CASE STUDY

Candiolo Cancer Institute (IRCC)

Flexible Data Models Provide Life-Saving Insights into Complex Cancer Research Data

INDUSTRY
Medicine / Life Sciences

GOAL
Develop a tool that provides insight to cancer data from multiple sources, tracks workflows and is accessible to external researchers

CHALLENGE
Relational databases didn't provide adequate flexibility for the multiple functions required

SOLUTION
Neo4j provides a flexible tool for uncovering complex relationships, modeling genomic domains and analyzing experimental procedures

RESULTS
- More efficient research data workflows
- A flexible database with better performance that provides new insights into cancer data

The IRCC relies on complex hierarchical data from a variety of sources to conduct advanced cancer research at the molecular level. Neo4j affords the organization with a way to manage and track this complex data to both gain insights and make it accessible to researchers around the world.

The Company

The Candiolo Cancer Institute (IRCC) is a scientific research non-profit based in Candiolo (Torino), Italy that makes significant contributions to the fight against cancer by both understanding its scientific basis and providing state-of-the-art diagnostic and therapeutic services. The core of the IRCC is at the interface of molecular biology and "precision" medicine and is funded by the Fondazione Piemontese per la Ricerca sul Cancro-Onlus.

The Challenge

The IRCC team performs molecular and biological tests on cancer samples that have been collected from hospitals around Europe. They needed to develop a laboratory information management system to track the data — such as the biological and molecular properties of the cancer samples — and the subsequent scientific procedures performed on these samples. This would feed a database used to analyze data and generate high-level biological hypotheses.

However, different types of structurally complex data tend to be hierarchical with intricate and frequently-changing relationships, which necessitated a number of integrated data models. Their initial tool — the relational database, MySQL — required a large number of JOINS and resulted in sluggish queries, as well as challenges with data integration and coherency.

Whatever tool the researchers chose also needed to be available to two distinct audiences: collaborators that were sharing their data with the IRCC, as well as other groups performing similar research who needed access to their software, all with the goal of working collectively to build cancer research knowledge.

This required a flexible, efficient tool that could organize and track cancer samples, as well as their molecular and biological features; serve as a data mining resource; and function as a database for tracking procedures.

"Our application relies on complex hierarchical data, which required a more flexible model than the one provided by the traditional relational database model," said Andrea Bertotti, MD, and the overall manager of the project.

The Strategy

Above all, the team knew they needed a flexible data model – which is best found in a graph database. They first downloaded and attempted to use OrientDB but found the software cumbersome and difficult to use. Next on the list was Neo4j.

Initially, IRCC applied Neo4j to a very focused goal: setting it up as a tool to maintain coherency between different relational databases. They began using it as a central hub to keep all the data from different relational databases in sync.

CASE STUDY

"Our application relies on complex heterogenous data, which required a more flexible model than the one provided by the traditional relational database model."

– Andrea Bertotti,
Medical Doctor,
Candiolo Cancer Institute

Instead of starting from scratch, the team began using the graph database as a place to collect all the data — including main entities — to ensure coherency across different existing databases. Now, every time an entity was created in one of the other databases, a copy would also be created in Neo4j.

This acted as a small project of what would later be implemented on a much larger scale. Over the next several years, their initial model became stratified with more complexity.

The Solution

IRCC has developed a production version of their database that relies on MySQL to store the legacy data and track entities, characteristics and laboratory procedures. This data is sent to Neo4j via scripts, and the database also continually imports data from publically-available resources.

They use MongoDB to store the raw, complex data and rely on Neo4j for all the rest: finding complex relationships, analyzing their experimental procedures, and modeling the genomic domain and complex semantics for genomic knowledge.

And while they initially tried to transpose the relational table models into the graph, they plan to remodel their database and use Neo4j as a more abstract layer to generate data models for each instance in order to integrate an abstract ontology that dictates relationships.

The Results

With Neo4j, the IRCC can now both get real insights into its data and share these insights with the international research community — all for the end of developing a cure for cancer.

The newfound flexibility of their database allows it to evolve and accommodate continually changing biological research and its findings, along with the ability to model relationships between concepts. And its faster queries provide better performance and allow the IRCC to model more complex relationships, gleaning more insight from their data than ever before.

And in addition to relying on Neo4j as its cancer-oriented biobank that contains detailed data; it also allows the team to track its workflow and share this and other data with researchers across the world.

About Neo4j

Neo4j is an internet-scale, native graph database that leverages connected data to help companies build intelligent applications that meet today's evolving challenges including machine learning and artificial intelligence, fraud detection, real-time recommendations and master data. As the #1 platform for connected data, Neo4j has over three million downloads, the world's largest graph developer community, and over thousands of graph-powered applications in production.

The world's most sophisticated organizations worldwide, from enterprises like Walmart, eBay, UBS, Cisco, HP, adidas and Lufthansa to hot startups like Medium, Musimap and Glowbl, use Neo4j to harness the connections in their data.

United Kingdom uk@neo4j.com
France ventes@neo4j.com
Scandinavia nordics@neo4j.com
DACH vertrieb@neo4j.com
Southern Europe southern-europe@neo4j.com