§0 Introduction

Qin Zhang
Self introduction: my research interests

- **Algorithms for Big Data:**
  - streaming/sketching algorithms;
  - algorithms on distributed data;
  - I/O-efficient algorithms;
  - data structures;

- **Complexity:**
  - communication complexity.

I am a theoretician, and occasionally work on databases and data mining
Self introduction: my research interests

- **Algorithms for Big Data:**
  streaming/sketching algorithms;
  algorithms on distributed data;
  I/O-efficient algorithms;
  data structures;

- **Complexity:**
  communication complexity.

I am a theoretician, and occasionally work on databases and data mining

You may ask: “why do you teach databases (and probably make our lives harder)”?
Self introduction: my research interests

- **Algorithms for Big Data**: streaming/sketching algorithms; algorithms on distributed data; I/O-efficient algorithms; data structures;
- **Complexity**: communication complexity.

I am a theoretician, and occasionally work on databases and data mining

You may ask: “why do you teach databases (and probably make our lives harder)”?

Hope you will not ask me again after this course :) I am learning together with you.
What does a typical undergrad database course cover?
How to represent data?

How to represent the data in the computer?

<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>Length</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Wars</td>
<td>1977</td>
<td>124</td>
<td>color</td>
</tr>
<tr>
<td>Mighty Ducks</td>
<td>1991</td>
<td>104</td>
<td>color</td>
</tr>
<tr>
<td>Wayne’s World</td>
<td>1992</td>
<td>95</td>
<td>color</td>
</tr>
</tbody>
</table>

OR?
How to represent the data in the computer?

<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>Length</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star Wars</td>
<td>1977</td>
<td>124</td>
<td>color</td>
</tr>
<tr>
<td>Mighty Ducks</td>
<td>1991</td>
<td>104</td>
<td>color</td>
</tr>
<tr>
<td>Wayne’s World</td>
<td>1992</td>
<td>95</td>
<td>color</td>
</tr>
</tbody>
</table>

OR?

OR?
**How to operate on data?**

Given the data, say, a set of tables, how to answer queries?

**Difficulty:** Queries may depend crucially on the data in all tables.

### Product

<table>
<thead>
<tr>
<th>PName</th>
<th>Price</th>
<th>Category</th>
<th>Manufacturer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gizmo</td>
<td>19.99</td>
<td>Gadgets</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>Powergizmo</td>
<td>29.99</td>
<td>Gadgets</td>
<td>GizmoWorks</td>
</tr>
<tr>
<td>SingleTouch</td>
<td>149.99</td>
<td>Photography</td>
<td>Canon</td>
</tr>
<tr>
<td>MultiTouch</td>
<td>203.99</td>
<td>Household</td>
<td>Hitachi</td>
</tr>
</tbody>
</table>

### Company

<table>
<thead>
<tr>
<th>cName</th>
<th>StockPrice</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>GizmoWorks</td>
<td>25</td>
<td>USA</td>
</tr>
<tr>
<td>Canon</td>
<td>65</td>
<td>Japan</td>
</tr>
<tr>
<td>Hitachi</td>
<td>15</td>
<td>Japan</td>
</tr>
</tbody>
</table>

**Q:** Find all products under price 200 manufactured in Japan?
How to operate on data? (cont.)

### SQL

```sql
SELECT x.PName, x.Price
FROM Product x, Company y
WHERE x.Manufacturer = y.CName
  AND y.Country = 'Japan'
  AND x.Price <= 200
```

### Relational Algebra

\[
\pi_{PName, \text{Price}} \\
(\sigma_{\text{Price} \leq 200 \land \text{Country} = \text{'Japan'}} (\text{Product} \bowtie_{\text{Manufacturer} = \text{CName}} \text{Company}))
\]
How to speed up the operation?

Relational operations can sometimes be computed much faster if we have precomputed a suitable data structure on the data. This is called **Indexing**.
Relational operations can sometimes be computed much faster if we have precomputed a suitable data structure on the data. This is called **Indexing**.

Most notably, two kinds of index structures are essential to database performance:

1. **B-trees**.
2. **External hash tables**.

For example, hash tables may speed up relational operations that involve finding all occurrences in a relation of a particular value.
How to make a good operation plan?

How to optimize the orders of the operations?

\[ R(A, B, C, D), S(E, F, G) \]

Find all pairs \((x, y), x \in R, y \in S\) such that
(1) \(x.D = y.E\), (2) \(x.A = 5\) and (3) \(y.G = 9\)

\[ \sigma_{A=5 \land G=9}(R \boxtimes_{D=E} S) = \sigma_{A=5}(R) \boxtimes_{D=E} \sigma_{G=9}(S) \]

Q: Use the LHS or RHS?
How to deal with transactions?

Transactions with the ideal ACID properties resolve the semantic problems that arise when many concurrent users access and change the same database.

- Atomicity (= recovery)
- Consistency
- Isolation (= concurrency control)
- Durability
How to deal with transactions?

Transactions with the ideal ACID properties resolve the semantic problems that arise when many concurrent users access and change the same database.

- Atomicity (= recovery)
- Consistency
- Isolation (= concurrency control)
- Durability

We will talk about how transactions are implemented using locking and timestamp mechanisms.

This knowledge is useful in database programming, e.g., it makes it possible in some cases to avoid (or reduce) rollbacks of transactions, and generally make transactions wait less for each other.
Database =

**Logic**
(express the query)

**Algorithm**
(solve the query)

**System**
(implementation)
Database =

Logic
(express the query)

Algorithm
(solve the query)

System
(implementation)

Implementation
(see B662 Database System and Internal Design)
Database =

**Logic**
(express the query)

- Data Representation, Relational Algebra, SQL (Datalog), etc.

**Algorithm**
(solve the query)

- Indexing, Query Optimization, Concurrency Control, etc.

**System**
(implementation)

**Implementation**
(see B662 Database System and Internal Design)
Summarize

Database =

**Logic**
(express the query)
Data Representation, Relational Algebra, SQL (Datalog), etc.

**Algorithm**
(solve the query)
Indexing, Query Optimization, Concurrency Control, etc.

And you need **math**!!

**System**
(implementation)

Implementation
(see B662 Database System and Internal Design)

Concept (our focus)
What’s more in this course?
Advanced topics

Beyond ”SQL, Relational Algebra, Data Models, Storage, Views and Indexing, Query Processing, Query Optimization, Transaction Recovery, Concurrency Control”

I will give you a taste of

1. **Data Privacy** (Data Suppression, Differential Privacy)
2. **External Memory a.k.a. I/O-Efficient** Algorithms (Sorting, List Ranking)
3. **Streaming** Algorithms (Sampling, Heavy Hitters, Distinct Elements)
4. **Data Integration / Cleaning** (Deduplication)
5. **MapReduce**
Other important topics in databases

More but probably will not cover

1. **Tree-based data models** e.g., XML
2. **Graph-based data models** e.g., RDF
3. **Spatial** databases
4. **Parallel and Distributed** databases
   partly covered in MapReduce
5. **Social Networks**
6. **Uncertainty** in databases
   etc.
Tentative course plan

Part 0 : Introductions

Part 1 & 2 : Basics
  – SQL, Relational Algebra
  – Data Models, Storage, Indexing

Part 3 : Optimization

Part 4 : Transactions

Part 5 : Data Privacy

Part 6 : I/O-Efficient Algorithms

Part 7 : Streaming Algorithms

Part 8 : Data Integration

Part 9 : MapReduce
Tentative course plan

Part 0 : Introductions

Part 1 & 2 : Basics
- SQL, Relational Algebra
- Data Models, Storage, Indexing

Part 3 : Optimization

Part 4 : Trasactions

Part 5 : Data Privacy

Part 6 : I/O-Efficient Algorithms

Part 7 : Streaming Algorithms

Part 8 : Data Integration

Part 9 : MapReduce

We will also have some student presentations at the end of the course
Resources

- Main reference book (we will go beyond this)
  - Database Systems: The Complete Book
    by Hector Garcia-Molina, Jeff Ullman
    and Jennifer Widom, 2nd Edition

- Other reference books (undergrad textbooks ... )
  - Database Management Systems
    by Ramakrishnan and Guhrke, 3rd Edition
  - Database System Concepts
    by UllSilberschatz, Korth and Sudarshan,
    6th Edition
Other reference books (cont.):

- **Readings in Database Systems “Red book”**
  Hellerstein and Stonebraker, eds., 4th Edition
  (Will be one of our readings)

- **Foundations of Databases: The Logical Level “Alice book”**
  by Abiteboul, Hull, Vianu

- **Concurrency Control and Recovery in Database Systems**
  by Bernstein, Hadzilacos, Goodman

Other reference books (cont.):

- **Algorithms and Data Structures for External Memory**
  by Vitter

- **Data Streams: Algorithms and Applications**
  by S. Muthukrishnan
Resources (cont.)

- Other reference books (cont.):
  - **Algorithms and Data Structures for External Memory**
    by Vitter
    

  - **Data Streams: Algorithms and Applications**
    by S. Muthukrishnan
    
    [http://www.cs.rutgers.edu/~muthu/stream-1-1.ps](http://www.cs.rutgers.edu/~muthu/stream-1-1.ps)

These are surely not enough, and sometimes dated. Do you want to learn more? Reading *original papers*! In fact, some of my slides are directly from VLDB tutorials.
Instructors

- Instructor: Qin Zhang
  Email: qzhangcs@indiana.edu
  Office hours: Tuesday 2:45pm-3:45pm
  (Lindley 215E temporary, then Lindley 430A)

- Associate Instructors:
  - Erfan Sadeqi Azer
  - Le Liu
  - Yifan Pan
  - Ali Varamesh
  - Prasanth Velamala

Office hours: Posted on course website
Assignments 50% : Three written assignments (each 10%). Solutions should be typeset in LaTeX (highly recommended) or Word.

And one reading assignment (20%) (next slide for details)

Selected/volunteer students will give presentations

Exams 50% : Mid-term (20%) and Final (30%).
Assignments 50% : Three written assignments (each 10%). Solutions should be typeset in LaTeX (highly recommended) or Word.

And one reading assignment (20%) (next slide for details)

Selected/volunteer students will give presentations

Exams 50% : Mid-term (20%) and Final (30%).

Use $A, B, \ldots$ for each item (assignments, exams). Final grade will be a **weighted average** (according to XX%).
One or a group of two read some (1 to $+\infty$) papers/surveys/articles and write a report (4 pages for one, and 8 pages for a group of two) on what **you think** of the articles you read (not just a repeat of what they have said).

Topics can be found in redbook http://redbook.cs.berkeley.edu/bib4.html, and more topics on the course website “More reading topics” (google the papers / surveys yourself; contact AI if you cannot find it).

Selected students/groups (volunteer first) will give 25mins talks (20mins presentation +5mins Q&A) in class. The best 1/3 individuals/groups will get a **bonus** in their final grades. A **penalty** will be given if you agree to give a talk but cannot do at the end, while the quality of the talk is irrelevant.
LaTeX: Highly recommended tools for assignments/reports

1. Read wiki articles:
   http://en.wikipedia.org/wiki/LaTeX
2. Find a good LaTeX editor.
3. Learn how to use it, e.g., read “A Not So Short Introduction to LaTeX 2e” (Google it)
Prerequisite

Participants are expected to have a background in algorithms and data structures. For example, have taken

1. C241 Discrete Structures for Computer Science
2. C343 Data Structures
3. B403 Introduction to Algorithm Design and Analysis

or equivalent courses, and know some basics of databases.
Is this a course good for my job hunting in industry?

Yes, if you get to know some advanced concepts in databases, that will certainly help.

But, this is a course on theoretical foundations of databases, but not designed for teaching commercially available techniques and not a programming language (SQL? PHP?) course, and not a “hands on” course (this is not a course for professional training; this is a graduate course in a major research university thus should be much more advanced)
Is this a course good for my job hunting in industry?

Yes, if you get to know some advanced concepts in databases, that will certainly help.

But, this is a course on **theoretical foundations of databases**, but **not designed for teaching commercially available techniques** and **not a programming language (SQL? PHP?) course**, and **not a “hands on” course** (this is not a course for professional training; this is a graduate course in a major research university thus should be much more advanced)

I haven’t taken B403 “Introduction to Algorithm Design and Analysis” or equivalent courses. Can I take the course? Or, will this course fit me?

Generally speaking, this is an advanced course. It will be difficult if you do not have enough background. You can take into consideration the touch-base exam.
Open / change your views of the world (of databases)
The goal of this course

Open / change your views of the world (of databases)

Seriously, it is not just SQL programming.

Read “The relational model is dead, SQL is dead, and I don't feel so good myself”
Big Data
Big Data

- Big data is everywhere
  - Walmart: over 2.5 petabytes of sales transactions
  - Google: an index of over 19 billion web pages
  - Facebook: over 40 billion of pictures
  - ...
Big Data

- Big data is everywhere
  - Walmart: over 2.5 petabytes of sales transactions
  - Google: an index of over 19 billion web pages
  - Facebook: over 40 billion of pictures
  - ...

- Magazine covers

Nature ’06  Nature ’08  CACM ’08  Economist ’10
Source and challenge

- Source
  - Retailer databases: *Amazon, Walmart*
  - Logistics, financial & health data: *Stock prices*
  - Social network: *Facebook, twitter*
  - Pictures by mobile devices: *iphone*
  - Internet traffic: *IP addresses*
  - New forms of scientific data: *Large Synoptic Survey Telescope*
Source and challenge

Source

- Retailer databases: Amazon, Walmart
- Logistics, financial & health data: Stock prices
- Social network: Facebook, twitter
- Pictures by mobile devices: iphone
- Internet traffic: IP addresses
- New forms of scientific data: Large Synoptic Survey Telescope

Challenge

- Volume
- Velocity
- Variety (Documents, Stock records, Personal profiles, Photographs, Audio & Video, 3D models, Location data, . . . )
Source and challenge

■ Source
  ● Retailer databases: *Amazon, Walmart*
  ● Logistics, financial & health data: *Stock prices*
  ● Social network: *Facebook, twitter*
  ● Pictures by mobile devices: *iphone*
  ● Internet traffic: *IP addresses*
  ● New forms of scientific data: *Large Synoptic Survey Telescope*

■ Challenge
  ● Volume
  ● Velocity
  ● Variety (Documents, Stock records, Personal profiles, Photographs, Audio & Video, 3D models, Location data, . . . )

} The main technical challenges
What does Big Data really mean?

- We don’t define Big Data in terms of TB, PB, EB, …
- The data is too big to fit in memory. What can we do?
What does Big Data really mean?

- We don’t define Big Data in terms of TB, PB, EB, …

- The data is too big to fit in memory. What can we do?
  - Store them in the disk, and read/write a block of data each time
What does Big Data really mean?

- We don’t define Big Data in terms of TB, PB, EB, ... 

- The data is too big to fit in memory. What can we do?
  - Store them in the disk, and read/write a block of data each time
  - Processing one by one as they come, and throw some of them away on the fly.
What does Big Data really mean?

- We don’t define Big Data in terms of TB, PB, EB, . . .

- The data is too big to fit in memory. What can we do?
  - Store them in the disk, and read/write a block of data each time
  - Processing one by one as they come, and throw some of them away on the fly.
  - Store in multiple machines, which collaborate via communication
What does Big Data really mean?

- We don’t define Big Data in terms of TB, PB, EB, ...  
  
- The data is too big to fit in memory. What can we do?
  
  - Store them in the disk, and read/write a block of data each time
  
  - Processing one by one as they come, and throw some of them away on the fly.

  - Store in multiple machines, which collaborate via communication

- RAM model does not fit

  - A processor and an infinite size memory

  - Probing each cell of the memory has a unit cost
Big Data:
A marketing buzzword??
Big Data:
A marketing buzzword??

A good reading topic
Popular models for big data
(see another slides)
Summary for the introduction

- We have discussed topics that will be covered in this course.

- We have introduced some models for big data computation.

- We have talked about the course plan and assessment.
Thank you!

Questions?

A few introductory slides are based on Rasmus Pagh’s slides
http://www.itu.dk/people/pagh/ADBT06/