Dataversity article – Data Lakes Get Smart with Semantic Graph Models

by Jennifer Zaino

Been swimming in a data lake recently? Perhaps not, as many companies still are just dipping their toes into these waters, as they become more familiar with the general idea of a data lake.

As research firm Gartner describes it, a data lake is:

“a collection of storage instances of various data assets additional to the originating data sources…[and whose purpose is to] present an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse).”

Even as enterprises consider the returns they may expect from diving deeper into data lakes, they’re now being exposed to another twist on the concept. Enter the smart data lake, also known as the semantic data lake. At DATAVERSITY’s® Smart Data Conference in San Jose in August, the issue came up in sessions including presentations by Cambridge Semantics CTO Sean Martin and Franz CEO Jans Aasman. Franz, in fact, notes

that it has copy written the term Semantic Data Lake, and points out that Gartner also has explained that data lakes need semantics in order to be usable by a broad set of users.

Dive into Smart / Semantic Data Lakes

What's the smart/semantic data lake all about? According to Cambridge VP of Solutions and Pre-Sales Ben Szekely, such a data lake is distinguished by its use of the flexibility and power of semantic graph models as the format for storing data at rest, front to back within the data lake.

The key benefit, he says, is how it improves upon the consumption of structured, semi-structured, and unstructured data for the average business analyst or other end user in search of critical insights. By defining and contextually connecting information in terms of the semantic graph model, data is immediately available and useable for discovery and analytics, he explains.

"You are taking data from different sources and establishing a well-defined linked graph of information to be analyzed and explored immediately by end users for quick value out of a data lake in ways that couldn't happen before," Szekely says. With traditional data lakes, little is done around preparing ingested data. Rather, diverse data sets in central repositories are left in a more or less untransformed state, so that it is not easy to join together different types of information for analysis.

Often as a result, business users or analysts wind up dealing with Hadoop-type de facto data silos, and are only able to gain insight in the context of exploring one at a time, he says. And, whenever they do want to make queries across the diversity of their data stores, they have to go to highly skilled experts to clean up, and manually link and prepare the data each time for each individual query job. "If you have to spend a lot of time trying to do that, you can't capture the

full value out of the data," he says – at least, not in a timely manner.

"The time to value of that approach is too long," he says, and users don't get the insights from those data lakes as fast as they want to.

That's a problem to industries that are today's prime candidates for leveraging data lakes – pharmaceuticals, life sciences, financials, and intelligence high among them. In those explosively information-laden sectors, the faster and easier it is to get analytics value out of data lakes, the better the chance an organization has of gaining a competitive advantage, saving money, or even pinpointing national security risks.

As an example, Cambridge VP of marketing John Rueter asks that you consider the fact that pharmaceuticals research and development organizations can spend more than $2 billion and more than a decade working to get a drug to market. If those researchers had access to smart/semantic data lake technology that contextually brings together more data from internal and external sources (online medical journals and so on), they could more seamlessly and expediently ask questions of this wider array of data than they could have before.

The resulting analysis could lead them to discover earlier on in a process that they'd be wise to accelerate or perhaps even abandon an effort to pursue developing a drug in a certain category. "The ramifications from a cost perspective could be quite significant," Rueter says.

Options in the Smart/Semantic Data Lake Space

Cambridge's Anzo Smart Data Lake takes the approach of enabling data flexibility from the start, by spending some time upfront linking and contextualizing data in a democratized way by end users, Szekely says, with proper governance.

It includes tools for loading data into Hadoop HDFS from any data source – structured or unstructured – into RDF graph models and transforming those sources for consumption; semantic models for describing the data in common business terms and linking and contextualizing the data across big data sets; and, other rules, methods and services to create a unified graph of different data sources that can be deployed at scale in a Cloud environment. For instance, a catalogue capability exists for tracking all the data. It also accommodates existing data lake environments, transforming their data into a smart data lake graph environment, Cambridge says.

Graph-aware front-end tools make it possible to search and discover data inside the smart data lake, subject to established access controls (as opposed to traditional Business Intelligence tools sitting on top of a data lake). Large sets of data can be brought up and queried in-memory thanks to a massively parallel, clustered in-memory SPARQL engine, called the Anzo Graph Query Engine, Szekely says. There's no need, however, for users to understand how to write SPARQL queries to navigate through the data.

Cambridge is working in partnerships with companies such as data providers in the pharmaceuticals and life sciences sector, integrating their information into smart data lakes it builds for some customers.

Semantic graph technology vendor Franz also is invested in the smart/semantic data lake space – and in ensuring that users can leverage its power without having to understand how to write SPARQL queries. "SPARQL is a pretty cool language, but it's difficult to write very complex SPARQL queries," says president and CEO Jans Aasman. That's why Franz created visual SPARQL, which enables users to graphically build queries starting with the data.

In collaboration with Montefiore Medical

Center, Intel, Cisco and Cloudera, Franz now has created a Semantic Data Lake platform initially for the U.S. hospital/health care vertical, which has to deal with the trend towards Accountable Care. The core of Accountable Care is that hospitals are being paid based on quality rather than quantity of treatment. That includes getting bonuses for keeping treatment costs below industry averages and being hit with penalties if those costs rise above those averages.

Aasman says that these organizations had a Big Data Science problem when it came to the Accountable Care issue. They have a lot of widely dispersed data in different formats and were having difficulties analyzing it to see how well they were performing on various government metrics for Accountable Care success, so that they would know where they needed to make improvements. Building data marts for each metric would be impractical and very expensive, he says, given the diverse data formats they must deal with.

The Semantic Data Lake platform it has created with its partners is based on AllegroGraph, Franz's Semantic Graph Database. Franz's Semantic Data Lake platform makes it possible for healthcare facilities to take their medical records and financial data and put it in Hadoop, together with every other relevant data set from within the hospital or outside of the hospital. Other internal data sets, Aasman says, can include tissue databases, clinical trial databases and the output of all sorts of medical devices. External data sets can include health exchanges, genome databases, drug interaction databases, and more. He adds that the Semantic Data Lake platform also stores a unified terminology system that combines more than eight different vocabularies and taxonomies used in healthcare, such as ICD9, ICD10, Mesh, Snomed, RxNorm, and UMLS.

All the data assets are stored as an RDF graph in Hadoop. This same graph is indexed with AllegroGraph so that data scientists can use other W3C mechanisms such as SPARQL to

query the semantic graph, Aasman explains.

With all this in place, “we can do computations of each of the 31 metrics the government uses to judge how well hospitals are supporting [Accountable Care standards],” he says, querying them via the visual SPARQL interface.

In addition to making it easy to handle complicated SPARQL queries in an easy manner in AllegroGraph and thereby in the Semantic Data Lake platform, the company also has built features into AllegroGraph that support capabilities that vendors are empowered to enable via the SPARQL 1.1 standard. For instance, SPARQL 1.1 “allows computed properties and it is up to the vendor to create certain computed properties,” Aasman says. A computed property (also known as property functions or magic properties) is a predicate that occurs in a SPARQL query pattern with matching semantics other than simple subgraph matching. The magic property capability has enabled Franz to build N-dimensional analysis into AllegroGraph. N-dimensional analysis refers to putting multiple dimensions into a single value, and AllegroGraph indexes those dimensions in a patented way that makes it possible to compute a query without conducting multiple joins among many values.

Regarding Franz’s Semantic Data Lake platform, hospitals can continue adding their own data as needed, or start their own Semantic Data Lake for other requirements. “We give them a set of APIs to make it easy to do analytics and templates on how to do analytics,” Aasman says. As a platform play, he sees opportunities in leveraging the underlying technology to other relevant segments, such as financials and life sciences.