CMSC424: Database Design
Introduction/Overview

Instructor: Amol Deshpande
amol@cs.umd.edu
Today

- Motivation: Why study databases? What is databases?
- Administrivia
  - Workload etc.
- Current Industry Outlook
- A typical DBMS at a glance
- No laptop use allowed in the class!!
Some To-Dos

- Sign up for Piazza!

- Set up the computing environment (project0), and make sure you can run Vagrant+VirtualBox, PostgreSQL, IPython, etc.

- Upcoming: Reading Homework 1, Project 1: SQL
Motivation: Data Overload

- Explosion of data, in pretty much every domain
  - Sensing devices and sensor networks that can monitor everything 24/7 from temperature to pollution to vital signs
  - Increasingly sophisticated smart phones
  - Internet, social networks makes it easy to publish data
  - Scientific experiments and simulations produce astronomical volumes of data
  - Internet of Things
  - **Dataification**: taking all aspects of life and turning them into data (e.g., what you like/enjoy turned into a stream of your "likes")

- How to handle that data? How to extract interesting actionable insights and scientific knowledge?

- Data volumes expected to get much worse
Four V’s of Big Data

- **Increasing data Volumes**
  - **Scientific data**: 1.5GB/genome -- can be sequenced in .5 hrs; LHC generates 100TB of data a day
  - 500M tweets per day (as of 2013)
  - As of 2012: 2.5 Exabytes of data created every day
  - EBay: Two data warehouses with 7.5PB and 40PB
  - Walmart: 583 terabytes of sales and inventory data
  - FICO monitors 2.5 billion active accounts worldwide

- **Variety:**
  - Structured data, spreadsheets, photos, videos, natural text, ...

- **Velocity**

- **Veracity**
Four V’s of Big Data

- Increasing data Volumes
- Variety
- Velocity
  - Sensors everywhere -- can generate tremendous volumes of "data streams"
  - Real-time analytics requires data to be consumed as fast as it is generated
- Veracity
  - How do you decide what to trust? How to remove noise? How to fill in missing values?
Big Data and Data Science to the Rescue

- Terms increasingly used synonymously: also data analytics, data mining, business intelligence
  - Loosely used for any process where interesting things are inferred from data
  - Google search: “How Big Data Will Change”
- Data scientist called the sexiest job of the 21st century
  - The term has becoming very muddled at this point
- Overhyped words
  - We are headed toward the trough of Disillusionment
No: Extracting insights and knowledge from data very important, and will continue to increase in importance
  ◦ Big data techniques are revolutionizing things in many domains like Education, Food Supply, Disease Epidemics, ...

But: it is not much different from what we, especially statisticians, have been doing for many years

What is different?
  ◦ Much more data is digitally available than was before
  ◦ Inexpensive computing + Cloud + Easy-to-use programming frameworks = Much easier to analyze it
  ◦ Often: large-scale data + simple algorithms > small data + complex algorithms
    • Changes how you do analysis dramatically
Motivation: Data Overload

- How do we do anything with this data?

- Where and how do we store it?
  - Disks are doubling every 18 months or so -- not enough
  - In many cases, the data is not actually recorded as it is; summarized first

- What if the disks crash?
  - Very common, especially with 10,000’s of disks

- How do we ensure “correctness”?
  - What if the system crashes in the middle of an ATM transaction?
    - Can’t have money disappearing
  - What happens when a million people try to buy tickets to <your favorite artist>’s concert at the same time?
Motivation: Data Overload

- What to do with the data? How to process/analyze it?
  - text search?
    - Very limited
  - “find the stores with the maximum increase in sales in last month”
    - We can’t expect the users to write Java programs
  - “how much time from here to Pittsburgh if I start at 2pm?”
    - Data is there; more will be soon (GPS, live traffic data)
    - Requires predictive capabilities
  - Increasing need to convert “information” to “knowledge”: Data mining
    - “How many DVDs should we order?” (Netflix)
    - Find videos with this type of an event (say car break-ins)
    - Mine the “blogs” to detect “buzz”
Motivation: Data Overload

- **Speed !!**
  - With TB’s of data, just finding something (even if you know what), is not easy
    - Reading a file with TB of data can take hours
  - Imagine a bank and millions of ATMs
    - How much time does it take you to do a withdrawal?
    - The data is not local

- **How do we guarantee the data will be there 10 years from now?**

- **Privacy and security !!!**
  - Every other day we see some database leaked on the web
  - How to make sure different users’ data is protected from each other
Why not use file systems?

- Drawbacks of using file systems to store data:
  - Data redundancy and inconsistency
    - Multiple file formats, duplication of information in different files
  - Difficulty in accessing data
    - Need to write a new program to carry out each new task
  - Data isolation — multiple files and formats
  - Integrity problems
    - Integrity constraints (e.g., account balance > 0) become “buried” in program code rather than being stated explicitly
    - Hard to add new constraints or change existing ones
Why not use file systems?

- Drawbacks of using file systems to store data:
  - Atomicity of updates
    - Failures may leave database in an inconsistent state with partial updates carried out
    - Example: Transfer of funds from one account to another should either complete or not happen at all
  - Concurrent access by multiple users
    - Concurrent access needed for performance
    - Uncontrolled concurrent accesses can lead to inconsistencies
      - Example: Two people reading a balance (say 100) and updating it by withdrawing money (say 50 each) at the same time
  - Security problems
    - Hard to provide user access to some, but not all, data
Today

- Motivation: Why study databases? What is databases?
- Administrivia
  - Workload etc.
- Current Industry Outlook
- A typical DBMS at a glance
- No laptop use allowed in the class!!
Today

- Motivation: Why study databases? What is databases?
  - Key Concept: Data Modeling
  - Key Concept: Data Abstraction
  - Database Design

- Administrivia
  - Workload etc.

- Current Industry Outlook

- A typical DBMS at a glance

- No laptop use allowed in the class!!
DBMSs to the Rescue

- Provide a systematic way to answer many of these questions...
- Aim is to allow easy management of high volumes of data
  - Storing, Updating, Querying, Analyzing ....

What is a Database?
- A large, integrated collection of (mostly structured) data
- Typically models and captures information about a real-world enterprise
  - Entities (e.g. courses, students)
  - Relationships (e.g. John is taking CMSC 424)
- Usually also contains:
  - Knowledge of constraints on the data (e.g. course capacities)
  - Business logic (e.g. pre-requisite rules)
  - Encoded as part of the data model (preferable) or through external programs
**DBMSs to the Rescue**

- Massively successful for *highly structured data*
  - Why? Structure in the data (if any) can be exploited for ease of use and efficiency
    - If there is no structure in the data, hard to do much
    - Contrast managing emails vs managing photos
  - Much of the data we need to deal with is highly structured
  - Some data is *semi-structured*
    - E.g.: Resumes, Webpages, Blogs etc.
  - Some has complicated structure
    - E.g.: Social networks
  - Some has no structure
    - E.g.: Text data, Video/Image data etc.
Structured vs Unstructured Data

- A lot of the data we encounter is structured
  - Some have very simple structures
    - E.g. Data that can be represented in tabular forms
  - Significantly easier to deal with
  - We will focus on such data for much of the class

<table>
<thead>
<tr>
<th>Account</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td>bname</td>
<td>cname</td>
</tr>
<tr>
<td>acct_no</td>
<td>cstreet</td>
</tr>
<tr>
<td>balance</td>
<td>ccity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Downtown</th>
<th>Jones</th>
<th>Main</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mianus</td>
<td>Smith</td>
<td>North</td>
</tr>
<tr>
<td>Perry</td>
<td>Hayes</td>
<td>Harrison</td>
</tr>
<tr>
<td>R.H</td>
<td>Curry</td>
<td>Rye</td>
</tr>
<tr>
<td></td>
<td>Lindsay</td>
<td>Pittsfield</td>
</tr>
<tr>
<td>A-101</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-215</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A-305</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td></td>
</tr>
<tr>
<td></td>
<td>700</td>
<td></td>
</tr>
<tr>
<td></td>
<td>400</td>
<td></td>
</tr>
<tr>
<td></td>
<td>350</td>
<td></td>
</tr>
</tbody>
</table>
Some data has a little more complicated structure

- E.g graph structures
  - Map data, social networks data, the web link structure etc

- Can convert to tabular forms for storage, but may not be optimal

- Queries often reason about graph structure
  - Find my “Erdos number”
  - Suggest friends based on current friends

- Growing importance in recent years in a variety of domains: Biological, social networks, web...
Structured vs Unstructured Data

- Increasing amount of data in a semi-structured format
  - XML – Self-describing tags (HTML ?)
  - Complicates a lot of things
  - We will discuss this toward the end

- A huge amount of data is unfortunately unstructured
  - Books, WWW
  - Amenable to pretty much only text search... so far
    - Information Retrieval research deals with this topic
  - What about Google search?
    - Google search is mainly successful because it uses the structure (in its original incarnation)

- Video ? Music ?
  - Can represent in DBMS‘s, but can’t really operate on them

circle size == page importance == pagerank
more incoming links → higher pagerank
incoming links from important pages → higher pagerank
What about a Database System?

- A DBMS is a software system designed to store, manage, facilitate access to databases.

  Provides:
  - Data Definition Language (DDL)
    - For defining and modifying the schemas
  - Data Manipulation Language (DML)
    - For retrieving, modifying, analyzing the data itself
  - Guarantees about correctness in presence of failures and concurrency, data semantics etc.

Common use patterns
- Handling transactions (e.g. ATM Transactions, flight reservations)
- Archival (storing historical data)
- Analytics (e.g. identifying trends, Data Mining)
Relational DBMS: SQL

- **SQL** (sequel): Structured Query Language

- **Data definition (DDL)**
  - `create table instructor (
    ID char(5),
    name varchar(20),
    dept_name varchar(20),
    salary numeric(8,2))`

- **Data manipulation (DML)**
  - Example: Find the name of the instructor with ID 22222
    ```sql
    select name
    from instructor
    where instructor.ID = '22222'
    ```
Logistics

- **Instructor:** Amol Deshpande
  - 3221 AV Williams Bldg
  - [amol@cs.umd.edu](mailto:amol@cs.umd.edu)
  - Class Webpage:
    - Off of [http://www.cs.umd.edu/~amol](http://www.cs.umd.edu/~amol),
    - Or [http://www.cs.umd.edu/class](http://www.cs.umd.edu/class)
    - Or through ELMS

- **Email to me:** write CMSC424 in the title
  - Piazza (public or private messages) much preferred

- **TAs:** Souvik Bhattacharjee, Hui Miao, Parth Desai
Logistics

- **Textbook:**
  - Database System Concepts
    - Sixth Edition
    - Abraham Silberschatz, Henry F. Korth, S. Sudarshan

- Lecture notes will be posted on the webpage

- Piazza
  - We will use this in place of a newsgroup
  - First resort for any questions
  - General announcements will be posted there
  - Register today!
Workload:

- 6 (individual) programming projects (30%)
  - 10 late days in total, no more than 4 for any project
- 2 midterms (25%), Final (25%)
- Reading homeworks (12%)
  - One every week (can get full credit with 12/14)
  - Assigned reading, simple questions on the reading (to ensure you read it) and homework on the previous week’s material
    - Readings will refer to the Sixth edition of the book
  - Expect to spend about 1.5-2 hours on each
    - With the exception of the ones for the cancelled classes
- Class participation (7%)
  - May do in-class activities later
- Meet the instructor (1%)
Logistics

- Project 1: SQL (out by tomorrow)
  - May want to get started on this soon since it covers the same stuff as the first reading homework

<table>
<thead>
<tr>
<th>Reading Homeworks Due</th>
<th>Projects Due</th>
<th>Midterms/Final</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>Tuesday</td>
<td>Wednesday</td>
</tr>
<tr>
<td>September</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>October</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>November</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>December</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>13</td>
</tr>
</tbody>
</table>
Logistics

- Grading
  - Approximate cut-offs
    - 80+: A
    - 70+: B
    - 60+: C
    - 60-: D/F

- Most had 40+ on non-exams last two times (out of 50)
  - Exams are usually somewhat harder (no curves)
  - We would enforce a minimum passing grade on the total exam score
Some To-Dos

- Sign up for Piazza!
- Set up the computing environment (project0), and make sure you can run Vagrant+VirtualBox, PostgreSQL, IPython, etc.
- Upcoming: Reading Homework 1 (Due next Wednesday), Project 1: SQL (Sept 16)