

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 \Rightarrow 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



Joomla!



COGNOS
AN IBM COMPANY



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



Joomla!



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
 - ...



Joomla!™



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



Why are Databases important?

Who produces DBs?

- Traditional Relational Database Systems is big business
 - IBM \Rightarrow DB2
 - Oracle \Rightarrow Oracle :-)
 - Microsoft \Rightarrow SQLServer
 - Open Source Systems \Rightarrow MySQL, PostgreSQL
- Emerging Distributed Systems with DB characteristics
 - Cloud storage and Key-value stores \Rightarrow Amazon S3, Google Big Table, ...
 - Big Data Analytics \Rightarrow 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



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) where used to create it?
 - who created it?
 - ...

Example

		result	
		shop	rev
t ₁		Migros	125
t ₂		Coop	25



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

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

Example

	shop	rev
t ₁	Migros	125
t ₂	Coop	25

↑

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

	shop	itemId
s ₁	Migros	1
s ₂	Migros	3
s ₃	Coop	3

↑ sales

	id	price
i ₁	1	100
i ₂	2	10
i ₃	3	25

↑ items



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?
 - ...

Definition (Data Provenance)

Information about the **origin** and **creation process** of data.

Example

	shop	rev
t ₁	Migros	125
t ₂	Coop	25

↑

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

	shop	itemId
s ₁	Migros	1
s ₂	Migros	3
s ₃	Coop	3

↑ sales

	id	price
i ₁	1	100
i ₂	2	10
i ₃	3	25

↑ items



Provenance Application - Query Debugging

Trace Source of Errors

- Incorrect query output
- Caused by which source data?

Example

result		
	shop	rev
t ₁	Migros	125
t ₂	Coop	25

↑

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

↑

sales

	shop	itemId
s ₁	Migros	1
s ₂	Migros	3
s ₃	Coop	3

↑

items

	id	price
i ₁	1	100
i ₂	2	10
i ₃	3	25



Provenance Application - Query Debugging

Trace Source of Errors

- Incorrect query output
- Caused by which source data?

Example

		result	
		shop	rev
t ₁		Migros	125
t ₂		Coop	25
	

↑

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

↑

sales

	shop	itemId
s ₁	Migros	1
s ₂	Migros	3
s ₃	Coop	3

↑

items

	id	price
i ₁	1	100
i ₂	2	10
i ₃	3	25



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>



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

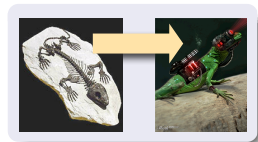
- Databases where old versions of updated or deleted rows are stored for later access
- SQL access: Give me the version of table R as it was at time t_0



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 ;-)



Data Exchange

Source Schema S

Target Schema T

WorksOn(Department,Project,City) **Projects**(PId, City, ManagerId)



M



IT	Web	Toronto
IT	Big Data	Chicago
Sales	Mobile	New York

NULL	Toronto	NULL
NULL	Chicago	NULL
NULL	New York	NULL

```
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



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?

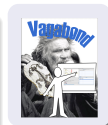
- Provide information that aids in debugging
- Allow for combination and filtering \Rightarrow Query language



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



Vagabond (Integration and Provenance)

Example

Source Schema S

WorksOn(Dep,Project,City)

IT	Web	Toronto
IT	Big Data	Chicago
Sales	Mobile	New York

Target Schema T

Projects(PId, City, ManagerId)

NULL	Toronto	NULL
NULL	Chicago	NULL
NULL	New York	NULL

```
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>
- 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>



Architecture Perm

