all these databases and put them into a unified event model?' and Montefiore solved that problem," Aasman says. "Think of it like a huge knowledgebase about diseases and proteins and genes and everything having to do about human physiology. We can take all these things in, and without even modifying them, load it into our system. This is a huge thing. There are other healthcare analytical companies that take these external database and put them into relational database schemas. But they lose all the power of reasoning and going higher and deep or lower into the taxonomy chains."

## Real World Impact

While data scientists may do the heavy data lifting behind the scenes to align the datasets for maximum impact, it will be the doctors and other medical practitioners who will put the insights into action. But patients are the ultimate beneficiaries.

"Just imagine that I could compute the likelihood of another symptom, and feed it back into the database," Aasman says. "The doctor opens up the electronic medical record, and one tab on system says, 'For this particular patient, this is most

likely thing that will happen to you in the future,' and the doctor can start planning for this."

The system could also be used to monitor for potential drug interactions when doctors are writing new prescriptions. "It would be really nice if the database warned you instantly. 'Hey this is a drug interaction,'" Aasman says. "Or it may say 'Wait a second, this is not officially in the drug database, but here's a clinical trial where people talk about potential side effect. We have integrated this in the same way the vocabulary system [was implemented]…The external databases are triples, so it's all the same data format."

Dr. Parsa Mirhaji, the Director of Clinical Research Informatics at the Montefiore Medical Center, discussed the work his team is doing during the recent HiMMS conference. "We have projects that are trying to build predictive models for patients that we anticipate will need [to go to the] ICU in the next 72 hours," Dr. Mirhaji says. "We really want to know who's going to be intubated in the next three days and preempt that."

Successfully preempting stays in the intensive care unit requires a lot of data, he adds. "Some of that is more static data than what you can get from labs and EMRs. Some come from devices you attach to the patient—monitors and devices in your clinical setting, which produce massive amount of information. Not only do you need to bring that in and integrate it but [you need to] create known alerts and notifications that you can distribute correctly among participants of the care and maintain that for future research."

## Looking Ahead

The Franz CEO agrees that technology can give us the upper hand in the medical field. "This is the whole promise of big data in healthcare," Aasman says. "Look at all the investments

in healthcare analytics last year and this year. IBM is spending billions of dollars. Everybody is spending lots of money, because with big data, you now can discover things that you could never, ever discover before about relationships."

The potential for healthcare analytics becomes richer when you bring other factors into the equation, such as a patient's race, gender, age, genetic makeup, physical health, and even their ZIP code. "And then link it to all the other people who have the same makeup—that's clearly something that doctors don't know yet, and that's the whole promise of healthcare analytics," he says.

Once the semantic data lake has been fully implemented at Montefiore Medical Center, the plan calls for the other six healthcare organizations in the NYC-CDRN to adopt it too. At that point, the lake will contain data on 2.5 million patients.