Who we are:
Database Research - Provenance, Integration, and more hot stuff

Boris Glavic

Department of Computer Science

September 24, 2013
Hi, I am Boris Glavic, Assistant Professor
Hi, I am Boris Glavic, Assistant Professor

I am a database guy!
Hi, I am Boris Glavic, Assistant Professor
I am a database guy!
I will tell you:
1) Why DBs are important
2) Why DBs are interesting
3) My Research
Why are Databases important?

What do DBs do?

1. Provide persistent storage
2. Efficient declarative access to data ⇒ Querying!
3. Protection from hardware/software failures
4. Safe concurrent access to data
Why are Databases important?

Who uses DBs?

- Most big software systems involve DBs!
  - Business Intelligence ⇒ E.g., IBM Cognos
  - Web based systems
Why are Databases important?

Who uses DBs?

- Most big software systems involve DBs!
  - Business Intelligence ⇒ E.g., IBM Cognos
  - Web based systems
- Also limited scale projects
  - Amarok
  - Your Web Content Management System

Every big company uses DBs to some extent
- banks
- insurance
- government agencies

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY
Why are Databases important?

<table>
<thead>
<tr>
<th>Who uses DBs?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Most big software systems involve DBs!</strong></td>
</tr>
<tr>
<td>- Business Intelligence ⇒ E.g., IBM Cognos</td>
</tr>
<tr>
<td>- Web based systems</td>
</tr>
<tr>
<td><strong>Also limited scale projects</strong></td>
</tr>
<tr>
<td>- Amarok</td>
</tr>
<tr>
<td>- Your Web Content Management System</td>
</tr>
<tr>
<td><strong>Every big company uses DBs to some extend</strong></td>
</tr>
<tr>
<td>- banks</td>
</tr>
<tr>
<td>- insurance</td>
</tr>
<tr>
<td>- government agencies</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>
Why are Databases important?

Who produces DBs?

- Traditional Relational Database Systems is big business
  - IBM ⇒ DB2
  - Oracle ⇒ Oracle :-)
  - Microsoft ⇒ SQLServer
  - Open Source Systems ⇒ MySQL, PostgreSQL

Emerging Distributed Systems with DB characteristics

- Cloud storage and Key-value stores ⇒ Amazon S3, Google Big Table, . . .

- Big Data Analytics ⇒ Hadoop, Google Map & Reduce, . . .
Why are Databases important?

Who produces DBs?

- Traditional Relational Database Systems is big business
  - IBM ⇒ DB2
  - Oracle ⇒ Oracle :-) 
  - Microsoft ⇒ SQLServer
  - Open Source Systems ⇒ MySQL, PostgreSQL
- Emerging Distributed Systems with DB characteristics
  - Cloud storage and Key-value stores ⇒ Amazon S3, Google Big Table, . . .
  - Big Data Analytics ⇒ Hadoop, Google Map & Reduce, . . .
Why are Databases interesting?
Why are Databases interesting?

Pragmatic Perspective

- Background in databases make you competitive in the job market ;-)}
Why are Databases interesting?

Systems and Theoretical Research

- Databases has a strong systems aspect
  - Hacking complex and large systems
  - Low-level optimizations
    - Cache-conscious algorithms
    - Exploit modern hardware
Why are Databases interesting?

Systems and Theoretical Research

- Databases has a strong systems aspect
  - Hacking complex and large systems
  - Low-level optimizations
    - Cache-conscious algorithms
    - Exploit modern hardware
- Databases have a strong theoretical foundation
  - Complexity of answering queries
  - Expressiveness of query languages
  - Cost of query evaluation
Why are Databases interesting?

Connection to many other CS fields

- Distributed Systems
  - Getting more and more important
- Compilers
- Modelling
- AI and Machine Learning
  - Data Mining
- Operating and File Systems
My Research

**Topics**

- **Data Provenance**
  - Where did my data come from?
- **Data Integration**
  - How to integrate data from different sources?
- **Data Stream Management**
  - How to query streaming data (sensors, stock analysis)?
- ...
Provenance in Databases

Given a piece of data

- How do we know . . .
  - which data it is derived from?
  - which transformations (SQL) were used to create it?
  - who created it?
  - . . .

Example

<table>
<thead>
<tr>
<th>shop</th>
<th>rev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migros</td>
<td>125</td>
</tr>
<tr>
<td>Coop</td>
<td>25</td>
</tr>
</tbody>
</table>
Given a piece of data

- How do we know...
  - which data it is derived from?
  - which transformations (SQL) were used to create it?
  - who created it?
  - ...

Example

Compute the revenue for each shop as sum of prices of items sold

```
SELECT shop, sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

<table>
<thead>
<tr>
<th>shop</th>
<th>rev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migros</td>
<td>125</td>
</tr>
<tr>
<td>Coop</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>shop</th>
<th>itemId</th>
</tr>
</thead>
<tbody>
<tr>
<td>s1</td>
<td>Migros</td>
</tr>
<tr>
<td>s2</td>
<td>Migros</td>
</tr>
<tr>
<td>s3</td>
<td>Coop</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>i1</td>
<td>100</td>
</tr>
<tr>
<td>i2</td>
<td>10</td>
</tr>
<tr>
<td>i3</td>
<td>25</td>
</tr>
</tbody>
</table>
Provenance in Databases

Given a piece of data
• How do we know . . .
  • which data it is derived from?
  • which transformations (SQL) where used to create it?
  • who created it?
  • . . .

Definition (Data Provenance)
Information about the origin and creation process of data.

Example

<table>
<thead>
<tr>
<th>shop</th>
<th>rev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migros</td>
<td>125</td>
</tr>
<tr>
<td>Coop</td>
<td>25</td>
</tr>
</tbody>
</table>

```
SELECT shop, 
  sum(price) AS rev 
FROM sales, items 
WHERE itemId = id 
GROUP BY shop
```

<table>
<thead>
<tr>
<th>shop</th>
<th>itemId</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migros</td>
<td>1</td>
</tr>
<tr>
<td>Migros</td>
<td>3</td>
</tr>
<tr>
<td>Coop</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
</tr>
</tbody>
</table>
Provenance Application - Query Debugging

Trace Source of Errors
- Incorrect query output
- Caused by which source data?

Example

```
SELECT shop, 
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

<table>
<thead>
<tr>
<th>shop</th>
<th>rev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migros</td>
<td>125</td>
</tr>
<tr>
<td>Coop</td>
<td>25</td>
</tr>
<tr>
<td>. . .</td>
<td>. .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>shop</th>
<th>itemId</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migros</td>
<td>1</td>
</tr>
<tr>
<td>Migros</td>
<td>3</td>
</tr>
<tr>
<td>Coop</td>
<td>3</td>
</tr>
<tr>
<td>. . .</td>
<td>. .</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>. .</td>
<td>. .</td>
</tr>
</tbody>
</table>

---

IIT College of Science and Letters
ILLINOIS INSTITUTE OF TECHNOLOGY
Provenance Application - Query Debugging

Trace Source of Errors
- Incorrect query output
- Caused by which source data?

Example

```
SELECT shop, 
    sum(price) AS rev
FROM sales, items
WHERE itemId = id
GROUP BY shop
```

<table>
<thead>
<tr>
<th>shop</th>
<th>rev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migros</td>
<td>125</td>
</tr>
<tr>
<td>Coop</td>
<td>25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>shop</th>
<th>itemId</th>
</tr>
</thead>
<tbody>
<tr>
<td>Migros</td>
<td>1</td>
</tr>
<tr>
<td>Migros</td>
<td>3</td>
</tr>
<tr>
<td>Coop</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>id</th>
<th>price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>25</td>
</tr>
</tbody>
</table>
**Provenance Extension of the Relational Model**
- Extended relational Database (PostgreSQL)
- On-demand generation of fine-grained provenance
- “Use SQL to generate and query the provenance of SQL”
- [http://cs.iit.edu/~dbgroup/research/perm.php](http://cs.iit.edu/~dbgroup/research/perm.php)

**Contributions**
- Different types of provenance
- Provenance for complex SQL features: Aggregation, Nested Subqueries, Set operations, ...
- Powerful query support for provenance and data (SQL)
- For large databases (Efficiency)
Provenance using Temporal Databases

Collaboration with Oracle

- Use temporal database techniques to compute provenance for
  - Past queries
  - Updates
  - Transactions

Temporal Databases

- Databases where old versions of updated or deleted rows are stored for later access
- SQL access: Give me the version of table $R$ as it was at time $t_0$
Native Database Provenance

Integrate Provenance into the Database Core

- New provenance-aware physical operators
- Provenance-aware query optimization
- Storing provenance data as queries

Potential Contributions

- Several orders of magnitude speed-up
- Small storage requirements
- Lots of coding fun ;-)
SELECT NULL AS PId, City, NULL AS ManagerId
FROM WorksOn;
Understanding and Debugging Data Exchange

- Complex multi-step, error-prone process
- Many sources of error:
  - Faulty source data
  - Incorrect transformations
- Hard to trace error source

How to help the user?
- Provide information that aids in debugging
- Allow for combination and filtering
  ⇒ Query language
Understanding and Debugging Data Exchange

- Complex multi-step, error-prone process
- Many sources of error:
  - Faulty source data
  - Incorrect transformations
- Hard to trace error source

How to help the user?
- Provide information that aids in debugging
- Allow for combination and filtering \(\Rightarrow\) Query language
Vagabond (Integration and Provenance)

**Vagabond**
- **Vagabond**: Generation, ranking, and visualization of explanations for errors
- **Input**: Set of attribute values in the target that are erroneous
- **Output**: Ranking of potential explanations for these errors

**Challenges**
- **Number of potential explanations**: Exponential
- **How to rank?** What is a ’good’ explanation?
- **How to generate explanations for large error sets?**
  Database-side processing
### Example

**Source Schema S**

<table>
<thead>
<tr>
<th>IT</th>
<th>Web</th>
<th>Toronto</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT</td>
<td>Big Data</td>
<td>Chicago</td>
</tr>
<tr>
<td>Sales</td>
<td>Mobile</td>
<td>New York</td>
</tr>
</tbody>
</table>

**Target Schema T**

<table>
<thead>
<tr>
<th></th>
<th>NULL</th>
<th>Toronto</th>
<th>NULL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NULL</td>
<td>Chicago</td>
<td>NULL</td>
</tr>
<tr>
<td></td>
<td>NULL</td>
<td>New York</td>
<td>NULL</td>
</tr>
</tbody>
</table>

```sql
SELECT NULL AS PId, City, NULL AS ManagerId
FROM WorksOn;
```
Questions?

Info
- **Homepage**: http://www.cs.iit.edu/~glavic/
- **DBGroup**: http://www.cs.iit.edu/~dbgroup/
- **Office**: 226 C

Open RA Positions
- Ph.D. RA positions in database research

Master Thesis and Graduate Research Projects
- [http://www.cs.iit.edu/~dbgroup/research/studentinfo.html](http://www.cs.iit.edu/~dbgroup/research/studentinfo.html)
- Ask me if you are interested

Short-term Undergraduate and Graduate Projects (CS 597)
- Good first step to get involved with research
- [http://www.cs.iit.edu/~dbgroup/research/studentinfo.html](http://www.cs.iit.edu/~dbgroup/research/studentinfo.html)
SELECT PROVENANCE * FROM ...

Parser & Analyzer

Postgres Parser

Postgres Analyser

Query Tree

Perm Module

Rewritten Query Tree

Perm Module

Optimizer

Executor

Query Plan

Execution Engine

Query Results

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>prov_a</th>
<th>prov_b</th>
</tr>
</thead>
<tbody>
<tr>
<td>123</td>
<td>'hello'</td>
<td>2.45</td>
<td></td>
</tr>
<tr>
<td>445</td>
<td>'test'</td>
<td>1.333</td>
<td></td>
</tr>
</tbody>
</table>