Giới thiệu NoSQL và Nhu cầu trong đời sống thực tế

Nhóm thực hiện:
Lê Trương Trọng Duy
Đinh Kim Quốc Khải

Tại sao thảo luận về chủ đề này?

- Kỷ nguyên của Big Data
- RDBMS - sự lựa chọn duy nhất?

Nội dung

1. Tại sao phải dùng NoSql?
2. NoSql là gì?
3. Ưu điểm và nhược điểm của NoSql.
4. Kiến trúc và phân loại theo cơ chế lưu trữ của NoSql.
5. So sánh giữa NoSql và RDBMS.
6. Một số khái niệm và nhu cầu của BigData trong thực tế.

- Yêu cầu cho việc lưu trữ.
- Xu hướng không dùng mô hình cơ sở dữ liệu quan hệ NoSQL.

Tương tác

- NoSQL sinh ra để khắc phục các vấn đề mà một cơ sở dữ liệu dạng RDBMS gặp phải.
- NoSQL sinh ra không phải để cạnh tranh với RDBMS mà là để đảm nhiệm những việc mà RDBMS chưa làm tốt.

VD

- Đếm và thống kê truy cập.
- Phân tích tâm lý thị trường chứng khoán sử dụng Google Trends.
- Phân loại hình ảnh Picase/Nhận dạng khuôn mặt Facebook.
- Hệ thống đặt chỗ cho du lịch dựa trên sở thích.
- Hệ chuẩn đoán y học.

Mục tiêu:

- Hiệu suất hoạt động cao với số lượng dữ liệu cực lớn.
- Bỏ qua việc thông dịch trong SQL cùng với những truy vấn rườm rà.

Các đặc tính tổng quan của NoSQL:

- Cách thiết kế dữ liệu khác với cơ sở dữ liệu truyền thống.
- Dữ liệu phi quan hệ:
  - Yếu tố quan trọng góp phần làm nên thành công cho NoSQL.
  - Không tuân theo các dạng chuẩn hóa mà cơ sở dữ liệu RDBMS đặt ra.

1. Tại sao chọn NoSQL?

- Cơ sở dữ liệu quan hệ được thiết kế cho những mô hình dữ liệu không quá lớn trong khi các dịch vụ mạng xã hội lại có một lượng lớn dữ liệu và cập nhật liên tục do số lượng người dùng quá nhiều.

Một số thống kê về dung lượng & nhu cầu lưu trữ:

+ Năm 2003 thế giới tạo ra 5 exabyte dữ liệu (5 tỷ Gigabyte).

+ Năm 2010, cứ 2 ngày thế giới lại tạo ra 5 exabyte dữ liệu.

+ Ước tính năm 2014, cứ 10 phút thế giới lại tạo ra chừng đó dữ liệu.

(Nguồn: Theo Eric Schmidt, CEO của Google)

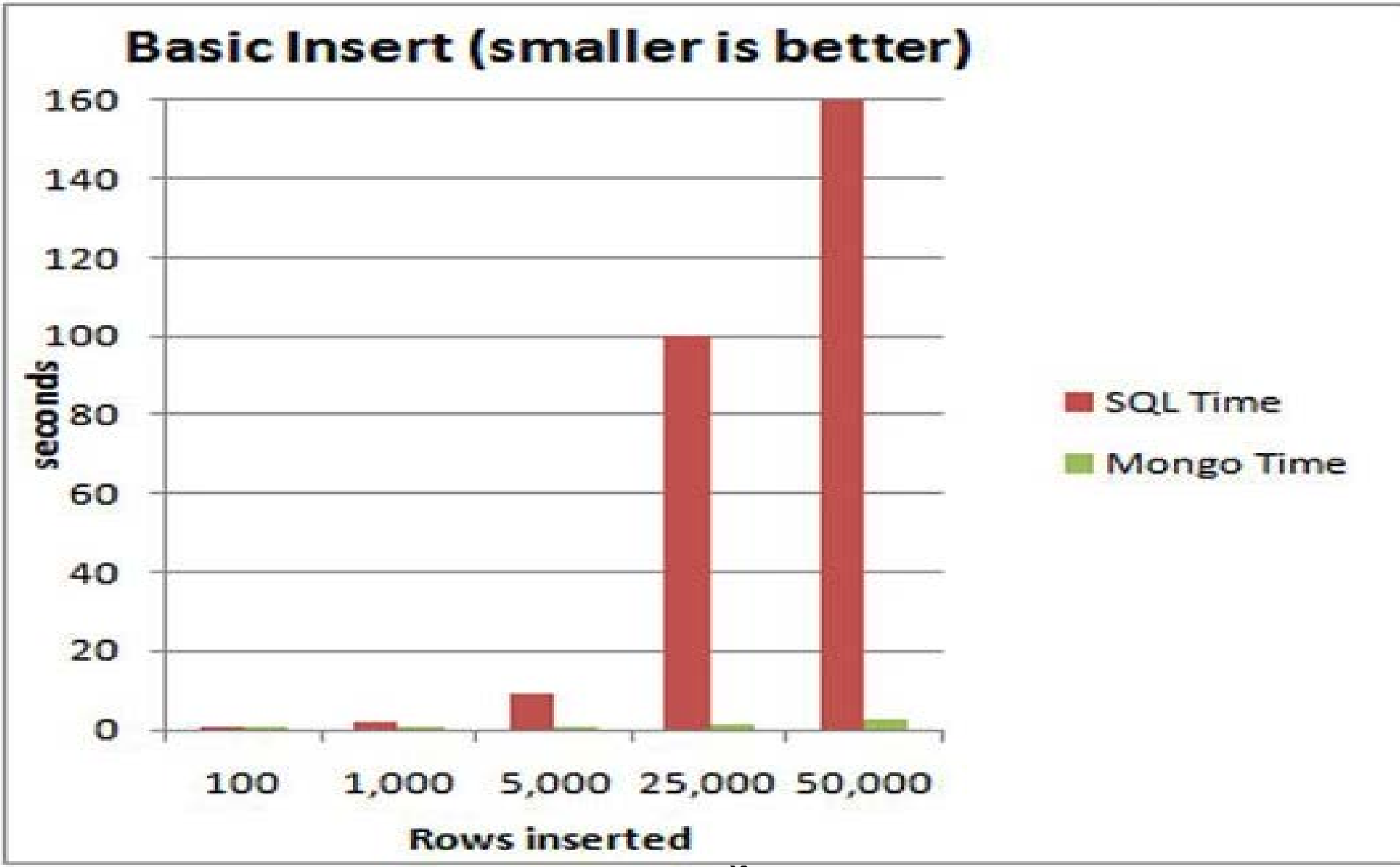
1. Tại sao chọn NoSQL?

- NoSQL sẽ tập trung giải quyết các vấn đề như:
  - Tốc độ thực thi.
  - Khả năng lưu trữ.
  - Các nghiệp vụ phức tạp (phân trang, đánh chỉ mục…)
  - Hạ thấp chi phí nếu so sánh với RDBMS truyền thống.

1. Tại sao chọn NoSQL?

- NoSQL vừa mang lại một giải pháp tốt hơn vừa tiết kiệm chi phí hơn
  - Hiệu suất làm việc tốt hơn.
  - Miễn phí.
  - Mang lại lợi ích to lớn.

So sánh hiệu năng



Nguồn: http://blog.adlegant.com/

2. NoSQL là gì?

NoSQL là:

- Một xu hướng cơ sở dữ liệu mà không dùng mô hình dữ liệu quan hệ để quản lý dữ liệu trong lĩnh vực phần mềm.
- Non-Relational (NoRel) -không ràng buộc.
- Thế hệ database kế tiếp của RDBMS.

Đặc điểm phân biệt

- Mở rộng theo chiều ngang.
- Lược đồ tự do (Schema-free).
- API đơn giản, hỗ trợ đánh chỉ mục tất cả các thuộc tính (Full Index Support).
- Giao tiếp ACID (ACID Transaction) là không cần thiết.

Xu hướng phát triển DB

Năm 2009, NoSQL đánh dấu thế hệ database mới:

- Distributed (phân tán).
- Non-relational (không ràng buộc).
- **Không giới hạn không gian dữ liệu.**

Các đặc tính khi làm việc với NoSQL

Khi làm việc với NoSQL ta sẽ gặp một số khái niệm sau:

- Fields: Columns
- Document: row
- Collection: table
- Key-value: cặp khóa - giá trị được dùng để lưu trữ dữ liệu trong NoSQL
- Cursor: tạm dịch là con trỏ. Sử dụng cursor để lấy dữ liệu từ database.

3. Ưu điểm và nhược điểm của NoSQL?

Ưu điểm của NoSql

- Hiệu suất hoạt động cao.
- Khả năng phân trang.
- NoSQL là nguồn mở.
- Việc mở rộng phạm vi là mềm dẻo.
- Tận dụng được việc cung cấp mềm dẻo của đám mây.
- NoSQL được các hãng lớn sử dụng.

Nhược điểm của NoSql

- Cấu trúc dữ liệu phi quan hệ.
- Chưa đủ kinh nghiệm cho các doanh nghiệp.
- Thiếu sự tinh thông.
- Những vấn đề về tính tương thích.

4. Kiến trúc và phân loại theo cơ chế lưu trữ của NoSql.

Kiến trúc của NoSql

- Mô hình lưu trữ tập dữ liệu theo cặp giá trị key-value.
- Distributed storage.
- Eventual consistency (nhất quán cuối).
- Vertical scalable (khả năng mở rộng chiều dọc).
- Horizontal scalable.

Phân loại cơ sở dữ liệu NoSql

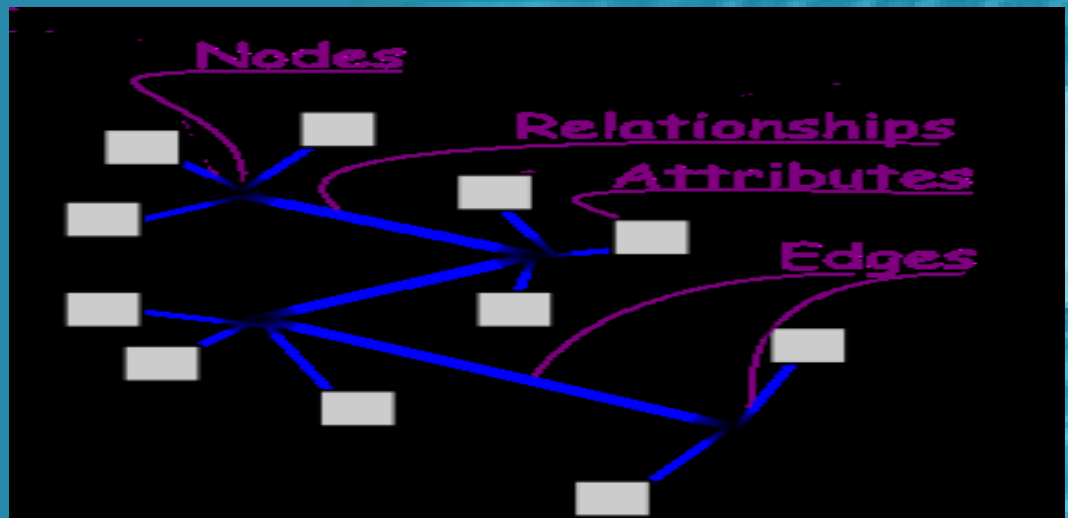
- Key-Value Store

Cơ sở dữ liệu NoSQL đơn giản nhất chính là Key/Value stores.

- Key/value cache in RAM
- Key/value save on disk
- Key Value Store

Phân loại cơ sở dữ liệu NoSql

- Key-value
- Column Families/Wide Column Store
- Document database
- Graph Database



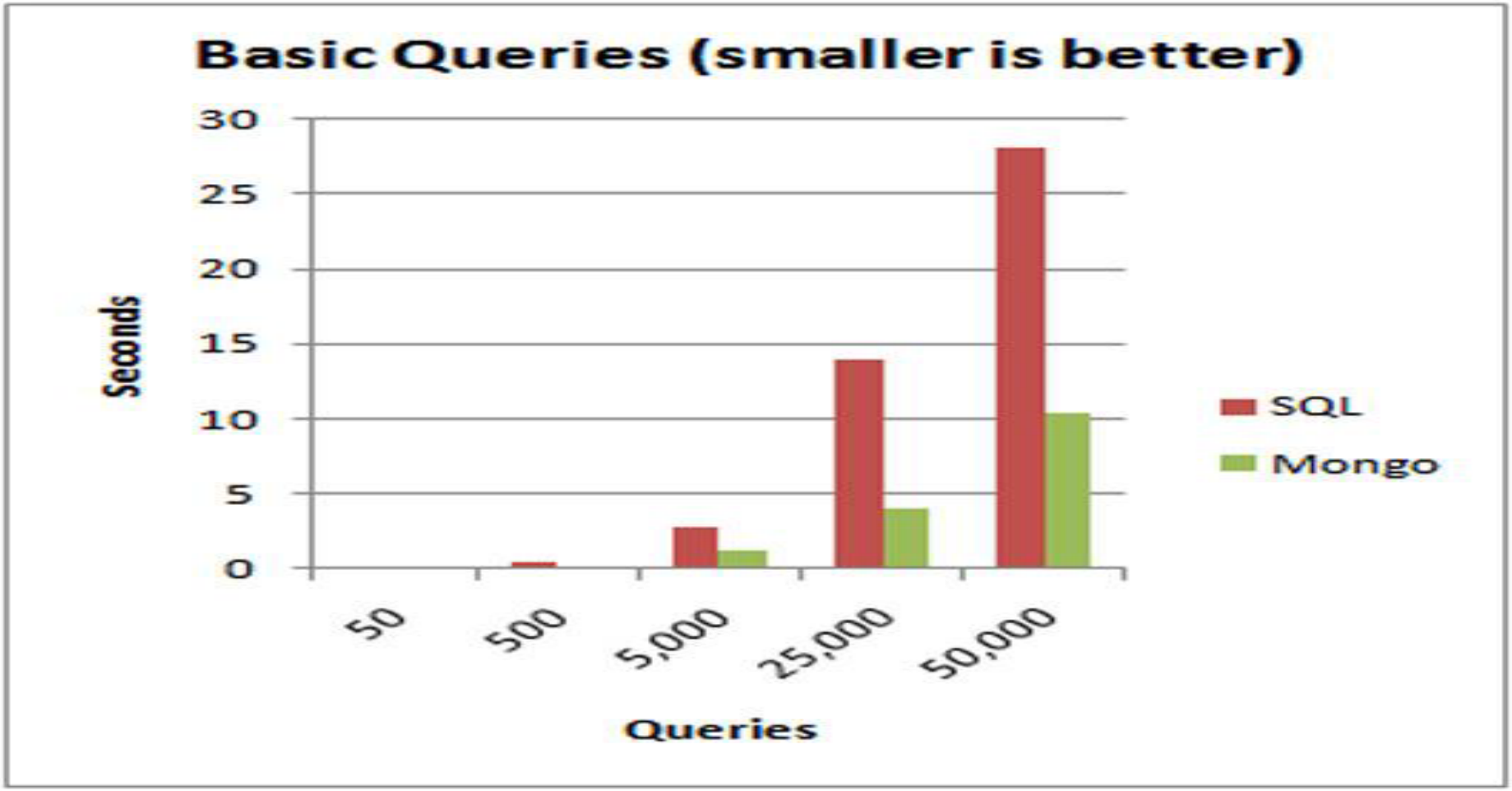
5. So sánh NoSQL với RDBMS

- Giải quyết các thiếu sót của RDBMS:
  - Các vấn đề như tốc độ thực thi, khả năng lưu trữ.
  - Các nghiệp vụ phức tạp (phân trang, đánh chỉ mục, …).
- Hiệu suất đạt được (biểu đồ dưới).

So sánh NoSQL với RDBMS

Number of Parallel Clients 5 Time in seconds

	Total Rows	Rows / client	SQL Time	Mongo Time	Sql Ops/sec	Mongo Ops/sec
Basic Query	50	10	0.1	0.08	500	625
with index	500	100	0.38	0.1	1,316	5,000
	5,000	1,000	2.8	1.2	1,786	4,167
	25,000	5,000	14	4	1,786	6,250
	50,000	10,000	28	10.4	1,786	4,808



So sánh NoSQL với RDBMS

Đặc điểm	CSDL quan hệ	NoSQL
Hiệu suất	Kém hơn SQL Relational giữa các table	Cực tốt Bỏ qua SQL Bỏ qua các ràng buộc dữ liệu
Khả năng mở rộng	Hạn chế về lượng.	Hỗ trợ một lượng rất lớn các node.
Hiệu suất đọc-ghi	Kém do thiết kế để đảm bảo sự vào/ra liên tục của dữ liệu	Tốt với mô hình xử lý lô và những tối ưu về đọc-ghi dữ liệu.
Thay đổi số node trong hệ thống	Phải shutdown cả hệ thống. Việc thay đổi số node phức tạp.	Không cần phải shutdown cả hệ thống. Việc thay đổi số node đơn giản, không ảnh hưởng đến hệ thống.

6. BigData

- Dữ liệu lớn có thể vừa là dữ liệu có cấu trúc, vừa là dữ liệu không có cấu trúc.
- Là thuật ngữ miêu tả sự gia tăng theo cấp lũy thừa của dung lượng dữ liệu, vượt quá khả năng vận chuyển, lưu trữ và phân tích.
- Thông qua khả năng mở rộng, các cơ sở dữ liệu này cũng quản lý cả dữ liệu không có cấu trúc.

- Value: Giá trị của dữ liệu.

Có bao nhiêu dữ liệu tin cậy khi những quyết định quan trọng cần được thực hiện trên số lượng lớn dữ liệu thu thập ở tỷ lệ cao.

- Volume: Dung lượng của dữ liệu.

IBM ước lượng, có 2.5 nhân 10 mũ 18 bytes (2,500,000,000,000,000,000) dữ liệu được tạo ra mỗi ngày

- Veracity: Tính chính xác của dữ liệu

- Velocity: Tốc độ cập nhật, kết nối của dữ liệu.

Nơi có tỷ lệ dữ liệu được gia tăng bởi vì băng thông mạng – điển hình như tỷ lệ gigabit ngày nay (gigE, 10G, 40G, 100G) được so sánh với tỷ lệ megabit.

- Variety: Sự đa dạng của hình thức, cấu trúc lưu trữ dữ liệu.

Bao gồm nhiều kiểu dữ liệu phi cấu trúc, như dòng hình ảnh kỹ thuật số (digital video streams), dữ liệu cảm biến, cũng như các file log nhật ký.

Big data giúp

- Không chỉ vấn đề kích cỡ và dung lượng của dữ liệu.
- Tiếp cận, chọn lọc nguồn dữ liệu, cung cấp thuật toán tối ưu.
- Phân tích, xử lý và khai thác thông tin nhằm phục vụ cho mục đích của con người.

Phân tích, xử lý và khai thác thông tin

- Đưa big-data trở thành smart-data (dữ liệu thông minh),
- Công cụ mạnh mẽ giúp chúng ta:
  - Nghiên cứu thói quen, tâm lý
  - Tương tác xã hội phức tạp của con người.

Smart data

Đề xuất giải pháp giúp người dùng quyết định hành vi của mình.

Điều đó không chỉ giúp doanh nghiệp thấy xu hướng người dùng mà còn giúp hỗ trợ những quyết định kinh doanh và quản lý rủi ro dễ dàng hơn.

Nhu cầu

- Các công ty startup.
- Các developer không cần lo Back-end (Cloud).
- Các trang web cần Realtime (Dịch vụ Database).
- Dịch vụ Re-seller (chủ yếu ở VN).
- Phân tích thị trường (Việt Nam đứng đầu khu vực Đông Nam Á về sức tăng Internet).

Ví dụ:

Các công ty ở nước ngoài:

- Amazon:
  - Xử lý hàng triệu thao tác back-end (người dùng và các công ty liên kết).
  - Năm 2005: Nắm 3 cơ sở dữ liệu 7.8 TB, 18.5 TB, and 24.7 TB.
- Walmart :1 triệu giao dịch/giờ, tạo ra 2.5 PB (=167xdữ liệu sách trong US Library of Congress).

BigData hiện nay

- Thể thao: Fadi Naoum, Phó Chủ tịch cấp cao của Đội tuyển Đức cho biết:

  "Big data là một nguồn lực đáng kinh ngạc cho các huấn luyện viên và cầu thủ để khoanh vùng thông tin theo từng hoàn cảnh, và cung cấp đầy đủ thông tin nhằm tối ưu hoá việc đào tạo và đề xuất chiến thuật."
- Quảng cáo và Cloud (Amazon, Google).

Thách thức

Chuyển đổi:

- Công cụ chuyển đổi hạn chế
- Đào tạo
- Bảo mật

Web References

- "NoSQL - Your Ultimate Guide to the Non - Relational Universe!"

http://nosql-database.org/links.html

- "NoSQL (RDBMS)"

http://en.wikipedia.org/wiki/NoSQL

- PODC Keynote, July 19, 2000. *Towards Robust. Distributed Systems*. Dr. Eric A. *Brewer*. Professor, UC Berkeley. Co-Founder & Chief Scientist, Inktomi. *www.eecs.berkeley.edu/~**brewer**/cs262b-2004/PODC-keynote.pdf*

- "Exploring CouchDB: A document-oriented database for Web applications", Joe Lennon, Software developer, Core International. http://www.ibm.com/developerworks/opensource/library/os-couchdb/index.html
- "Graph Databases, NOSQL and Neo4j" Posted by Peter Neubauer on May 12, 2010 at:

  http://www.infoq.com/articles/graph-nosql-neo4j
- "Cassandra vs MongoDB vs CouchDB vs Redis vs Riak vs HBase comparison", Kristóf Kovács. http://kkovacs.eu/cassandra-vs-mongodb-vs-couchdb-vs-redis

- “Distinguishing Two Major Types of Column-Stores” Posted by Daniel Abadi onMarch 29, 2010 http://dbmsmusings.blogspot.com/2010/03/distinguishing-two-major-types-of_29.html
- “MapReduce: Simplified Data Processing on Large Clusters”, Jeffrey Dean and Sanjay Ghemawat, December 2004. http://labs.google.com/papers/mapreduce.html
- “Scalable SQL”, ACM Queue, Michael Rys, April 19, 2011 http://queue.acm.org/detail.cfm?id=1971597

- “a practical guide to noSQL”, Posted by Denise Miura on March 17, 2011 at http://blogs.marklogic.com/2011/03/17/a-practical-guide-to-nosql/
- http://www.code4life.vn/2013/03/mongodb-nosql-la-gi.html
- “Brewer's CAP Theorem” posted by Julian Browne, January 11, 2009.

http://www.julianbrowne.com/article/viewer/brewers-cap-theorem

Questions?

The end