Getting Biologists off ACID

Ryan Verdon

3/13/12

Outline

- Thesis Idea
- Specific database
- Effects of losing ACID
- What is a NoSQL database
- Types of NoSQL databases with examples

Thesis Idea

- Amount of biological data is growing exponentially
- Hardware cannot keep up with the growth

Database Survey

- In 2003, a survey was conducted of 111 biological databases.
- 96 (i.e. 87%) of the 111 considered databases have Hypertext references to other databases,
- 40 to 44 (i.e. 36% to 40%) are implemented as **flat files**,
- 41 (or 42) (i.e. 37%) are implemented using a **relational database management system**,
- 7 (i.e. 6%) use an object database management system,
- 3 (.i.e. 3%) use an object-relational database management system,
- and all databases collect data from different sources.

Specific Database

Flybase

- Primary biology database on the insect family Drosophilidae (fruit fly)
- Important research from fruit flies includes
  - Genes
  - Recombination
  - Signaling networks (important for major diseases)
  - Stem cells
  - Growth control

Info on the Flybase database

- Single server relational database
- Over 135 tables
- Uses Chado schema
- Gives data a “type”
- Creates a graph that relates the different types

Effects of losing ACID

ACID

- Atomicity
- Consistency
- Isolation
- Durability

CAP Theorem

- Consistency
- Availability
- Partition tolerance

BASE consistency model

- **B**asically **A**vailable
- **S**oft state
- **E**ventually consistent
- Given enough time where no changes are made, all replicas will see the same data

NoSQL

What is a NoSQL database

- Any non-relational database
  - Hierarchical
  - Graph
  - Object oriented
  - Etc.

Why were NoSQL databases created

- To overcome limitations of relational databases
  - Predefined layout
  - Scaling
  - SQL
  - Large feature set

Major types of NoSQL databases

- Key-value stores
- Column-oriented databases
- Document based stores

Key-value stores

- Stored values are indexed for retrieval by keys
- Can store unstructured or structured data
- Similar to DHTs

Example: Dynamo

- Created by Amazon, only used internally
- Highly available even in the face of continual failures
- Amazon realized many services only need primary-key access
  - Best seller lists
  - Shopping carts
  - Session management

Column-oriented databases

- Contain extendable columns of closely related data
- Can greatly benefit from compression

Example: HBase

- Apache open-source project
- Based off of Google’s Bigtable database
- Ties in closely with Hadoop and map-reduce

Document based stores

- Data stored as a collection of documents
- Documents can have any number of fields and any length
- Documents are accessed via a unique key
- Capabilities of the query language heavily depend on the implementation

Example: MongoDB

- Started by 10gen
- Supports
  - Replication
  - Map-reduce
  - Sharding
- Used by lots of large companies including
  - Disney
  - Craigslist

Questions

?