Graph Databases
Overview and Applications

By Rodger Lepinsky

University of Winnipeg
April 29, 2013
Overview

• Literature search from blogs, online articles, company websites, videos, twitter

• Private research

• Only a little in the academic realm

• Originally intended to approach companies.
Rodger Lepinsky – Formal Training

• Bachelor of Commerce (Honours)
  – Asper - University of Manitoba

• Bachelor of Applied Computer Science
  – University of Winnipeg

• Passed Chartered Financial Analyst (CFA) Level 1 exam (pass rate: 33% to 40%)
Rodger Lepinsky – RDBMS

• RDBMS Expert
• DB Architecture, Design, Development, Warehousing, Tuning, DB Administration
• Working with databases since 1992
• With enterprise Oracle, SQL Server, Sybase databases since 1995
• Oracle User Groups Presentations:
  – High Speed Database Tuning
  – Cartesian products
• Technical Blog: rodgersnotes.Wordpress.com
• Twitter: @rodgernotes
The RDBMS world is changing rapidly.
Big Data

Volume, velocity, variety of data

Often machine generated:
Internet logs/analytics
Sensors in machines like modern jets

Online gaming companies: ½ terabyte of new data, daily
Google’s Paper – IEEE 2009

• The Unreasonable Effectiveness of Data: Alon Halevy, Peter Norvig, and Fernando Pereira

• “simple models and a lot of data trump more elaborate models based on less data.”

• “simple n-gram models or linear classifiers based on millions of specific features perform better than elaborate models that try to discover general rules.”
“each n-gram sequence from a corpus of billions or trillions of words”
Big Data, NOSQL Databases

- NOSQL: Not Only SQL
- Also called New SQL

- Four main types of NOSQL Databases:
  - Key Value
  - Column
  - Document
  - Graph Database
NOSQL DB – Key Value

• Works like a simple hashtable
• Tools: Memcached, Amazon’s Dynamo, Project Voldemort, Riak, Redis
• Twitter, StackOverFlow, Instagram, Youtube, Wikipedia
• Use: Store user information, like Session, Profiles, Preferences, Shopping Cart
• Drawback: Can’t query by value. No relationships. No rollbacks.
NOSQL – Column Databases

• Store data in column families. Ie. Person is usually queried by name or id, not salary.
• Tools: Cassandra, Hbase
• Ebay, Instagram, NASA, Twitter, Facebook, Yahoo
• Use: Logging, and Blogging. Tags, categories, posts in different column families.
• Drawback: No ACID transactions
• (Column databases are used in data warehouses)
NOSQL – Document Databases

• Store data as documents using XML, JSON or JSONB
• Tools: MongoDB, CouchDB, RavenDB
• SAP, Codecademy, Foursquare, NBC News

• Use: No fixed schema. Store different info.
• Drawback: Does not support transactions between documents
NOSQL – Graph Databases

• Store data as graphs, not rows and columns.
• Tools: Neo4J, Infinite Graph, OrientDB
• Linked In, Facebook, Google, NSA
• Use: with data that is connected.
• Not all data can be modeled in graph. I.e. Spreadsheets rows and columns are better in RDBMS.
RDBMS Data Structure

• Rows and columns, like a spreadsheet
• Rows added/deleted, and columns updated frequently
• Table structures never change without a conscious decision
• Unlike programs, Relational DB Design is rarely refined

• Result: Awful DB designs are put into production, and huge amounts of code required to make them work.

• See: DB Design Mistakes To Avoid by Lepinsky

http://rodgersnotes.wordpress.com/2010/09/14/database-design-mistakes-to-avoid/
Four tables
Many rows in each table
Graph DB Data Structure

• Nodes/Vertices, and Edges
• Adding or modifying Nodes or Edges changes the structure
• Structure constantly changing, as nodes and edges are inserted and updated.
Each row becomes a node

Many nodes

Rows in M:N tables become an edge between nodes

New nodes can be inserted at will. News events
RDBMS vs Graph

- **RDBMS:**
  - Good fit for static data structures, that do not change much
  - Ubiquitous in business.

- **Graph:**
  - Good for semi- or unstructured data
  - Fits complex and dynamic data better
  - Assumption: the relationships are as important as the records

http://www.zdnet.com/facebook-neo4j-7000009866/

Copyright Rodger Lepinsky 2014
RDBMS vs Graph

- **RDBMS:**
  - Join and query multiple tables to see relationship
  - Retrieve rows and columns

- **Graph:**
  - Query nodes and edges
  - Edges are the relationship
  - Relationship (edge) is labelled
  - Queries can return edges only

http://www.zdnet.com/facebook-neo4j-7000009866/
Lorenzo Alberton:

SQL has historically been unable to express recursive functions needed to maintain the transitive closure of a graph without an auxiliary table.
BIOIT Problem

2003 – BIOIT conference

How to model (DNA) molecules in RDBMS?

http://www.nasa.gov/audience/foreducators/postsecondary/features/F_Space_Radiation_Project_pr.htm
Tree Structures

- Difficult to do represent or use in RDBMS
- Easy in Graph DB

- Lorenzo Alberton: Trees In The Database, Attempts to represent trees in RDBMS/SQL.
- 128 slides, but still no simple or complete solution.
My First Graph Problem

• DBA_Objects are created in a tree structure.
• Type, used in a table, used in a view, view used by multiple procedures.
• You can have a single procedure, reading 15 tables: pyramid
• Can have one table, Customer, or Error_Log, used by many procedures: inverted pyramid.
• What’s the order of operations to build the objects?
• N factorial?
Oracle’s DBA_Dependencies only refers one level up, or down.

Many recursive reads required to see the whole structure, and correct order of operations.

SQL output: more like directory structure.

Ultimate problem: SQL output is in rows and columns. But object structure is actually a tree.

Software to solve problem by Yuri Slutsky:

http://www.samtrest.com/
DBA_Objects

- SQL output: single object found in multiple places and branches in the output. No clear order of operations.

- OBJECT_LVL_OBJID ROWNUM
  -------------------------------------------------------------
  - PACKAGE BODY  APPS.HR_DELETE   1   278801   1
  - PACKAGE BODY  APPS.HR_DELETE   1   278801   2
  - PACKAGE BODY  APPS.HR_DELETE   1   278801   3
  - SYNONYM  PUBLIC.USER_CATALOG   2   1167   4
  - VIEW  SYS.USER_CATALOG   3   1166   5
  - VIEW  SYS.USER_CATALOG   3   1166   6
    - VIEW  SYS._CURRENT_EDITION_OBJ   4   3270113   7
    - VIEW  SYS._CURRENT_EDITION_OBJ   4   3270113   8
  - PACKAGE BODY  APPS.HR_DELETE   1   278801   9
  - SYNONYM  PUBLIC.DBMS_SQL   2   2328   10
    - PACKAGE  SYS.DBMS_SQL   3   2327   11
    - PACKAGE  SYS.DBMS_SQL   3   2327   12
    - PACKAGE  SYS.DBMS_SQL   3   2327   13
      - PACKAGE  SYS.UTL_IDENT   4   3291213   14

DBA_OBJECTS as a Graph (Subset)
DBA_OBJECTS as a Graph (Gephi)

http://rodgersnotes.wordpress.com/2013/08/06/visualizing-fifty-thousand-dba_objects/
Graph Structures

Directed

Undirected
Applications For Graphs

• Where the data model is connected:
  • social
  • telecommunications
  • logistics
  • master data management
  • bioinformatics
  • fraud detection
Applications For Graphs

• Data Connections and complex interrelationships:
  • network management
  • content management
  • property and asset management
  • relationship management (CRM, ERM),
  • Not only does an association between nodes state that a relationship exists, but also describes how.
  •
  • Most of the data inside of the enterprise is very complex: Key/value stores may not work.

Copyright Rodger Lepinsky 2014
Applications For Graphs

• Aggregate data stores (key-value, column, document db) - solve problems related to atomic intelligence

• Graph databases - leverage connected intelligence

Application For Graphs

- social networking
- logistics networks (for package routing)
- financial transaction graphs (for detecting fraud)
- telecommunications networks
- ad optimization
- recommendation engines
- bioinformatics
Application For Graphs

- Social
- Recommendations
- Geo
- Logistics Networks: for package routing, finding shortest Path
- Financial Transaction Graphs: for fraud detection
- Master Data Management
- Bioinformatics: Era7 to relate complex web of information that includes genes, proteins and enzymes
- Authorization and Access Control: Adobe Creative Cloud, Telenor
Applications For Graphs

- friend-of-friend
- shortest path
- Gartner: “five richest big data sources on the Web:”
- social graph
- intent graph
- consumption graph
- interest graph
- mobile graph
Organizations Using Graph Databases

- Facebook
- LinkedIn
- Google
- Cisco
- Mozilla (Firefox)
- T-Mobile
- NSA – US National Security Agency
Social Network Analysis

• Facebook became one of the most prominent technology companies in the world by understanding that the relationships connecting people are just as important as the people themselves.

• Linked IN: Relationships matter
Facebook

- Facebook’s Graph Search feature contains billions of nodes and trillions of edges (understood to be in the low trillions)

- Facebook users are generating more than 500 terabytes of new data every day.
Facebook User’s Network of Connections
Social Network Research Study

- Leskovic and Horvitz - 2008
- Analyzed Whole of Microsoft Messenger System
- 30 billion conversations
- 240 million people
- Mean: 125 conversations per person
Social Network Research Study

- Social network
- 180 million nodes
- 1.3 billion undirected edges
- Graph is well connected and robust to node removal
- Average path length among messenger users: 6.6
- "Six degrees of separation"
Use Case:

History of philosophy

Each philosopher is a node in the network.

Edges represent lines of influence.

SPARQL - language to query the semantic web

Queries information that is structured in triples subject-relationship-object
The social network between characters in Homer’s Odyssey is remarkably similar to real social networks today. Suggests the story is based, at least in part, on real events.
Social Network of Alexander The Great

Impact on the network of the two conspiracies

http://www.academia.edu/2153390/The_Social_network_of_Alexander_the_Great_Social_Network_Analysis_in_Ancient_History
Diane H. Cline, Ph.D. University of Cincinnati
World Events

- US diplomatic relations with Algiers
- Network of parties involved
- 1785 to 1800

- Green: Algiers
- Red: United States
- Purple: England
- Light blue: Tripoli
- Darker blue: France
- Light purple: Spain
- Yellow: Portugal
- Orange: Sweden
Marketing Intelligence

Are there any conflicts of interest in our proposal?

Who could refer or introduce me to Larry Ellison?
## Financial Exposure

### Entity Details - Stellent Inc

#### Details for **Stellent Inc**

<table>
<thead>
<tr>
<th>Entity Details</th>
<th>Relation Details</th>
</tr>
</thead>
</table>

- **General Employment (18)**
- **Company Ownership (25)**

- **Ownership (2)**
  - Alan Menkes
  - Applied Signal Technology Inc
  - Daniel P Ryan
  - Dimensional
  - Frank A Radichel
  - Gamco Investors Inc
  - Gregg A Waldon
  - Ken H H Holec
  - **MARK RUPORT**
  - Optika Inc (2)
  - Oracle
  - Oracle Corp
  - Oracle Systems Corp
  - PRICE T ROWE GROUP Inc
  - Philip E Soran
  - Raymond A Tucker
The network of the top borrowers.
http://www.nature.com/srep/2012/120802/srep00541/full/srep00541.html
Corporate Ownership

• Can be very complex.

• Corporate Ownership can actually be circular:
  • A owns B. B owns C. C owns stock in A.

• Accounting rules: conglomerates must aggregate intra-company sales.
Corporate Ownership – TransUnion Canada

TRANSUNION OF CANADA, INC. One of 17 subsidiaries registered in Canada

CONTROL CHAIN: GOLDMAN SACHS GROUP, INC. > TransUnion Netherlands II, B.V. > TRANSUNION OF CANADA, INC.
GSCP VI PARALLEL NORTH HOLDING CORP.

Control Chain: Goldman Sachs Group, Inc. > GS Capital Partners VI PIA Fund, L.P. > GS Capital Partners VI Parallel, L.P. > GSCP VI Parallel North Holdings S.r.l. > GSCP VI Parallel North Holding Corp.

One of 17 subsidiaries registered in Canada.
Goldman Sachs Group – As A Tree

http://opencorporates.com/companies/us_de/2923466/network
Goldman Sachs Group – As A Network
Google And Graphs

• Social networks: graphs that describe relationships among people.

• Transportation routes: create a graph of physical connections among geographical locations

• Paths of disease outbreaks form a graph

• Games among soccer teams

• Computer network topologies

• Citations among scientific papers

• Internet / World Wide Web: documents are vertices and links are edges.
Google And Graphs

- Pregel: Google’s other data-processing infrastructure

- Google: MapReduce (Hadoop) is used for 80% of all the data processing needs: indexing web content, clustering engines for Google News, Google Trends, processing satellite imagery, language model processing for statistical machine translation, data backup and restore.

- The other 20% is handled by a lesser known infrastructure called “Pregel” which is optimized to mine relationships from “graphs”.

http://www.royans.net/arch/pregel-gogles-other-data-processing-infrastructure/
Google And Graphs

• Google extracts more than 200 signals from the web graph: language of webpages, number and quality of other pages pointing to it.

• Google: scalable infrastructure, named Pregel, to mine a wide range of graphs. In Pregel, programs are expressed as a sequence of iterations.

• PageRank, for example, takes only about 15 lines of code.
Security – National Security Agency

• NSA Application: determine who else is in contact with suspected terrorists
• Stores tens of petabytes of data
• Internal system, built on top of Hadoop

• Accumulo is able to process:
  • 4.4-trillion-node, 70-trillion-edge graph.
• Human brains:
  • 100 billion nodes/vertices, 100 trillion edges

http://gigaom.com/2013/06/06/heres-how-the-nsa-analyzes-all-that-call-data/
Graph500 Experiment

**Graph500 Huge class — scale 42**
- $2^{42}$ (4.40 trillion) vertices
- $2^{46}$ (70.4 trillion) edges
- 1 Petabyte

**Cluster specs**
- 1200 nodes
- 2 Intel quad-core per node
- 48 GB RAM per node
- 7200 RPM sata drives

- **Huge** problem is 19.5x more than cluster memory
- linear performance from 1 trillion to 70 trillion edges...
- despite multiple hardware failures!

![Graph500 Scalability Benchmark](http://www.pdl.cmu.edu/SDI/2013/slides/big_graph_nsa_rd_2013_56002v1.pdf)
National Security Agency - NSA

- Largest supercomputer installations do not have enough memory to process the Brain Graph (3 PB)!

- Electrical power cost
- At 10 cents per kilowatt-hour — $7 million per year

<table>
<thead>
<tr>
<th>Class</th>
<th>Scale</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toy</td>
<td>26</td>
<td>17 GB</td>
</tr>
<tr>
<td>Mini</td>
<td>29</td>
<td>140 GB</td>
</tr>
<tr>
<td>Small</td>
<td>32</td>
<td>1 TB</td>
</tr>
<tr>
<td>Medium</td>
<td>36</td>
<td>17 TB</td>
</tr>
<tr>
<td>Large</td>
<td>39</td>
<td>140 TB</td>
</tr>
<tr>
<td>Huge</td>
<td>42</td>
<td>1.1 PB</td>
</tr>
</tbody>
</table>
Fraud – Cheque Kiting Scheme

http://photos.cleveland.com/plain-dealer/2012/02/19fvisualejpg.html
Bernie Madoff Corporate Network

Diagram of companies feeding money to Bernie Madoff

http://twinkle_toes_engineering.home.comcast.net/~twinkle_toes_engineering/ponzi_madoff.htm
Shortest Paths - Trains

• “In 2007, a colleague and I used Java with Oracle 9i to implement Dijkstra’s Algorithm. Our “MapQuest for Trains” application would route a rail train over various right-of-ways while minimizing cost. The cost was a function of distance, fuel surcharge, and obstacles. The task to route a train from Los Angeles to Chicago had a grotesquely long response time. Nobody wanted their applications deployed on our nodes because we spiked the servers!”

• Solved by using a Neo4J graph database as the underlying storage
Shortest Paths - Trains
Social Network Prediction Engine

• Problem: predict which blog posts a WordPress user would ‘like’ based on prior user activity and blog content
Results:

Nearly 50% of all new likes are from blogs one ‘edge’ from the user.

A distance of 3 edges/likes traversed – encompasses 90% of all new likes.

http://www.overkillanalytics.net/kaggles-wordpress-challenge-the-like-graph/
Graph DB on the Market

Traversal processing speed

In memory graphs
(not studied in this article)

Graph databases

*DEX

Large-scale graph processing tools

Apache Hama
Apache Giraph

OrientDB

Neo4j

JUNG (Java Universal Network/Graph)
Visualization

Andrei Kashcha:

Uses VivaGraphJS, google app engine, U of F sparse matrices

http://www.yasiv.com/graphs
Visualization

- Tools in the browser (Neo4J, Linkurious, D3, Keylines)
- Gephi, on the desktop
- With Excel: nodexl.codeplex.com
- Nathan Yau, Flowing Data (more for R)
Slowing demand in Oracle

From 3% to 2% in about 7 years

Job Trends

Strong demand growth in

“Data Science”
“Big Data”
“R statistics”

Although, fewer jobs overall

http://www.indeed.com/jobtrends?q=%22Data+Science%22&l=New+York%2C+NY
Data Science

• New profession
• Expertise involves:
  • Computer and software
  • Math and Statistics
  • Data (often Big Data)
  • Subject Matter Domain Knowledge
• Find significant inferences, trends
• Add value to the organization
• Jingjing’s thesis
Canadian Universities

- Queen’s
  Master’s Degree in Management Analytics (Business)

- University of Toronto
  Certificate: Management of Enterprise Data Analytics

- York University, Toronto, Ontario
  Master of Science in Business Analytics

- University of Ottawa
  Master in Electronic Business Technologies

- Simon Fraser:
  Master’s program in Big Data
US Universities

- Arizona State University
- Bentley University, Waltham, Mass.
- Carnegie Mellon University, Pittsburgh, Pa.
- Columbia University, New York, N.Y.
- DePaul University, Chicago, Ill.
- Drexel University, Philadelphia, Pa.
- Fordham University, New York, N.Y.
- Harvard University, Cambridge, Mass.
- Louisiana State University, Baton Rouge, La.
- Massachusetts Institute of Technology, Cambridge, Mass.
- New York University, New York, N.Y.
- North Carolina State University, Raleigh, N.C.
- Northwestern University, Evanston, Ill.
- Purdue University, Lafayette, Ind.
- Rutgers University, New Brunswick, N.J.
- University of San Francisco, San Francisco, Cal.
- Stanford University, Stanford, Calif.
- University of California at Berkeley, Berkeley, California
- University of Southern California, Los Angeles, California
- University of Cincinnati, Cincinnati, Ohio
- University of Connecticut, Graduate Learning Center, Hartford, Conn.
- University of Illinois, Champaign, Ill.
- University of Tennessee, Knoxville, Tenn.
US University Degrees

- Master of Business Administration
- Master of Business Administration, Business Analytics
- Master of Business Administration, specialization In Business Analytics
- Master of Business And Science degree in Operations Research and Business Analytics
- Master of Engineering
- Master of Information and Data Science
- Master of Information Systems Management, Business Intelligence and Data Analytics.
- Master of Science (MS), Applied Urban Science and Informatics
- Master of Science In Analytics
- Master of Science in Business Analytics
- Master of Science in Business Analytics and Project Management
- Master of Science in Computer Science - Data Science
- Master of Science In Computer Science, Specialization in Information Management and Analytics
- Master of Science In Marketing Analytics
- Master of Science in Predictive Analytics
- Master of Science in Statistics: Analytics Concentration
- Masters of Science in Computational Science and Engineering
- Masters of Science in Computer Science, Machine Learning

MOOC

- Massive Open Online Courses
- Coursera.com
- MIT
- Code Academy
- Khan Academy

https://en.wikipedia.org/wiki/Massive_Open_Online_Course

Copyright Rodger Lepinsky 2014
Coursera

• Johns Hopkins: Data Science Specialization
• The Data Scientist’s Toolbox
• R Programming
• Getting and Cleaning Data
• Exploratory Data Analysis
• Reproducible Research
• Statistical Inference
• Regression Models
• Practical Machine Learning
• Developing Data Products
• Capstone Project
Coursera

- Core concepts in data analysis: getting started
- National Research University - Higher School of Economics (HSE), Russia

- Duke University:
- Irrational Behavior – Dan Ariely
Questions