Graph Database and Neo4j

谷文栋

2013-01-12 @北京

谷文栋

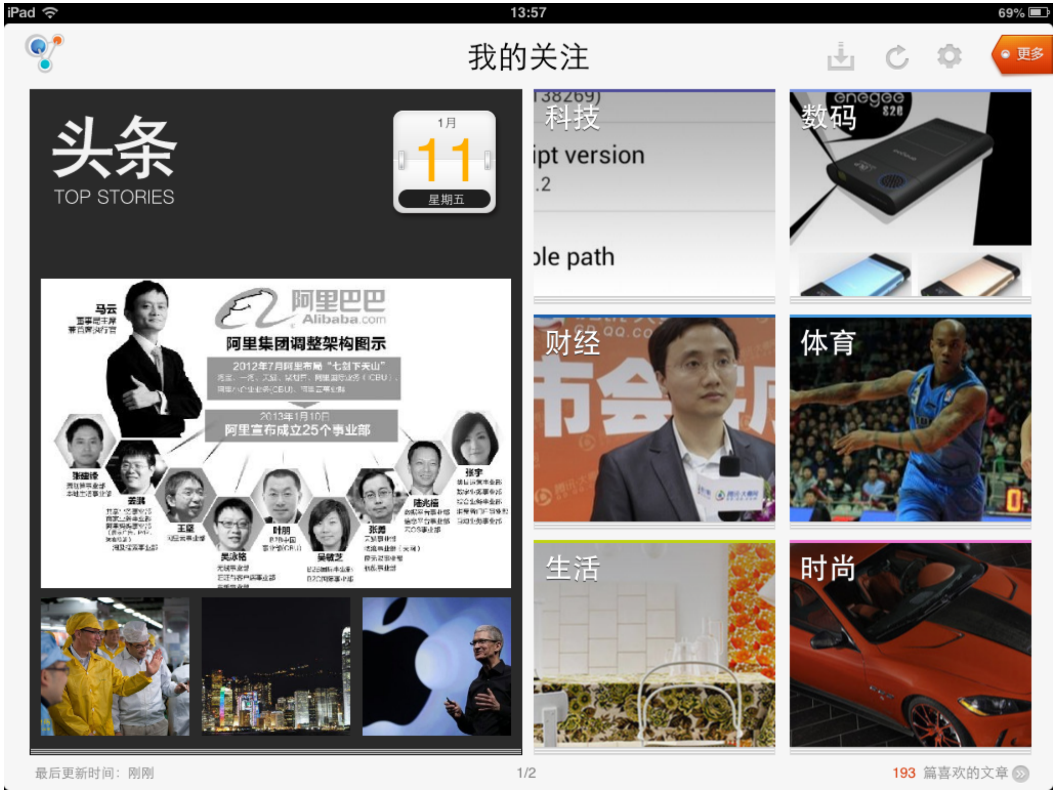
- 指阅 - 联合创始人 & CTO
  - http://zhiyue.me/
- ResysChina 推荐技术社区发起人
  - http://resyschina.com/2012/
- Beyond Search 博客作者
  - http://guwendong.com/
- 常用网名 @clickstone

指阅

http://zhiyue.me/

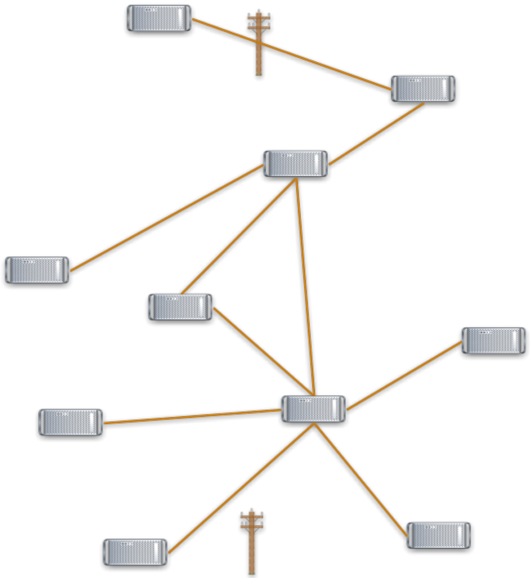
你的个性化杂志

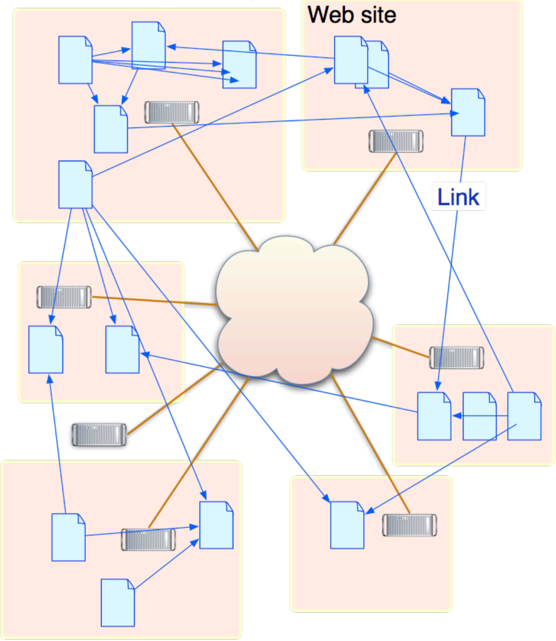
- 信息基因
- 主题
- 聚合/筛选
- 个性化

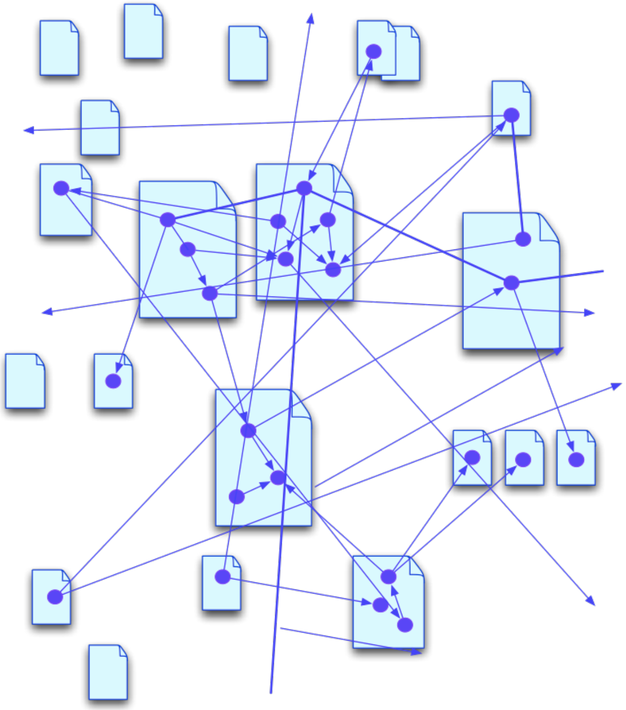


互联网三阶段

- It's not the wires
  -- it's the computers
- It’s not the computers
  -- it's the documents
- It’s not the documents
  -- it's the Things







Source: Tim Burners-Lee, Director, W3C
http://www.w3.org/DesignIssues/Abstractions.html

基因工程

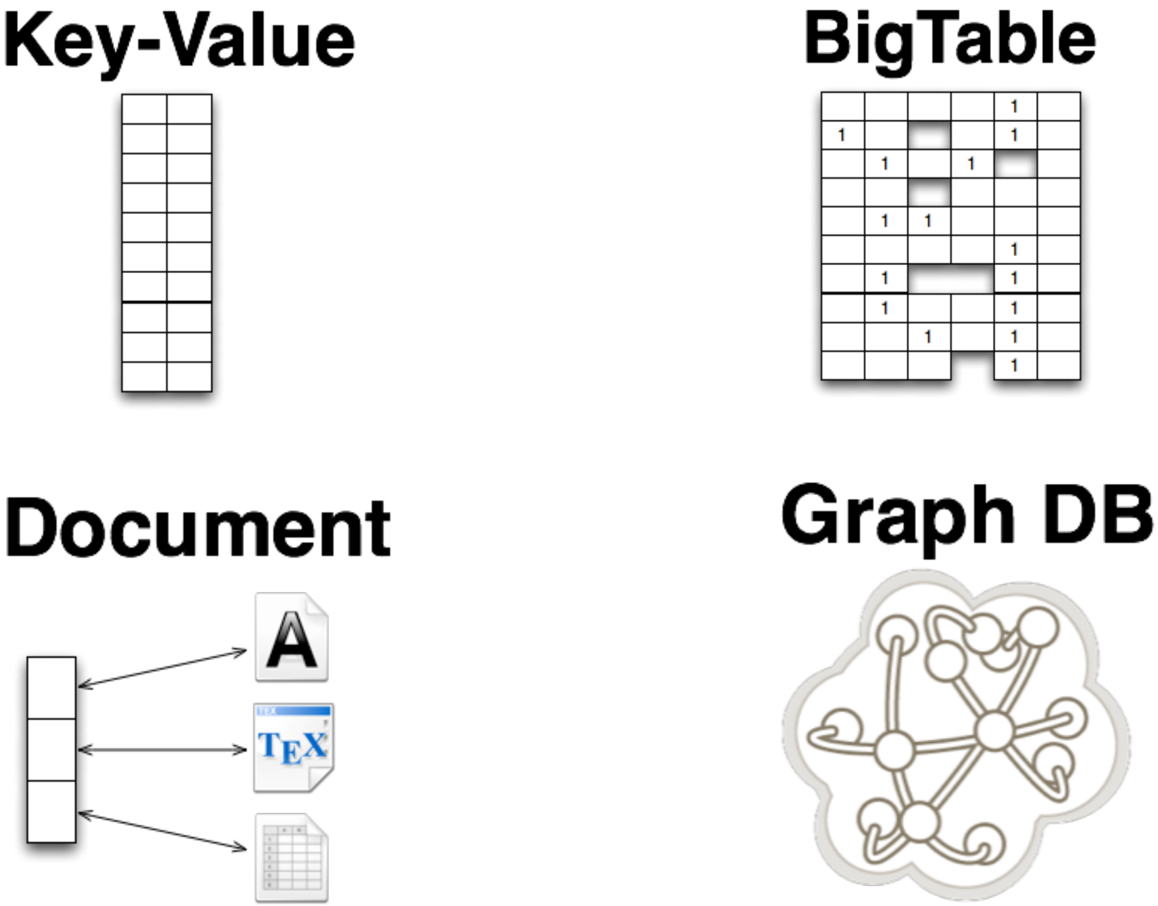
- Freebase - An entity graph of people, places and things
- Google Knowledge Graph
- Pandora – 音乐基因工程
- Jinni – 电影基因工程

NoSQL

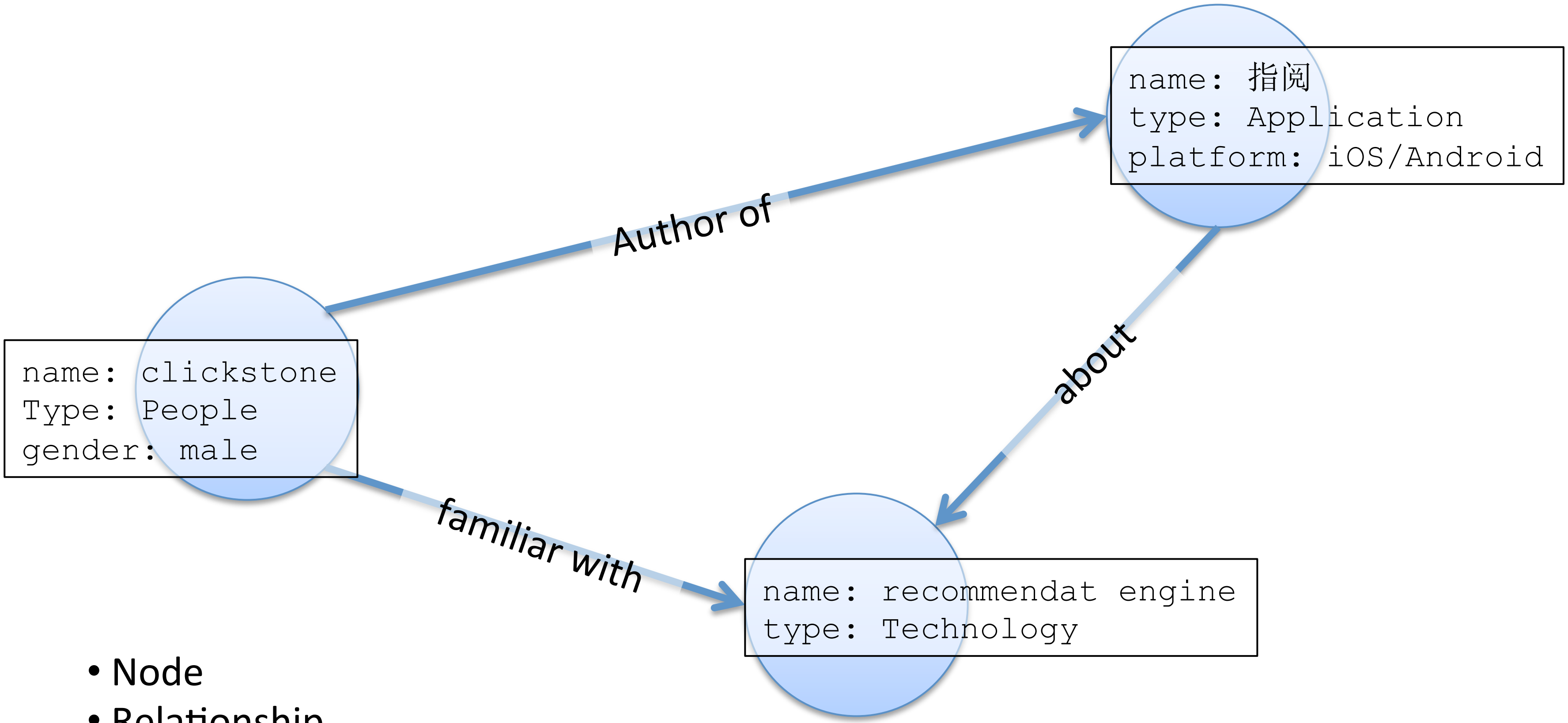
is simply

Not Only SQL

Four NoSQL categories



Property Graph Model



name: 指阅
type: Application
platform: iOS/Android

Author of

name: clickstone
Type: People
gender: male

about

familiar with

name: recommendat engine
type: Technology

- Node
- Relationship
- Property

Graph Databases

- Property Data Model:
  - Nodes with properties
  - Relationships with properties
- Examples:
  - Neo4j
  - FlockDB(Twitter)
  - Google Pregel
  - AllegroGraph, SonesGraphDB, OrientDB, InfiniteGraph …

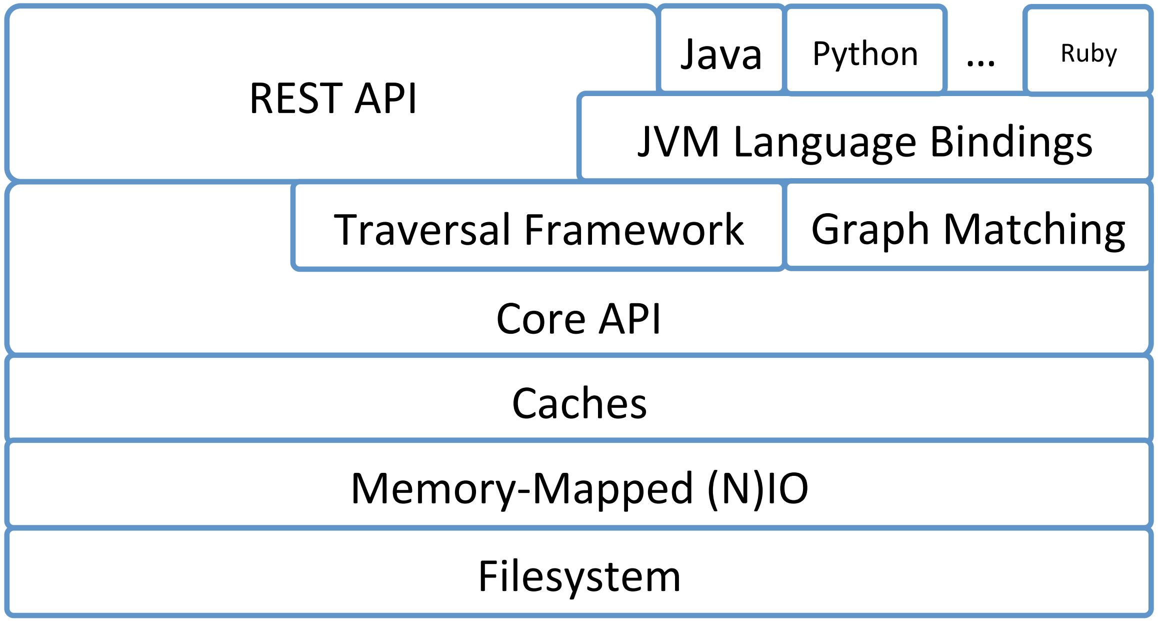
What’s Neo4j

- It’s is a Graph Database
- Embeddable and server
- Full ACID transactions
  - don’t mess around with durability, ever.
- Schema free, bottom-up data model design

More on Neo4j

- Neo4j is stable
  - In 24/7 operation since 2003
- Neo4j is under active development
- High performance graph operations
  - Traverses 1,000,000+ relationships / second on commodity hardware

Neo4j Logical Architecture



REST API

Java

Python

…

Ruby

JVM Language Bindings

Traversal Framework

Graph Matching

Core API

Caches

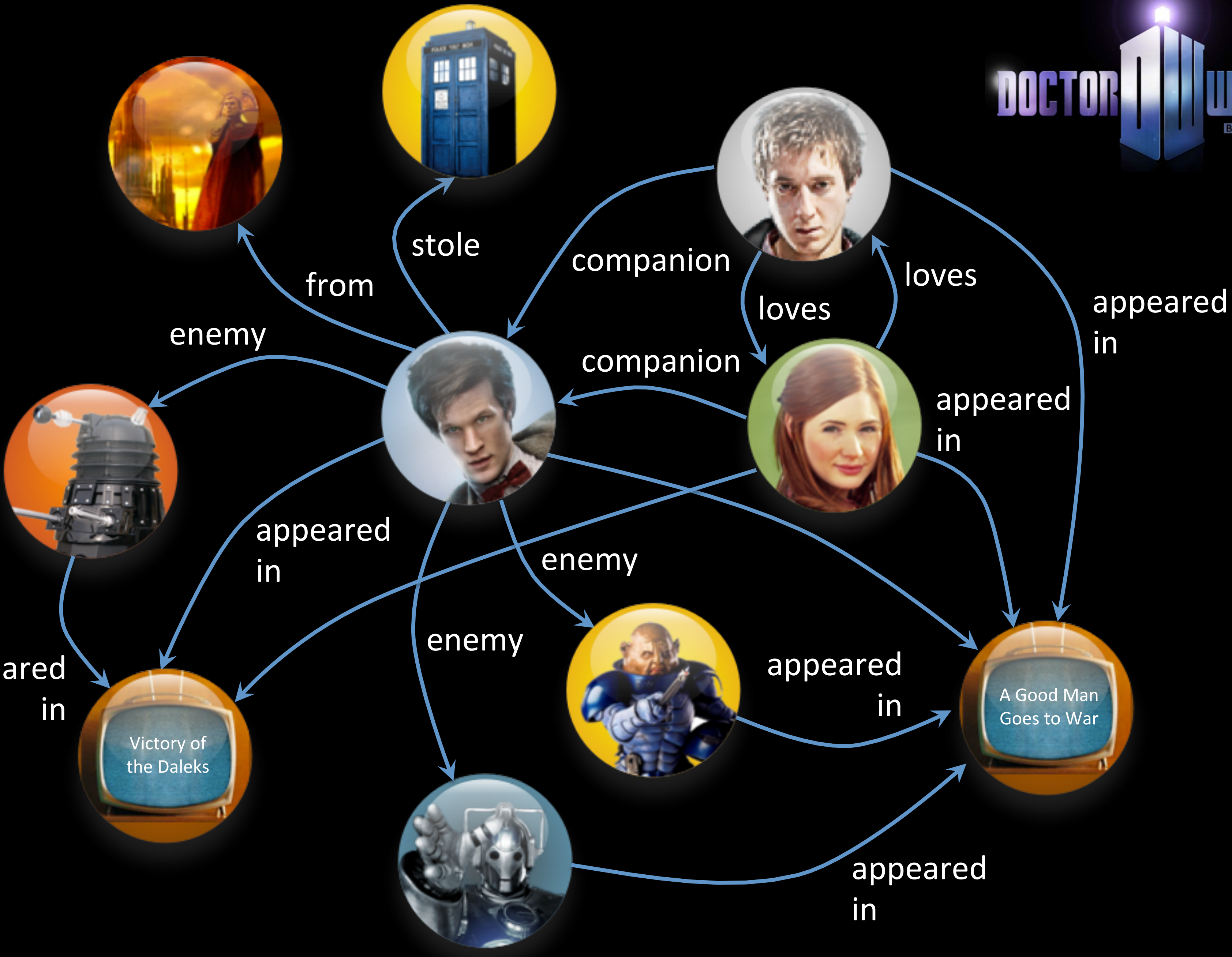
Memory-Mapped (N)IO

Filesystem

Remember, there’s NOSQL

So how do we query it?

DOCTOR WHO
BBC



stole

from

companion

loves

loves

enemy

companion

appeared in

appeared in

appeared in

enemy

enemy

appeared in

appeared in

Victory of the Daleks

A Good Man Goes to War

appeared in

Data access is *programmatic*

- Through the Java APIs
  - – JVM languages have bindings to the same APIs
    - JRuby, Jython, Clojure, Scala…
- Managing nodes and relationships
- Indexing / lucene
- Traversing
- Path finding
- Pattern matching

What is Cypher?

- Declarative graph pattern matching language
  - “SQL for graphs”
  - Tabular results
- Cypher is evolving steadily
  - Syntax changes between releases
- Supports queries
  - Including aggregation, ordering and limits
  - Mutating operations in product roadmap

Example Query

- The top 5 most frequently appearing companions:

```
start doctor=node:characters(name = 'Doctor')
match (doctor)<-[:COMPANION_OF]-(companion)
      -[:APPEARED_IN]->(episode)
return companion.name, count(episode)
order by count(episode) desc
limit 5
```

Start node from index

Subgraph pattern

Accumulates rows by episode

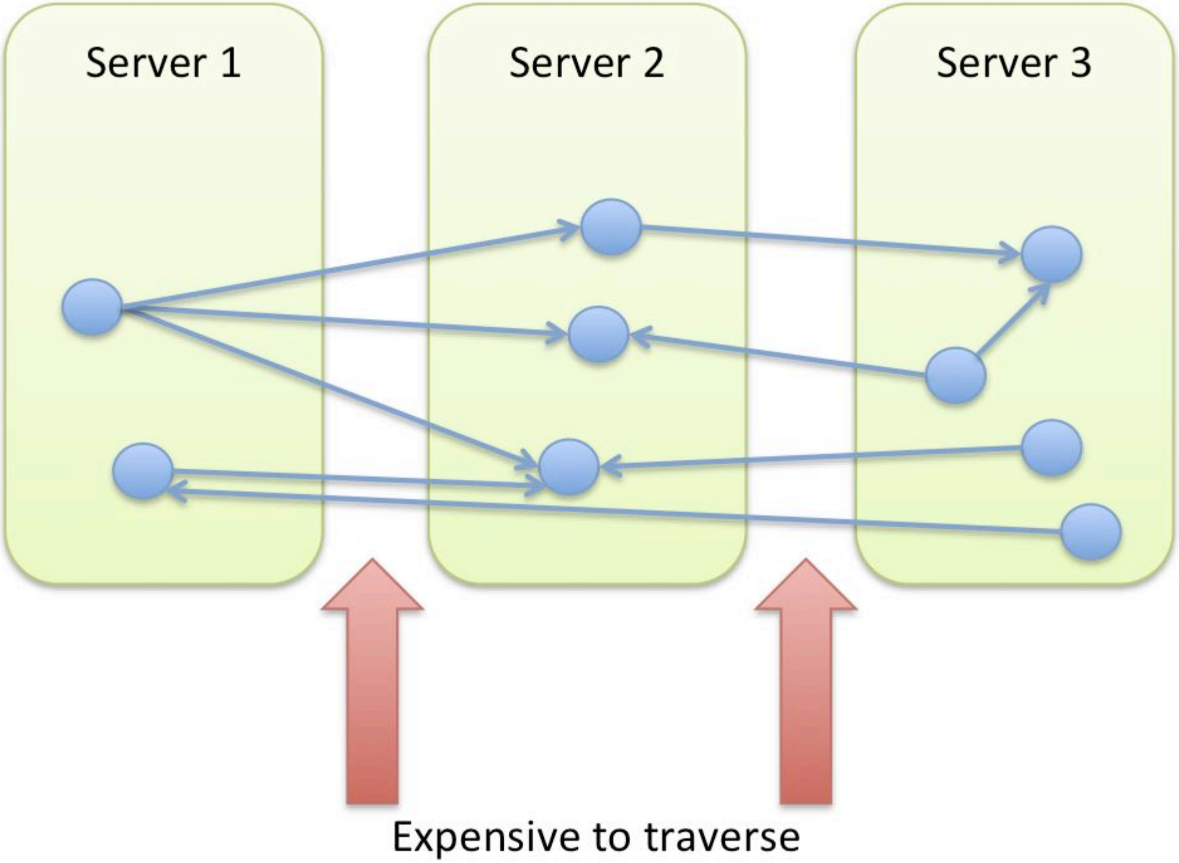
Limit returned rows

Results

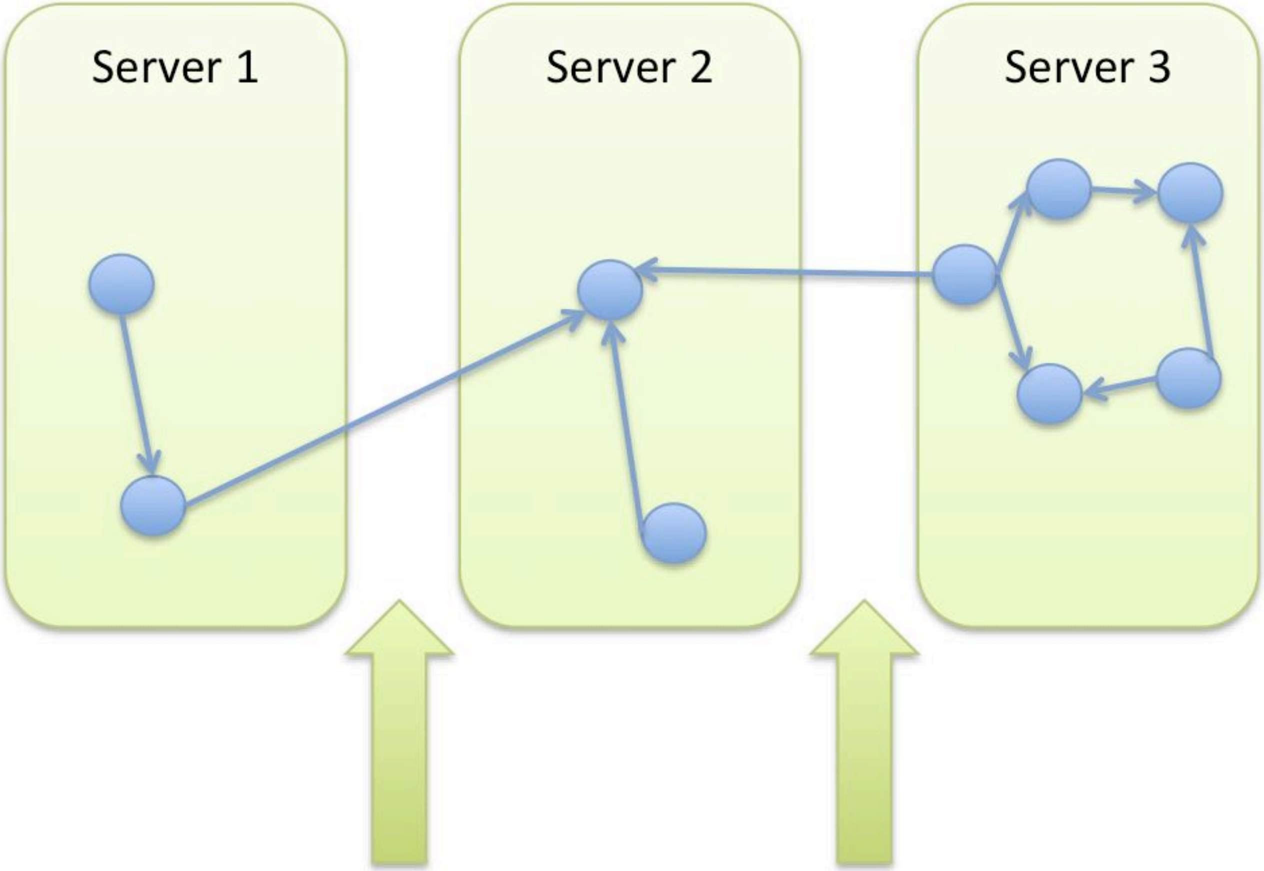
```
+-----------------------------------+
| companion.name   | count(episode) |
+-----------------------------------+
| Rose Tyler       | 30             |
| Sarah Jane Smith | 22             |
| Jamie McCrimmon  | 21             |
| Amy Pond         | 21             |
| Tegan Jovanka    | 20             |
+-----------------------------------+
| 5 rows, 49 ms                     |
+-----------------------------------+
```

Scaling graphs is **<u>hard</u>**

Chatty Network



Minimal Point Cut



Fewer expensive traversals

Domain-specific sharding

- Eventually (Petabyte) level data cannot be replicated practically
- Need to shard data across machines
- **Remember: no perfect algorithm exists**

- But we humans sometimes have *domain insight*

Pros and Cons

- Strengths
  - Powerful data model
    - don’t excuse you from design
  - Fast
    - For connected data, can be many orders of magnitude faster than RDBMS
- Weaknesses:
  - Sharding
    - Though they *can* scale reasonably well
    - And for some domains you can shard too!

Applications of Graph Databases and Traversal Engines

IT互联网领域的影响力分析

1. 从挑选一些种子用户开始。
2. 获取种子用户的关注列表和微博内容。
3. 构造图：

   1）A关注B，添加一条边A->B；

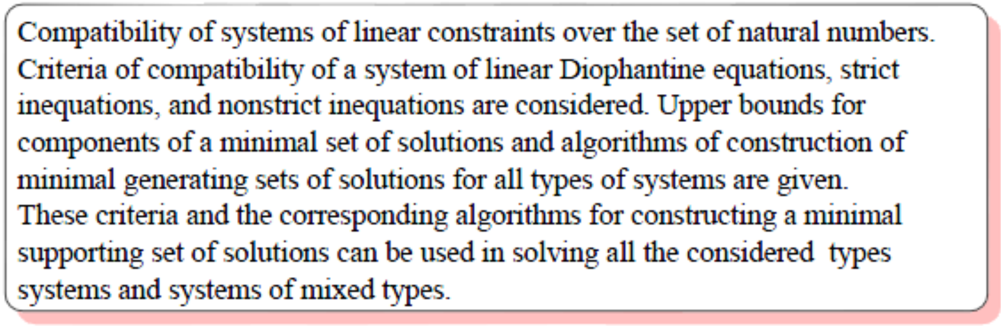
   2）A转发B，添加一条边A->B；如果同时存在关注和转发关系，会影响到边的权重。

   3）通过关注和转发识别出种子用户之外的新用户，重复1）和2）。
4. HITS / SPEAR

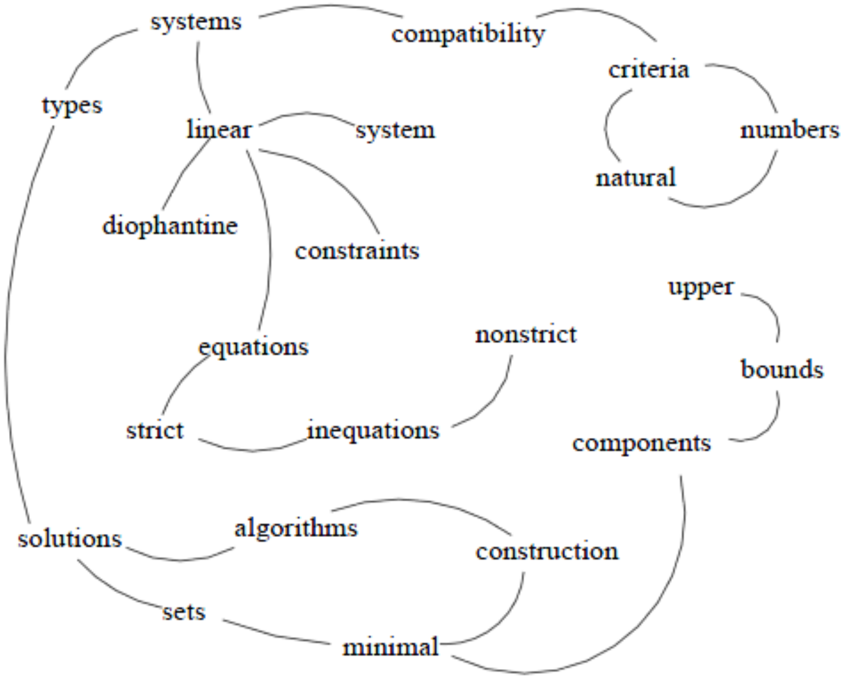
http://weibo.com/1641544424/eB2mq72b8ai

指阅的信息基因技术

- Text as Graph
- TextRank
  - TextRank: Bringing Order into Texts



Compatibility of systems of linear constraints over the set of natural numbers. Criteria of compatibility of a system of linear Diophantine equations, strict inequations, and nonstrict inequations are considered. Upper bounds for components of a minimal set of solutions and algorithms of construction of minimal generating sets of solutions for all types of systems are given. These criteria and the corresponding algorithms for constructing a minimal supporting set of solutions can be used in solving all the considered types systems and systems of mixed types.



个性化阅读

- 每天产生的信息非常多
- 信息的生命周期短暂
- 同质化问题严重

- 人的兴趣变化琢磨不定
- 挑战用户习惯
- 缺乏存在感与互动

Value in Relationships
用“关联”的视角去思考问题

Thanks!

欢迎大家下载 **指阅**

https://itunes.apple.com/cn/app/id450737500?mt=8