CS520
Data Integration, Warehousing, and Provenance

Course Info

IIT DBGroup

Boris Glavic
http://www.cs.iit.edu/~glavic/
http://www.cs.iit.edu/~cs520/
http://www.cs.iit.edu/~dbgroup/

Outline

0) Course Info
1) Introduction
2) Data Preparation and Cleaning
3) Schema mappings and Virtual Data Integration
4) Data Exchange
5) Data Warehousing
6) Big Data Analytics
7) Data Provenance

What is information integration?
• Combination of data and content from multiple sources into a common format
  – Completeness
  – Correctness
  – Efficient

Why Information Integration?
• Data is already available, right?
• …, but
• Heterogeneity
  – Structural
    • Data model (relational, XML, unstructured)
    • Schema (if there)
  – Semantic
    • Naming and identity conflicts
    • Data conflicts
  – Syntactic
    • Interfaces (web form, query language, binary file)

Why Information Integration?
• Autonomy
  – Sources may not give you unlimited access
    • Web form only support a fixed format of queries
    • Does not allow access to unlimited amounts of data
  – Source may not be available all the time
    • Naming and identity conflicts
    • Data conflicts
  – Data, schema, and interfaces of sources may change
    • Potentially without notice

About me
“Real World” Examples?

- **Portal websites**
  - Flight websites (e.g., Expedia) gather data from multiple airlines, hotels

- **Google News**
  - Integrates information from a large number of news sources

- **Science:**
  - Biomedical data source

- **Business**
  - Warehouses: integrate transactional data

---

Example Integration Problem [1]

- Integrate stock ticker data from two web services A and B
  - **Service A:** Web form (Company name, year)
  - **Service B:** Web form (year)

---

Example Integration Problem [2]

- **Service A:**
  - `<Stock>`
    - `<Company>` IBM</Company>
    - `<DollarValue>`155.8</DollarValue>
    - `<Month>`12</Month>
  - `<Stock>`

- **Service B:**
  - `<Stock>`
    - `<Company>`International Business Machines</Company>
    - `<Date>`2014-08-01</Date>
    - `<Value>`106.8</Value>
    - `<Currency>`Euro</Currency>

---

Example Integration Problem [2]

- **Service A:**
  - `<Stock>`
    - `<Company>`IBM</Company>
    - `<DollarValue>`155.8</DollarValue>
    - `<Month>`12</Month>
  - `<Stock>`

- **Service B:**
  - `<Stock>`
    - `<Company>`International Business Machines</Company>
    - `<Date>`2014-08-01</Date>
    - `<Value>`106.8</Value>
    - `<Currency>`Euro</Currency>

---

Example Integration Problem [3]

- **Service A:**
  - `<Stock>`
    - `<Company>`IBM</Company>
    - `<DollarValue>`155.8</DollarValue>
    - `<Month>`12</Month>
  - `<Stock>`

- **Service B:**
  - `<Stock>`
    - `<Company>`International Business Machines</Company>
    - `<Date>`2014-08-01</Date>
    - `<Value>`106.8</Value>
    - `<Currency>`Euro</Currency>

---

Example Integration Problem [2]

- **Global Schema**
  - `<Stock>`
    - `<Company>`IBM</Company>
    - `<Value>`</Value>
    - `<Month>`</Month>
    - `<Year>`</Year>

---

Example Integration Problem [2]

- SQL interface for integrated service
  ```sql
  SELECT month, value
  FROM ticker
  WHERE year = 2014
  AND cmp = 'IBM'
  ```

- **Service A:** (IBM, 2014)
- **Service B:** (2014)
Example Integration Problem [4]

- For web service A we can either
  - Get stocks for IBM in all years
  - Get stocks for all companies in 2014
  - Get stocks for IBM in 2014
- Trade-off between amount of processing that we have to do locally, amount of data that is shipped, …

Example Integration Problem [5]

- Service A: (IBM, 2014)
- Service B: (2014)

Example Integration Problem [6]

- Service A:
  `<Stock>
   <Company>IBM</Company>
   <DollarValue>155.8</DollarValue>
   <Month>12</Month>
  </Stock>

- Service B:
  `<Stock>
   <Company>International Business Machines</Company>
   <Date>2014-12-01</Date>
   <Value>106.8</Value>
   <Currency>Euro</Currency>
  </Stock>

Example Integration Problem [7]

- IBM vs. Integrated Business Machines

Example Integration Problem [8]

- Granularity of time attribute
  - Month vs. data
- What if both services return different values (after adapting granularity)
  - Average?
  - Median?
  - Trust-based?

Example Integration Problem [9]

- Return final results:
  `<Stock>
   <Month>01</Month>
   <Value>105</Value>
 </Stock>

- `<Stock>
   <Month>12</Month>
   <Value>107</Value>
 </Stock>`
### Why hard?

- **System challenges**
  - Different platforms (OS/Software)
  - Efficient query processing over multiple heterogeneous systems
- **Social challenges**
  - Find relevant data
  - Convince people to share their data
- **Heterogeneity of data and schemas**
  - A problem that even exists if we use same system

**Often called AI-complete**

- Meaning: “It requires human intelligence to solve the problem”
- Unlikely that general completely automated solutions will exist
- So why do we still sit here
  - There exist automated solutions for relevant less general problems
  - Semi-automated solutions can reduce user effort (and may be less error prone)

### AI completeness

- Yes, but still why is this problem really so hard?
  - **Lack of information**: e.g., the attributes of a database schema have only names and data types, but no computer interpretable information on what type of information is stored in the attribute
  - **Undecidable computational problems**: to decide whether a user query can be answered from a set of sources that provide different views on the data requires **query containment** checks which are undecidable for certain query types

### Relevant less general problems

- **Data cleaning**
  - Clean dirty data before integration
  - Conformance with a set of constraints
  - Deal with missing and outlier values
- **Entity resolution**
  - Determine which objects from multiple dataset represent the same real world entity
- **Data fusion**
  - Merge (potentially conflicting) data for the same entity
- **Schema matching**
  - Given two schemas determine which elements store the same type of information
- **Schema mapping**
  - Describe the relationships between schemas
    - Allows us to rewrite queries written against one schema into queries of another schema
    - Allows us to translate data from one schema into
- **Virtual data integration**
  - Answer queries written against a **global mediated schema** by running queries over **local sources**
- **Data exchange**
  - Map data from one schema into another
- **Warehousing: Extract, Transform, Load**
  - Clean, transform, fuse data and load it into a data warehouse to make it available for analysis
Relevant less general problems

• **Integration in Big Data Analytics**
  – Often “pay-as-you-go”:
    • No or limited schema
    • Engines support wide variety of data formats

• **Provenance**
  – Information about the origin and creation process of data
  – Very important for integrated data
    • E.g., “from which data source is this part of my query result”

Webpage and Faculty

• **Course Info**
  – Course Webpage: [http://cs.iit.edu/~cs520](http://cs.iit.edu/~cs520)
    • Used for announcements
    • Use it to discuss with me, TA, and fellow students

• **Faculty**
  – Boris Glavic ([http://cs.iit.edu/~glavic](http://cs.iit.edu/~glavic))
  – Email: bgavic@iit.edu
  – Phone: 312.567.5205
  – Office: Stuart Building, room 226C
  – Office Hours: Mondays, 12pm-1pm (and by appointment)

TAs

• **TAs**
  – TBA

Workload and Grading

• **Exams (60%)**
  – Final

• **Homework Assignments** (preparation for exams!)
  – Practice theory for final exam
  – Practice the tools we discuss in class

• **Literature Review (40%)**
  – In groups of 2 students
  – Topics will be announced soon
  – You have to read a research paper
  – Papers will be assigned in the first few weeks of the course
  – You will give a short presentation (15min) on the topic in class
  – You will write a report summarizing and criticizing the paper (up to 4 pages)

Course Objectives

• Understand the problems that arise with querying heterogeneous and autonomous data sources

• Understand the differences and similarities between the data integration/exchange, data warehouse, and Big Data analytics approaches

• Be able to build parts of a small data integration pipeline by “glueing” existing systems with new code

Course Objectives cont.

• Have learned formal languages for expressing schema mappings

• Understand the difference between virtual and materialized integration (data integration vs. data exchange)

• Understand the concept of data provenance and know how to compute provenance
Fraud Policies

• All work has to be original!
  – Cheating = 0 points for review/exam
  – Possibly E in course and further administrative sanctions
  – Every dishonesty will be reported to office of academic honesty

• Late policy:
  – -20% per day
  – You have to give your presentation to pass the course!
  – No exceptions!

Fraud Policies cont.

• Literature Review:
  – Every student has to contribute in both the presentation and report!
  – Don’t let others freeloard on you hard work!
    • Inform me or TA immediately

Reading and Prerequisites

• Textbook: Doan, Halevy, and Ives.
  – Principles of Data Integration, 1st Edition
  – Morgan Kaufmann
  – Publication date: 2012
  – Prerequisites:
    • CS 425

Additional Reading

• Papers assigned for literature review
• Optional: Standard database textbook

Outline

0) Course Info
1) Introduction
2) Data Preparation and Cleaning
3) Schema mappings and Virtual Data Integration
4) Data Exchange
5) Data Warehousing
6) Big Data Analytics
7) Data Provenance
1. Introduction

Overview

- Topics covered in this part
  - Heterogeneity and Autonomy
  - Data Integration Tasks
  - Data Integration Architectures (Methods)
  - Some Formal Background (sorry!)

1.1 System Heterogeneity

- Hardware/Software
  - Different hardware capabilities of sources
  - Different protocols, binary file formats, …
  - Different access control mechanism
- Interface Heterogeneity
  - Different interfaces for accessing data from a source
    - HTML forms
    - XML-Web services
    - Declarative language

1.1 Heterogeneity + Autonomy

- Taxonomy of Heterogeneity

1.1.1 System Heterogeneity

- Hardware/Software
  - Different hardware capabilities of sources
  - Mobile phone vs. server: Cannot evaluate cross-product of two 1GB relations on a mobile phone
  - Different protocols, binary file formats, …
  - Order information stored in text files: line ending differs between Mac/Window/Linux, character encoding
- Different access control mechanism
  - FTP-access to files: public, ssh authentication, …
1.1 System Heterogeneity

- Interface Heterogeneity
  - Different interfaces for accessing data from a source
    - HTML forms
    - Services (SOA)
    - Declarative language
    - Files
    - Proprietary network protocol
    - ...

- Interface Heterogeneity – Expressiveness
  - Keyword-search vs. query language
  - **Predicates:** equality (=), inequality (<, !=)
  - **Logical connectives:** conjunctive (AND), disjunctive (OR), negation
  - **Complex operations:** aggregation, quantification
  - **Limitations:** restriction to particular tables, predicates, fixed queries with parameters, ...

- Interface Heterogeneity – Examples
  - Google search (+/-, site:, intitle:, filetype:)
  - SQL
  - Web-form (with DB backend?)
1.1 System Heterogeneity

• Interface Heterogeneity – Examples
  – Email-client

1.1 System Heterogeneity

• Problems with interface heterogeneity
  – Global query language is more powerful
    • User queries may not be executable
    • Integration system has to evaluate part of the query
  – Bound parameters are incompatible with query
    • User query may not be executable

1.1 System Heterogeneity

• Example: more expressive global language
  – SQL with one table
    • books (title, author, year, isbn, genre)
  – Web form for books about history shown below
  – What problems do may arise translating user queries?

1.1 System Heterogeneity

• Integration system has to process part of the query

1.1 System Heterogeneity

• Query requires multiple requests

1.1 System Heterogeneity

• Query cannot be answered

1.1 System Heterogeneity

• Example: more expressive global language
  – SQL with one table
    • books (title, author, year, isbn, genre)
  – Web form for books about history shown below
  – What problems do may arise translating user queries?
1.1 Heterogeneity + Autonomy

- Taxonomy of Heterogeneity

1.1 Structural Heterogeneity

- Data model
  - Different semantic/expressiveness
  - Different structure

- Schema
  - Integrity constraints, keys
  - Schema elements:
    - use attribute or separate relations
  - Structure:
    - e.g., normalized vs. denormalized relational schema

1.1 Structural Heterogeneity

- Example: data model
  - Relational model
  - XML model
  - JSON
  - OO

- Person and their addresses

1.1 Structural Heterogeneity

- Schema
  - Modeling choices
    - Relation vs. attribute
    - Attribute vs. value
    - Relation vs. value
  - Naming
  - Normalized vs. denormalized (relational concept)
  - Nesting vs. reference
1.1 Structural Heterogeneity

• **Relation-relation conflicts**
  – Naming conflicts
    • Relations with different name representing the same data *(synonym)*
    • Relations with same name representing different information *(homonym)*
  – Structural conflicts
    • Missing attributes
    • Many-to-one
    • Missing, but derivable attributes
  – Integrity constraint conflicts

---

Example: Conflicts between relations

- `Person(id, firstname, lastname, male, female)`
- `Manager(id, name, gender, age)`

---

Example: Conflicts between attributes

<table>
<thead>
<tr>
<th>SSN</th>
<th>FirstName</th>
<th>LastName</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>Peter</td>
<td>Schmeter</td>
<td>30</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>Hans</td>
<td>Glanz</td>
<td>NULL</td>
</tr>
</tbody>
</table>

---

Example: Conflicts between attributes

```sql
<table>
<thead>
<tr>
<th>SSN</th>
<th>FirstName</th>
<th>LastName</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>3333333333</td>
<td>Peter</td>
<td>Schmeter</td>
<td>30</td>
</tr>
<tr>
<td>3333339999</td>
<td>Hans</td>
<td>Glanz</td>
<td>1</td>
</tr>
</tbody>
</table>
```

---

1.1 Structural Heterogeneity

• **Attribute-attribute conflicts**
  – Naming conflicts
    • Attributes with different name representing the same data *(synonym)*
    • Attributes with same name representing different information *(homonym)*
  – Default value conflict
  – Integrity constraint conflicts
    • Datatype
    • Constraints restricting values
1.1 Structural Heterogeneity

- Normalized vs. denormalized
  - E.g., relational model: Association between entities can be represented using multiple relations and foreign keys or one relation

1.1 Structural Heterogeneity

- Nested vs. flat
  - Association between entities can be represented using nesting or references (previous slides)

1.1 Structural Heterogeneity

- Problems caused by schema heterogeneity
  - Unified access to multiple schemas or integrate schemas into new schema
    - Schema level: schema mapping, model management operators, schema languages
    - Data Level: virtual data integration, data exchange, warehousing (ETL)

1.1 Heterogeneity + Autonomy

- Taxonomy of Heterogeneity

1.1 Semantic Heterogeneity

- Semantic Heterogeneity
  - Naming Conflicts
  - Identity Conflicts (Entity resolution)
  - Value Conflicts (Data Fusion)

1.1 Semantic Heterogeneity

- Naming Conflicts
  - Ontological (concepts)
    - Birds vs. Animals
  - Synonyms
    - Surname vs. last name
  - Homonyms
  - Units
    - Gallon vs. liter
  - Values
    - Manager vs. Boss
1.1 Semantic Heterogeneity

• Ontological concepts
  – Relationships between concepts
    • $A = B$ - Equivalence
    • $A \subseteq B$ - Inclusion
    • $A \cap B$ - Overlap
    • $A \neq B$ - Disjunction

Example

- Equivalence: Human vs Homo sapiens
- Inclusion: Bird vs Animal
- Overlap: Salamander vs aquatic lifeform
- Disjunction: Fish vs Mammal

1.1 Semantic Heterogeneity

• Naming concepts (synonyms)
  • Different words with same meaning

Example

- Person (Title, Name, Salary)
  - Name (First, Last, Age)

1.1 Semantic Heterogeneity

• Naming concepts (homonyms)
  • Same words with different meaning

Example

- Person (First, Last, Name)
  - Movie (Title, Year)

1.1 Semantic Heterogeneity

• Naming concepts (units)

Example

- Person (Title, Name, Salary)
  - $5$
  - CAD

1.1 Semantic Heterogeneity

• Identity Conflicts
  – What is an object?
    • E.g., multiple tuples in relational model
  – Central question:
    • Does object A represent the same entity as B
  – This problem has been called
    • Entity resolution
    • Record linkage
    • Deduplication
    • …
1.1 Semantic Heterogeneity

- Identity Conflicts

<table>
<thead>
<tr>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM (300000000 USA)</td>
</tr>
<tr>
<td>IBM (International Business Machines Corporation, USA)</td>
</tr>
</tbody>
</table>

1.1 Semantic Heterogeneity

- Value Conflicts
  - Objects representing the same entities have conflicting values for semantically equivalent attributes
    - We have to identify that these objects are the same entity first!
  - Resolving such conflicts require **Data Fusion**
    - Pick value from conflicting values
    - Numerical methods: e.g., average
    - Preferred value
    - …

1.1 Autonomy

- How autonomous are data sources
  - One company
    - Can enforce, e.g., schema and software
  - …
  - The web
    - Website decides
      - Interface
      - Determines access restrictions and limits
      - Availability
      - Format
      - Query restrictions
      - …

1.2 Data integration tasks

- Cleaning and prepreparation
- Entity resolution
- Data Fusion
- Schema matching
- Schema mapping
- Query rewrite
- Data translation

1.3 Data integration architectures

- Virtual data integration
- Data Exchange
- Peer-to-peer data integration
- Datawarehousing
- Big Data analytics

1.4 Formal Background

- Query Equivalence
  - Complexity for different query classes
- Query Containment
  - Complexity for different query classes
- Datalog
  - Recursion + Negation
- Integrity Constraints
  - Logical encoding of integrity constraints
- Similarity Measures/Metrics
1.4 Integrity constraints

- You know some types of integrity constraints already
  - Functional dependencies
    - Keys are a special case
  - Foreign keys
    - We have not really formalized that

- Other types are
  - Conditional functional dependencies
    - E.g., used in cleaning
  - Equality-generating dependencies
  - Multi-valued dependencies
  - Tuple-generating dependencies
  - Join dependencies
  - Denial constraints
  - …

- How to manage all these different types of constraints?
  - Has been shown that these constraints can be expressed in a logical formalism.
  - Formulas which consist of relational and comparison atoms. Variables represent values
    - $R(x, y, z)$
    - $x = y$

- Types of constraints we will use a lot
  - Tuple-generating dependencies (tgds)
    - Implication with conjunction of relational atoms
      - $\forall \bar{x}: \phi(\bar{x}) \rightarrow \exists \bar{y}: \psi(\bar{x}, \bar{y})$
  - Equality-generating dependencies (egds)
    - Generalizes keys, FDs
      - $\forall \bar{x}: \phi(\bar{x}) \rightarrow \land_{k=1}^{n} x_{i_k} = x_{j_k}$

- What is datalog?
  - Prolog for databases (syntax very similar)
  - A logic-based query language
  - Queries (Program) expressed as set of rules
    - $Q(\bar{x}) : -R_1(x_1), \ldots, R_n(x_n)$.
    - One Q is specified as the answer relation (the relation returned by the query)
1.4 Datalog - Intuition

- A Datalog rule
  \[ Q(\bar{x}) : \neg R_1(x_1), \ldots, R_n(x_n). \]
- For all bindings of variables in the right-hand side (RHS) that makes the RHS true (conjunction) return bindings of \( \bar{x} \)

Example

\[ Q(\text{Name}) : \neg \text{Person}(\text{Name}, \text{Age}). \]
Return names of persons

1.4 Datalog - Syntax

- A Datalog program is a set of datalog rules
  - Optionally a distinguished answer predicate
- A Datalog rule is
  \[ Q(\bar{x}) : \neg R_1(x_1), \ldots, R_n(x_n). \]
- X’s are lists of variables and constants
- Ri’s are relation names
- Q is a relation name

1.4 Datalog - Terminology

- Left-hand side of a rule is called its head
- Right-hand side of a rule is called its body
- Relation are called predicates
- \( R(\bar{x}) \) is called an atom
- An instance \( I \) of a database is the data
- The active domain \( \text{adom}(I) \) of an instance \( I \) is the set of all constants that occur in \( I \)
  \[ Q(\bar{x}) : \neg R_1(x_1), \ldots, R_n(x_n). \]

Example

\[ Q(\text{Name}) : \neg \text{Person}(\text{Name}, \text{Age}). \]
N, A are variables
Q(N) is an atom
Person and Q are predicates

\begin{array}{|c|c|}
\hline
\text{Name} & \text{Age} \\
\hline
\text{peter} & 34 \\
\hline
\text{bob} & 45 \\
\hline
\end{array}

Activate domain
\( \text{adom}(I) = \{ \text{peter, bob, 34, 45} \} \)

1.4 Datalog - Terminology

- Intensional vs. extensional
  - Extensional database (eddb)
    - What we usually call database
  - Intensional database (idb)
    - Relations that occur in the head of rules (are populated by the query)
    - Usually we assume that these do not overlap
    \[ Q(\bar{x}) : \neg R_1(x_1), \ldots, R_n(x_n). \]

1.4 Datalog - Safety

- A datalog program is safe if all its rules are safe
- A rule is safe if all variables in \( \bar{x} \) occur in at least one \( x_i \)
  \[ Q(\bar{x}) : \neg R_1(x_1), \ldots, R_n(x_n). \]

Example

\[ Q(\text{Name}) : \neg \text{Person}(\text{Name}, \text{Age}). \] (safe)
\[ Q(\text{Name}, \text{Sal}) : \neg \text{Person}(\text{Name}, \text{Age}). \] (unsafe)
1.4 Datalog - Semantics

• The instance of an idb predicate Q in a datalog program for an edb instance I contains all facts that can be derived by applying rules with Q in the head
• A rule derives a fact Q(c) if we can find a binding of variables of the rule to constants from adom(I) such that x is bound to c and the body is true

\[ Q(\bar{x}) : \neg R_1(x_1), \ldots, \neg R_n(x_n). \]

1.4 Datalog - Semantics

Example

\[ \text{Q}(N) := \text{Person}(N,A). \]

\[ \begin{array}{c|c}
\text{N} & \text{A} \\
\hline
\text{peter} & 34 \\
\text{bob} & 34 \\
\end{array} \]

Activate domain
adom(I) = \{peter, bob, 34\}

1.4 Datalog

• Different flavors of datalog
  – Conjunctive query
    • Only one rule
    • Expressible as Select-project-join (SPJ) query in relational algebra
  – Union of conjunctive queries
    • Also allow union
    • SPJ + set union in relational algebra
    • Rules with the same head in Datalog
  – Conjunctive queries with inequalities
    • Also allow inequalities, e.g., \(<\)

1.4 Datalog

• Different flavors of datalog
  – Recursion
    • Rules may have recursion:
      – E.g., head predicate in the body
    • Fix point semantics based on immediate consequence operator
  – Negation (first-order queries)
    • Negated relational atoms allowed
    • Require that every variable used in a negated atom also occurs in at least one positive atom (safety)
  – Combined Negation + recursion
    • Stronger requirements (stratification)

1.4 Datalog

Example

Relation \text{hops(A,B)} storing edges of a graph.

\[ \begin{array}{c}
\text{Q}_{\text{hop}}(x,z) : \text{hop}(x,y), \text{hop}(x,z). \\
\text{Q}_{\text{reach}}(x,y) : \text{hop}(x,y). \\
\text{Q}_{\text{reach}}(x,z) : \text{Q}_{\text{reach}}(x,y), \text{Q}_{\text{reach}}(y,z). \\
\text{Q}_{\text{node}}(x) : \text{hop}(x,y). \\
\text{Q}_{\text{node}}(x) : \text{hop}(x,y). \\
\end{array} \]
1.4 Datalog

Example
Relation \( \text{hops}(A, B) \) storing edges of a graph.

\( Q_{\text{node}}(x) : \text{hop}(x, y). \)
\( Q_{\text{node}}(x) : \text{hop}(y, x). \)
\( Q_{\text{notReach}}(x, y) : Q_{\text{node}}(x), Q_{\text{node}}(y), \) not \( Q_{\text{reach}}(x, y). \)

1.4 Containment and Equivalence

Definition: Query Equivalence
Query \( Q \) is equivalent to \( Q' \) iff for every database instance \( I \) both queries return the same result

\[ Q \equiv Q' \iff \forall I : Q(I) = Q'(I) \]

Definition: Query Containment
Query \( Q \) is contained in query \( Q' \) iff for every database instance \( I \) the result of \( Q \) is contained in the result of \( Q' \)

\[ Q \subseteq Q' \iff \forall I : Q(I) \subseteq Q'(I) \]

1.4 Complexity of Eq. and Cont.

<table>
<thead>
<tr>
<th>Set semantics</th>
<th>Relational Algebra</th>
<th>Conjunctive Queries (CQ)</th>
<th>Union of Conjunctive Queries (UCQ)</th>
<th>Monotone Queries/ CQ*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Evaluation</td>
<td>PSPACE-complete</td>
<td>NP-complete</td>
<td>NP-complete</td>
<td>NP-complete</td>
</tr>
<tr>
<td>(Combined Complexity)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query Evaluation</td>
<td>LOGSPACE (that means in P)</td>
<td>LOGSPACE (that means in P)</td>
<td>LOGSPACE (that means in P)</td>
<td>LOGSPACE (that means in P)</td>
</tr>
<tr>
<td>(Data Complexity)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Query Equivalence</td>
<td>Undecidable</td>
<td>NP-complete</td>
<td>NP-complete</td>
<td>( \Pi_2^P )-complete</td>
</tr>
<tr>
<td>Query Containment</td>
<td>Undecidable</td>
<td>NP-complete</td>
<td>NP-complete</td>
<td>( \Pi_2^P )-complete</td>
</tr>
</tbody>
</table>
1.4 Complexity of Eq. and Cont.

<table>
<thead>
<tr>
<th>Bag semantics</th>
<th>Relational Algebra</th>
<th>Conjunctive Queries (CQ)</th>
<th>Union of Conjunctive Queries (UCQ)</th>
<th>Monotone Queries/CDQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Equivalence</td>
<td>Undecidable</td>
<td>Equivalent to graph isomorphism</td>
<td>It is in PSPACE; lower-bound unknown</td>
<td></td>
</tr>
<tr>
<td>Query Containment</td>
<td>Undecidable</td>
<td>Open Problem</td>
<td>Undecidable</td>
<td>T1^2-complete</td>
</tr>
</tbody>
</table>

1.4 Containment Mappings

- NP-completeness for set semantics CQ and UCQ for the containment, evaluation, and equivalence problems is based on reducing these problems to the same problem
  - [Chandra & Merlin, 1977]
- Notational Conventions:
  - head(Q) = variables in head of query Q
  - body(Q) = atoms in body of Q
  - vars(Q) = all variable in Q

1.4 Boolean Conjunctive Queries

- A conjunctive query is boolean if the head does not have any variables
  - Q() :- hop(x,y), hop(y,z)
  - We will use Q :- … as a convention for Q() :- …
  - What is the result of a boolean query
    - Empty result {}, e.g., no hop(x,y), hop(y,z)
    - If there are tuples matching the body, then a tuple with zero attributes is returned {}|
  - We interpret {} as false and {} as true
  - Boolean query is essentially an existential check

1.4 Boolean Conjunctive Queries

- BCQ in SQL

```
Example
Hop relation: Hop(A,B)
Q :- hop(x,y)
SELECT EXISTS (SELECT * FROM hop)
Note: in Oracle and DB2 we need a FROM clause
```

```
Example
SELECT CASE WHEN EXISTS (SELECT * FROM hop)
    THEN 1 ELSE 0
END AS x
FROM dual;
Notes:
- Oracle and DB2 FROM not optional
- Oracle has no boolean datatype
```

1.4 Boolean Conjunctive Queries

- BCQ in SQL

```
Example
Q :- hop(x,y), hop(y,z)
SELECT EXISTS
    (SELECT * FROM hop l, hop r
    WHERE l.B = r.A)
```
1.4 Containment Mappings

- How to check for containment of CQs (set)

**Definition: Variable Mapping**

A variable mapping $\psi$ from query $Q$ to query $Q'$ maps the variables of $Q$ to constants or variables from $Q'$.

**Definition: Containment Mapping**

A containment mapping from query $Q$ to $Q'$ is a variable mapping $\psi$ such that:

$\psi(\text{head}(Q)) = \text{head}(Q')$

$\forall R(x_i) \in \text{body}(Q) : \psi(x_i) \in \text{body}(Q')$

Theorem: Containment Mapping and Query Containment

Query $Q$ is contained in query $Q'$ iff there exists a containment mapping $\psi$ from $Q'$ to $Q$.

Example

$Q_1(u, z) : R(u, z)$.
$Q_2(x, y) : R(x, y)$.

$Q_1 \rightarrow Q_2 : \psi(u) = x, \psi(z) = y$

$Q_2 \rightarrow Q_1 : \psi(x) = u, \psi(y) = z$

Example

$Q_1(a, b) : R(a, b), R(c, b)$.
$Q_2(x, y) : R(x, y)$.

Do containment mappings exist?

$Q_1 \rightarrow Q_2 : \text{none exists}$

$Q_2 \rightarrow Q_1 : \psi(x) = a, \psi(y) = b$

Example

$Q_1(a, b) : R(a, b), R(c, b)$.
$Q_2(x, y) : R(x, y)$.

$Q_1 \rightarrow Q_2 : \psi(a) = x, \psi(b) = y, \psi(c) = y$

$Q_2 \rightarrow Q_1 : \psi(x) = a, \psi(y) = b$
1.4 Containment Background

- It was shown that query evaluation, containment, equivalence as all reducible to homomorphism checking for CQ
  - Canonical conjunctive query Q^I for instance I
    - Interpret attribute values as variables
    - The query is a conjunction of all atoms for the tuples
    - I = {hop(a,b), hop(b,c)} -> Q^I: hop(a,b), hop(b,c)
  - Canonical instance I^Q for query Q
    - Interpret each conjunct as a tuple
    - Interpret variables as constants
    - Q :- hop(a,a) -> I^Q = {hop(a,a)}

1.4 Containment Mappings

Example

\[ Q_1() : R(a,b), R(c,b) \]
\[ Q_2() : R(x,y) \]
\[ Q_2 \rightarrow Q_1 : \Psi(x)=a, \Psi(y)=b \]
\[ D = \{R(1,1), R(1,2)\} \]
\[ Q_1(D) = \{(1,1), (1,2)\} \]
\[ \phi(a)=1, \phi(b)=2, \phi(c)=1 \]
\[ \Psi \phi(x)=1, \Psi \phi(y)=2 \]

1.4 Containment Background

- Containment Mapping <-> Containment
- Proof idea (boolean queries)
  - (if direction)
    - Assume we have a containment mapping Q_1 to Q_2
    - Consider database D
    - Q_2(D) is true then we can find a mapping from vars(Q_2) to D
    - Compose this with the containment mapping and prove that this is a result for Q_1

Example

\[ Q_1() : R(a,b), R(c,b) \]
\[ Q_2() : R(x,y) \]
\[ Q_2 \rightarrow Q_1 : \Psi(x)=a, \Psi(y)=b \]
\[ I^{Q_2} = \{(a,b), (c,b)\} \]
\[ Q_2(I^{Q_2}) = \{()\} \]
\[ \phi(x)=a, \phi(y)=b \]
\[ \phi \text{ is our containment mapping } \Psi \]

1.4 Containment Background

- If you are not scared and want to know more:
  - Look up Chandra and Merlins paper(s)
  - The text book provides a more detailed overview of the proof approach
  - Look at the slides from Phokion Kolaitis excellent lecture on database theory
    - https://classes.soe.ucsc.edu/cmps277/Winter10/
1.4 Containment Background

• A more intuitive explanation why containment mappings work
  – Variable naming is irrelevant for query results
  – If there is a containment mapping Q to Q’
    • Then every condition enforced in Q is also enforced by Q’
    • Q’ may enforce additional conditions

1.4 Containment Mappings

Example

\[ Q_1(1): R(a,b), R(c,b). \]
\[ Q_2(1): R(x,y). \]
\[ Q_2 \Rightarrow Q_1: \Psi(x) = a, \Psi(y) = b \]

If there exists tuples \( R(a,b) \) and \( R(c,b) \)
in \( R \) that make \( Q_1 \) true, then we take \( R(a,b) \)
to fulfill \( Q_2 \).

1.4 Containment Background

• From boolean to general conjunctive queries
  – Instead of returning true or false, return bindings of variables
  – Recall that containment mappings enforce that the head is mapped to the head
  – \( \Rightarrow \) same tuples returned, but again Q’ s condition is more restrictive

1.4 Similarity Measures

• Problem faced by multiple integration tasks
  – Given two objects, how similar are they
  – E.g., given two attribute names in schema matching, given two values in data fusion/entity resolution, …

1.4 Similarity Measures

• Object models
  – Multidimensional (feature vector model)
    • Object is described as a vector of values - one for each dimension out of a given set of dimensions
    • E.g., Dimensions are gender (male/female), age (0-120), and salary (0-1,000,000). An example object is \[ \text{[male, 80, 70, 000]} \]
  – Strings
    • E.g., how similar is “Poeter” to “Peter”
  – Graphs and Trees
    • E.g., how similar are two XML models
1.4 Similarity Measures

Definition: Similarity Measure
Function \( d(p, q) \) where \( p \) and \( q \) are objects, that returns a real score with
- Non-negative: \( d(p, q) \geq 0 \)
- Symmetry: \( d(p, q) = d(q, p) \)
- Identity of indiscernibles: \( d(p, q) = 0 \) if \( p = q \)
- Triangle inequality: \( d(p, q) + d(q, r) \geq d(p, r) \)

- Interpretation: the lower the score the "more similar" the objects are
- We require \( d(p, p) = 0 \), because nothing can be more similar to an object than itself
- Note: often scores are normalized to the range \([0, 1]\)

Example
- String equality: \( d(p, q) = 0 \) if \( p = q \)
- Euclidean distance: \( d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \)
- Edit distance: \( d(p, q) = \text{minimum number of single character insertions, deletions, replacements to transform } p \text{ into } q \)

Definition: Metric
Function \( d(p, q) \) where \( p \) and \( q \) are objects, that returns a real score with
- Non-negative: \( d(p, q) \geq 0 \)
- Symmetry: \( d(p, q) = d(q, p) \)
- Identity of indiscernibles: \( d(p, q) = 0 \) if \( p = q \)
- Triangle inequality: \( d(p, q) + d(q, r) \geq d(p, r) \)

- Metric is a stricter definition
- Which of the previous similarity measure is a metric?
  - All of them!

Why do we care whether \( d \) is a metric?
- Some data mining algorithms only work for metrics
  - E.g., some clustering algorithms such as k-means
  - E.g., clustering has been used in entity resolution
- Metric spaces allow optimizations of some methods
  - E.g., Nearest Neighborhood-search: find the most similar object to an object \( p \). This problem can be efficiently solved using index structures that only apply to metric spaces

Summary
- Heterogeneity
  - Types of heterogeneity
  - Why do they arise?
  - Hint at how to address them
- Autonomy
- Data Integration Tasks
- Data Integration Architectures
- Background
  - Datalog + Query equivalence/containment + Similarity + Integrity constraints
Outline

0) Course Info
1) Introduction
2) Data Preparation and Cleaning
   3) Schema matching and mapping
   4) Virtual Data Integration
   5) Data Exchange
   6) Data Warehousing
   7) Big Data Analytics
   8) Data Provenance
2. Data Preparation and Cleaning

Boris Glavic
http://www.cs.iit.edu/~glavic/
http://www.cs.iit.edu/~cs520/
http://www.cs.iit.edu/~dbgroup/

2. Overview

- Topics covered in this part
  - Causes of Dirty Data
  - Constraint-based Cleaning
  - Outlier-based and Statistical Methods
  - Entity Resolution
  - Data Fusion

2. Causes of “Dirty” Data

- Manual data entry or result of erroneous integration (cont.)
  - Redundancy:
    - (ID: 1, City: Chicago, Zip: 60616)
    - (ID: 2, City: Chicago, Zip: 60616)
  - Inconsistent references to entities
    - Dept. of Energy, DOE, Dep. Of Energy, …

2. Cleaning Methods

- Enforce Standards
  - Applied in real world
  - How to develop a standard not a fit for this lecture
  - Still relies on no human errors
- Constraint-based cleaning
  - Define constraints for data
  - “Make” data fit the constraints
- Statistical techniques
  - Find outliers and smoothen or remove
    - E.g., use a clustering algorithm

2. Overview

- Topics covered in this part
  - Causes of Dirty Data
  - Constraint-based Cleaning
  - Outlier-based and Statistical Methods
  - Entity Resolution
  - Data Fusion

2.1 Cleaning Methods

- Constraint-based cleaning
  - Choice of constraint language
  - Detecting violations to constraints
  - Fixing violations (automatically?)

2.1 Constraint Languages

- First work focused on functional dependencies (FDs)
- Extensions of FDs have been proposed to allow rules that cannot be expressed with FDs
  - E.g., conditional FDs only enforce the FD is a condition is met
    - Finer grained control, e.g., zip -> city only if country is US
- Constraints that consider master data
  - Master data is highly reliable data such as a government issued zip, city lookup table

2.1 Example Constraints

<table>
<thead>
<tr>
<th>SSN</th>
<th>zip</th>
<th>city</th>
<th>name</th>
<th>boss</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
<td>Gert</td>
<td>50,000</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
<td>NULL</td>
<td>40,000</td>
</tr>
<tr>
<td>333-333-5599</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
<td>Hans</td>
<td>10,000</td>
</tr>
<tr>
<td>333-333-6666</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
<td>NULL</td>
<td>1,000,000</td>
</tr>
<tr>
<td>333-355-4343</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcom</td>
<td>Hans</td>
<td>20,000</td>
</tr>
</tbody>
</table>

- The zip code uniquely determines the city
- Nobody should earn more than their direct superior
- Salaries are non-negative
2.1 Example Constraints

<table>
<thead>
<tr>
<th>SSN</th>
<th>ZIP</th>
<th>CITY</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
</tr>
<tr>
<td>333-333-5599</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
</tr>
<tr>
<td>333-333-6666</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
</tr>
<tr>
<td>333-355-4343</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcom</td>
</tr>
</tbody>
</table>

2.1 Example Constraints

Example: Constraints Languages

<table>
<thead>
<tr>
<th>SSN</th>
<th>ZIP</th>
<th>CITY</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
</tr>
<tr>
<td>333-333-5599</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
</tr>
<tr>
<td>333-333-6666</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
</tr>
<tr>
<td>333-355-4343</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcom</td>
</tr>
</tbody>
</table>

\[ \text{C}_1: \text{The zip code uniquely determines the city} \]
\[ \forall \neg (E(x, y, z, u, v, w) \land E(x', y', z', u', v', w')) \land x = x' \land y \neq y') \]
\[ \text{C}_2: \text{Nobody should earn more than their direct superior} \]
\[ \forall \neg (E(x, y, z, u, v, w) \land E(x', y', z', u', v', w') \land v = u' \land w > w') \]
\[ \text{C}_3: \text{Salaries are non-negative} \]
\[ \forall \neg (E(x, y, z, u, v, w) \land w < 0) \]

2.1 Constraint based Cleaning Overview

- Define constraints
- Given database D
  - 1) Detect violations of constraints
    - We already saw example of how this can be done using queries. Here a bit more formal
  - 2) Fix violations
    - In most cases there are many different ways to fix the violation by modifying the database (called solution)
      - What operations do we allow: insert, delete, update
      - How do we choose between alternative solutions

2.1 Constraint Repair Problem

Definition: Constraint Repair Problem

Given set of constraints \( \Sigma \) and an database instance \( I \) which violates the constraints find a clean instance \( I' \) so that \( I' \) fulfills \( \Sigma \)

- This would allow us to take any \( I' \) – E.g., empty for FD constraints
- We do not want to loose the information in \( I \) (unless we have to)
- Let us come back to that later

2.1 Example Constraints

Example: Constraint Violations

<table>
<thead>
<tr>
<th>SSN</th>
<th>ZIP</th>
<th>CITY</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
</tr>
<tr>
<td>333-333-5599</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
</tr>
<tr>
<td>333-333-6666</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
</tr>
<tr>
<td>333-355-4343</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcom</td>
</tr>
</tbody>
</table>

\[ \text{C}_1: \text{zip \rightarrow city} \]
2.1 Example Constraints

Example: Constraint Violations

<table>
<thead>
<tr>
<th>SSN</th>
<th>Zip</th>
<th>City</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
</tr>
<tr>
<td>333-333-5599</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
</tr>
<tr>
<td>333-333-6666</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
</tr>
<tr>
<td>333-355-4343</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcom</td>
</tr>
</tbody>
</table>

How to repair?

- **Deletion:**
  - remove some conflicting tuples
  - quite destructive

- **Update:**
  - modify values to resolve the conflict
  - equate RHS values (city here)
  - disequate LHS value (zip)

Example: Constraint Repair

<table>
<thead>
<tr>
<th>SSN</th>
<th>Zip</th>
<th>City</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
</tr>
<tr>
<td>333-333-5599</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
</tr>
<tr>
<td>333-333-6666</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
</tr>
<tr>
<td>333-355-4343</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcom</td>
</tr>
</tbody>
</table>

How to repair:

- **Deletion:**
  - remove some conflicting tuples
  - quite destructive

- **Update:**
  - modify values to resolve the conflict
  - equate RHS values (city here)
  - disequate LHS value (zip)

2.1 Constraint based Cleaning Overview

- **How to repair?**
- **Deletion:**
  - remove some conflicting tuples
  - quite destructive
- **Update:**
  - modify values to resolve the conflict
  - equate RHS values (city here)
  - disequate LHS value (zip)
- **Insertion?**
  - Not for FDs, but e.g., FKs

2.1 Detecting Violations

- **Principle of minimality**
  - Choose repair that minimally modifies database
  - Motivation: consider the solution that deletes every tuple

Most update approaches **equate RHS** because there is usually no good way to choose LHS values unless we have **master data**

- E.g., update zip to 56423 or 52456 or 22322 ...

2.1 Detecting Violations

- **Given FD A -> B on R**
  - Recall logical representation
  - Forall X, X': R(X) and R(X') and A=A' -> B=B'
  - Only violated if we find two tuples where A=A', but B != B'
  - In datalog:
    - Q0: R(X), R(X'), A=A', B!=B'
  - In SQL:
    - SELECT EXISTS (SELECT *
                        FROM R x, R y
                        WHERE A=A' AND B>B')
2.1 Example Constraints

Example: SQL Violation Detection

Relation: Person(name, city, zip)
PSI: zip \rightarrow city
Violation Detection Query
SELECT EXISTS (SELECT * FROM Person x, Person y WHERE x.zip = y.zip AND x.city \neq y.city)
To know which tuples caused the conflict:
SELECT * FROM Person x, Person y WHERE x.zip = y.zip AND x.city \neq y.city

2.1 Fixing Violations

- Principle of minimality
  - Choose solution that minimally modifies the database
- Updates:
  - Need a cost model
- Deletes:
  - Minimal number of deletes

2.1 Constraint Repair Problem

Definition: Constraint Repair Problem (restated)
Given set of constraints \( \Sigma \) and an database instance \( I \) which violates the constraints find a clean instance \( I' \) (does not violate the constraints) with cost \( (I, I') \) being minimal

- Cost metrics that have been used
  - Deletion + Insertion
    \( \Delta(I, I') = (I - I') \cup (I' - I) \)
    - S-repair: minimize measure above under set inclusion
    - C-repair: minimize cardinality
  - Update
    - Assume distance metric \( d \) for attribute values

2.1 Cost Metrics

- Deletion + Insertion
  \( \Delta(I, I') = (I - I') \cup (I' - I) \)
  - S-repair: minimize measure above under set inclusion
  - C-repair: minimize cardinality
- Update
  - Assume single relation \( R \) with uniquely identified tuples
  - Assume distance metric \( d \) for attribute values
  - Schema(\( R \)) \rightarrow attributes in schema of relation \( R \)
  - \( t' \) is updated version of tuple \( t \)
  - Minimize: \( \sum_{t \in R} \sum_{A \in \text{Schema}(R)} d(t.A, t'.A) \)

2.1 Naïve FD Repair Algorithm

- Update
  - Assume single relation \( R \) with uniquely identified tuples
  - Assume distance metric \( d \) for attribute values
  - Schema(\( R \)) \rightarrow attributes in schema of relation \( R \)
  - \( t' \) is updated version of tuple \( t \)
  - Minimize: \( \sum_{t \in R} \sum_{A \in \text{Schema}(R)} d(t.A, t'.A) \)
  - We focus on this one
  - This is NP-hard
    - Heuristic algorithm

- FD Repair Algorithm: 1. Attempt
  - For each FD \( X \rightarrow Y \) in \( \Sigma \) run query to find pairs of tuples that violate the constraint
  - For each pair of tuples \( t \) and \( t' \) that violate the constraint
    - update \( t.Y \) to \( t'.Y \)
      - choice does not matter because cost is symmetric, right?
2.1 Constraint Repair

Example: Constraint Repair

<table>
<thead>
<tr>
<th>SSN</th>
<th>zip</th>
<th>city</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
</tr>
<tr>
<td>3334</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
</tr>
<tr>
<td>3335</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
</tr>
<tr>
<td>3336</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
</tr>
<tr>
<td>3337</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcom</td>
</tr>
</tbody>
</table>

2.1 Problems with the Algorithm

**FD Repair Algorithm: 1. Attempt**
- For each FD $X \rightarrow Y$ in $\Sigma$ run query to find pairs of tuples that violate the constraint
- For each pair of tuples $t$ and $t'$ that violate the constraint: $t.X = t'.X$ and $t.Y \neq t'.Y$
  - update $t.Y$ to $t'.Y$
  - choice does not matter because cost is symmetric, right?
  - Our updates may cause new violations!

**FD Repair Algorithm: 2. Attempt**
- $I' = I$
  - 1) For each FD $X \rightarrow Y$ in $\Sigma$ run query to find pairs of tuples that violate the constraint
  - 2) For each pair of tuples $t$ and $t'$ that violate the constraint: $t.X = t'.X$ and $t.Y \neq t'.Y$
    - update $t.Y$ to $t'.Y$
    - choice does not matter because cost is symmetric, right?
    - 3) If we changed $I'$ goto 1)
  - May never terminate
2.1 Problems with the Algorithm

- FD Repair Algorithm: 2. Attempt
  - Even if we succeed the repair may not be minimal. There may be many tuples with the same X values
    - They all have to have the same Y value
    - Choice which to update matters!

- FD Repair Algorithm: 3. Attempt
  - Equivalence Classes
    - Keep track of sets of cells (tuple, attribute) that have to have the same values in the end (e.g., all Y attribute values for tuples with same X attribute value)
    - These classes are updated when we make a choice
    - Choose Y value for equivalence class using minimality, e.g., most common value
  - Observation
    - Equivalence Classes may merge, but never split if we only update RHS of all tuples with same X at once
    - We can find an algorithm that terminates

2.1 Constraint Repair

Example: Constraint Repair

<table>
<thead>
<tr>
<th>SSN</th>
<th>Zip</th>
<th>City</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
</tr>
<tr>
<td>333-333-5599</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
</tr>
<tr>
<td>333-333-6666</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
</tr>
<tr>
<td>333-355-4343</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcolm</td>
</tr>
</tbody>
</table>

Changes: set t1.city = Chicago
Not as cheap: set t4.city and t5.city = New York

2.1 Consistent Query Answering

- As an alternative to fixing the database which requires making a choice we could also leave it dirty and try to resolve conflicts at query time
  - Have to reason over answers to the query without knowing which of the possible repairs will be chosen
  - Intuition: return tuples that would be in the query result for every possible repair

A Cost-Based Model and Effective Heuristic for Repairing Constraints by Value Modification
2.1 Constraint Repair

Example: Constraint Repair

<table>
<thead>
<tr>
<th>SSN</th>
<th>Zip</th>
<th>City</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>New York</td>
<td>Peter</td>
</tr>
<tr>
<td>333-333-9999</td>
<td>60615</td>
<td>Chicago</td>
<td>Gert</td>
</tr>
<tr>
<td>333-333-5599</td>
<td>60615</td>
<td>Schaumburg</td>
<td>Gertrud</td>
</tr>
<tr>
<td>333-333-6666</td>
<td>60616</td>
<td>Chicago</td>
<td>Hans</td>
</tr>
<tr>
<td>333-355-4343</td>
<td>60616</td>
<td>Chicago</td>
<td>Malcom</td>
</tr>
</tbody>
</table>

Changes: t1.city = Chicago
Not so change: set t4.city and t5.city = New York

2.2 Statistical and Outlier

• Assumption
  – Errors can be identified as outliers

• How do we find outliers?
  – Similarity-based:
    • Object is dissimilar to all (many) other objects
    • E.g., clustering, objects not in cluster are outliers
  – Some type of statistical test:
    • Given a distribution (e.g., fitted to the data)
    • How probable is it that the point has this value?
    • If low probability -> outlier

2.3 Entity Resolution

• Entity Resolution (ER)
  • Alternative names
    – Duplicate detection
    – Record linkage
    – Reference reconciliation
    – Entity matching
    – …
2.3 Entity Resolution

- Similarity based on similarity of attribute values
  - Which distance measure is appropriate?
  - How do we combine attribute-level distances?
  - Do we consider additional information?
    - E.g., foreign key connections
    - How similar should duplicates be?
    - E.g., fixed similarity threshold
    - How to guarantee transitivity of \( E \)
      - E.g., do this afterwards

2.3 Entity Resolution – Distance Measures

- Edit-distance
  - Measures similarity of two strings
  - \( d(s,s') = \) minimal number of insert, replace, delete operations (single character) that transform \( s \) into \( s' \)
  - Is symmetric (actually a metric)
    - Why?

Example:

<table>
<thead>
<tr>
<th>SSN</th>
<th>Zip</th>
<th>City</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>Chicago</td>
<td>Peter</td>
</tr>
<tr>
<td>3333333333</td>
<td>IL</td>
<td>60616</td>
<td>Petre</td>
</tr>
</tbody>
</table>

Example: Per attribute similarity

<table>
<thead>
<tr>
<th>SSN</th>
<th>Zip</th>
<th>City</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>333-333-3333</td>
<td>60616</td>
<td>Chicago</td>
<td>Peter</td>
</tr>
<tr>
<td>3333333333</td>
<td>IL</td>
<td>60616</td>
<td>Petre</td>
</tr>
</tbody>
</table>

Definition: Edit Distance

Given two strings \( s, s' \) we define the edit distance \( d(s,s') \) as the minimum number of single-character insert, replacements, deletions that transforms \( s \) into \( s' \).

Example:

\( \text{NEED} \rightarrow \text{STREET} \)

Trivial solution: delete all chars in \( \text{NEED} \), then insert all chars in \( \text{STREET} \)

\[ d(\text{NEED}, \text{STREET}) = 10 \]

Example:

\( \text{NEED} \rightarrow \text{STREET} \)

Minimal solution:
- insert \( T \)
- replace \( N \) with \( E \)
- replace \( D \) with \( T \)

\[ d(\text{NEED}, \text{STREET}) = 4 \]
2.3 Entity Resolution

- **Principal of optimality**
  - Best solution of a subproblem is part of the best solution for the whole problem

- **Dynamic programming algorithm**
  - $D(i,j)$ is the edit distance between prefix of len $i$ of $s$ and prefix of len $j$ of $s'$
  - $D(\text{len}(s), \text{len}(s'))$ is the solution
  - Represented as matrix
  - Populate based on rules shown on the next slide

---

**Recursive definition**

- $D(i,0) = i$
  - Cheapest way of transforming prefix $s[i]$ into empty string is by deleting all $i$ characters in $s[i]$
- $D(0,j) = j$
  - Same holds for $s'[j]$
- $D(i,j) = \min \{ D(i-1,j) + 1 \quad D(i,j-1) + 1 \quad D(i-1,j-1) + d(i,j) \}$
  - with $d(i,j) = 1$ if $s[i] \neq s'[j]$ and 0 else

---

Example:

<table>
<thead>
<tr>
<th>NEED -&gt; STREET</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

Example:

<table>
<thead>
<tr>
<th>NEED -&gt; STREET</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

---

Example:

<table>
<thead>
<tr>
<th>NEED -&gt; STREET</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>

Example:

<table>
<thead>
<tr>
<th>NEED -&gt; STREET</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>E</td>
</tr>
<tr>
<td>D</td>
</tr>
</tbody>
</table>
2.3 Entity Resolution

Example:

NEED -> STREET

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

2.3 Entity Resolution – Distance Measures

- Other sequence-based measures for string similarity
  - Needleman-Wunsch
    - Missing character sequences can be penalized differently from character changes
  - Affine Gap Measure
    - Limit influence of longer gaps
    - E.g., Peter Friedrich Mueller vs. Peter Mueller
  - Smith-Waterman Measure
    - More resistant to reordering of elements in the string
    - E.g., Prof. Franz Mueller vs. F. Mueller, Prof.

- Other sequence-based measures for string similarity
  - Jaro-Winkler
    - Consider shared prefixes
    - Consider distance of same characters in strings
    - E.g., Johann vs. ojhann vs. ohannj
  - See textbook for details!
2.3 Entity Resolution – Distance Measures

• Token-set based measures
  – Split string into tokens
    • E.g., single characters
    • E.g., words if string represents a longer text
  – Potentially normalize tokens
    • E.g., word tokens replace word with its stem
      – Generating, generated, generates are all replaced with generate
    – Represent string as set (multi-set) of tokens

Jaccard-Measure

– \( B_s = \text{Tok}(s) = \text{token set of string } s \)
– Jaccard measures relative overlap of tokens in two strings
  • Number of common tokens divided by total number of tokens

\[
d_{\text{jacc}}(s, s') = \frac{|B_s \cap B_{s'}|}{|B_s \cup B_{s'}|}
\]

Other set-based measures

– TF/IDF: term frequency, inverse document frequency
  • Take into account that certain tokens are more common than others
  • If two strings (called documents for TF/IDF) overlap on uncommon terms they are more likely to be similar than if they overlap on common terms
    – E.g., the vs. carbon nanotube structure
2.3 Entity Resolution

- **TF/IDF**: term frequency, inverse document frequency
  - Represent documents as feature vectors
    - One dimension for each term
    - Value computed as frequency times IDF
      - Inverse of frequency of term in the set of all documents
  - Compute cosine similarity between two feature vectors
    - Measure how similar they are in term distribution (weighted by how uncommon terms are)
    - Size of the documents does not matter
      - See textbook for details

2.3 Entity Resolution

- **Entity resolution**
  - Concatenate attribute values of tuples and use string similarity measure
    - Loose information encoded by tuple structure
    - E.g., [Gender: male, Salary: 9000]
      - -> "Gender: male, Salary: 9000"
      - or -> "male, 9000"
  - Combine distance measures for single attributes
    - Weighted sum or more complex combinations
      - E.g., \( d(t, t') = w_1 \times d_1(A, A') + w_2 \times d_2(B, B') \)
    - Use quadratic distance measure
      - E.g., earth-movers distance

2.3 Entity Resolution

- **Weighted linear combination**
  - Say tuples have \( n \) attributes
  - \( w_i \) predetermined weight of an attribute
  - \( d_i(t, t') \): similarity measure for the \( i \)th attribute
    \[
    d(t, t') = \sum_{i=0}^{n} w_i \times d_i(t, t')
    \]
  - Tuples match if \( d(t, t') > \beta \) for a threshold \( \beta \)

Example: Weighted sum of attribute similarities

<table>
<thead>
<tr>
<th>ID</th>
<th>SSN</th>
<th>Zip</th>
<th>City</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>333-333-3333</td>
<td>60616</td>
<td>Chicago</td>
<td>Peter</td>
</tr>
<tr>
<td>2</td>
<td>3333333333</td>
<td>IL</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assumption: SSNs and names are most important, city and zip are not very predictive

\( w_{SSN} = 0.4, w_{zip} = 0.05, w_{city} = 0.15, w_{name} = 0.4 \)

\[
\begin{align*}
    d(t, t') &= 0.4 \times 1 + 0.05 \times 0.8 + 0.15 \times 0 + 0.4 \times 0.6 \\
            &= 0.4 + 0.04 + 0 + 0.24 \\
            &= 0.68
\end{align*}
\]
2.3 Entity Resolution

- Entity resolution
  - Rule-based approach
  - Learning-based approaches
  - Clustering-based approaches
  - Probabilistic approaches to matching
  - Collective matching

Rule-based approach

- Collection (list) of rules
  - if \( d_{name}(t, t') < 0.6 \) then unmatched
  - if \( d_{zip}(t, t') = 1 \) and \( t\.country = USA \) then matched
  - if \( t\.country \neq t'.country \) then unmatched

- Advantages
  - Easy to start, can be incrementally improved
- Disadvantages
  - Lot of manual work, large rule-bases hard to understand

Learning-based approach

- Build all pairs \((t, t')\) for training dataset
  - Represent each pair as feature vector from, e.g., similarities
  - Train classifier to return \{match, no match\}

- Advantages
  - Automated
- Disadvantages
  - Requires training data

Clustering-based approach

- Apply clustering method to group inputs
  - Typically hierarchical clustering method
  - Clusters now represent entities
    - Decide how to merge based on similarity between clusters

- Advantages
  - Automated, no training data required
- Disadvantages
  - Choice of cluster similarity critical
2.3 Entity Resolution

- Entity resolution
  - Rule-based approach
  - Learning-based approaches
  - Clustering-based approaches
  - Probabilistic approaches to matching
  - Collective matching
    - See text book

2.4 Data Fusion

- Data Fusion = how to combine (possibly conflicting) information from multiple objects representing the same entity
  - Choose among conflicting values
    - If one value is missing (NULL) choose the other one
    - Numerical data: e.g., median, average
    - Consider sources: have more trust in certain data sources
    - Consider value frequency: take most frequent value
    - Timeliness: latest value

2. Overview

- Topics covered in this part
  - Causes of Dirty Data
  - Constraint-based Cleaning
  - Outlier-based and Statistical Methods
  - Entity Resolution
  - Data Fusion

Outline

0) Course Info
1) Introduction
2) Data Preparation and Cleaning
3) Schema matching and mapping
4) Virtual Data Integration
5) Data Exchange
6) Data Warehousing
7) Big Data Analytics
8) Data Provenance
3. Why matching and mapping?

• **Problem: Schema Heterogeneity**
  – Sources with different schemas store overlapping information
  – Want to be able to translate data from one schema into a different schema
    • Data warehousing
    • Data exchange
  – Want to be able to translate queries against one schema into queries against another schema
    • Virtual data integration

• Why both mapping and matching
  – Split complex problem into simpler subproblems
    • Determine matches and then correlate with constraint information into mappings
  – Some tasks only require matches
    • E.g., matches can be used to determine attributes storing the same information in data fusion
  – Mappings are naturally a generalization of matchings

3. Why matching and mapping?

• **Problem: Schema Heterogeneity**
  – We need to know how elements of different schemas are related!
  
  – **Schema matching**
    • Simple relationships such as attribute name of relation person in the one schema corresponds to attribute lastname of relation employee in the other schema
  
  – **Schema mapping**
    • Also model correlations and missing information such as links caused by foreign key constraints

3. Overview

• Topics covered in this part
  – **Schema Matching**
    – Schema Mappings and Mapping Languages
3.1 Schema Matching

- **Problem: Schema Matching**
  - Given two (or more schemas)
  - For now called **source** and **target**
  - Determine how elements are related
    - Attributes are representing the same information
      - `name` vs. `lastname`
    - Attribute can be translated into an attribute
      - `MonthlySalary * 12` vs. `YearlySalary`
    - 1-1 matches vs. M-N matches
      - `name` vs. `lastname`
      - `name` vs. `concat(firstname, lastname)`

- **Why is this hard?**
  - **Insufficient information**: schema does not capture full semantics of a domain
  - **Schemas can be misleading**:
    - E.g., attributes are not necessarily descriptive
    - E.g., finding the right way to translate attributes not obvious

- **What information to consider?**
  - Attribute names
    - or more generally element names
  - Structure
    - E.g., belonging to the same relation
  - Data
    - Not always available
  - Need to consider multiple types to get reasonable matching quality
    - Single types of information not predictable enough

- **Example: Types of Matching**

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>City</th>
<th>Office-contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>1</td>
<td>Chicago</td>
<td>(312) 123-456</td>
</tr>
<tr>
<td>Alice</td>
<td>3</td>
<td>Chicago</td>
<td>(123) 345-678</td>
</tr>
<tr>
<td>Bob</td>
<td>3</td>
<td>New York</td>
<td>(987) 654-321</td>
</tr>
</tbody>
</table>

  Based on element names we could match Office-contact to both Office-phone and Office-address.
  Based on data we could match Office-contact to both Office-phone and Home-phone.

- **Typical Matching System Architecture**
  - Determine actual matches
  - Use constraints to modify similarity matrix
  - Combine individual similarity matrices
  - Each matcher uses one type of information to compute similarity matrix
3.1 Schema Matching

• **Matcher**
  - Input: Schemas
  - Maybe also data, documentation
  - Output: Similarity matrix
  - Storing value [0,1] for each pair of elements from the source and the target schema

3.1 Schema Mapping

Example: Types of Matching

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Office-phone</th>
<th>Office-address</th>
<th>Home-phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Person</td>
<td>Person</td>
<td>Person</td>
<td>Person</td>
</tr>
<tr>
<td>Name</td>
<td>Address</td>
<td>Office-phone</td>
<td>Office-address</td>
<td>Home-phone</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Address</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.4</td>
</tr>
<tr>
<td>Id</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>City</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Office-contact</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

3.1 Schema Matching

• **Name-Based Matchers**
  - String similarities measures
    - E.g., Jaccard and other measure we have discussed
  - Preprocessing
    - Tokenization?
    - Normalization
      - Expand abbreviations and replace synonyms
    - Remove stop words
      - In, and, the

3.1 Schema Matching

• **Data-Based Matchers**
  - Determine how similar the values of two attributes are
  - Some techniques
    - Recognizers
      - Dictionaries, regular expressions, rules
    - Overlap matcher
      - Compute overlap of values in the two attributes
    - Classifiers

3.1 Schema Matching

• **Recognizers**
  - Dictionaries
    - Countries, states, person names
  - Regular expression matchers
    - Phone numbers: `{\+}\d{2} {\(}\d{3}{\) } \d{3} \d{4}

3.1 Schema Matching

• **Overlap of attribute domains**
  - Each attribute value is a token
  - Use set-based similarity measure such as Jaccard

• **Classifier**
  - Train classifier to identify values of one attribute A from the source
    - Training set are values from A as positive examples and values of other attributes as negative examples
  - Apply classifier to all values of attributes from target schema
    - Aggregate into similarity score
3.1 Schema Matching

**Combiner**
- **Input:** Similarity matrices
- **Output:** Similarity matrix

**Constraint Enforcer**
- **Input:** Similarity matrix
- **Output:** Similarity matrix

**How to search match combinations**
- Full search
  - Exponentially many combinations potentially
- Informed search approaches
  - A* search
  - Local propagation
  - Only local optimizations

**Example: Constraints**

**Constraint 1:** An attribute matched to `source.cust-phone` has to get a score of 1 from the phone regex matcher.

**Constraint 2:** Any attribute matched to `source.fax` has to have fax in its name.

**Constraint 3:** If an attribute is matched to `source.firstname` with score > 0.9 then there has to be another attribute from the same target table that is matched to `source.lastname` with score > 0.9.
3.1 Schema Matching

- **A* search**
  - Given a search problem
    - Set of states: start state, goal states
    - Transitions about states
    - Costs associated with transitions
    - Find cheapest path from start to goal states
  - Need admissible heuristics \( h \)
    - For a path \( p \), \( h \) computes lower bound for any path from start to goal with prefix \( p \)
  - Backtracking best-first search
    - Choose next state with lowest estimated cost
    - Expand it in all possible ways

- **Algorithm**
  - Data structures
    - Keep a priority queue \( q \) of states sorted on \( f(n) \)
    - Initialize with start state
    - Keep set \( v \) of already visited nodes
    - Initially empty
  - While \( q \) is not empty
    - pop state \( s \) from head of \( q \)
    - If \( s \) is goal state return
    - Foreach \( s' \) that is direct neighbor of \( s \)
      - If \( s' \) not in \( v \)
      - Compute \( f(s') \) and insert \( s' \) into \( q \)

- **Application to constraint enforcing**
  - Source attributes: \( A_1 \) to \( A_n \)
  - Target attributes: \( B_1 \) to \( B_m \)
  - States
    - Vector of length \( n \) with values \( B_i \) or * indicating that no choice has not been taken
    - \([B_1, *, *, B_3]\)
    - Initial state
    - \([*, *, *, *]\)
    - Goal states
    - All states without *

- **Match Selector**
  - Input: Similarity matrix
    - Output of the individual matchers
  - Output: Matches

- **Match Selection**
  - Merge similarity matrices produced by the matchers into single matrix
  - Typical strategies
    - Average, Minimum, Max
    - Weighted combinations
    - Some script
3.1 Schema Matching

- Many-to-many matchers
  - Combine multiple columns using a set of functions
    - E.g., concat, +, currency exchange, unit exchange
  - Large or even unlimited search space
  - Need method that explores interesting part of the search space
  - Specific searchers
    - Only concatenation of columns (limit number of combinations, e.g., 2)

3. Overview

- Topics covered in this part
  - Schema Matching
  - Schema Mappings and Mapping Languages

3.2 Schema Mapping

Example: Matching Result

<table>
<thead>
<tr>
<th>Person</th>
<th>Name</th>
<th>Address</th>
<th>City</th>
<th>Office-contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>1</td>
<td>Chicago</td>
<td>312</td>
<td>12345343453</td>
</tr>
<tr>
<td>Alice</td>
<td>3</td>
<td>Chicago</td>
<td>312</td>
<td>7777777777</td>
</tr>
<tr>
<td>Bob</td>
<td>3</td>
<td>New York</td>
<td>465</td>
<td>123123123</td>
</tr>
</tbody>
</table>

Assume: We have data in the source as shown above
What data should we create in the target? Copy values based on matches?
How do we choose values for attributes that do not have a match?
How do we combine data from different source tables?

3.2 Schema Mapping

- Matches do not determine completely how to create the target instance data! (Data Exchange)
  - How do we choose values for attributes that do not have a match?
  - How do we combine data from different source tables?
- Matches do not determine completely what the answers to queries over a mediated schema should be! (Virtual Data Integration)

3.2 Schema Mapping

- Schema mappings
  - Generalize matches
  - Describe relationship between instances of schemas
  - Mapping languages
    - LAV, GAV, GLAV
  - Mapping as Dependencies: tuple-generating dependencies
- Mapping generation
  - Input: Matches, Schema constraints
  - Output: Schema mappings
3.2 Schema Mapping

- Instance-based definition of mappings
  - Global schema $G$
  - Local schemas $S_1$ to $S_n$
  - Mapping $M$ can be expressed as for each set of instances of the local schemas what are allowed instances of the global schema
    - Subset of $(I_G \times I_1 \times \ldots \times I_n)$
  - Useful as a different way to think about mappings, but not a practical way to define mappings

3.2 Schema Mapping

- Certain answers
  - Given mapping $M$ and $Q$
  - Instances $I_1$ to $I_n$ for $S_1$ to $S_n$
  - Tuple $t$ is a certain answer for $Q$ over $I_1$ to $I_n$
    - If for every instance $I_G$ so that $(I_G \times I_1 \times \ldots \times I_n)$ in $M$, then $t \in Q(I_G)$

3.2 Schema Mapping

- Languages for Specifying Mappings
- Describing mappings as inclusion relationships between views:
  - Global as View (GAV)
  - Local as View (LAV)
  - Global and Local as View (GLAV)
- Describing mappings as dependencies
  - Source-to-target tuple-generating dependencies (st-tgds)

3.2 Schema Mapping

- Describing mappings as inclusion relationships between views:
  - Global as View (GAV)
  - Local as View (LAV)
  - Global and Local as View (GLAV)
- Terminology stems from virtual integration
  - Given a global (or mediated, or virtual) schema
  - A set of data sources (local schemas)
  - Compute answers to queries written against the global schema using the local data sources

3.2 Schema Mapping

- Excursion Virtual Data Integration
  - More in next section of the course

3.2 Schema Mapping

- Global-as-view (GAV)
  - Express the global schema as views over the local schemata
  - What query language do we support?
    - CQ, UCQ, SQL, …?
  - Closed vs. open world assumption
    - Closed world: $R = Q(S_1, \ldots, S_n)$
      - Content of global relation $R$ is defined as the result of query $Q$
        over the sources
    - Open world: $R \supseteq Q(S_1, \ldots, S_n)$
      - Relation $R$ has to contain the result of query $Q$, but may contain additional tuples
3.2 Schema Mapping

**Example: Types of Matching**

Local Schema
- Person
- Address
- Phone

Global Schema
- Person
- Address
- Phone
- Email

Person(X',Y',Z',A',B') = Person(X,Y), Address(Y,Z,A)

Since heads of LHS and RHS have to be the same we can use simpler notation without the head of the view expression:

Person(X,Z,A, NULL, NULL) = Person(X,Y), Address(Y,Z,A)

Example:)Types)of)Matching)

Local Schema
- Person
- Name
- Address
- Phone

Global Schema
- Person
- Name
- Address
- Office-phone
- Home-phone

Consider switching local and global schema

Person(X, NULL) = Person(X,Y,Z,A,B)
Address(NULL,Y,Z) = Person(X,Y,Z,A,B)

• **Global-as-view (GAV)**
  - **Solutions (mapping M)**
    - Unique solutions (1 solution!)
    - Intuitively, execute queries over local instance that produced global instance

• **Global-as-view (GAV)**
  - **Answering Queries**
    - Simply replace references to global tables with the view definition

  Mapping R(X,Y) = S(X,Y), T(Y,Z)

  Q(X) :- R(X,Y)

  Rewrite into

  Q(X) :- S(X,Y), T(Y,Z)

• **Local-as-view (LAV)**
  - Express the local schema as views over the global schemata
  - What query language do we support?
    - CQ, UCQ, SQL, …?
  - **Closed vs. open world** assumption
    - Closed world: \( S_i = Q(G) \)
      - Content of local relation \( S_i \) is defined as the result of query \( Q \) over the sources
    - Open world: \( S_i \supseteq Q(G) \)
      - Local relation \( S_i \) has to contain the result of query \( Q \), but may contain additional tuples
### 3.2 Schema Mapping

#### Example: Types of Matching

<table>
<thead>
<tr>
<th>Local Schema</th>
<th>Global Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Person</td>
</tr>
<tr>
<td>Name</td>
<td>Name</td>
</tr>
<tr>
<td>Address</td>
<td>Address</td>
</tr>
<tr>
<td>City</td>
<td>Office-phone</td>
</tr>
<tr>
<td>Office-contact</td>
<td>Home-phone</td>
</tr>
</tbody>
</table>

\[
\text{Local (S)}(X, Y, Z) = \text{Global (G)}(X, Y, Z, A, B)
\]

#### Example: Types of Matching

- **Local-as-view (LAV)**
- **Solutions (mapping M)**
  - May be many solutions
- **Local-as-view (GAV)**
  - **Answering Queries**
    - Need to find equivalent query using only the views (this is a hard problem, more in next course section)
  - Mapping \( S(X,Z) = R(X,Y), T(Y,Z) \)
  - \( Q(X) : R(X,Y) \)
  - Rewrite into ???
    - Need to come up with missing values
    - Give up query equivalence?
- **Global-Local-as-view (GLAV)**
  - Express both sides of the constraint as queries
  - What query language do we support?
    - CQ, UCQ, SQL, …?
  - Closed vs. open world assumption
    - Closed world: \( Q'(G) = Q(S) \)
    - Open world: \( Q'(G) \not\subset Q(S) \)
3.2 Schema Mapping

- **Local-as-view (GLAV) Discussion**
  - Kind of best of both worlds (almost)
  - Complexity of query answering is the same as for LA V
  - Can address the lost correlation and missing values problems we observed using GAV and LAV

3.2 Schema Mapping

- **Source-to-target tuple-generating dependencies (st-tgds)**
  - Local way of expressing GLAV mappings
    \[ \forall \vec{x} : \phi(\vec{x}) \rightarrow \exists \vec{y} : \psi(\vec{x}, \vec{y}) \]
  - Equivalence to a containment constraint:
    \[ Q'(G) \supseteq Q(S) \]

Example: Types of Matching

<table>
<thead>
<tr>
<th>Local Schema</th>
<th>Global Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Person</td>
</tr>
<tr>
<td>Address</td>
<td>Address</td>
</tr>
<tr>
<td></td>
<td>Office-address</td>
</tr>
</tbody>
</table>

\[ \forall x, y, z, a : \text{Person}(x, y) \wedge \text{Address}(y, z, a) \rightarrow \exists b, c : \text{Person}(x, z, a, b, c) \]

3.2 Schema Mapping

- **Generating Schema Mappings**
  - **Input**: Schemas (Constraints), matches
  - **Output**: Schema mappings
  - **Ideas**:
    - Schema matches tell us which source attributes should be copied to which target attributes
    - Foreign key constraints tell us how to join in the source and target to not lose information

3.2 Schema Mapping

- **Clio**
  - Clio is a data exchange system prototype developed by IBM and University of Toronto researchers
  - The concepts developed for Clio have been implemented in IBM InfoSphere Data Architect
  - Clio does matching, mapping generation, and data exchange
    - For now let us focus on the mapping generation

3.2 Schema Mapping

- **Clio Mapping Generation Algorithm**
  - **Inputs**: Source and Target schemas, matches
  - **Output**: Mapping from source to target schema
  - Note, Clio works for nested schemas such as XML too not just for relational data.
    - Here we will look at the relational model part only
3.2 Schema Mapping

- **Clio Algorithm Steps**
  - 1) Use foreign keys to determine all reasonable ways of joining data within the source and the target schema
    - Each alternative of joining tables in the source/target is called a logical association
  - 2) For each pair of source-target logical associations: Correlate this information with the matches to determine candidate mappings

- **Clio Algorithm: 1) Find logical associations**
  - This part relies on the chase procedure that first introduced to test implication of functional dependencies (’77)
  - The idea is that we start use a representation of foreign keys are inclusion dependencies (tgds)
    - There are also chase procedures that consider edges (e.g., PKs)
  - Starting point are all single relational atoms
    - E.g., R(X,Y)

3.2 Schema Mapping

- **Chase step**
  - Works on tableau: set of relational atoms
  - A chase step takes one tgd t where the LHS is fulfilled and the RHS is not fulfilled
    - We fulfill the tgd t by adding new atoms to the tableau and mapping variables from t to the actually occurring variables from the current tableau
  - **Chase**
    - Applying the chase until no more changes
    - Note: if there are cyclic constraints this may not terminate

- **Clio Algorithm: 2) Generate Candidate Mappings**
  - For each pair of logical association \(A_S\) in the source and \(A_T\) in the target produced in step 1
  - Find the matches that are covered by \(A_S\) and \(A_T\)
    - Matches that lead from an element of \(A_S\) to an element from \(A_T\)
  - If there is at least one such match then create mapping by equating variables as indicated by the matches and create st-tgd with \(A_S\) in LHS and \(A_T\) in RHS

Outline

0) Course Info
1) Introduction
2) Data Preparation and Cleaning
3) Schema matching and mapping
4) Virtual Data Integration
5) Data Exchange
6) Data Warehousing
7) Big Data Analytics
8) Data Provenance
4. Virtual Data Integration

4.1 Query Answering with Views

- How to compute query $Q$ over global schema based on source schemas only?
  - What language is used to express mappings?
  - What language do we allow for $Q$?
  - What language(s) can we use to query local sources?
  - What language can we use to compute $Q$ from query results returned by local sources?
  - How to deal with incompleteness?
4. Query Answering with Views

* Problems
  - How to determine whether query can be answered at all?
  - Given a rewriting of the query using views, how do we know it is correct?
  - What to do if views can only return some of the query results?

Motivating Example (Part 1)

\[ Q(T,Y,D) : \neg Movie(T,Y,G), Y \geq 1950, G = \text{"comedy"} \]
\[ Director(I,D), Actor(I,D) \]
\[ V_1(T,Y,D) : \neg Movie(T,Y,G), Y \geq 1940, G = \text{"comedy"} \]
\[ Director(I,D), Actor(I,D) \]
\[ V_2(T,Y,D) \implies V_1(T,Y,D), Y \geq 1950 \]

Containment is enough to show that \( V_2 \) can be used to answer \( Q \).

Motivating Example (Part 2)

\[ Q(T,Y,D) : \neg Movie(T,Y,G), Y \geq 1950, G = \text{"comedy"} \]
\[ Director(I,D), Actor(I,D) \]
\[ V_1(T,Y,D) : \neg Movie(T,Y,G), Y \geq 1940, G = \text{"comedy"} \]
\[ Director(I,D), Actor(I,D) \]
\[ Containment \] does not hold, but intuitively, \( V_2 \) and \( V_3 \) are useful for answering \( Q \).

\[ Q^*(T,Y,D) : \neg V_1(T,Y), V_3(I,D) \]

How do we express that intuition?

Answering queries using views!

Problem Definition

Input: Query \( Q \)

View definitions: \( V_1, \ldots, V_n \)

A rewriting: a query \( Q' \) that refers only to the views and interpreted predicates (comparisons)

An equivalent rewriting of \( Q \) using \( V_1, \ldots, V_n \): a rewriting \( Q' \), such that \( Q' \Leftrightarrow Q \)
Naïve approach

- **Given Q and views**
  - Randomly combine views into a query Q'
  - Check equivalence of Q' and Q
  - If Q' is equivalent we are done
  - Else repeat

- **Why is this not good?**
  - There are infinitely many ways of combining views
    - E.g., V, V x V, V x V x V, ...
  - We are not using any information in the query

Motivating Example (Part 3)

- Movie(ID,title,year,genre)
- Director(ID,director)
- Actor(ID,actor)

Q(T,Y,D):
- Movie(I,T,Y,G), Y ≥ 1950, G = "comedy"
- Director(I,D)
- Actor(I,D)

V₁(I,T,Y) :
- Movie(I,T,Y,G), Y ≥ 1960, G = "comedy"

V₂(I,D) :
- Director(I,D), Actor(ID,D)

Q'''(T,Y,D) :
- V₁(I,T,Y), V₂(I,D)

maximally-contained rewriting

Maximally-Contained Rewritings

**Input:** Query Q

**Rewriting query language L**

*View definitions: V₁,...,Vₙ*

Q' is a maximally-contained rewriting of Q given V₁,...,Vₙ and L if:

1. Q' ∈ L,
2. Q' ⊆ Q, and
3. there is no Q'' in L such that Q'' ⊆ Q and Q'' ⊂ Q''

### Why again?

Global Schema

Local Schema 1

Local Schema 2

Local Schema n

Mappings

LAV/GLAV!

Exercise: which of these views can be used to answer Q?

Q(T,Y,D):
- Movie(I,T,Y,G), Y ≥ 1950, G = "comedy"
- Director(I,D), Actor(I,D)

V₁(I,T,Y) :
- Movie(I,T,Y,G), Y ≥ 1950, G = "comedy"

V₂(I,D) :
- Director(I,D), Actor(ID,D)

V₃(T,Y) :
- Movie(I,T,Y,G), Y ≥ 1950, G = "comedy"

V₄(I,T,Y) :
- Movie(I,T,Y,G), Y ≥ 1950, G = "comedy"

V₅(I,T) :
- Movie(I,T,Y,G), Y ≥ 1940, G = "comedy"

Other use-cases

- **Query optimization with materialized views**
  - Need equivalent rewritings
  - Implemented in many commercial DBMS
  - Here interest is cost: how to speed-up query processing by using materialized views
Algorithms for answering queries using views

- **Step 1**: we’ll bound the space of possible query rewritings we need to consider (no comparisons)
- **Step 2**: we’ll find efficient methods for searching the space of rewritings
  - Bucket Algorithm, MiniCon Algorithm
- **Step 2b**: we consider “logical approaches” to the problem:
  - The Inverse-Rules Algorithm

Bounding the Rewriting Length

**Theorem**: if there is an equivalent rewriting, there is one with at most \( n \) subgoals.

Query: \( Q(\overline{X}) : \neg p_1(\overline{X}_1), \ldots, \neg p_n(\overline{X}_n) \)

Rewriting: \( Q(\overline{X}) : \neg V(\overline{X}_1), \ldots, \neg V_m(\overline{X}_m) \)

Expansion: \( Q'(\overline{X}) : \neg g_1^{s_1}, \ldots, g_k^{s_k}, \ldots, g_j^{s_j} \)

Proof: Only \( n \) subgoals in \( Q \) can contribute to the image of the containment mapping \( \phi \).

Complexity Result

[LMSS, 1995]

- Applies to queries with no interpreted predicates.
- Finding an equivalent rewriting of a query using views is NP-complete
  - Need only consider rewritings of query length or less.
- Maximally-contained rewriting:
  - Union of all conjunctive rewritings of length \( n \) or less.

The Bucket Algorithm

**Key idea**:
- Create a bucket for each subgoal \( g \) in the query.
- The bucket contains views that contribute to \( g \).
- Create rewritings from the Cartesian product of the buckets (select one view for each goal)

- **Step 1**: assign views with renamed vars to buckets
- **Step 2**: create rewritings, refine them, until equivalent/all contained rewriting(s) are found

The Bucket Algorithm

**Step 1 Intuition**
- A view can only be used to provide information about a goal \( R(X) \) if it has a goal \( R(Y) \)
  - \( Q(X) : \neg R(X,Y) \)
  - \( V(X) : \neg S(X,Y) \)
- If the query goal contains variables that are in the head of the query, then the view is only useful if it gives access to these values (they are in the head)
  - \( Q(X) : \neg R(X,Y) \)
  - \( V(X) : \neg S(X,Y), R(Y, Z) \)
Bucket Algorithm in Action

\[ Q(ID, \text{Dir}) \rightarrow \text{Movie}(ID, \text{title}, \text{year}, \text{genre}), \text{Revenues}(ID, \text{amount}), \text{Director}(ID, \text{dir}) \]

**View atoms that can contribute to Movie:**

\[ V_1(ID, \text{year}''), V_2(ID, A'), V_3(ID, D', \text{year}') \]

Buckets and Cartesian Product

<table>
<thead>
<tr>
<th>Movie(ID, title, year, genre)</th>
<th>Revenues(ID, amount)</th>
<th>Director(ID, dir)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_1(ID, \text{year}) )</td>
<td>( V_2(ID, \text{year}') )</td>
<td>( V_4(ID, \text{Dir}, \text{year}') )</td>
</tr>
<tr>
<td>( V_3(ID, A') )</td>
<td>( V_3(ID, \text{amount}) )</td>
<td>( V_3(ID, \text{Dir}, \text{year}') )</td>
</tr>
<tr>
<td>( V_4(ID, D', \text{year}) )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consider first candidate rewriting: first V1 subgoal is redundant, and V1 and V4 are mutually exclusive.

\[ q'_1(ID, \text{dir}) : \neg V_1(ID, \text{year}''), V_2(ID, A'), V_3(ID, D', \text{year}') \]

Next Candidate Rewriting

<table>
<thead>
<tr>
<th>Movie(ID, title, year, genre)</th>
<th>Revenues(ID, amount)</th>
<th>Director(ID, dir)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V_1(ID, \text{year}) )</td>
<td>( V_2(ID, \text{year}') )</td>
<td>( V_4(ID, \text{Dir}, \text{year}') )</td>
</tr>
<tr>
<td>( V_3(ID, A') )</td>
<td>( V_3(ID, \text{amount}) )</td>
<td>( V_3(ID, \text{Dir}, \text{year}') )</td>
</tr>
<tr>
<td>( V_4(ID, D', \text{year}) )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ q'_2(ID, \text{dir}) : \neg V_3(ID, A'), V_2(ID, \text{amount}), V_4(ID, \text{dir}, \text{year}'), \text{amount} \geq 100M \]

The Bucket Algorithm: Summary

- Cuts down the number of rewriting that need to be considered, especially if views apply many interpreted predicates.
- The search space can still be large because the algorithm does not consider the interactions between different subgoals.
- See next example.

The MiniCon Algorithm

\[ Q(\text{title}, \text{year}, \text{dir}) : \rightarrow \text{Movie}(ID, \text{title}, \text{year}, \text{genre}), \text{Director}(ID, \text{dir}), \text{Actor}(ID, \text{dir}) \]

\[ V_3(D,A) : \rightarrow \text{Director}(I,D), \text{Actor}(I,A) \]

**Intuition:** The variable \( l \) is not in the head of \( V_3 \), hence \( V_3 \) cannot be used in a rewriting. **MiniCon** discards this option early on, while the Bucket algorithm does not notice the interaction.
MinCon Algorithm Steps

• 1) Create MiniCon descriptions (MCDs):
  – Homomorphism on view heads
  – Each MCD covers a set of subgoals in the query with a set of subgoals in a view

• 2) Combination step:
  – Any set of MCDs that covers the query subgoals (without overlap) is a rewriting
  – No need for an additional containment check!

MiniCon Descriptions (MCDs)

An atomic fragment of the ultimate containment mapping

\[ Q(\text{title, year, dir}): \neg \text{Movie}(\text{ID}, \text{title, year, genre}), \]
\[ \quad \quad \text{Director}(\text{ID}, \text{dir}), \text{Actor}(\text{ID}, \text{act}) \]
\[ V(\text{I, D, A}): \neg \text{Director}(\text{I, D}), \text{Actor}(\text{I, A}) \]

MCD:
  \[ \text{ID} \rightarrow \text{I} \]
  \[ \text{dir} \rightarrow \text{D} \]
  \[ \text{act} \rightarrow \text{A} \]
  covered subgoals of \( Q \): \{2,3\}

MCDs: Detail 1

\[ Q(\text{title, year, dir}): \neg \text{Movie}(\text{ID}, \text{title, year, genre}), \]
\[ \quad \quad \text{Director}(\text{ID}, \text{dir}), \text{Actor}(\text{ID}, \text{dir}) \]
\[ V(\text{I, D, A}): \neg \text{Director}(\text{I, D}), \text{Actor}(\text{I, A}) \]

Need to specialize the view first:
\[ V'(\text{I, D, D}): \neg \text{Director}(\text{I, D}), \text{Actor}(\text{I, D}) \]

MCD:
  \[ \text{ID} \rightarrow \text{I} \]
  \[ \text{dir} \rightarrow \text{D} \]
  covered subgoals of \( Q \): \{2,3\}

MCDs: Detail 2

\[ Q(\text{title, year, dir}): \neg \text{Movie}(\text{ID}, \text{title, year, genre}), \]
\[ \quad \quad \text{Director}(\text{ID}, \text{dir}), \text{Actor}(\text{ID}, \text{dir}) \]
\[ V(\text{I, D, D}): \neg \text{Director}(\text{I, D}), \text{Actor}(\text{I, D}), \]
\[ \quad \quad \text{Movie}(\text{I, T, Y, G}) \]

Note: the third subgoal of the view is not included in the MCD.

MCD:
  \[ \text{ID} \rightarrow \text{I} \]
  \[ \text{dir} \rightarrow \text{D} \]
  covered subgoals of \( Q \) still: \{2,3\}

Inverse-Rules Algorithm

• A "logical" approach to AQUV
• Produces maximally-contained rewriting in polynomial time
  – To check whether the rewriting is equivalent to the query, you still need a containment check.
• Conceptually simple and elegant
  – Depending on your comfort with Skolem functions...

Inverse Rules by Example

Given the following view:
\[ V_1(\text{I, T, Y, G}): \neg \text{Movie}(\text{I, T, Y, G}), \text{Director}(\text{I, D}), \text{Actor}(\text{I, D}) \]

And the following tuple in \( V_1 \):
\[ V_1(79, \text{Manhattan, 1979, Comedy}) \]

Then we can infer the tuple:
\[ \text{Movie}(79, \text{Manhattan, 1979, Comedy}) \]

Hence, the following ‘rule’ is sound:
\[ \text{IN}_1: \text{Movie}(\text{I, T, Y, G}) \rightarrow V_1(\text{I, T, Y, G}) \]
Skolem Functions

\[ V_7(I,T,Y,G) : \neg \text{Movie}(I,T,Y,G), \text{Director}(I,D), \text{Actor}(I,D) \]

Now suppose we have the tuple
\[ V_7(79, \text{Manhattan}, 1979, \text{Comedy}) \]

Then we can infer that there exists some director. Hence, the following rules hold (note that they both use the same Skolem function):

\[ \text{IN}_2: \text{Director}(I,f_1(I,T,Y,G)) \leftarrow V_7(I,T,Y,G) \]
\[ \text{IN}_3: \text{Actor}(I,f_1(I,T,Y,G)) \leftarrow V_7(I,T,Y,G) \]

Inverse Rules in General

\[ Q_2(\text{title}, \text{year}, \text{genre}) : \neg \text{Movie}(\text{ID}, \text{title}, \text{year}, \text{genre}) \]

Given Q2, the rewriting would include:
\[ \text{IN}_1, \text{IN}_2, \text{IN}_3, Q_2. \]

Given input: \[ V_7(79, \text{Manhattan}, 1979, \text{Comedy}) \]

Inverse rules produce:
\[ \text{Movie}(79, \text{Manhattan}, 1979, \text{Comedy}), \]
\[ \text{Director}(79, f_1(79, \text{Manhattan}, 1979, \text{Comedy})), \]
\[ \text{Actor}(79, f_1(79, \text{Manhattan}, 1979, \text{Comedy})), \]
\[ \text{Movie}(\text{Manhattan}, 1979, \text{Comedy}) \]
(the last tuple is produced by applying \( Q_2 \)).

Comparing Algorithms

- Bucket algorithm:
  - Good if there are many interpreted predicates
  - Requires containment check. Cartesian product can be big
- MiniCon:
  - Good at detecting interactions between subgoals

Algorithm Comparison (Continued)

- Inverse-rules algorithm:
  - Conceptually clean
  - Can be used in other contexts (see later)
  - But may produce inefficient rewritings because it “undoes” the joins in the views (see next slide)
- Experiments show MiniCon is most efficient.
  - Even faster:
    Konstantinidis, G. and Ambite, J.L, Scalable query rewriting: a graph-based approach. SIGMOD ’11

Inverse Rules Inefficiency

Example

Query and view:
\[ Q(X,Y) : \neg e_1(X,Z), e_2(Z,Y) \]
\[ V(A,B) : \neg e_1(A,C), e_2(C,B) \]

Inverse rules:
\[ e_1(A,f_1(A,B)) : \neg V(A,B) \]
\[ e_2(f_1(A,B),B) : \neg V(A,B) \]

Now we need to re-compute the join...

View-Based Query Answering

- Maximally-contained rewritings are parameterized by query language.
- More general question:
  - Given a set of view definitions, view instances and a query, what are all the answers we can find?
- We introduce certain answers as a mechanism for providing a formal answer.
Consider the two views:
\[ V_8(\text{dir}): \neg \text{Movie}(ID, \text{dir}, \text{actor}) \]
\[ V_9(\text{actor}): \neg \text{Movie}(ID, \text{dir}, \text{actor}) \]

And suppose the extensions of the views are:
\[ V_8: \{\text{Allen}, \ \text{Copolla}\} \]
\[ V_9: \{\text{Keaton}, \ \text{Pacino}\} \]

There are multiple databases that satisfy the above view definitions: (we ignore the first argument of \text{Movie} below)

DB1. \{\{\text{Allen, Keaton}, \ \{\text{Coppola, Pacino}\}\} \]
DB2. \{\{\text{Allen, Pacino}, \ \{\text{Coppola, Keaton}\}\} \]

If we ask whether Allen directed a movie in which Keaton acted, we can’t be sure. Certain answers are those true in all databases that are consistent with the views and their extensions.

---

**Certain Answers: Formal Definition**  
**[Open-world Assumption]**

- Given:
  - Views: \( V_1, \ldots, V_n \)
  - View extensions \( v_1, \ldots, v_n \)
  - A query \( Q \)
- A tuple \( t \) is a certain answer to \( Q \) under the open-world assumption if \( t \in Q(D) \) for all databases \( D \) such that:
  - \( V_i(D) \subseteq v_i \) for all \( i \).

---

**Certain Answers: Example**

\[ V_8(\text{dir}): \neg \text{Director}(ID, \text{dir}) \]
\[ V_9(\text{actor}): \neg \text{Actor}(ID, \text{actor}) \]
\[ Q(\text{dir}, \text{actor}): \neg \text{Director}(ID, \text{dir}), \text{Actor}(ID, \text{actor}) \]

Under closed-world assumption:
- single DB possible \( \Rightarrow \) \{\text{Allen, Keaton}\}

Under open-world assumption:
- no certain answers.

---

**Certain Answers: Formal Definition**  
**[Closed-world Assumption]**

- Given:
  - Views: \( V_1, \ldots, V_n \)
  - View extensions \( v_1, \ldots, v_n \)
  - A query \( Q \)
- A tuple \( t \) is a certain answer to \( Q \) under the open-world assumption if \( t \in Q(D) \) for all databases \( D \) such that:
  - \( V_i(D) = v_i \) for all \( i \).

---

**The Good News**

- The MiniCon and Inverse-rules algorithms produce all certain answers
  - Assuming no interpreted predicates in the query (ok to have them in the views)
  - Under open-world assumption
  - Corollary: they produce a maximally-contained rewriting

---
In Other News...

- Under closed-world assumption finding all certain answers is co-NP hard!

**Proof:** encode a graph - $G = (V,E)$

- $v_1(X) : -\text{color}(X, Y)$  \quad $h(V_1) = V$
- $v_2(Y) : -\text{color}(X, Y)$  \quad $h(V_2) = \{\text{red, green, blue}\}$
- $v_3(X, Y) : -\text{edge}(X, Y)$  \quad $h(V_3) = E$

$q() : -\text{edge}(X, Y), \text{color}(X, Z), \text{color}(Y, Z)$

$q$ has a certain tuple iff $G$ is not 3-colorable

Interpreted Predicates

- In the views: no problem (all results hold)
- In the query $Q$:
  - If the query contains interpreted predicates, finding all certain answers is co-NP-hard even under open-world assumption
  - Proof: reduction to CNF.

Outline

0) Course Info
1) Introduction
2) Data Preparation and Cleaning
3) Schema matching and mapping
4) Virtual Data Integration
5) Data Exchange
6) Data Warehousing
7) Big Data Analytics
8) Data Provenance
5. Data Exchange

Virtual Data Integration
- Never materialize instances for the global schema
- Data of global schema only “visible” through queries

Data Exchange
- Materialize instance of global instance
- We call it the “target schema”
- Based on information from an instance of the local schema
- We call this the “source schema”

Data Exchange Problem Statement
- Input:
  - Given a source and a target schema
  - + instance of the source schema
  - + set of schema mappings (here st-tgds)
- Output:
  - Instance of the target schema that fulfills constraints

Example: Types of Matching

∀x, y, z, a : Person(x, y) ∧ Address(y, z, a) → ∃b, c : Person(x, z, a, b, c)
5.1 Data Exchange Setting

**Definition: Data Exchange Setting**

Data Exchange setting is a tuple \((S, T, I, \Sigma)\)
- **Source Schema** `S`
- **Target Schema** `T`
- **Instance** `I`
- **Mappings** `\Sigma` from `S` to `T`

5.1 Data Exchange Solutions

**Example: Solutions**

<table>
<thead>
<tr>
<th>Person</th>
<th>Name</th>
<th>Address</th>
<th>Id</th>
<th>City</th>
<th>Office-contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>1</td>
<td>1 Chicago</td>
<td>(312) 123 4343</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alice</td>
<td>2</td>
<td>2 Chicago</td>
<td>(312) 555 7777</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>3</td>
<td>3 New York</td>
<td>(465) 123 1234</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(\forall x, y, z, a: \text{Person}(x, y) \land \text{Address}(y, z, a) \rightarrow \exists b, c: \text{Person}(x, z, a, b, c)\)

Can we come up with a solution?

5.1 Number of Solutions

- **How many solutions exists?**
  - Depends on how whether we use existentially quantified variables in the mappings?
    - i.e., do we have attributes for which we have to invent values?
  - What attribute values do we allow?
    - Surely values from the source instance (active domain)
    - NULL?
      - Need multiple NULL values as placeholders for missing values that have to be the same
  - Note that this is the open-world assumption
    - there are infinitely many solutions (if domains infinite)

- **Target instance domain**
  - Consider a universe `U`
    - Source instance can only use values from `U`
  - Consider an infinite set `N` of labeled nulls
    - Target instance can use these as placeholders for missing values
5.1 Data Exchange Solutions

**Example: Multiple Solutions**

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Office Phone</th>
<th>Office Address</th>
<th>Home Phone</th>
<th>Home Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>Chicago</td>
<td>(312) 123 456</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alice</td>
<td>Chicago</td>
<td>(312) 567 890</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bob</td>
<td>New York</td>
<td>(465) 123 456</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Example: Solution generality**

Consider query:

\[ Q(n) :\exists P(n,a,op,oa,hp), \text{oa} = \text{Hometown} \]

**Definition: Universal solution**

Given data exchange setting \((S,T,I,\Sigma)\), an instance \(J\) of \(T\) is called an universal solution for a source instance \(I\) if it is a solution and for every other solution \(J'\) hold that:

- There exists a homomorphism from \(J\) to \(J'\)

**Definition: Homomorphism**

A homomorphism \(h\) from instance \(J\) to instance \(J'\) maps the constants and nulls of \(J\) to the constants and nulls of \(J'\) and fulfills the following conditions:

- Constants are mapped onto themselves: \(h(c) = c\)
- Every tuple \(R(a_1, \ldots, a_n)\) in \(J\) is mapped to a tuple in \(J'\):
  \[ R(h(a_1), \ldots, h(a_n)) \]

5.1 Certain answers (… again)

- **Have multiple solutions**
  - Define certain answers for queries as before
  - Every tuple \(t\) so that \(t\) is in the result of query \(Q\) over any valid solution \(J\)

- **What’s new?**
  - Want to materialize an instance so that computing certain answers over this instance is easy
    - Not immediately clear that this actually possible

5.1 Universal solutions

- **Universal solution**
  - Want a solution that is as general as possible
  - We call such most general solutions universal solutions
  - How do we know whether it is most general
    - We can map the tuples in this solution to any other less general solution by replacing unspecified values (labelled nulls) with actual data values

- **Query answering with universal solutions**
  - For UCQs: run query over universal instance
  - Remove tuples with labelled nulls
  - Result are the certain answers!
5.1 Data Exchange Solutions

Example: Solution generality

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Office phone</th>
<th>Office address</th>
<th>Home phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>Chicago</td>
<td>(312) 301-450</td>
<td></td>
<td>111-3456</td>
</tr>
<tr>
<td>Alice</td>
<td>Chicago</td>
<td>(312) 305-350</td>
<td></td>
<td>111-350</td>
</tr>
<tr>
<td>Bob</td>
<td>New York</td>
<td>(465) 301-123</td>
<td></td>
<td>111-360</td>
</tr>
</tbody>
</table>

Above is universal solution.

How to map to below non-universal solution?

Replace generic labelled Nulls with values:

X: Hometown, Y: 111X322X3454, Z: other town,

Example: Solution generality

<table>
<thead>
<tr>
<th>Name</th>
<th>Address</th>
<th>Office phone</th>
<th>Office address</th>
<th>Home phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>Chicago</td>
<td>(312) 301-450</td>
<td></td>
<td>111X322X3454</td>
</tr>
<tr>
<td>Alice</td>
<td>Chicago</td>
<td>(312) 305-350</td>
<td></td>
<td>111-350</td>
</tr>
<tr>
<td>Bob</td>
<td>New York</td>
<td>(465) 301-123</td>
<td></td>
<td>111-360</td>
</tr>
</tbody>
</table>

5.2 Computing Solutions

• Note
  – Schema mappings (st-tgds) are tuple-generating dependencies
  – What other tgd’s do we know
    • Foreign keys
  – How did we solve violations to FKs?
    • The chase!
  – Chase produces universal solution!

• Can we use a database system to compute solutions?
  – Yes, systems such as Clio generate queries that compute universal solutions!
    • SQL
    • Java
    • XSLT (for XML docs)

• Generating Executable Transformations
  – How to preserve semantics of labeled nulls
    • n = n’ true if we have the same labeled null only
    • n = n’ if one is a constant and the other one is a labeled null

5.2 Skolem Functions

• Skolem functions for labeled nulls
  – For each existential variable in a tgd we create a new skolem function
  – What should be the arguments of the function?
    • Naive: all universally quantified variables
    • Better: only relevant ones

Example: Skolem functions
5.2 Skolem Functions

Example: Skolem Functions

∀ a, b, c, d, e : Person(a, b, c, d, e) → ∃ f, gPerson(a, f, g) ∧ Address(f, b, c)

Introduce Skolem functions sk1 and sk2 for f and g.

What arguments to choose for sk1 and sk2?

E.g., f should be fixed for a certain address and should not depend on the person.

Example: Skolem Functions

∀ a, b, c, d, e : Person(a, b, c, d, e) → ∃ f, gPerson(a, f, g) ∧ Address(f, b, c)

5.2 Skolem Functions

• Clio Schema Graph Algorithm
  • Nodes
    – Create a graph with one node for every target attribute and one node for every target relation
    – Also add nodes for source attribute if they are copied to the target according to the mapping
  • Edges
    – Edges between a relation and its attributes
    – Edges between target attributes that use the same variable
    – Edges between source attributes and target attributes if they use the same variable

Example: Skolem Functions

∀ a, b, c, d, e : Person(a, b, c, d, e) → ∃ f, gPerson(a, f, g) ∧ Address(f, b, c)
5.2 Skolem Functions

Example: Skolem Functions

∀a, b, c, d, e : Person(a, b, c, d, e) → ∃f, gPerson(a, f, g) ∧ Address(f, h, c)

5.1 Data Exchange Solutions

Example: Skolem Functions

∀a, b, c, d, e : Person(a, b, c, d, e) → ∃f, gPerson(a, f, g) ∧ Address(f, h, c)

5.2 Skolem Functions

- Clio Schema Graph Algorithm
- Skolem functions
  - Derive skolem function arguments from the schema graph annotations of an element

Example: Skolem Functions

∀a, b, c, d, e : Person(a, b, c, d, e) → ∃f, gPerson(a, f, g) ∧ Address(f, h, c)

For variable f (id, address) we assign sk1(a, b, c)
For variable g (age) we assign sk2(a, b, c)

5.2 Executable Transformations

Example: Skolem Functions

∀a, b, c, d, e : Person(a, b, c, d, e) → ∃f, gPerson(a, f, g) ∧ Address(f, h, c)

Select SQL:

```
SELECT name || address || office-phone AS address,
       office-phone AS age
FROM Person
```

5.3 Recap Data Exchange Steps

- Schema Matching
- Generate Schema Mappings
  - Use constraints
- Generate Executable Transformations
  - SQL, XSLT, XQuery
  - Skolems for missing value
- Run Transformations over source instance to generate target instance
  - Universal solution
5.3 Comparison with virtual integration
- Pay cost upfront instead of at query time
- Making decisions early vs. at query time
  - When generating a solution
  - Caution: bad decisions stick!
- **Universal solutions** allow efficient computation of certain types of queries using, e.g., SQL

Outline
0) Course Info
1) Introduction
2) Data Preparation and Cleaning
3) Schema matching and mapping
4) Virtual Data Integration
5) Data Exchange
6) **Data Warehousing**
7) Big Data Analytics
8) Data Provenance
6. Data Warehousing

- **Problem:** Data Analysis, Prediction, Mining
  - Example: Walmart
  - Transactional databases
    - Run many “cheap” updates concurrently
    - E.g., each store has a database storing its stock and sales
  - Complex Analysis over Transactional Databases?
    - Want to analyze across several transactional databases
    - E.g., compute total Walmart sales per month
    - Distribution and heterogeneity
    - Want to run complex analysis over large datasets
    - Resource consumption of queries affects normal operations on transactional databases

- **Solution:**
  - **Performance**
    - Store data in a different system (the datawarehouse) for analysis
    - Bulk-load data to avoid wasting performance on concurrency control during analysis
  - **Heterogeneity and Distribution**
    - Preprocess data coming from transactional databases to clean it and translate it into a unified format before bulk-loading

6. Overview

- **The multidimensional datamodel (cube)**
  - Multidimensional data model
  - Relational implementations
- **Preprocessing and loading (ETL)**
- **Query language extensions**
  - ROLL UP, CUBE, …
- **Query processing in datawarehouses**
  - Bitmap indexes
  - Query answering with views
  - Self-tuning
6. Multidimensional Datamodel

• Analysis queries are typically aggregating lower level facts about a business
  – The revenue of Walmart in each state (country, city)
  – The amount of toy products in a warehouse of a company per week
  – The call volume per zip code for the Sprint network
  – …

6. Commonality among these queries:
  – At the core are facts: a sale in a Walmart store, a toy stored in a warehouse, a call made by a certain phone
  – Data is aggregated across one or more dimensions
    • These dimensions are typically organized hierarchically: year – month – day – hour, country – state - zip
  – Example
    – The revenue (sum of sale amounts) of Walmart in each state

6. Example 2D

<table>
<thead>
<tr>
<th>Toy</th>
<th>2014</th>
<th>2015</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Puppet</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Fishing rod</td>
<td>11</td>
<td>13</td>
</tr>
<tr>
<td>Model</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Books</td>
<td></td>
<td></td>
</tr>
<tr>
<td>King Lear</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

6. Data cubes

• Given n dimensions
  – E.g., product type, location, time
• Given m measures
  – E.g., number of sales, items in stock, …
• A datacube (datahypercube) is an n-dimensional datastructure that maps values in the dimensions to values for the m measures
  – Schema: D1, …, Dn, M1, …, Mn
  – Instance: a function
    \[ \text{dom}(D_1) \times \ldots \times \text{dom}(D_n) \rightarrow \text{dom}(M_1) \times \ldots \times \text{dom}(M_m) \]
6. Dimension Example

- **Location**
  - **Levels**: location, state, city

  **Schema**
  - **Location**
  - **state**
  - **City**

  **Instance**
  - **Illinois**
  - **Wisconsin**
  - **Chicago**
  - **Schaumburg**
  - **Madison**
  - **Whitewater**

6. Dimension Schema

- **Schema of a Dimension**
  - A set \( D = \{ D_1, \ldots, D_n \} \)
  - These correspond to the levels
  - A partial order \( \rightarrow \) over \( D \) which represents parent-child relationships in the hierarchy
  - These correspond to upward edges in the hierarchy
  - \( \text{Top}_D \) is larger than anything else
    - For every \( D_i \), \( D_i \rightarrow \text{Top}_D \)
  - There exists \( D_{\text{min}} \) which is smaller than anything else
    - For every \( D_i \), \( D_{\text{min}} \rightarrow D_i \)

6. Dimension Schema Example

- **Schema of Location Dimension**
  - Set of categories \( D = \{ \text{location, state, city} \} \)
  - Partial order
    - \( \{ \text{city} \rightarrow \text{state}, \text{city} \rightarrow \text{location}, \text{state} \rightarrow \text{location} \} \)
  - \( \text{Top}_D = \text{location} \)
  - \( D_{\text{min}} = \text{city} \)

6. Remarks

- In principle there does not have to exist an order among the elements at one level of the hierarchy
  - E.g., cities
- Hierarchies do not have to be linear

6. Cells, Facts, and Measures

- Each cell in the cube corresponds to a combination of elements from each dimension
  - Facts are non-empty cells
  - Cells store measures
- Cube for a combination of levels of the dimension

**Facts**

- Targets of analytics
  - E.g., revenue, #sales, #stock
- A fact is uniquely defined by the combination of values from the dimensions
  - E.g., for dimensions time and location
    - Revenue in Illinois during Jan 2015
- **Granularity**: Levels in the dimension hierarchy corresponding to the fact
  - E.g., state, month
Facts (Event vs. Snapshot)

- **Event Facts**
  - Model real-world events
  - E.g., Sale of an item

- **Snapshot Facts**
  - Temporal state
  - A single object (e.g., a book) may contribute to several facts
  - E.g., number of items in stock

Measures

- **A measure** describes a fact
  - May be derived from other measures

- **Two components**
  - **Numerical value**
  - **Formula** (optional): how to derive it
    - E.g., \( \text{avg}(\text{revenue}) = \frac{\text{sum}(\text{revenue})}{\text{count}(\text{revenue})} \)

- We may associate multiple measures to each cell
  - E.g., number of sales and total revenue

Measures - Granularity

- Similar to facts, measures also have a granularity
- **How to change granularity of a measure?**
- **Need algorithm to combine measures**
  - **Additive measures**
    - Can be aggregated along any dimension
  - **Semi-additive/non-additive**
    - Cannot be aggregated along some/all dimensions
    - E.g., snapshot facts along time dimension
      - Number of items in stock at Jan + Feb + … = items in stock during year
      - Median of a measure

Design Process (after Kimball)

- **Comparison to classical relational modeling**
  - **Analysis driven**
    - No need to model all existing data and relationships relevant to a domain
    - Limit modeling to information that is relevant for predicted analytics
  - **Redundancy**
    - Tolerate redundancy for performance if reasonable
      - E.g., in dimension tables to reduce number of joins

Design Process – Steps

- 1) **Select relevant business processes**
  - E.g., order shipping, sales, support, stock management
- 2) **Select granularity**
  - E.g., track stock at level of branches or regions
- 3) **Design dimensions**
  - E.g., time, location, product, …
- 4) **Select measures**
  - E.g., revenue, cost, #sales, items in stock, #support requests

Design Process Example

- **Coffee shop chain**
  - **Processes**
    - Sell coffee to customers
    - Buy ingredients from suppliers
    - Ship supplies to branches
    - Pay employees
    - HR (hire, advertise positions, …)
  - **Which process is relevant to be analysed to increase profits?**
Design Process Example

1) Selecting process(es)
   - sell coffee to customers

2) Select granularity
   - Single sale?
   - Sale per branch/day?
   - Sale per city/year?

Design Process Example

1) Selecting process(es)
   - sell coffee to customers

2) Select granularity
   - Sale of type of coffee per branch per day
   - Sufficient for analysis
      - Save storage

3) Determine relevant dimensions
   - Location
   - Time
   - Product, …

Design Process Example

1) Selecting process(es)
   - sell coffee to customers

2) Select granularity
   - Sale of type of coffee per branch per day

3) Determine relevant dimensions
   - Location (country, state, city, zip, shop)
   - Time (year, month, day)
   - Product (type, brand, product)

4) Select measures
   - cost, revenue, profit?

Relational representation

How to model a datacube using the relational datamodel

We start from
   - Dimension schemas
   - Set of measures
Star Schema

• A data cube is represented as a set of dimension tables and a fact table
• Dimension tables
  – For each dimension schema $D = (D_1, ..., D_k, Top_D)$ we create a relation
  – $D_i (PK_i, D_{i1}, ..., D_{ik})$
  – Here $PK_i$ is a primary key, e.g., $D_{min}$
• Fact table
  – $F (FK_1, ..., FK_m, M_1, ..., M_n)$
  – Each $FK_i$ is a foreign key to $D_i$
  – Primary key is the combination of all $FK_i$

Star Schema - Remarks

• Dimension tables have redundancy
  – Values for higher levels are repeated
• Fact table is in 3NF
• $Top_D$ does not have to be stored explicitly
• Primary keys for dimension tables are typically generated (surrogate keys)
  – Better query performance by using integers

Snowflake Schema

• A data cube is represented as a set of dimension tables and a fact table
• Dimension tables
  – For each dimension schema $D = (D_1, ..., D_k, Top_D)$ we create a relation multiple relations connected through FKs
  – $D_i (PK_i, A_{i1}, ..., A_{il}, FK_j)$
  – $A_{il}$ is a descriptive attribute
  – $FK_j$ is foreign key to the immediate parent(s) of $D_i$
• Fact table
  – $F (FK_1, ..., FK_m, M_1, ..., M_n)$
  – Each $FK_i$ is a foreign key to $D_i$
  – Primary key is the combination of all $FK_i$

Snowflake Schema - Remarks

• Avoids redundancy
• Results in much more joins during query processing
• Possible to find a compromise between snowflake and star schema
  – E.g., use snowflake for very fine-granular dimensions with many levels

Snowflake Schema - Example

• Coffee chain example

6. Extract-Transform-Load (ETL)

• The preprocessing and loading phase is called extract-transform-load (ETL) in datawarehousing
• Many commercial and open-source tools available
• ETL process is modeled as a workflow of operators
  – Tools typically have a broad set of build-in operators: e.g., key generation, replacing missing values, relational operators,
  – Also support user-defined operators
6. Extract-Transform-Load (ETL)

- Some ETL tools
  - Pentaho Data Integration
  - Oracle Warehouse Builder (OWB)
  - IBM Infosphere Information Server
  - Talend Studio for Data Integration
  - CloverETL
  - Cognos Data Manager
  - Pervasive Data Integrator
  - ...

6. Operators supported by ETL

- Many of the preprocessing and cleaning operators we already know
  - Surrogate key generation (like creating existentials with skolems)
  - Fixing missing values
    - With default value, using trained model (machine learning)
  - Relational queries
    - E.g., union of two tables or joining two tables
  - Extraction of structured data from semi-structured data and/or unstructured data
  - Entity resolution, data fusion

6. ETL Process

- Operators can be composed to form complex workflows

6. Typical ETL operators

- Elementizing
  - Split values into more fine-granular elements
- Standardization
- Verification
- Matching with master data
- Key generation
- Schema matching, Entity resolution/ Deduplication, Fusion

6. Typical ETL operators

- Control flow operators
  - AND/OR
  - Fork
  - Loops
  - Termination
    - Successful
    - With warning/errors

6. Typical ETL operators

- Elementizing
  - Split non 1NF data into individual elements
- Examples
  - name: “Peter Gertsen” -> firstname: “Peter”, lastname: “Gertsen”
  - Address: “10 W 31st, Chicago, IL 60616” -> street = “10 W 31st”, city = “Chicago”, state = “IL”, zip = “60616”
6. Typical ETL operators

• **Standardization**
  – Expand abbreviation
  – Resolve synonyms
  – Unified representation of, e.g., dates

• **Examples**
  – “IL” -> “Illinois”
  – “m/w”, “M/F” -> “male/female”
  – “Jan”, “01”, “January”, “january” -> “January”
  – “St” -> “Street”, “Dr” -> “Drive”, …

• **Verification**
  – Same purpose as constraint based data cleaning but typically does not rely on constraints, but, e.g.,
    regular expression matching

• **Examples**
  – Phone matches “[0-9]{3}-[0-9]{3}-[0-9]{4}”
  – For all t in Tokens(product description), t exists in
    English language dictionary

6. Typical ETL operators

• **Matching master data (lookup)**
  – Check and potentially repair data based on
    available master data

• **Examples**
  – E.g., using a clean lookup table with (city.zip) replace
    the city in each tuple if the pair (city.zip) does not occur
    in the lookup table

6. Metadata management

• As part of analysis in DW data is subjected to a
  complex pipeline of operations
  – Sources
  – ETL
  – Analysis queries

• -> important, but hard, to keep track of what
  operations have been applied to data and from
  which sources it has been derived

• Need metadata management
  • Including provenance (later in this course)

6. Querying DW

• **Targeted model (cube vs. relational)**
  – Design specific language for datacubes
  – Add suitable extensions to SQL

• **Support typical analytical query patterns**
  – Multiple parallel grouping criteria
    • Show total sales, subtotal per state, and subtotal per city
    • -> three subqueries with different group-by in SQL
  – Windowed aggregates and ranking
    • Show 10 most successful stores
    • Show cumulative sales for months of 2016
      – E.g., the result for Feb would be the sum of the sales for Jan +
        Feb

• **Targeted model (cube vs. relational)**
  – Design specific language for datacubes
    • MDX
  – Add suitable extensions to SQL
    • GROUPING SETS, CUBE, …
    • Windowed aggregation using OVER(), PARTITION BY,
      ORDER BY, window specification
    • Window functions
      – RANK, DENSE_RANK
6. Cube operations

• **Roll-up**
  – Move from fine-granular to more coarse-granular in one or more dimensions of a datacube
    - E.g., sales per (city, month, product category) to Sales per (state, year, product category)

• **Drill-down**
  – Move from coarse-granular to more fine-granular in one of more dimensions
    - E.g., phonecalls per (city, month) to phonecalls per (zip, month)

6. Cube operations

• **Drill-out**
  – Add additional dimensions
    - special case of drill-down starting from TopD in dimension(s)
    - E.g., sales per (city, product category) to Sales per (state, year, product category)

• **Drill-in**
  – Remove dimension
    - special case for roll-up move to TopD for dimension(s)
    - E.g., phonecalls per (city, month) to phonecalls per (month)

6. SQL Extensions

• **Slice**
  – Select data based on restriction of the values of one dimension
    - E.g., sales per (city, month) -> sales per (city) in Jan

• **Dice**
  – Select data based on restrictions of the values of multiple dimensions
    - E.g., sales per (city, month) -> sales in Jan for Chicago and Washington DC

6. SQL Extensions

• **Recall that grouping on multiple sets of attributes is hard to express in SQL**
  – E.g., give me the total sales, the sales per year, and the sales per month
  – Practice

6. GROUPING SETS

• **GROUP BY GROUPING SETS ((set_1), ..., (set_n))**
  – Explicitly list sets of group by attributes
  – Semantics:
    - Equivalent to UNION over duplicates of the query each with a group by clause GROUP BY set,
    - Schema contains all attributes listed in any set
    - For a particular set, the attribute not in this set are filled with NULL values
6. GROUPING SETS

SELECT quarter,
city,
product_typ,
SUM(profit) AS profit
FROM facttable F, time T, location L, product P
WHERE
F.TID = T.TID AND F.LID = L.LID AND F.PID = P.PID
GROUP BY GROUPING SETS
( (quarter, city),
(quarter, product_typ))

<table>
<thead>
<tr>
<th>quarter</th>
<th>city</th>
<th>product_typ</th>
<th>profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010 Q1</td>
<td>Books</td>
<td>8347</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Books</td>
<td>7836</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Gardening</td>
<td>12300</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Chicago</td>
<td>12344</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Seattle</td>
<td>124345</td>
<td></td>
</tr>
</tbody>
</table>

6. GROUPING SETS

• Problem:
  – How to distinguish between NULLs based on grouping sets and NULL values
    in a group by column?

GROUP BY GROUPING SETS
( (quarter, city),
quarter, product_typ,
quarter, product_typ, city)

<table>
<thead>
<tr>
<th>quarter</th>
<th>city</th>
<th>product_typ</th>
<th>profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010 Q1</td>
<td>Books</td>
<td>8347</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Books</td>
<td>7836</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Gardening</td>
<td>12300</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Chicago</td>
<td>12344</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Seattle</td>
<td>124345</td>
<td></td>
</tr>
</tbody>
</table>

6. GROUPING SETS

• Solution:
  – GROUPING predicate
  – GROUPING(A) = 1 if grouped on attribute A, 0 else

GROUP BY GROUPING SETS
( (quarter, city),
quarter, product_typ,
quarter, product_typ, city)

<table>
<thead>
<tr>
<th>quarter</th>
<th>city</th>
<th>product_typ</th>
<th>profit</th>
<th>gp_prd</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010 Q1</td>
<td>Books</td>
<td>8347</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Books</td>
<td>7836</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Gardening</td>
<td>12300</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Chicago</td>
<td>12344</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012 Q2</td>
<td>Seattle</td>
<td>124345</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. GROUPING SETS

• Combining GROUPING SETS

GROUP BY A, B
= GROUP BY GROUPING SETS ((A,B))

GROUP BY GROUPING SETS ((A,B), (A,C), (A))
= GROUP BY A, GROUPING SETS ((B), (C), ())

GROUP BY GROUPING SETS ((A,B), (B,C),
GROUPING SETS ((D,E), (D))
= GROUP BY GROUPING SETS ((A,B,D,E), (A,B,D), (B,C,D,E), (B,C,D))

6. CUBE

• GROUP BY CUBE (set)
• Group by all 2^n subsets of set

GROUP BY CUBE (A,B,C)
= GROUP BY GROUPING SETS (
  (),
  (A), (B), (C),
  (A,B), (A,C), (B,C),
  (A,B,C)
6. CUBE

• GROUP BY ROLLUP \((a_1, \ldots, a_n)\)
• Group by all prefixes
• Typically different granularity levels from single dimension hierarchy, e.g., year-month-day
  – Database can often find better evaluation strategy

\[
\text{GROUP BY ROLLUP } (a_1, b, c) \\
= \text{GROUP BY GROUPING SETS} \\
\{ \\
(a, b, c), \\
(a, b), \\
a, \\
\} \\
\]

6. OVER clause

• Agg OVER (partition-clause, order-by, window-specification)
• New type of aggregation and grouping where
  – Each input tuple is paired with the aggregation result for the group it belongs to
  – More flexible grouping based on order and windowing
  – New aggregation functions for ranking queries
    – e.g., RANK(), DENSE_RANK()

\[
\text{GROUP BY ROLLUP } (a, b, c) = \text{GROUP BY GROUPING SETS} \\
\{ \\
(a, b, c), \\
(a, b), \\
(a), \\
\} \\
\]

6. GROUPING SETS

SELECT year, month, city, profit
SUM(profit) OVER () AS ttl
FROM sales
• For each tuple build a set of tuples belonging to the same window
  – Compute aggregation function over window
  – Return each input tuple paired with the aggregation result for its window
• OVER() = one window containing all tuples

\[
\begin{array}{cccc}
\text{year} & \text{month} & \text{city} & \text{profit} \\
2010 & 1 & Chicago & 10 \\
2010 & 2 & Chicago & 5 \\
2010 & 3 & Chicago & 20 \\
2011 & 1 & Chicago & 45 \\
2010 & 1 & New York & 12 \\
\end{array}
\]

6. GROUPING SETS

SELECT year, month, city, profit
SUM(profit) OVER (PARTITION BY year) AS ttl
FROM sales
• PARITION BY
  – only tuples with same partition-by attributes belong to the same window
• LIKE GROUP BY

\[
\begin{array}{cccc}
\text{year} & \text{month} & \text{city} & \text{profit} \\
2010 & 1 & Chicago & 10 \\
2010 & 2 & Chicago & 5 \\
2010 & 3 & Chicago & 20 \\
2011 & 1 & Chicago & 45 \\
2010 & 1 & New York & 12 \\
\end{array}
\]
6. GROUPING SETS

```
SELECT year, month, city
SUM(profit) OVER (ORDER BY year, month) AS ttl
FROM sales
```

- **ORDER BY**
  - Order tuples on these expressions
  - Only tuples which are \(\leq\) to the order as the current tuple belong to the same window
- E.g., can be used to compute an accumulate total

<table>
<thead>
<tr>
<th>year</th>
<th>month</th>
<th>city</th>
<th>profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>1</td>
<td>Chicago</td>
<td>10</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>Chicago</td>
<td>5</td>
</tr>
<tr>
<td>2010</td>
<td>3</td>
<td>Chicago</td>
<td>20</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>Chicago</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>New York</td>
<td>18</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>New York</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
</tbody>
</table>

6. GROUPING SETS

```
SELECT year, month, city
SUM(profit) OVER (PARTITION BY year ORDER BY month) AS ttl
FROM sales
```

- **ORDER BY**
  - Order tuples on these expressions
  - Only tuples which are \(\leq\) to the order as the current tuple belong to the same window
- E.g., can be used to compute an accumulate total

<table>
<thead>
<tr>
<th>year</th>
<th>month</th>
<th>city</th>
<th>profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>1</td>
<td>Chicago</td>
<td>10</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>Chicago</td>
<td>5</td>
</tr>
<tr>
<td>2010</td>
<td>3</td>
<td>Chicago</td>
<td>20</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>Chicago</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>New York</td>
<td>18</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>New York</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
</tbody>
</table>

6. GROUPING SETS

```
SELECT year, month, city
SUM(profit) OVER (PARTITION BY year ORDER BY month) AS ttl
FROM sales
```

- **ORDER BY**
  - Order tuples on these expressions
  - Only tuples which are \(\leq\) to the order as the current tuple belong to the same window
- E.g., can be used to compute an accumulate total

<table>
<thead>
<tr>
<th>year</th>
<th>month</th>
<th>city</th>
<th>profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>1</td>
<td>Chicago</td>
<td>10</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>Chicago</td>
<td>5</td>
</tr>
<tr>
<td>2010</td>
<td>3</td>
<td>Chicago</td>
<td>20</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>Chicago</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>New York</td>
<td>18</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>New York</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
</tbody>
</table>

6. GROUPING SETS

```
SELECT year, month, city
SUM(profit) OVER (PARTITION BY year ORDER BY month)
AS ttl
FROM sales
```

- **Combining PARTITION BY and ORDER BY**
  - First partition, then order tuples within each partition

<table>
<thead>
<tr>
<th>year</th>
<th>month</th>
<th>city</th>
<th>profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>1</td>
<td>Chicago</td>
<td>10</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>Chicago</td>
<td>5</td>
</tr>
<tr>
<td>2010</td>
<td>3</td>
<td>Chicago</td>
<td>20</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>Chicago</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>New York</td>
<td>18</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>New York</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
</tbody>
</table>

6. GROUPING SETS

```
SELECT year, month, city
SUM(profit) OVER (ORDER BY year, month)
AS ttl
FROM sales
```

- **Explicit window specification**
  - Requires ORDER BY
  - Determines which tuples “surrounding” the tuple according to the sort order to include in the window

<table>
<thead>
<tr>
<th>year</th>
<th>month</th>
<th>city</th>
<th>profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>1</td>
<td>Chicago</td>
<td>10</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>Chicago</td>
<td>5</td>
</tr>
<tr>
<td>2010</td>
<td>3</td>
<td>Chicago</td>
<td>20</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>Chicago</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>New York</td>
<td>18</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>New York</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
</tbody>
</table>
6. GROUPING SETS

```sql
SELECT year, month, city
    SUM(profit) OVER (ORDER BY year, month
                        RANGE BETWEEN 1 PRECEDING
                        AND 1 FOLLOWING) AS ttl
FROM sales
```

- **Explicit window specification**
  - Requires **ORDER BY**
  - Determines which tuples “surrounding” the tuple according to the sort order to include in the window

### Example Table

<table>
<thead>
<tr>
<th>Year</th>
<th>Month</th>
<th>City</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>1</td>
<td>Chicago</td>
<td>10</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
<td>Chicago</td>
<td>5</td>
</tr>
<tr>
<td>2010</td>
<td>3</td>
<td>Chicago</td>
<td>20</td>
</tr>
<tr>
<td>2011</td>
<td>1</td>
<td>Chicago</td>
<td>45</td>
</tr>
<tr>
<td>2010</td>
<td>1</td>
<td>New York</td>
<td>12</td>
</tr>
</tbody>
</table>

### MDX

- **Multidimensional expressions (MDX)**
  - Introduced by Microsoft
  - Query language for the cube data model
  - SQL-like syntax
    - Keywords have different meaning
    - MDX queries return a multi-dimensional report
      - 2D = spreadsheet
      - 3D or higher, e.g., multiple spreadsheets

### MDX Query

- **Basic Query Structure**
  ```sql
  SELECT <axis-spec>, ...  
  FROM <cube-spec>, ...  
  WHERE ( <select-spec> )
  ```

- **Note!**
  - Semantics of SELECT, FROM, WHERE not what you would expect knowing SQL

### MDX - SELECT

- **Select specifies dimensions in result and how to visualize**
  - ON COLUMNS, ON ROWS, ON PAGES, OR SECTIONS, OR CRAFTERS
  - Every dimension in result corresponds to one dimension in the cube
  - Set of concepts from this dimensions which may be from different levels of granularity
6. MXD - SELECT

- Specify concepts from dimensions
  - List all values as set, e.g., \{[2010], [2011]\}
  - Not necessarily from same level of hierarchy (e.g., mix years and months)

- Language constructs for accessing parents and children or members of a level in the hierarchy
  - \textsc{children}: all direct children
    - E.g., [2010].\textsc{children} = \{[2010 Jan], ..., [2010 Dec]\}
  - \textsc{parent}: the direct parent
    - E.g., [2010 Jan].\textsc{parent} = [2010]
  - \textsc{members}: all direct children
    - E.g., Time.Years.\textsc{members} = \{[1990], [1991], ..., [2016]\}
  - \textsc{lastchild}: last child (according to order of children)
    - E.g., [2010].\textsc{lastchild} = [2010 Dec]
  - \textsc{nextmember}: right sibling on same level
    - E.g., [2010].\textsc{nextmember} = [2011]
  - \([a]:[b]\): all members in interval between \(a\) and \(b\)

6. MXD - SELECT

- Nesting of sets: \textsc{crossjoin}
  - Project two dimensions into one
  - Forming all possible combinations

\[
\text{SELECT} \ \text{crossjoin}\{\ 
\text{Chicago, Schaumburg} \} , \ 
\{\ [2010], [2011] \}\ 
\text{on rows} \ 
\text{on columns} \ 
\text{from PhoneCallsCube} \ 
\text{where \{Measures\}.numCalls}\ 
\]

6. MXD - SELECT

- Conditional selection of members: \textsc{filter}
  - One use members that fulfill condition
  - E.g., condition over aggregation result

\[
\text{SELECT} \ \text{filter}\{\ [2010].\text{children}, \ 
(Sprint, numCalls) > (ATT, numCalls)\} \ 
\text{on rows} \ 
\text{on columns} \ 
\text{from PhoneCallsCube} \ 
\text{where \{Measures\}.numCalls}\ 
\]

6. Query Processing in DW

- Large topic, here we focus on two aspects
  - Partitioning
  - Query answering with materialized views

6. Partitioning

- \textsc{partitioning} splits a table into multiple fragments that are stored independently
  - E.g., split across X disks, across Y servers

- \textsc{vertical partitioning}
  - Split columns across fragments
    - E.g., \textit{R} = \{A,B,C,D\}, fragment \textit{F}1 = \{A,B\}, \textit{F}2 = \{C,D\}
    - Either add a row id to each fragment or the primary key to be able to reconstruct

- \textsc{horizontal partitioning}
  - Split rows
  - Hash vs. range partitioning
6. Partitioning

• Why partitioning?
  – Parallel/distributed query processing
    • read/write fragments in parallel
    • Distribute storage load across disks/servers
  – Avoid reading data that is not needed to answer a query
    • Vertical
      – Only read columns that are accessed by query
    • Horizontal
      – only read tuples that may match queries selection conditions

6. Partitioning

• Vertical Partitioning
  – Fragments $F_1$ to $F_n$ of relation $R$ such that
    • $Sch(F_1) \cup Sch(F_2) \cup \ldots \cup Sch(F_n) = Sch(R)$
    • Store row id or PK of $R$ with every fragment
    • Restore relation $R$ through natural joins

<table>
<thead>
<tr>
<th>Name</th>
<th>Salary</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>12,000</td>
<td>45</td>
<td>M</td>
</tr>
<tr>
<td>Alice</td>
<td>24,000</td>
<td>34</td>
<td>F</td>
</tr>
<tr>
<td>Bob</td>
<td>20,000</td>
<td>22</td>
<td>M</td>
</tr>
<tr>
<td>Gertrud</td>
<td>50,000</td>
<td>55</td>
<td>F</td>
</tr>
<tr>
<td>Pferdegert</td>
<td>14,000</td>
<td>23</td>
<td>M</td>
</tr>
</tbody>
</table>

6. Partitioning

• Horizontal Partitioning
  – Range partitioning on attribute $A$
    • Split domain of $A$ into intervals representing fragments
    • E.g., tuples with $A = 15$ belong to fragment $[0,20]$
  – Fragments $F_1$ to $F_n$ of relation $R$ such that
    • $Sch(F_1) = Sch(F_2) = \ldots = Sch(F_n) = Sch(R)$
    • $R = F_1 \cup \ldots \cup F_n$

<table>
<thead>
<tr>
<th>Name</th>
<th>Salary</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>12,000</td>
<td>45</td>
<td>M</td>
</tr>
<tr>
<td>Alice</td>
<td>24,000</td>
<td>34</td>
<td>F</td>
</tr>
<tr>
<td>Bob</td>
<td>20,000</td>
<td>22</td>
<td>M</td>
</tr>
<tr>
<td>Gertrud</td>
<td>50,000</td>
<td>55</td>
<td>F</td>
</tr>
<tr>
<td>Pferdegert</td>
<td>14,000</td>
<td>23</td>
<td>M</td>
</tr>
</tbody>
</table>

6. Partitioning

• Horizontal Partitioning
  – Hash partitioning on attribute $A$
    • Split domain of $A$ into $x$ buckets using hash function
    • E.g., tuples with $h(A) = 3$ belong to fragment $F_3$
    • $Sch(F_1) = Sch(F_2) = \ldots = Sch(F_n) = Sch(R)$
    • $R = F_1 \cup \ldots \cup F_n$

<table>
<thead>
<tr>
<th>Name</th>
<th>Salary</th>
<th>Age</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>12,000</td>
<td>45</td>
<td>M</td>
</tr>
<tr>
<td>Alice</td>
<td>24,000</td>
<td>34</td>
<td>F</td>
</tr>
<tr>
<td>Bob</td>
<td>20,000</td>
<td>22</td>
<td>M</td>
</tr>
<tr>
<td>Gertrud</td>
<td>50,000</td>
<td>55</td>
<td>F</td>
</tr>
<tr>
<td>Pferdegert</td>
<td>14,000</td>
<td>23</td>
<td>M</td>
</tr>
</tbody>
</table>

Outline

0) Course Info
1) Introduction
2) Data Preparation and Cleaning
3) Schema matching and mapping
4) Virtual Data Integration
5) Data Exchange
6) Data Warehousing
7) Big Data Analytics
8) Data Provenance
3. Big Data Analytics

- Big Topic, big Buzzwords ;-)  
- Here
  - Overview of two types of systems
    - Key-value/document stores
    - Mainly: Bulk processing (MR, graph, …)
  - What is new compared to single node systems?
  - How do these systems change our approach to integration/analytics
  - Schema first vs. Schema later
  - Pay-as-you-go

3. Big Data Overview

- 2) Overview of systems and how they achieve scalability
  - Bulk processing
    - MapReduce, Shark, Flink, Hyracks, …
    - Graph: e.g., Giraph, Pregel, …
  - Key-value/document stores = NoSQL
    - Cassandra, MongoDB, Memcached, Dynamo, …
  - Fault tolerance
    - Replication
    - Handling stragglers
  - Load balancing
    - Partitioning
    - Shuffle
3. Big Data Overview

• 3) New approach towards integration
  – Large clusters enable directly running queries over semi-structured data (within feasible time)
    • Take a click-stream log and run a query
  – One of the reasons why pay-as-you-go is now feasible
    • Previously: designing a database schema upfront and designing a process (e.g., ETL) for cleaning and transforming data to match this schema, then query
    • Now: start analysis directly, clean and transform data if needed for the analysis

3. Big Data Overview

• 3) New approach towards integration
  – Advantage of pay-as-you-go
    • More timely data (direct access)
    • More applicable if characteristics of data change dramatically (e.g., yesterdays ETL process no longer applicable)
  – Disadvantages of pay-as-you-go
    • Potentially repeated efforts (everybody cleans the click-log before running the analysis)
    • Lack of meta-data may make it hard to
      – Determine what data to use for analysis
      – Hard to understand semantics of data

3. Big Data Overview

• Scalable systems
  – Performance of the system scales in the number of nodes
    • Ideally the per node performance is constant independent of how many nodes there are in the system
    • This means: having twice the number of nodes would give us twice the performance
  – Why scaling is important?
    • If a system scales well we can “throw” more resources at it to improve performance and this is cost effective

3. Big Data Overview

• What impacts scaling?
  – Basically how parallelizable is my algorithm
    • Positive example: problem can be divided into subproblems that can be solved independently without requiring communication
      – E.g., array of 1-billion integers $[i_1, \ldots, i_{1,000,000}]$ add 3 to each integer. Compute on n nodes, split input into n equally sized chunks and let each node process one chunk
    • Negative example: problem where subproblems are strongly intercorrelated
      – E.g., Context Free Grammar Membership: given a string and a context free grammar, does the string belong to the language defined by the grammar.

3. Big Data – Processing at Scale

• New problems at scale
  – DBMS
    • running on 1 or 10’s of machines
    • running on 1000’s of machines
  • Each machine has low probability of failure
    – If you have many machines, failures are the norm
    – Need mechanisms for the system to cope with failures
      • Do not loose data
      • Do not use progress of computation when node fails
    – This is called fault-tolerance

3. Big Data – Processing at Scale

• New problems at scale
  – DBMS
    • running on 1 or 10’s of machines
    • running on 1000’s of machines
  • Each machine has limited storage and computational capabilities
    – Need to evenly distribute data and computation across nodes
      • Often most overloaded node determine processing speed
    – This is called load-balancing
3. Big Data – Processing at Scale

- Building distributed systems is hard
  - Many pitfalls
    - Maintaining distributed state
    - Fault tolerance
    - Load balancing
  - Requires a lot of background in
    - OS
    - Networking
    - Algorithm design
    - Parallel programming

- Hard to debug
  - Even debugging a parallel program on a single machine is already hard
    - Non-determinism because of scheduling: Race conditions
    - In general hard to reason over behavior of parallel threads of execution
  - Even harder when across machines
    - Just think about how hard it was for you to first program with threads/processes

3. Big Data – Why large scale?

- Datasets are too large
  - Storing a 1 Petabyte dataset requires 1 PB storage
    - Not possible on single machine even with RAID storage
  - Processing power/bandwidth of single machine is not sufficient
    - Run a query over the facebook social network graph
      - Only possible within feasible time if distributed across many nodes

3. Big Data – User’s Point of View

- How to improve the efficiency of distributed systems experts
  - Building a distributed system from scratch for every store and analysis task is obviously not feasible!
- How to support analysis over large datasets for non distributed systems experts
  - How to enable somebody with some programming but limited/no distributed systems background to run distributed computations

3. Big Data – Abstractions

- Solution
  - Provide higher level abstractions
- Examples
  - MPI (message passing interface)
    - Widely applied in HPC
    - Still quite low-level
  - Distributed file systems
    - Make distribution of storage transparent
  - Key-value storage
    - Distributed store/retrieval of data by identifier (key)
- More Examples
  - Distributed table storage
    - Store relations, but no SQL interface
  - Distributed programming frameworks
    - Provide a, typically, limited programming model with automated distribution
  - Distributed databases, scripting languages
    - Provide a high-level language, e.g., SQL-like with an execution engine that is distributed
3. Distributed File Systems

- Transparent distribution of storage
  - Fault tolerance
  - Load balancing?

- Examples
  - HPC distributed filesystems
    - Typically assume a limited number of dedicated storage servers
    - GPFS, Lustre, PVFS
  - “Big Data” filesystems
    - Google file system, HDFS

3. HDFS

- Hadoop Distributed Filesystem (HDFS)
  - Architecture
    - One nodes storing metadata (name node)
    - Many nodes storing file content (data nodes)
  - Filestructure
    - Files consist of blocks (e.g., 64MB size)
  - Limitations
    - Files are append only

3. HDFS

- Name node
  - Stores the directory structure
  - Stores which blocks belong to which files
  - Stores which nodes store copies of which block
  - Detects when data nodes are down
  - Clients communicate with the name node to gather FS metadata

3. HDFS

- Data nodes
  - Store blocks
  - Send/receive file data from clients
  - Send heart-beat messages to name node to indicate that they are still alive
  - Clients communicate data nodes for reading/writing files

3. Distributed FS Discussion

- What do we get?
  - Can store files that do not fit onto single nodes
  - Get fault tolerance
  - Improved read speed (caused on replication)
  - Decreased write speed (caused by replication)

- What is missing?
  - Computations
3. Frameworks for Distributed Computations

- **Problems**
  - Not all algorithms do parallelize well
  - How to simplify distributed programming?

- **Solution**
  - Fix a reasonable powerful, but simple enough model of computation for which scalable algorithms are known
  - Implement distributed execution engine for this model and make it fault tolerant and load-balanced

---

3. MapReduce

- **Data Model**
  - Sets of key-value pairs \( \{(k_1, v_1), \ldots, (k_n, v_n)\} \)
  - **Key** is an identifier for a piece of data
  - **Value** is the data associated with a key

- **Programming Model**
  - We have two higher-level functions map and reduce
    - Take as input a user-defined function that is applied to elements in the input key-value pair set
    - Complex computations can be achieved by chaining map-reduce computations

---

3. MapReduce Datamodel

- **Data Model**
  - Sets of key-value pairs \( \{(k_1, v_1), \ldots, (k_n, v_n)\} \)
  - **Key** is an identifier for a piece of data
  - **Value** is the data associated with a key

- **Examples**
  - Document \( d \) with an id
    - \((\text{id}, d)\)
  - Person with name, salary, and SSN
    - \((\text{SSN}, \text{"name, salary"})\)
3. MapReduce Computational Model

- **Example**
  - **Input**: Set of (state, population) pairs one for each city in the state
  - **Task**: compute the total population per state

- **Reduce function**
  - \( f: (\text{state}, [p_1, \ldots, p_n]) \rightarrow \{\text{state}, \text{SUM}(p_1, \ldots, p_n)\} \)

- **Application of f through map**
  - **Input**: \(\{(\text{illinois}, 3), (\text{illinois}, 1), (\text{oregon}, 15)\}\)
  - **Output**: \(\{(\text{illinois}, 4), (\text{oregon}, 15)\}\)

3. MapReduce Workflows

- **Workflows**
  - Computations in MapReduce consists of map phases followed by reduce phases
    - The input to the reduce phase is the output of the map phase
    - Complex computations may require multiple map-reduce phases to be chained together

3. MapReduce Implementations

- **MapReduce**
  - Developed by google
  - Written in C
  - Runs on top of GFS (Google’s distributed filesystem)

- **Hadoop**
  - Open source Apache project
  - Written in Java
  - Runs on-top of HDFS

3. Hadoop

- **Anatomy of a Hadoop cluster**
  - **Job tracker**
    - Clients submit MR jobs to the job tracker
    - Job tracker monitors progress
  - **Task tracker aka workers**
    - Execute map and reduce jobs

- **Job**
  - **Input**: files from HDFS
  - **Output**: written to HDFS
  - Map/Reduce UDFs

- **Fault tolerance**
  - **Handling stragglers**
    - Job tracker will reschedule jobs to a different worker if the worker falls behind too much with processing
  - **Materialization**
    - Inputs are read from HDFS
    - Workers write results of map jobs assigned to them to local disk
    - Workers write results of reduce jobs to HDFS for persistence
3. Hadoop – MR Job

- Job tracker sends jobs to task tracker on worker nodes
- Try to schedule map jobs on nodes that store the chunk processed by a job
- Job tracker monitors progress

- Each mapper reads its chunk from HDFS, translates the input into key-value pairs and applies the map UDF to every (k,v)
- Outputs are written to disk with one file per reducer hashing on key
- Job tracker may spawn additional mappers if mappers are not making progress

3. Combiners

- Certain reduce functions lend themselves to pre-aggregation
  - E.g., SUM(revenue) group by state
  - Can compute partial sums over incomplete groups and then sum up the pre-aggregated results
  - This can be done at the mappers to reduce amount of data send to the reducers
- Supported in Hadoop through a user provided combiner function
  - The combiner function is applied before writing the mapper results to local disk
3. Example code – Word count

- https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

```java
public static class MapReduceExample implements Mapper<K, Text, Text, IntWritable> {
    public void map(K key, Text value, OutputCollector<Text, IntWritable> output, Reporter report) {
        int count = 1;
        output.collect(value, new IntWritable(count));
    }
}
```

3. Systems/Languages on top of MapReduce

- Pig
  - Scripting language, compiled into MR
  - Akin to nested relational algebra
- Hive
  - SQL interface for warehousing
  - Compiled into MR
  - ...

```sql
FROM (MAP doc.text USING 'python wc_mapper.py' AS (word, cnt))
FROM docs
CLUSTER BY word
REDUCE word, cnt USING 'python wc_reduce.py';
```

3. Hive Architecture
3. Hive Datamodel

- **Tables**
  - Attribute-DataType pairs
  - User can instruct Hive to partition the table in a certain way

- **Datatypes**
  - Primitive: integer, float, string
  - Complex types
    - Map: Key->Value
    - List
    - Struct
  - Complex types can be nested

- **Example**: `CREATE TABLE t1(st string, fl float, li list<map<string, struct<p1:int, p2:int>>);`

- **Implementation**
  - Tables are stored in HDFS
  - Serializer/Deserializer - transform for querying

3. Hive - Query Processing

- Compile HiveQL query into DAG of map and reduce functions.
  - A single map/reduce may implement several traditional query operators
    - E.g., filtering out tuples that do not match a condition (selection) and filtering out certain columns (projection)
    - Hive tries to use the partition information to avoid reading partitions that are not needed to answer the query
      - For example
        - table instructor(name, department) is partitioned on department
        - `SELECT name FROM instructor WHERE department = 'CS'`
        - This query would only access the partition of the table for department 'CS'

3. Operator implementations

- **Join implementations**
  - **Broadcast join**
    - Send the smaller table to all nodes
    - Process the other table partitioned
      - Each node finds all the join partners for a partition of the larger table and the whole smaller table
  - **Reduce join (partition join)**
    - Use a map job to create key-value pairs where the key is the join attributes
    - Reducer output joined rows

3. Example plan

Spark

- MR uses heavy materialization to achieve fault tolerance
  - A lot of I/O
- **Spark**
  - Works in main memory (where possible)
  - Inputs and final outputs stored in HDFS
  - Recomputes partial results instead of materializing them - resilient distributed datasets (RDD)
  - **Lineage**: Need to know from which chunk a chunk was derived from and by which computation

Summary

- Big data storage systems
- Big data computation platforms
- Big data “databases”
- How to achieve scalability
  - Fault tolerance
  - Load balancing
- Big data integration
  - Pay-as-you-go
  - Schema later
<table>
<thead>
<tr>
<th>Outline</th>
</tr>
</thead>
<tbody>
<tr>
<td>0) Course Info</td>
</tr>
<tr>
<td>1) Introduction</td>
</tr>
<tr>
<td>2) Data Preparation and Cleaning</td>
</tr>
<tr>
<td>3) Schema matching and mapping</td>
</tr>
<tr>
<td>4) Virtual Data Integration</td>
</tr>
<tr>
<td>5) Data Exchange</td>
</tr>
<tr>
<td>6) Data Warehousing</td>
</tr>
<tr>
<td>7) Big Data Analytics</td>
</tr>
<tr>
<td>8) Data Provenance</td>
</tr>
</tbody>
</table>