IBM Corporation

Resolution Aware Query Answering for Business Intelligence

Presenter:	Ling Wang	(IBM Silicon Valley Lab)
Authors:	Yannis Sismanis	(IBM Almaden Research Center)
	Ling Wang	(IBM Silicon Valley Lab)
	Ariel Fuxman	(Microsoft Research Center)
	Peter J. Haas	(IBM Almaden Research Center)
	Berthold Reinwald	(IBM Almaden Research Center)

Outline

- Motivation & Background
  - Business Intelligence application requirements and entity resolution
  - Traditional entity-resolution approach and its drawbacks
- **R**esolution **A**ware **Q**uery **A**nswering (RAQA)
  - Conceptual approach
  - Efficient algorithms using existing DBMS
  - Experimental results
- Q&A

Motivation – Traditional data cleaning for BI

- Business Intelligence (BI) applications require:
  - Grouping business entities by common attributes (e.g., city, year)
  - Aggregating measures of interest (e.g., sales amount)
- However, data is often …
  - From heterogeneous sources, lacking overall data integrity
  - Uncertain and inconsistent
- Current systems clean data before BI queries are executed
  - Entity resolution at load time (during ETL phase)
  - Is this the best approach?

Motivation – Traditional Entity Resolution for BI

- Entity discovery and resolution

Entity discovery → Entity Resolution

Op.System	CustId	Name	City	State
1	C_1	John	San Francisco	CA
2	C_2	Johnny	San Jose	CA
1	C_3	Jack	San Francisco	CA
1	C_4	William	San Francisco	CA
2	C_5	Bill	San Jose	CA

(a) Customer Data

EntityID	CustID
e_1	C_1
e_1	C_2
e_2	C_3
e_3	C_4
e_3	C_5

(a) Entity Mapping Table

EntityID	City	State
e_1	San Francisco	CA
e_2	San Francisco	CA
e_3	San Francisco	CA

(b) Possible Entity Resolution

Loss of information & High expense

- Answer BI query after data cleaning

Op.System	TransID	CustID	Sales
1	Tr_1	C_1	$15
1	Tr_2	C_1	$5
2	Tr_3	C_2	$30
2	Tr_4	C_2	$20
1	Tr_5	C_3	$30
1	Tr_6	C_4	$90
2	Tr_7	C_5	$25
2	Tr_8	C_5	$15

(b) Transaction Data

EntityId	City	State
e_1	San Jose	CA
e_2	San Francisco	CA
e_3	San Francisco	CA

Dangerous… what if e1 is actually from San Jose ?

City	State	Sum(Sales)
San Francisco	CA	$230

(c) Possible Sum of Sales Grouped by City, State

Uncertainty is IGNORED, which leads to RISK

Resolution-Aware Query Answer (RAQA)

- Databases are maintained in an unresolved state

Op.System	CustId	Name	City	State
1	C_1	John	San Francisco	CA
2	C_2	Johnny	San Jose	CA
1	C_3	Jack	San Francisco	CA
1	C_4	William	San Francisco	CA
2	C_5	Bill	San Jose	CA

(a) Customer Data

EntityID	CustID
e_1	C_1
e_1	C_2
e_2	C_3
e_3	C_4
e_3	C_5

(a) Entity Mapping Table

EntityID	City	State
e_1	San Francisco	CA

EntityID	City	State
e_1	San Francisco	CA

EntityID	City	State
e_1	San Francisco	CA

EntityId	City	State
e_1	San Jose	CA
e_2	San Francisco	CA
e_3	San Francisco	CA

- The answer to a BI query reflects the data inconsistency

Op.System	TransID	CustID	Sales
1	Tr_1	C_1	$15
1	Tr_2	C_1	$5
2	Tr_3	C_2	$30
2	Tr_4	C_2	$20
1	Tr_5	C_3	$30
1	Tr_6	C_4	$90
2	Tr_7	C_5	$25
2	Tr_8	C_5	$15

(b) Transaction Data

City	State	strict range	status
San Francisco	CA	[$30,$230]	*guaranteed*
San Jose	CA	[$70,$200]	*non-guaranteed*

(a) Grouped by City, State

At least one possible resolution where no entity is in (San Jose, CA),
So that aggregation result is undefined

All other resolutions fall into the indicated range

Benefits of RAQA

- No information loss
- No expensive entity-resolution
- Risk-management
  - Query result contains information about data uncertainty
    e.g. Small range => **high quality** integrated data
  - Multi-RAQA analysis of uncertainty
    e.g. Large SUM range + small COUNT range
    => **small** group of **large-valued** transactions with uncertain attributes

Conceptual Model of RAQA

Op.System	CustId	Name	City	State
1	C_1	John	San Francisco	CA
2	C_2	Johnny	San Jose	CA
1	C_3	Jack	San Francisco	CA
1	C_4	William	San Francisco	CA
2	C_5	Bill	San Jose	CA

(a) **Customer Data**

EntityID	CustID
e_1	C_1
e_1	C_2
e_2	C_3
e_3	C_4
e_3	C_5

(a) Entity Mapping Table

Op.System	TransID	CustID	Sales
1	Tr_1	C_1	$15
1	Tr_2	C_1	$5
2	Tr_3	C_2	$30
2	Tr_4	C_2	$20
1	Tr_5	C_3	$30
1	Tr_6	C_4	$90
2	Tr_7	C_5	$25
2	Tr_8	C_5	$15

(b) Transaction Data

TransID	EntityID	Sales
Tr_1	e_1	$15
Tr_2	e_1	$5
Tr_3	e_1	$30
Tr_4	e_1	$20
Tr_5	e_2	$30
Tr_6	e_3	$90
Tr_7	e_3	$25
Tr_8	e_3	$15

TABLE IV

RESOLVED FACT TABLE

Possible resolution:

EntityId	City	State
e_1	San Francisco	CA
e_2	San Francisco	CA
e_3	San Francisco	CA

EntityId	City	State
e_1	San Francisco	CA
e_2	San Francisco	CA
e_3	San Jose	CA

EntityId	City	State
e_1	San Jose	CA
e_2	San Francisco	CA
e_3	San Francisco	CA

EntityId	City	State
e_1	San Jose	CA
e_2	San Francisco	CA
e_3	San Jose	CA

San Francisco	CA	$230

San Francisco	CA	$100
San Jose	CA	$130

San Francisco	CA	$160
San Jose	CA	$70

San Francisco	CA	$30
San Jose	CA	$200

City	State	strict range	status
San Francisco	CA	[$30,$230]	*guaranteed*
San Jose	CA	[$70,$200]	*non-guaranteed*

(a) Grouped by City, State

RAQA Algorithm

- Step 1: Entity Aggregation

Op.System	CustId	Name	City	State
1	C_1	John	San Francisco	CA
2	C_2	Johnny	San Jose	CA
1	C_3	Jack	San Francisco	CA
1	C_4	William	San Francisco	CA
2	C_5	Bill	San Jose	CA

(a) Customer Data

EntityID	CustID
e_1	C_1
e_1	C_2
e_2	C_3
e_3	C_4
e_3	C_5

a) Entity Mapping Table

EntityID	G	status
e_1	San Francisco, CA	inconsistent
e_1	San Jose, CA	inconsistent
e_2	San Francisco, CA	consistent
e_3	San Francisco, CA	inconsistent
e_3	San Jose, CA	inconsistent

Op.System	TransID	CustID	Sales
1	Tr_1	C_1	$15
1	Tr_2	C_1	$5
2	Tr_3	C_2	$30
2	Tr_4	C_2	$20
1	Tr_5	C_3	$30
1	Tr_6	C_4	$90
2	Tr_7	C_5	$25
2	Tr_8	C_5	$15

(b) Transaction Data

EntityID	ecount(m)	esum(m)	emin(m)	emax(m)
e_1	4	$70	$5	$30
e_2	1	$30	$30	$30
e_3	3	$130	$15	$90

- Step 2: RAQA for aggregation function

SUM()

Lower	all-inconsistent, all-nonneg.	$\min_{I^+}$ esum *
	all-inconsistent, some-negative	$\sum_{I^-}$ esum *
	otherwise	$\sum_{C \cup I^-}$ esum
Upper	all-inconsistent, all-negative	$\max_{I^-}$ esum *
	all-inconsistent, some-nonneg.	$\sum_{I^+}$ esum *
	otherwise	$\sum_{C \cup I^+}$ esum

EntityID	G	esum(m)	status
e_1	San Jose, CA	$70	inconsistent
e_3	San Jose, CA	$130	inconsistent

G	Lower	Upper	Status
San Jose, CA	$70	$200	non-guaranteed

Experimental Evaluation

- TPC-H benchmark SF – Scale Factor ED – Entity Duplication IP – Inconsistency percentage

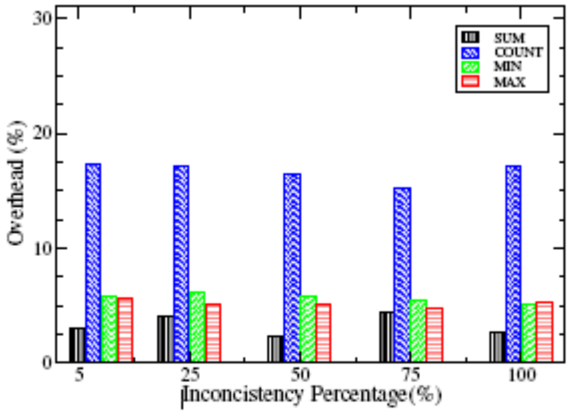

Fig. 1. Effect of inconsistency percentage (SF=9GB, ED=2)

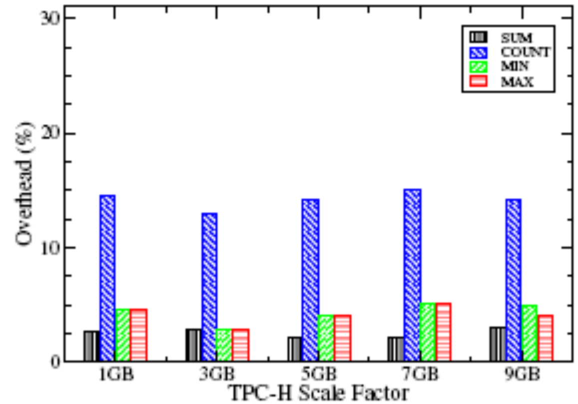

Fig. 2. Effect of scale factor (IP=25%, ED=2)

Overhead is acceptable over a wide range of conditions

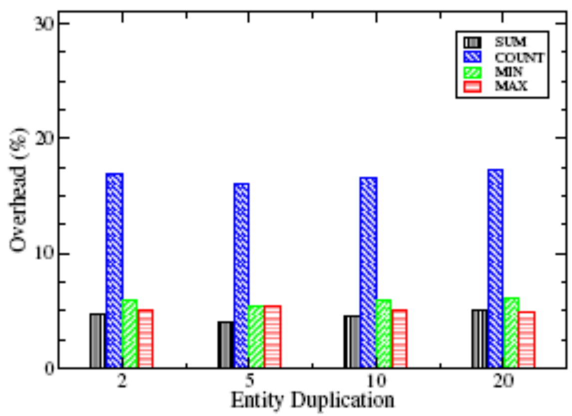

Fig. 3. Effect of entity duplication (IP=25%, SF=9GB)

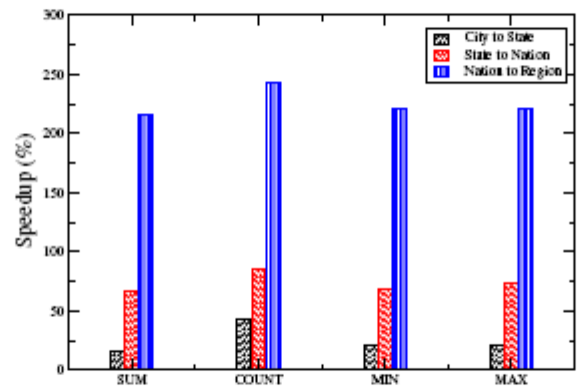

Fig. 4. Benefit of Roll-Up (IP=25%, SF=9GB)

Can exploit classical performance benefits of Roll-Up

Contributions

- Enhanced the OLAP model with resolution-aware aggregations and their semantics
  - Eliminates ETL costs
  - Exposes and quantifies uncertainty at user level, for risk assessment
- Group-by queries:
  - Efficient algorithms for all core aggregation functions
  - Implemented in traditional RDBMS via SQL queries
  - Immediately and widely applicable
- Rollup queries:
  - Based on aggregation result
  - No access to original data
  - Dramatic performance benefits
- Performance:
  - Only 5% overhead for the majority of queries
  - Insensitive to DB size and the degree of inconsistency

Q&A

- Thank you very much for your attention!

- Questions…

Related Work

- Probabilistic database (probDB):
  Probability distribution over possible query results
  - ✓ Sharper picture of data uncertainty
  - – Cannot handle OLAP operations (e.g., roll-up) efficiently
  - – Cannot easily implement on top of a traditional DBMS
  - – Affected by uncertainty

- Query in Inconsistent Database
  - ✓ Query in inconsistent database
  - – Does not focus on aggregation queries.

[1] **A. Fuxman, E. Fazli, and R. J. Miller, "Efficient management of inconsistent databases," in SIGMOD, 2005, pp. 155–166.**

[2] A. Fuxman and R. J. Miller, "First-Order Query Rewriting for Inconsistent Databases," in ICDT, 2005, pp. 337–351.