Use of NoSQL Graph databases to model diverse biological data

Detecting correlations between entities and forming theories

Sakib Adnan
Introduction to NoSQL

• NOSQL (Not Only SQL): A wide category for a group of database solutions which don’t follow the relational data model, built for processing large sets of data.

• **4 different types:**
  • Key-value databases: Every database item is stored as a key-value pair.
  • Wide-column databases: Uses tables, rows and columns but column names and format can change from row to row within the same table.
  • Document databases: Pairs each key with a data structure known as a document.
  • Graph databases: Databases are modeled into graphs, using nodes and edges.
Graph Databases

- Graph databases model various kinds of relationships between entities using nodes and edges (representing relationships), and both capable of having their own attributes. This is known as the Property Graph Model.
- **Node**: Represents biological entities (protein, genes, drugs, diseases).
- **Edge (relationship)**: labeled and directed edge that represents relationships between nodes.
- **Property (attribute)**: On both nodes and edges.
Neo4j and Cypher (GraphDB Query language)

• The industry standard for modelling biological data.
• Open-source Java based graph database management software.
• Queried using the Cypher Query Language to query and modify the database.
• Helpful features for large data sets such as large page cache memory, which allows for more data to be stored in memory.
• Allows memory to be used as heap memory in a Java Virtual Machine (JVM).
Correlating biological entities

- Graph databases used to represent highly connected, non-uniformly distributed, unpredictable data.
- Appropriate for biological data. To investigate various underlying processes involved in disease pathology.
- Integrate various heterogeneous data sets, such as complex relationships between viruses, candidate genes, proteins, drugs, their effects and interactions.
- Genetic and epigenetic factors on colon cancer.
- Association of asthma with multiple phenotypes and similarities to other respiratory diseases.
Protein and other entity interaction

- Nodes, their interactions, relationships help us discover correlations between them.
- Finding the best scoring path in weighted graphs yield the most effective drugs.
- Inferences made to create disease networks, discovering relationships between genes and diseases, effect of drugs on certain proteins or lack thereof.

Fig: Data Model for Protein against other entities
Easy modeling of biological entities

- Nodes: Proteins (blue), Drugs (red), Asthma (Yellow).
- Edges: drug-target associations, red.
- Sequence similarity relationships, grey.

Fig: Drugs that target proteins that are related to asthma.
Dense, Complex networks

Fig: Protein, drug and various other entity graph with even more associations.
Neo4j structure and optimization

Fig: Diagram of underlying structure of Neo4j
**Neo4j structure and optimization (continued)**

- **Bottom layer**: file open limit optimization; Neo4j often produces many small and random reads when querying data.

- **Middle layer**: page cache sizing; if all, or at least most, of the graph data files from a hard disk are cached into memory, it will reduce disk access and result in optimal performance.

- **Top layer**: heap sizing; it is beneficial to set a large heap space to support various query operations. OS, operating system; JVM, Java Virtual Machine.

- **Disk access log**: Logical transaction logs for system and data recovery can also be increased.
Advantages of using Neo4j

- For research purposes, with proper funding main memory can be configured in the order of terabytes/petabytes.
- Neo4j allows flexible schema structure. Supports extension to billions of nodes, edges and attributes, unlike traditional RDBMS with fixed schema.
- Allows searching for complex relationships, as well as a fast search and retrieval for relationships among multiple nodes.
- Uses efficient graph traversal algorithms to find relationships among connected nodes to find correlations (e.g. Djikstra’s, Bellman-Ford).
Use over traditional RDBMS

- RDBMS useful for well-structured data with pre-defined columns.
- Fixed schema that cannot be edited once a table was created.
- Does not store relationships amongst heterogeneous data, rather infers relationships among them by using queries.
- Although graph structures can be used in RDBMS, there are query performance implications for recursive structures like file trees.
- Cost implications: Operations over relationships between entities of a network in RDBMS results in expensive “join” operations.
Implications for the future

• For biological research, hardware required to run graph databases is very expensive.
• With decreasing volatile memory costs, graph databases are becoming more efficient.
• Neo4j has been updated to find shortest path between two nodes 2441x faster than PostgreSQL, under certain conditions.
• Multiple different open source graph databases available now such as OrientDB, MapGraph, etc. which are being increasingly used for molecular biology.
Questions?