Big Data

Sunny S Chung
A Brief History

Relational database management systems

Time

1975-1985
1985-1995
1995-2005
2005-2010
2020

Let us first see what a relational database system is
Data Management

User/Application

Query
Query
Query

Data

DataBase Management System (DBMS)
**Example: At a Company**

Query 1: Is there an employee named “Nemo”?  
Query 2: What is “Nemo’s” salary?  
Query 3: How many departments are there in the company?  
Query 4: What is the name of “Nemo’s” department?  
Query 5: How many employees are there in the “Accounts” department?

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>DeptID</th>
<th>Salary</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Nemo</td>
<td>12</td>
<td>120K</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Dory</td>
<td>156</td>
<td>79K</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>Gill</td>
<td>89</td>
<td>76K</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>Ray</td>
<td>34</td>
<td>85K</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>IT</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>Accounts</td>
<td></td>
</tr>
<tr>
<td>89</td>
<td>HR</td>
<td></td>
</tr>
<tr>
<td>156</td>
<td>Marketing</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
DataBase Management System (DBMS)

High-level Query Q

Answer

DBMS

Translates Q into best execution plan for current conditions, runs plan

Data
Example: Store that Sells Cars

Owners of Honda Accords who are <= 23 years old

Join (Cars.OwnerID = Owners.ID)

Filter (Make = Honda and Model = Accord)

Filter (Age <= 23)

Cars

<table>
<thead>
<tr>
<th>Make</th>
<th>Model</th>
<th>OwnerID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honda</td>
<td>Accord</td>
<td>12</td>
</tr>
<tr>
<td>Toyota</td>
<td>Camry</td>
<td>34</td>
</tr>
<tr>
<td>Mini</td>
<td>Cooper</td>
<td>89</td>
</tr>
<tr>
<td>Honda</td>
<td>Accord</td>
<td>156</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

 Owners

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>Nemo</td>
<td>22</td>
</tr>
<tr>
<td>34</td>
<td>Ray</td>
<td>42</td>
</tr>
<tr>
<td>89</td>
<td>Gill</td>
<td>36</td>
</tr>
<tr>
<td>156</td>
<td>Dory</td>
<td>21</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
DataBase Management System (DBMS)

High-level Query Q

Answer

DBMS

Keeps data safe and correct despite failures, concurrent updates, online processing, etc.

Data

Translates Q into best execution plan for current conditions, runs plan
A Brief History

Relational database management systems

Time

1975-1985
1985
1985-1995
1995
1995-2005
2005
2005-2010
2010
2020

Assumptions and requirements changed over time

- Semi-structured and unstructured data (Web)
- Hardware developments
- Developments in system software
- Changes in data sizes
Big Data: How much data?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- Facebook has 36 PB of user data + 80-90 TB/day (6/2010)
- CERN’s LHC: 15 PB a year (any day now)
- LSST: 6-10 PB a year (~2015)

640K ought to be enough for anybody.

From http://www.umiacs.umd.edu/~jimmylin/
eBay Analytics Technology Highlights

>50 TB/day of new, incremental data

>100k data elements

>150^10 new records/day

>50 PB/day Processed

>50k chains of logic

>5000 business users & analysts

Active/Active

turning over a TB every 5 seconds

24x7x365 Always online

Millions of queries/day

99.98+% Availability

Near-Real-time

From: http://www.cs.duke.edu/smdb10/
NEW REALITIES

The quest for knowledge used to begin with grand theories.

Now it begins with massive amounts of data.

Welcome to the Petabyte Age.

From: http://db.cs.berkeley.edu/jmh/
• Greenplum parallel DB •
  • 42 Sun X4500s (“Thumper”) each
    with:
    • 48 500GB drives
    • 16GB RAM
    • 2 dual-core Opterons

• Big and growing
  • 200 TB data (mirrored)
  • Fact table of 1.5 trillion rows
  • Growing 5TB per day
  • 4-7 Billion rows per day

Also extensive use of R and Hadoop

Yahoo! runs a 4000 node Hadoop cluster (probably the largest).
Overall, there are 38,000 nodes running Hadoop at Yahoo!

From: http://db.cs.berkeley.edu/jmh/
As reported by FAN, Feb, 2009
A SCENARIO FROM FAN

How many female WWF fans under the age of 30 visited the Toyota community over the last 4 days and saw a Class A ad?

How are these people similar to those that visited Nissan?

Open-ended question about statistical densities (distributions)

From: http://db.cs.berkeley.edu/jmh/
SQL or MapReduce
Sequential code in a variety of languages
- Perl
- Python
- Java
- R
- Mix and Match!

From: http://db.cs.berkeley.edu/jmh/
The Next Gen = Cloud Computing

I can develop and deploy quickly in a Cloud!

I can do my job from anywhere!

We can Back Up our data center in the cloud!

What is important to learn

- Principles of query processing (35%)
  - Indexes
  - Query execution plans and operators
  - Query optimization
- Data storage (15%)
  - Databases Vs. Filesystems (Google/Hadoop Distributed FileSystem)
  - Data layouts (row-stores, column-stores, partitioning, compression)
- Scalable data processing (40%)
  - Parallel query plans and operators
  - Systems based on MapReduce
  - Scalable key-value stores
  - Processing rapid, high-speed data streams
- Concurrency control and recovery (10%)
  - Consistency models for data (ACID, BASE, Serializability)
  - Write-ahead logging