

Will Graph Database Technology Uncover Pharma's Hidden Insights?

Neo Technology's Emil Eifrem looks at how life science researchers can probe large datasets efficiently and expose new insights with the unique power of graph technology

By definition, medical research is about dealing with large quantities of data. That's even more true in the leading edges of the life sciences where tackling the thorny issues in genomics and personalised treatments has to take place at the petabyte and increasingly the exabyte, rather the mere megabyte and gigabyte level, in terms of promising dataset size.

With Big Data one also has to assume that's going to be less structured information. So Big Data life science research demands we go beyond the simplistic managing, analysing and storage of data – which means we have to take a hard look at the tools we've historically used to model data, namely SQL and relational database technology.

That's because traditional relational database methods struggle with the sheer volume, as well as the unstructured nature, of the data we want to work on. But a realistic answer is emerging in the shape of graph databases. An approach opened up by social web giants Google, LinkedIn and Facebook, graph database technology is an approach based on the insight that when it comes to data, it's the *relationships* that are all-important and of interest to the researcher.

The power of graph databases recently came to the world's attention due it being behind the world's largest investigative journalism project, The Panama Papers [<https://panamapapers.icij.org>] – a probe of a 2.6 terabyte dataset that would not have been possible using traditional database means, says the global team of researchers that mounted it.

Limitations of SQL

In fact, the Panama Papers is the biggest data-based investigation so far conducted, and the body that led it, the ICIJ, has stated that, "It wasn't until we picked up a graph database that we started to really grasp the potential of the data."

Investigators like the ICIJ are among many who say that graph databases are not only great at handling a lot of data, but are uniquely able to uncover the patterns difficult to detect using SQL-based rdbms (or even other approaches, like NoSQL and Hadoop).

The reason is that the high-volume, highly-linked dataset the Panama Papers represent is too hard for SQL to easily work with. That's an issue of fundamental architecture. Relational databases model the world as a set of tables and columns, carrying out complex joins and self-joins to action queries when the data becomes more inter-related. But the interesting questions we need to ask in areas like the life sciences are technically challenging to construct and expensive to run in this form, while making them work in synchronous time is a challenge, with performance degrading as data volume increases.

The relational data model also doesn't match our mental visualisation of the application (technically defined as "object-relational impedance mismatch"). As people, we instinctively like to model connections between data elements visually, creating an intuitive model as we whiteboard a problem – so attempting to take a data model based on relationships and forcing it into a tabular framework, the way a data platform like Oracle asks us to do, creates a needless disconnect that can waste time and potentially miss useful patterns and leads.

Relationships in Data Conceal Possible Insights

By contrast, the USP of graph database technology lies in discovering relationships between data points and understanding them – and at huge scale. That's why it is ideal at allowing the medical researcher to uncover hidden patterns when they are looking at the hard problems out there, like new molecule research and big clinical trial work.

Medical and Pharma Use Cases of Graph Technology Emerging with Greater Rapidity

Tim Williamson, a data scientist at Monsanto whose role focuses on coming up with ways to help the firm get better research inferences out of genomic datasets, is a graph database convert.

He says that his team had previously managed its data problems in a classical, relational way. However, "Every question we would want to ask needed lots of real-time analysis to be run and it would take us seconds to minutes to hours to perform one round of analysis, which doesn't scale."

Monsanto is continually researching the best possible plant varieties and what genetic traits cause them to thrive in different climatic and environmental conditions. There are genetic problems that rely on being able to treat a dataset as an ancestor family tree. Williamson discovered that these family tree datasets naturally have a graph structure – and it was really easy to write graph queries instead.

Now, he says, analysis that used to take minutes or hours took seconds. "This was really cool, because then we could do it across everything; I could ask the same question of several million objects instead of one at a time."

That frees Williamson and his team to make abstractions around important genetic features, such as which plant ancestors continually form the most productive cross-breeds. That means the biotech leader is starting to get much quicker at spotting which plant varieties are the most productive, so the firm can spend its resources researching those rather than less promising options.

Another great graph medical research example is the EU FP7 HOMAGE consortium [www.homage-hf.eu], which is all about early detection and prevention of heart failure. Working with a dataset from 45,000 patients from 22 cohort studies covering patient characteristics, clinical parameters such as medical history, electrocardiograms and biochemical measurements, all this great data is now being connected with existing biomedical knowledge in public databases to create an analysis platform so implicit and explicit

relationships can be more easily exposed and explored. That really helps, as this is complex; the graph database for just one heart failure network analysis platform contains over 130,000 nodes and seven million relationships alone, for instance.

One more use case is provided by Stephan Reiling, Senior Scientist at Novartis Institute for Biomedical Research. Novartis has created a large graph database of all kinds of heterogeneous biological data, which his team is combining with text mining results. Currently with half a billion relationships in the database, with plans to triple this number, Novartis is looking to create one database to better understand biology and how they can use this scientific knowledge to develop the next generation of medicines.

"There is a huge amount of biological data available, along with incredible data sources," he says. "By really bringing all this data together, for the first time we can say, 'I want

to find compounds that are similar to this compound that have annotations about this disease.'

"To have the flexibility to navigate all of these data sources is really powerful."

Discovering the Unknowns at Big Data Scale

As integrating data and knowledge in life sciences involves the modelling of an incomplete and ever-changing model of how our bodies work, we need a better way, then, for modelling this complexity.

And as our knowledge improves, these models continuously need to change – to take just one example, large human DNA sequences we originally thought were just 'junk' turn out to be an important component of our genomes.

So everything is connected – and those connections change depending on context, time and environmental triggers. To be able to deal with this reality, researchers need tools

Sartorius Stedim Biotech introduces new BIOSTAT STR® bioreactor range

- New generation of bioreactors combined with new Flexsafe STR® bags offer a fully scalable, single-use system
- Systems enable significant reduction of bioprocess development timelines and costs

Sartorius Stedim Biotech (SSB), a leading international supplier for the biopharmaceutical industry, announced the launch of its next generation BIOSTAT STR®, a fully scalable, single-use bioreactor family based on a conventional stirred-tank design. This new bioreactor range featuring upgraded hardware and software, as well as a fully integrated, new design of Flexsafe STR® single-use bags, ensures quick and easy bioprocess scale-up of biologics and vaccines.

The BIOSTAT STR® bioreactors are equipped with an improved stainless steel bag holder for userfriendly installation of the single-use Flexsafe STR® bag. The bioreactor series consisting of five systems in different sizes offers working volumes from 12.5 L to 2,000 L. Because the bioreactors are designed with the same geometries as SSB's ambr® 250 mini bioreactor, the industry gold standard for scale-down, linear scale-up and

process transfer from 250 mL to 2,000 L can be achieved in weeks rather than months using this innovative technology platform.

The new bioreactors and the newly introduced single-use Flexsafe STR® bags are a perfect match. The bags with advanced single-use sensor solutions are manufactured from a proprietary, robust multilayer S80 polyethylene film. The formulation of resins and additives for the film are fully characterized, and extrusion process parameters are controlled within established ranges, providing consistent batch-to-batch extractable and leachable profiles. The fully self-contained design of the new bioreactor with its single-use bag prevents product cross-contamination, saving time in set-up, validation, clean-in-place procedures and sterilize-in-place operations.

Besides using bags of the highest purity, the system is designed for efficient oxygen transfer, mixing and CO₂ stripping. These combined features ensure excellent cell culture performance with reproducible high-density growth of even sensitive cell lines. As a result, BIOSTAT STR® bioreactors are ideal for achieving very high cell densities in continuous processes and for safe manufacture of vaccines and recombinant proteins in cGMP environments.

The new bioreactor range provides greater flexibility in bioprocess control and data acquisition as software connectivity has been upgraded to allow integrated control by either BioPAT® MFCS software or commonly used third-party industrial distributed control systems (DCS), such as Emerson DeltaV™ or Siemens SIMATIC PCS7.



Complete BIOSTAT STR® series of single-use bioreactors.

Product News



SSB's new BIOSTAT STR® 500L and single-use Flexsafe STR® bag

Dr. Thorsten Adams, Director of Product Management for Fermentation Technologies at Sartorius Stedim Biotech, stated: "Direct linear scalability is crucial for ensuring the efficiency and costeffectiveness of bioprocess development campaigns. Compared with conventional stainless steel vessels, our next generation BIOSTAT STR® bioreactors in combination with the ambr® 250 technology will help reduce process development timelines significantly. These highly scalable, single-use bioreactors are an intelligent, low-risk bioprocess development solution for use in multi-product facilities, as well as at contract manufacturing organizations, for the production of biologics and vaccines."

Contact:
Sartorius Stedim Biotech GmbH
August-Spindler-Str.11
7079 Göttingen
www.sartorius.com/biostat-str
www.connect-upstream.com

 **sartorius stedim**
biotech

adept at discovering ‘the unknown unknowns’ at mass scale in order to make those serendipitous discoveries and which are able to handle dynamic and constantly evolving data.

Clearly, graph databases are helping life science make real advances. Could it help your life sciences project, too?

Cross-Disciplinary Is Where Science Happens Best

Neo Technology’s Petra Selmer explains how exposure to graph databases changed her career – and put her on a mission to explain the approach’s time-saving and complexity-defeating benefits

I’m a software engineer by training who has worked on life sciences and healthcare-related projects at a medical research centre and a major university. I did a long stint developing software for clinical trials and national medical audits, combining NHS and private healthcare patient datasets, for the Intensive Care National Audit & Research Centre. I also worked at a multi-participant academic project centered at University College London’s Wellcome Trust Centre for Neuroimaging. In both cases, I was tasked with helping on the data management side of the projects.

What I found: medical data is *very* heterogeneous. It can range from cell-level, hugely detailed data to macro-scale disease network tracking, often in the same research. Very often, you want to link the two ends of the spectrum, as that’s where the interesting results tend to lie, but it can be a real challenge to model that.

Biologists and physicists, to name but two fields of scientific endeavour, think they inhabit different worlds, but often they are looking at the

same thing, just from different perspectives. What’s frustrating is the breakthroughs will often come from the interface between the two sides, or multiple sides. Cross-disciplinary working across all scientific fields is the place where the best science happens, many believe.

What hampers this collaboration is that IT is not seen as a useful participant in these conversations. Working out the smartest way to model the complexity discussed and to find ways for multiple actors to find ways to engage with the same information doesn’t happen enough. Scientists love data, but they are a bit casual about the modelling of it, so you may have a lot of Master’s degree students and post docs scrabbling around with Excel and basic approaches to getting the dataset into a viable shape to be looked at. That’s not efficient.

The key thing is that the relationships in data conceal the actual breakthroughs. I think that’s worth elaborating, and it needs recording. Research is about exploring the unknowns. The easier it is to expose that darkness to light, the better the chances of success. Data provenance is vital here; for example – it’s crucial to know exactly where data came from and what might be affecting it, as more and more researchers need to back up their claims through the veracity of their data (and repeatability of their experiments). Meanwhile, visualising complexity, being able to use the amazing human pattern recognition system we call ‘sight’, is also a real aid.

In the course of working on these projects, I became convinced graph database technology could help. In the clinical trials project, we were getting confused by the complexity of participant and stakeholder identities – they had multiple roles we needed to be aware of. It was only when I modelled these areas in a graph database that we made any headway. Mapping all this in Excel and SQL, as we’d been battling to do, just wasn’t effective.

A similar leap forward came with the neuroscience work. We had to

collect a wide range of attributes for people in various concurrent projects, which meant a lot of shared participants whose identity and relationships needed to be carefully and accurately tracked. I tried to do this in a relational database – and it proved to be a disaster! But using graph was a quantum leap in terms of clarity, depth and ease of maintenance.

That experience – and what I hear from other researchers and data modellers in life sciences – has convinced me of two things. One: graph database technology, thanks to its innate ability to model complexity, scale and connections, is the only data tool for serious medical research.

And two, positive help from developers to build graph-based data structures for research would make a huge difference, as well as allow a lot of highly-trained specialists get their hands-on information in a form they can work with a lot sooner.



Emil Eifrem

CEO and co-founder of Neo Technology. Previously CTO of Sweden’s Windh AB, where he headed up the development of highly complex information architectures for enterprise content management systems, Emil famously sketched out what today is known as the property graph model on a flight to Mumbai in 2000. Since then, Emil has devoted his professional life to building and evangelising graph databases, and jokes that, as a result, he plans to save the world through graphs and own Larry Ellison’s yacht by the end of the decade. He is a frequent conference speaker and a well-known author and blogger on NoSQL and graph databases, as well as co-author of the agreed Bible on graph databases, O’Reilly’s Graph Databases.

Email: emil@neotechnology.com

