HealthITAnalytics article – Semantic Big Data Lakes Can Support Better Population Health

September 08, 2015 – The healthcare industry is currently holding its breath as it waits to see how CMS will adjust the final Stage 3 meaningful use rule in response to stakeholder complaints about the maturity of EHRs and big data analytics technologies, but not all organizations are standing stock still during this waiting period.

As healthcare providers navigate the treacherous transitional waters of Stage 2 and try to predict how future regulations will shape their actions, the need to lay the groundwork for advanced population health management and accountable care is only becoming clearer.

No matter what the outcome of debates about the future course of the EHR Incentive Programs, one thing remains abundantly clear for organizations of all shapes and sizes: advancements in healthcare big data analytics will not be driven solely by rules and mandates, but by the pressing financial need to

collect, corral, understand, and leverage information in order to refine and expand population health management techniques.

Developing the underlying architecture for value-based reimbursement, namely a strong framework for population health management, data governance, and big data analytics, is becoming a top priority for a growing number of providers looking to get a head start on the new realities of healthcare reform.

These organizations, like Montefiore Medical Center, are looking for cutting edge analytics tools which won't just help them meet the clinical and financial stresses of today's environment, but will also prepare them for the uncertain paths ahead.

In order to position Montefiore for success in the unpredictable future, Parsa Mirhaji, MD, PhD, Associate Professor of Systems and Computational Biology and the Director of Clinical Research Informatics at the Albert Einstein College of Medicine and Montefiore Medical Center-Institute for Clinical Translational Research, turned to one of the most promising developments in big data analytics: semantic data lakes.

A few steps beyond the traditional data warehouse, which requires a relatively narrow set of parameters in order to accept, compare, and retrieve information, semantic data lakes allow for unprecedented flexibility.

Data lakes, built using graph database technology, may soon allow clinicians to access sophisticated clinical decision support using natural language queries, thanks to their unique ability to synthesize and normalize disparate datasets and draw conclusions from seemingly unrelated pieces of information.

The potential for improving the quality of patient care is enormous, Mirhaji said in an interview

with *HealthITAnalytics.com*, as the technology evolves to meet the full spectrum of as-yet-untold demands for detailed risk stratification, predictive analytics, and patient safety.

"From the standpoint of an accountable care organization (ACO), where we really need to cover the full spectrum of health data, we need to capture and represent everything that will give clinicians a precise understanding of an individual's care and wellness," Mirhaji said. "On one hand, that includes diving into clinical genetics, molecular medicine, and biomarkers."

"On the other hand, patients interact with their environments and with each other in a community setting, which makes it very important to look at population health management and community care at the other end of the spectrum. For an ACO, it's all about the coordination of care within different communities."

But healthcare providers cannot look to coordinate care in the community if they do not have an organized method for keeping their own house in order. From EHRs to research results to financial data and patient demographics, big data is everywhere in the typical healthcare organization, and each type of data may be locked into its own individualized analytics architecture.

Not only is it expensive and time-consuming to craft separate infrastructures for each category of information, but it prevents data scientists from drawing actionable insights from cross-pollinated datasets.

"We don't have the time and resources to build silos and specialized systems for specific needs," Mirhaji said. "So we did a rigorous analysis of the use cases we need to cover, starting from personalized precision medicine and moving all the way up to population health management."

"We asked ourselves what functional competencies and technical

competencies we need in order to support all of these. Where do we need to make investments, and what are the properties of the technologies that we need to support more than one use-case at a time? A lot of our problems could only be solved by graph databases. Other technologies we looked into were not really prepared to address this long spectrum of requirements."

A standard relational database can help many organizations meet a number of their goals, Mirhaji says, but they include some inherent limitations. "Relational databases require a very fine structure that you have to plan out before you can use it – you have to frame your problems in a very specific way," he explained. Within that frame, you can do wonderful things, but you have to pre-coordinate your schema before you start investing in application development and data management."

"The problem with that is that you have to predict all future-use cases," he continued. "And the costs of changing your mind or your requirements are huge. And that's why you end up with these data silos. You end up with different architectures for different problems, because you have to box the problem before you begin."

In contrast, flexibility and adaptability are built into the fabric of graph databases, which use cognitive computing techniques to help draw connections across datasets that may be vastly different in size, detail, or scope.

"You don't have to predict the future," says Mirhaji. "You can start from where you are, from exactly where you are, based on the kinds of needs that you have right now with the confidence that it will grow into the dimensions and directions as your organization wants to grow."

What allows graph databases to operate with such a high level of fluidity? It's the way that data points are identified,

codified, and linked within the system, explained Dr. Jans Aasman, CEO of Franz, Inc., which has worked with Montefiore to develop its big data capabilities.

Semantic data lakes are structured in a fundamentally different way than relational databases. In a semantic graph system, each element is given a standardized, unique identifier, which allows the database to link separate concepts and generate complex insights the way a human brain does.

“You can, for example, say, ‘Jans is a person,’” Aasman said. “That statement includes two things: a discrete person and a discrete concept. ‘Jans,’ the person, links to ‘man,’ the concept. Now we can say, ‘Jans lives in San Francisco.’ We know San Francisco is a place that is part of California. California is a State that is part of the USA, which is a Country. Now your system knows that ‘Jans lives in State of California in the USA’ without you ever explicitly saying it, or pre-coordinating for it.”

“We can also say, ‘Jans works on semantic data lakes.’ So now I have about five different links, which we call ‘triples,’” he continued. “Now I can say, ‘Parsa Mirhaji is a man. He lives in New York City, which is part of New York State, which is part of the USA. He works on semantic data lakes.’”


Jans Aasman


Parsa Mirhaji

“So without ever directly describing the schema, we can add these triples, or statements, to the database, which will index it all in the most optimal way. Now we can ask more complex queries, such as ‘give me all the cities in which people work on semantic data lakes.’”

“From a natural language point of view, we can all see in our heads a picture of how the nodes were connected, starting with

‘Jans’ and moving through the connection points until we get to ‘Parsa,’” Aasman said. “These concepts are all linked together. That’s just one small example of using nodes and triples to create a graph.”

But that is only the beginning of what this type of technology can do, Aasman says. “The important thing about semantic graphs is that every node is not just a simple word, but it’s actually a Unique Resource Identifier (URI) that can globally identify and contain data on a whole concept.”

“For example, when we talk about aspirin, we have a standardized URI for the concept of ‘aspirin.’ We might want to use a clinical trial database that talks about aspirin, and a disease database, and a side effect database – we will all use the same URI for the same concept across all those datasets.”

“That means we can load up all these databases into a single location, and suddenly we can start making connections across disparate sources because they all took care to describe ‘aspirin’ as a standardized element shared between them.”

New data elements can be added and curated to enrich existing datasets or create more capabilities, allowing an organization to include previously unavailable metadata or load fresh data sources into the system to generate more sophisticated insights. The database grows organically, Mirhaji says, allowing providers to shift their perspective or focus on developing highly detailed data in a specific area depending on the necessary use cases.

“A typical relational database is like planning a city,” Mirhaji said. “Before you can build anything, you really have to pre-coordinate every single street, every single building; all the pipelines for water and electricity. A graph database is more like a forest. No one plans a forest. No one has control over how different trees will grow. Plants are big

and small and different from each other, but they do all follow the same basic framework: there is a stem and leaves and roots that may have connections to each other."

"Each tree can have very, very detailed structure of their own, but as in a graph database, they are all connected. And you can use this interconnection to find exactly what you want without having to pre-coordinate everything. As long as every plant follows the same basic shape, every trunk can give birth to branches, and in as needed basis and just in time, branches grow other branches, and so forth."

These capabilities can be especially important for organizations that may handle very complex cases and rare diseases, or simply experience something out of the ordinary that might initially give a physician pause. When it comes to patient safety issues involving medications, graph databases can help providers take a precision approach to solving an unusual adverse reaction or forestalling a large-scale event originating from a specific manufacturer or dosage problem.

"You might have a code that tells you what medication was given to a specific patient," Mirhaji conjectured. "That code can also be mapped to the National Drug Codes, which will tell you who made that medication and the specific formulation that went into it."

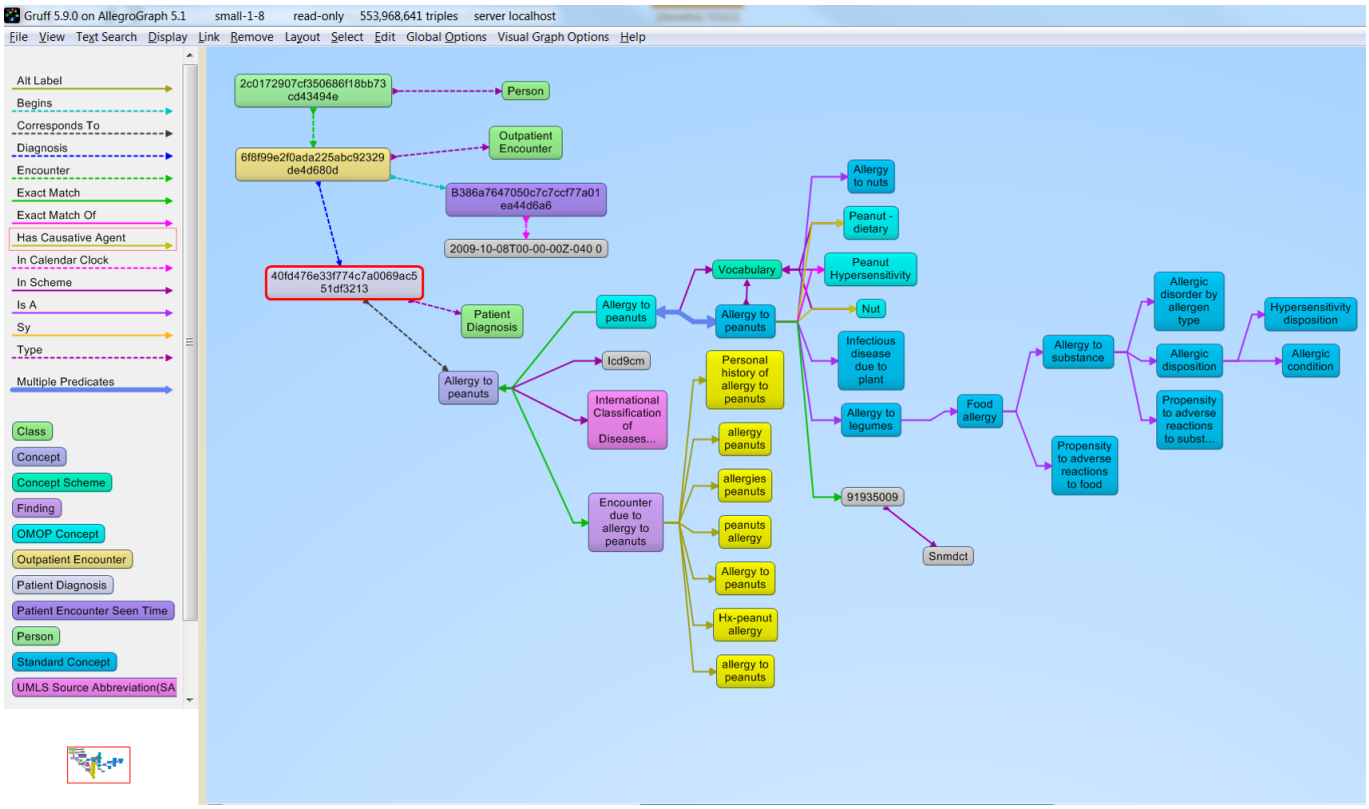
"And then there are pharmaceutical knowledge bases that have the chemical structure of those medications. And then there are databases at the FDA that have data on known complications or the ingredients in those medications. Then there are clinical trials that are actually doing observations and trials on the effects of those medications or ingredients on different diseases and clinical problems."

"Now, imagine that all of these datasets are different on their own," he said. "Each is its own type of tree, as it were. Different people with different backgrounds are

building them. You can't predict how each developer is going to make changes."

"But you want to be able to combine them to ask a complicated question, such as, 'How many patients currently admitted to the hospital have been given a medication that contains ingredients tested by a clinical trial that may produce this specific complication, and what company is marketing the formulation used by those patients?'"

"So now we have combined five different databases together to produce a query that resembles very much like how a clinician thinks," added Mirhaji. "As a clinician, I may think that I have given my patient a medication that might explain why he has a new rash on his face; and I may wonder if this has a precedence in a clinical trial. I may want to know if all formulations of the medicine in the market have this complication or they happen more or less in products of a certain company."



Researchers have not yet refined the ability to input such a query into a simple text-based interface linked to a clinical

decision support window in an EHR. But when the capability is fully developed, clinicians will have a powerful ability to get answers to on-the-fly questions that require little in the way of specialized knowledge about how the database works.

"As a clinician, I probably don't want to know the schema behind this data," Mirhaji says. "I just want to know if there is a clinical trial somewhere that addresses the potential impact of some ingredient in a medication that might be producing an adverse effect."

"I want a machine to do it for me. I don't want my clinicians to have to know and think about whether the data exists and in what format and how exactly if it can be connected to some other piece of data, before they can even start asking a question. I want my clinicians to think freely and get the answers they need."

Before that can happen, however, graph database technology needs to address some of its biggest challenges. Standardizing elements across different healthcare terminologies, such as ICD-10, LOINC, CPT, and SNOMED, is problematic for most health IT applications, even the most advanced EHR systems. The same basic interoperability concerns apply to graph databases, Aasman and Mirhaji acknowledged.

Additionally, in order to use a semantic data lake for meaningful population health management, users must also be able to get answers to queries that may include multiple events that take place at different times, in specific locations, or in certain sequences.

The system must be able to incorporate temporal reasoning that arranges events in relation to one another, Mirhaji explains. "Some events happen within the other, one after the other, or overlap in duration of the other. There are meaningful inferences to be made if you know how exactly these events are

temporally arranged."

"Geospatial relationships are another thing that require a specialized approach," he added. "Combining time and space is especially important for community-based population health management. Behavioral data, mobility questions, and health disparities all require tracking where and when an event takes place. We cultivate our data in such a way that it knows how to account for the temporal and geospatial relationships between events."

Much of this cultivation happens at the hands of dedicated human curators, who help to ensure that information entering the system is complete and accurate, as well as detailed and uniform enough to mesh with existing datasets.

"We have developed a methodology for curating and harmonizing all of this data," says Mirhaji. "It's a process and a tool that does almost 60 to 65 percent of the job when it comes to connecting and identifying the elements, capturing and mapping their terminologies, and resolving their redundancies. One of the big problems we face is the issue of making sure that we have a reliable master patient index, and that we can accurately identify events by patient."

"The tool, which we call the 'interrogator,' helps us annotate the data with many different meta-elements. That helps us deal with the quality aspect of the data. When you start creating such a big data forest, you really want to make sure that you can pull only relevant data – not all related data – to answer a specific query. So we have created quality measures for our data, and we can annotate that, as well."

"We need human curators to describe data in terms of what the data means at the source. And that's all they do. They do not define structure; they do not map to a schema. They just describe data as it is at the source: the data has these timestamps; it has these controls; this quality; these are the

people that can access this data. Our curators do all of these annotations, and then once it's done, the future imports will be almost automatic."

As cognitive computing tools evolve and become increasingly sophisticated, organizations like Montefiore hope to closely integrate natural language query capabilities into everyday clinical care. The possibilities for extracting actionable population health management and precision medicine insights from this vast and malleable treasure trove of big data are nearly infinite – if developers and data scientists find intuitive and user-friendly ways of deploying their tools into clinical workflows.

Data forests will continue to grow in complexity and usefulness as developers pursue the most impactful methods of using big data analytics to further strategic goals like accountable care and population health. With personalized care, predictive analytics, and tailored insights on the wish-list for the majority of healthcare providers, semantic graph databases may provide an intriguing avenue forward into the uncharted waters of data-driven, quality care.