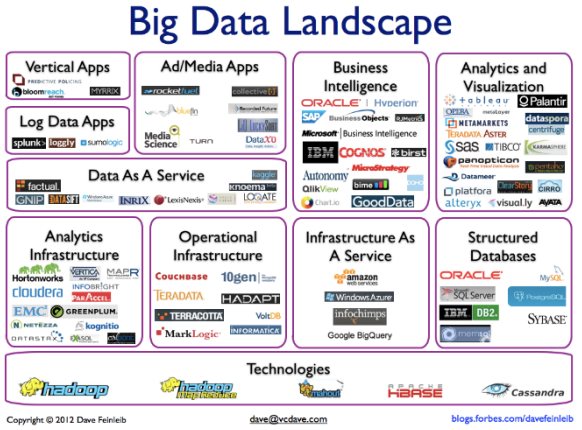


Big data and NoSQL databases

Seminar on big data management

Lecturer: Jiaheng Lu

Spring 2016

Information on preparing Presentation and Report

Goals for presentation and report are different:

1. Presentation: Let the audience to understand your topic;

2. Report: Show your own critical thinking and new ideas.

Contents of Presentation (Length: 35-40 minutes)

- 1. Introduction: please make a clear introduction
- 1.1 Why you are interested in this topic: what kind of problems do you hope to solve?
- 1.2 How had the problem been studied before?
- 1.3 What is the application of this problem for big data?
-
- 2. Related works:
- 2.1 Make sure you leave sufficient time to present all related prior work. Do not assume that the audience knows the prior work,
- 2.2 Present it on an intuitive level.
-

Contents of Presentation (Cont.)

-
- 3 Main algorithms and contributions
- 3.1 Show the main solutions of the paper(s).
- 3.2 Present it with examples. The examples are quite important for understanding.
-
- 4. Your own comments and conclusion
- 4.1 Present your own comments about the paper(s)
- 4.2 It would be very good to identify the weak points of the paper(s) after your critical thinking.
-

Contents of Report (6-8 pages, Single column)

- 1. What are the research problems?
- 2. What are the strengths of the paper(s)?
- 3. What are the main weaknesses of the paper(s)?
- 4. If you were to solve this problem, what would you do?
- 5. Why do you like/dislike the paper(s)?
- 6. Conclusion and summary of your report.

Opponent

- Carefully listen to the presentation
- Ask questions after the presentation
- Complete an opponent assessment form and submit it to the teacher after the presentation

- Big data and NoSQL databases

Data storage and history

Before-1950s Data was stored as paper records

Lot of time was wasted. e.g. when searching.
Therefore inefficient.

Magnetic tapes and hard disk

- 1950s and early 1960s: Data processing using magnetic tapes for storage
- Late 1960s and 1970s: Hard disks allow direct access to data
-
- Data stored in files

Drawbacks of file system

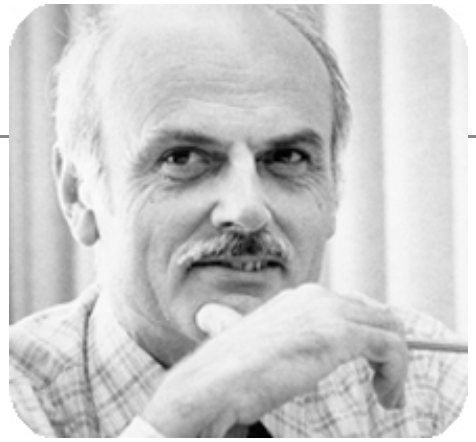
- Each program has its own data format
- Programs are written in different languages, and so cannot easily access each other's files.
- Any new requirement needs a new program

Database Approach

- 1960’s Network databases
- 1970’s Relational databases
- 1990’s Object-oriented and object-relational
- 1995+ XML, Mobile, GeoDB, Embedded DB
- 2005+ NoSQL DB, NewSQL DB

History of databases: Turing awards

1973 Charles W. Bachman



1981 Edgar F. Codd

1998 Jim Gray

2014 Michael Stonebraker

History of databases: Turing awards

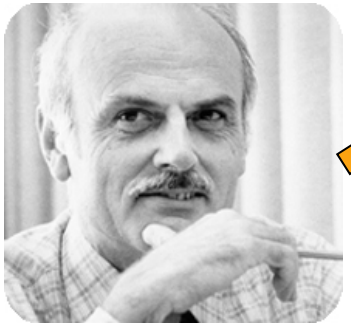
Object-relational model, column stores,…Modern databases

Distributed databases and transaction

Relational databases

Network databases

1973 Charles W. Bachman



1981 Edgar F. Codd

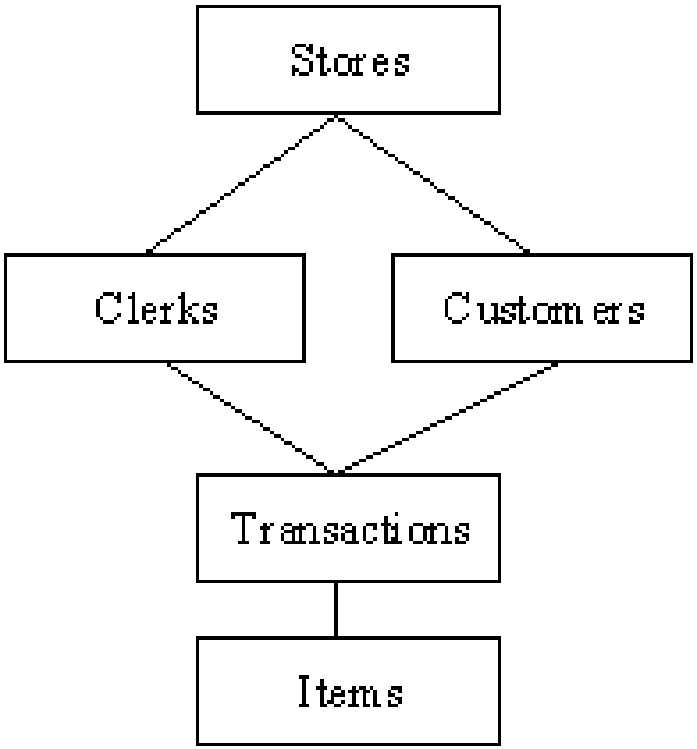
1998 Jim Gray

2014 Michael Stonebraker

Network Model

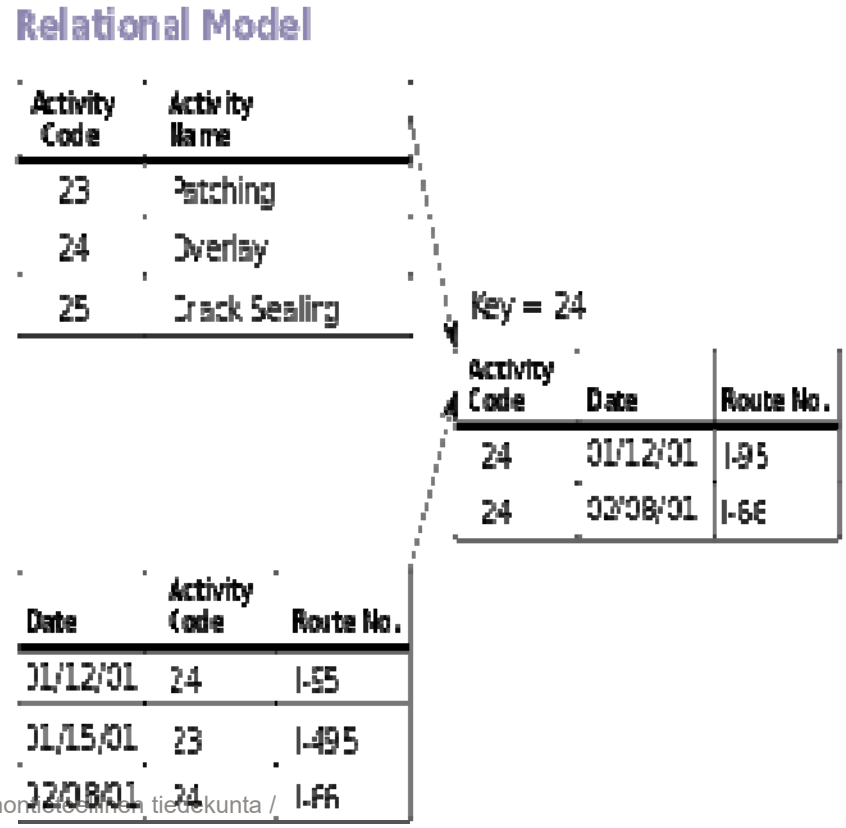
Physical file pointers are used to model the relations between files

Most suitable for large databases with well-defined queries and well-defined applications



Relational model

- E. F Codd introduced the relational model in 1970

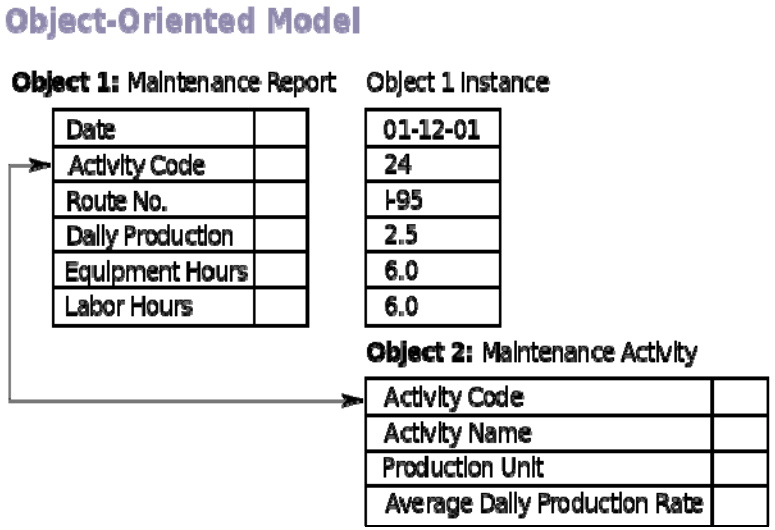


Relational model

- Support relational algebra and operations
- Data and program are separated
- Improved data sharing and better integration
- DB2, Oracle and SQL server are the most prominent commercial DBMS products

Object oriented data model (1990's)

- The purpose of OODBMS is to store object-oriented programming objects in a database without having to transform them into relational format



Object-relational model

- Extend the relational data model by including object orientation
- Allow attributes of tuples to have complex types, including non-atomic values such as nested relations

Big Data Challenge

Big Data Challenges

- 1-2 billion people on the Internet
- Cisco estimates annual Internet traffic will reach 677 exabytes by 2013

- Google processes 1TB an hour
- eBay processing 80TB a day
- Facebook 12PB cluster, adding 10TB a day

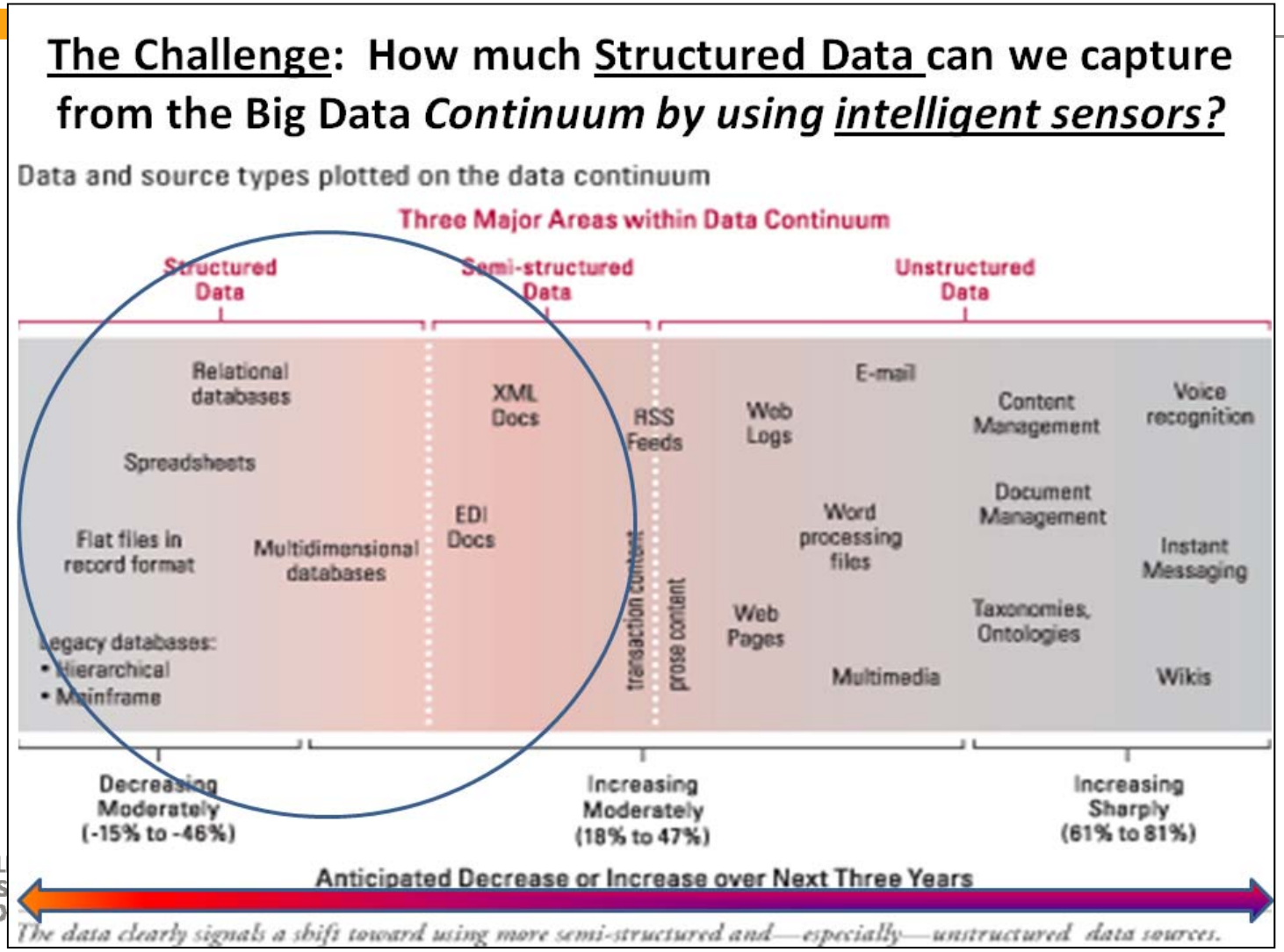
- 85 million Tweets per day
- 500 million Facebook users

5V's of big data

- Volume
  - TB → PB → EB
- Variety
  - Text, audio, video
- Velocity
  - Real time Operational / Analytic Applications
- Value
  - Extract Value from big data, complex Analytics
- Veracity
  - Biases, noise and abnormality in data.

Limitation for relational databases(1)

Different Types of Data: Data Variety



Limitation for relational databases(2)

- What are Big Analytics
  - Not only simple “group by” aggregation，But also
    - Machine leaning, artificial intelligence
    - Data mining、natural language processing
    - Social network analysis and search
    - ……

What are Big Analytics

Aster Data works on Graph

What is Big Data Analytics: Example 1

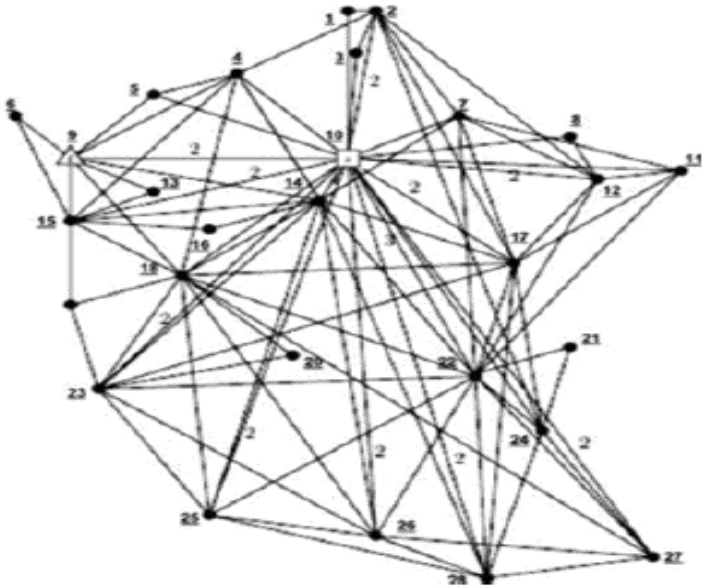
Deep graph analysis

Business Problem

- *Retailer*: How can we design marketing, packaging, and promotions to target key segments?
- *Telco:* What are the common calling patterns for a specific user group?

Analytics Problem

- What are the most important clusters and interconnections?
- What are the patterns within a cluster or set of interconnections?



- *Difficult to express* in SQL
- Requires *repeated iterations* through data

Limitation for relational databases(3)

- Design for relational data, but not suitable for
  - Graph data，Geo-spatial data，unstructured data
- Limited Scalability
  - No RDBMS has been deployed onto a cluster of more than 1000 nodes
- Separation of Data Storage and Data Analytics
  - Data migration
  - Difficulty for parallel
  - ……

Limitation for relational databases(4)

- Extending relational database
  - Relational table sharding
    - Depending on the program
    - Data size increase, need resharding
  - De-normalization for relational table to improve the performance
    - Increase more redundancy data
    - Increase the cost to maintain data consistence

Relational databases cannot solve those challenges. We need new types of databases

NoSQL databases

NoSQL DEFINITION:

- Next Generation Databases mostly addressing some of the points: being non-relational, distributed, open-source and horizontally scalable
- Non-SQL or Not only SQL
- Watch a video about NoSQL from Jens Dittrich:
- Say No! No! and No! CIDR 2013

Types and examples of NoSQL databases

Types	Examples
Column	Accumulo, Cassandra, Druid, HBase, Vertica
Document	HyperDex, Lotus Notes, MarkLogic, MongoDB, OrientDB, Qizx, RethinkDB
Key-value:	Aerospike, CouchDB, Dynamo, FairCom c-treeACE, FoundationDB, HyperDex, MemcacheDB, MUMPS, Oracle NoSQL databases
Graph	Allegro, InfiniteGraph, MarkLogic, Neo4J, OrientDB, Virtuoso, Stardog
Multi-model	Alchemy Database, ArangoDB, CortexDB, FoundationDB, MarkLogic, OrientDB

Column stores

- A column-oriented DBMS is a database management system (DBMS) that stores data tables as sections of columns of data rather than as rows of data.

- This column-oriented DBMS has advantages for data warehouses, clinical data analysis, customer relationship management (CRM) systems, and library card catalogs, and other ad hoc inquiry systems

Example of column stores

RowId	EmpId	Name	Age
1	123	Anna	34
2	456	Mikko	30
3	789	Emilia	44

Row-oriented storage:
1:123,Anna,34; 2:456,Mikko,30;3:789,Emilia,44

Column-oriented storage:
123:1,456:2,789:3; Anna:1, Mikko:2,Emilia:3;34:1,30:2,44:3

Key-value stores

- Key-value (KV) stores use the associative array as their fundamental data model.

- In this model, data is represented as a collection of key-value pairs, such that each possible key appears at most once in the collection.

Example of Key-value stores

RowId	EmpId	Name	Age
1	123	Anna	34
2	456	Mikko	30
3	789	Emilia	44

1: (123,Anna,34); 2: (2,456,Mikko,30); 3: (789,Emilia,44)

Insertion of a column and a record in Key-value stores

RowId	EmpId	Name	Age	Salary
1	123	Anna	34	
2	456	Mikko	30	
3	789	Emilia	44	
4	147	Joha	28	3000

1: (123,Anna,34); 2: (2,456,Mikko,30); 3: (789,Emilia,44);
4: (147,Joha,28,3000)

Document store

- The central concept of a document store is the notion of a "document".
- Encodings in use include XML, YAML, and JSON as well as binary forms like BSON.
- Documents are addressed in the database via a unique key that represents that document.

Example of document store

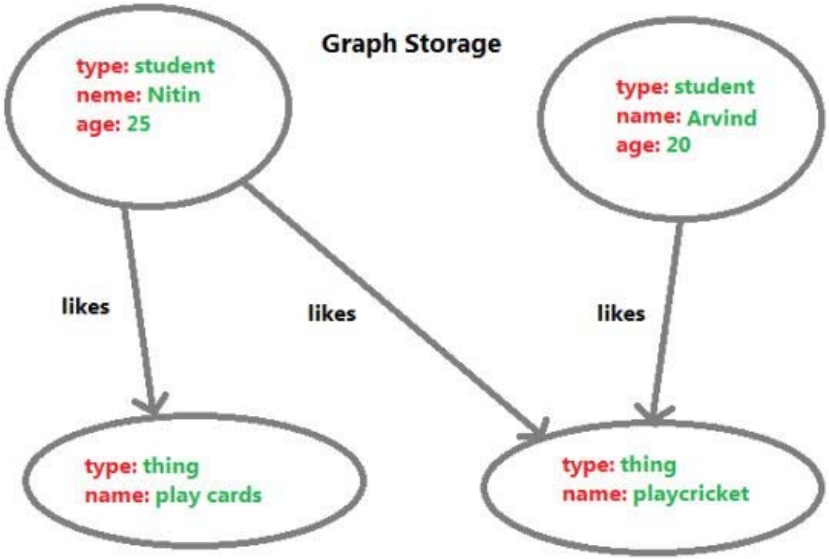
University of Helsinki
Yliopistonkatu 4,
00100 Helsinki
Finland

XML: <contact>
<company> Universtiy of Helsinki </company>
<address> Yliopistonkatu 4 </address >
<city>Helsinki</city>
<zip> 00100 </zip>
<country>Finland</country>
</contact>

JSON: “contact": {
“company": "Universtiy of Helsinki",
" address ": " Yliopistonkatu 4 ",
“city": " Helsinki ",
“zip": "00100“,
“country”:”Finland”
},

Graph stores

- Designed for graph data
- Applications: social relations, public transport links, road maps or network topologies, etc.



Multi-model stores

- Support multiple data models against a single, integrated backend: Document, graph, relational, and key-value models are examples of data models

Database	Key-value	SQL	Document	Graph	Object	Transaction
OrientDB	Yes	Yes	Yes	Yes	Yes	Full ACID, even distributed
CouchDB	Yes	Yes	Yes	No	Yes	
Marklogic	Yes	Yes	Yes	Yes	No	Full ACID

Summary

- Relational databases is very successful to manage table and relational data, but it has limitations for managing big data.
- NOSQL databases is a general term, which includes five types of data stores.
- NOSQL database are starting to gain market traction