ENTREPÔTS, REPRÉSENTATION & INGÉNIERIE des CONNAISSANCES

ACM Fifteenth International Workshop On Data Warehousing and OLAP

DOLAP 2012

Colocated with
ACM CIKM 2012

Maui, Hawaii, USA
November 2, 2012

Benchmarking Summarizability Processing in XML Warehouses with Complex Hierarchies

By Chantola KIT
Marouane HACHICHA
Jérôme DARMONT

Lyon 1 — UNIVERSITÉ LUMIÈRE LYON 2 — UNIVERSITÉ DE LYON — UNIVERSITÉ DE LYON

Outline

- Introduction
- Background
- Benchmark Specification
- Experimental Demonstration
- Conclusion and Future Work

Introduction

- Decision Making:

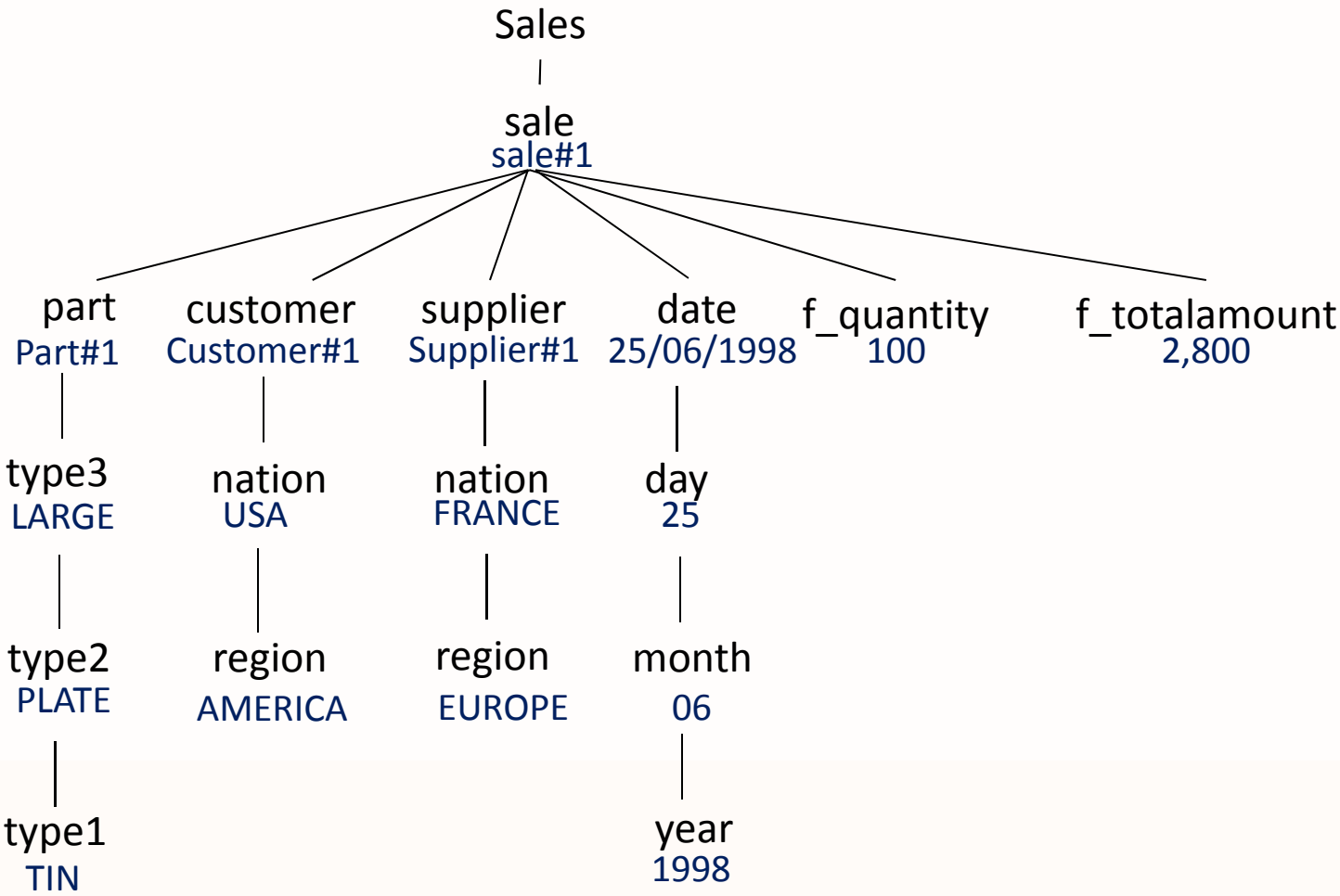
1. **Business Intelligence (BI)** is famed for complex analysis
   - OLAP is a notable BI tool for multi-dimensional analysis
2. **DWs**: collection of historical and concurrent data
   - XML is widely used to represent **complex hierarchical data**

Introduction (Cont.)

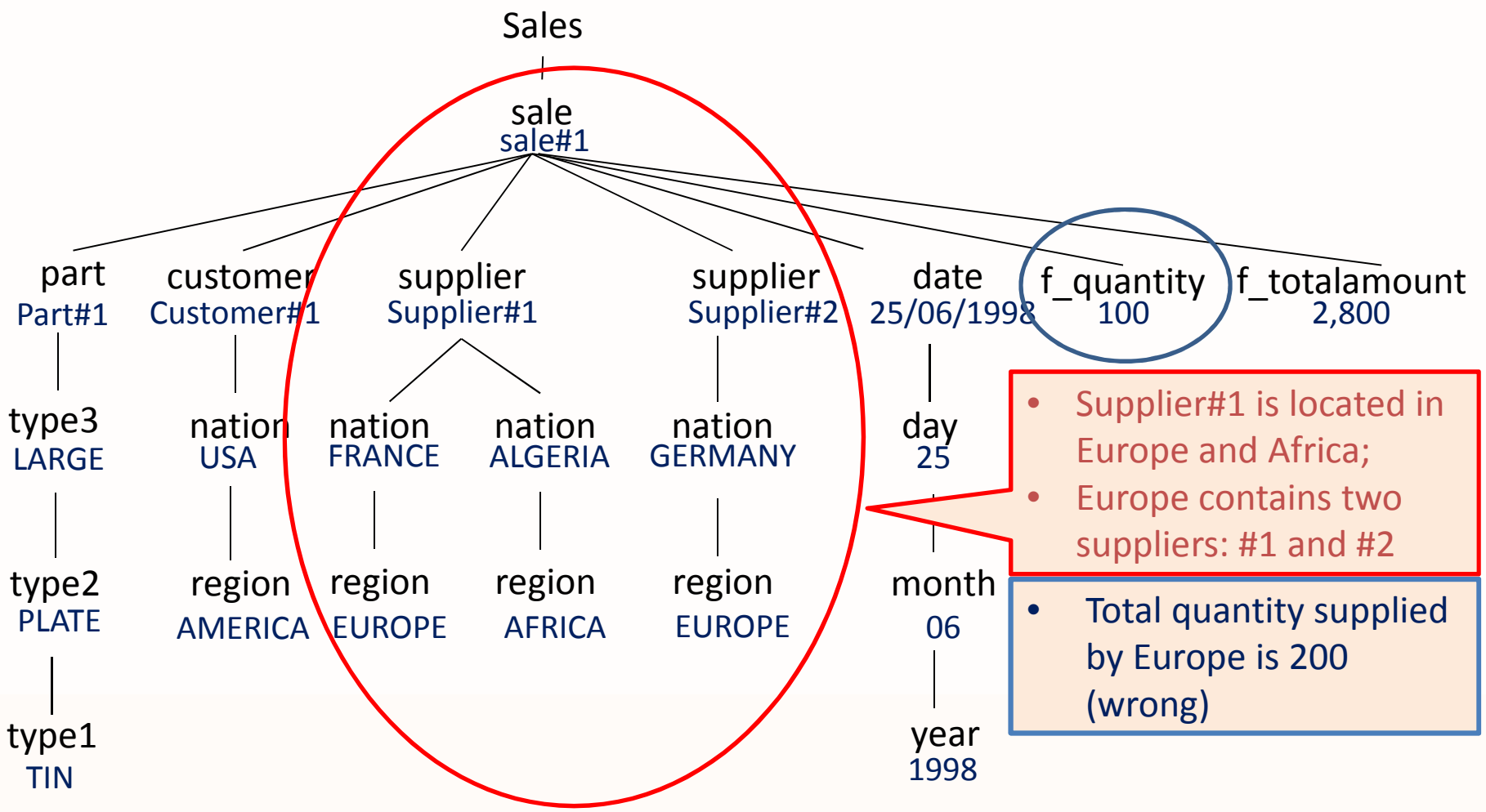
Effectiveness of Summarizability processing on complex hierarchies

- Benchmarks are used to support performance evaluation
- Existing XML data warehouse benchmark: XWeB
  - Complex hierarchies are not scalable

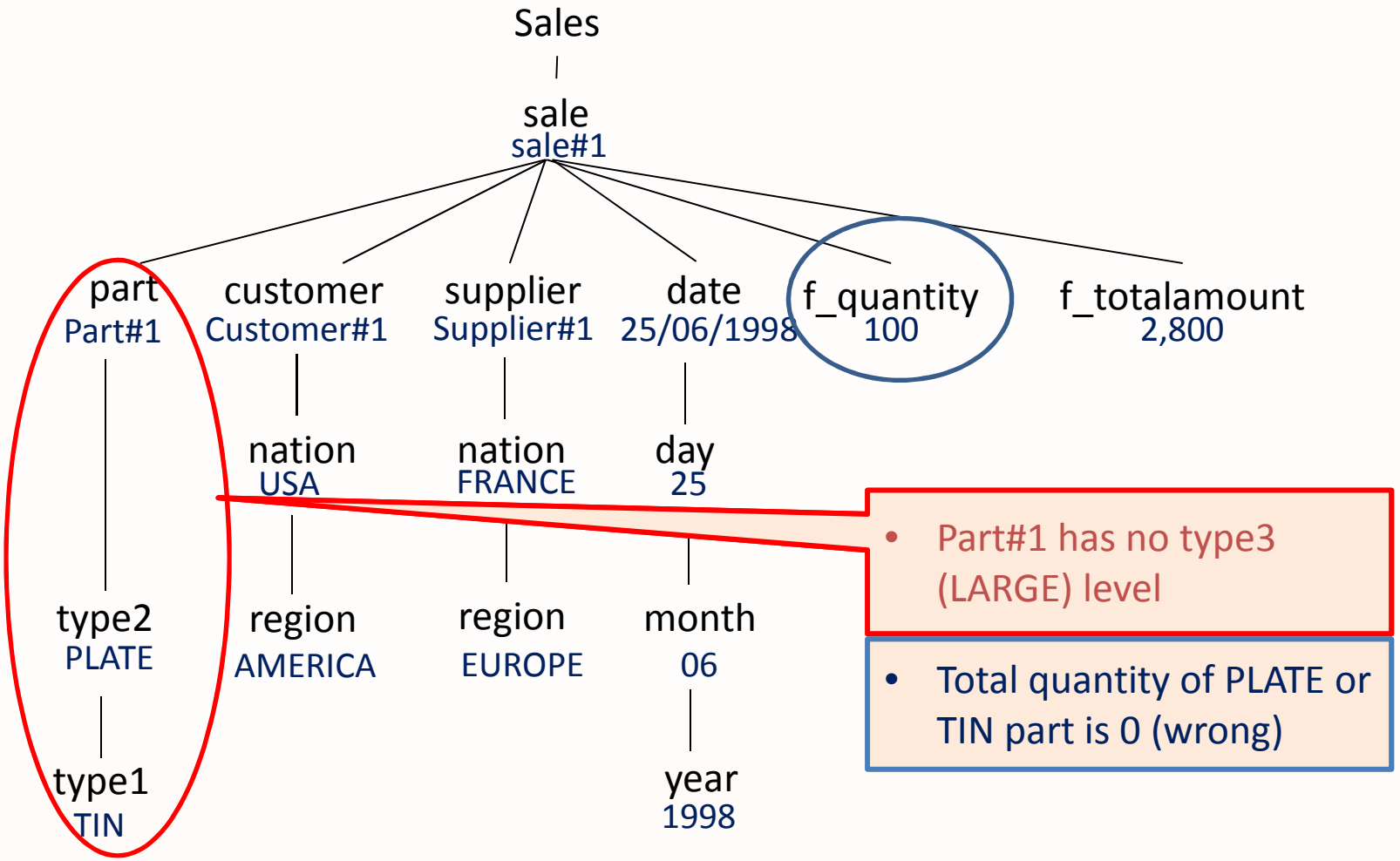
XML Data Example



Non-Strict Hierarchies



Incomplete Hierarchies



Related Work

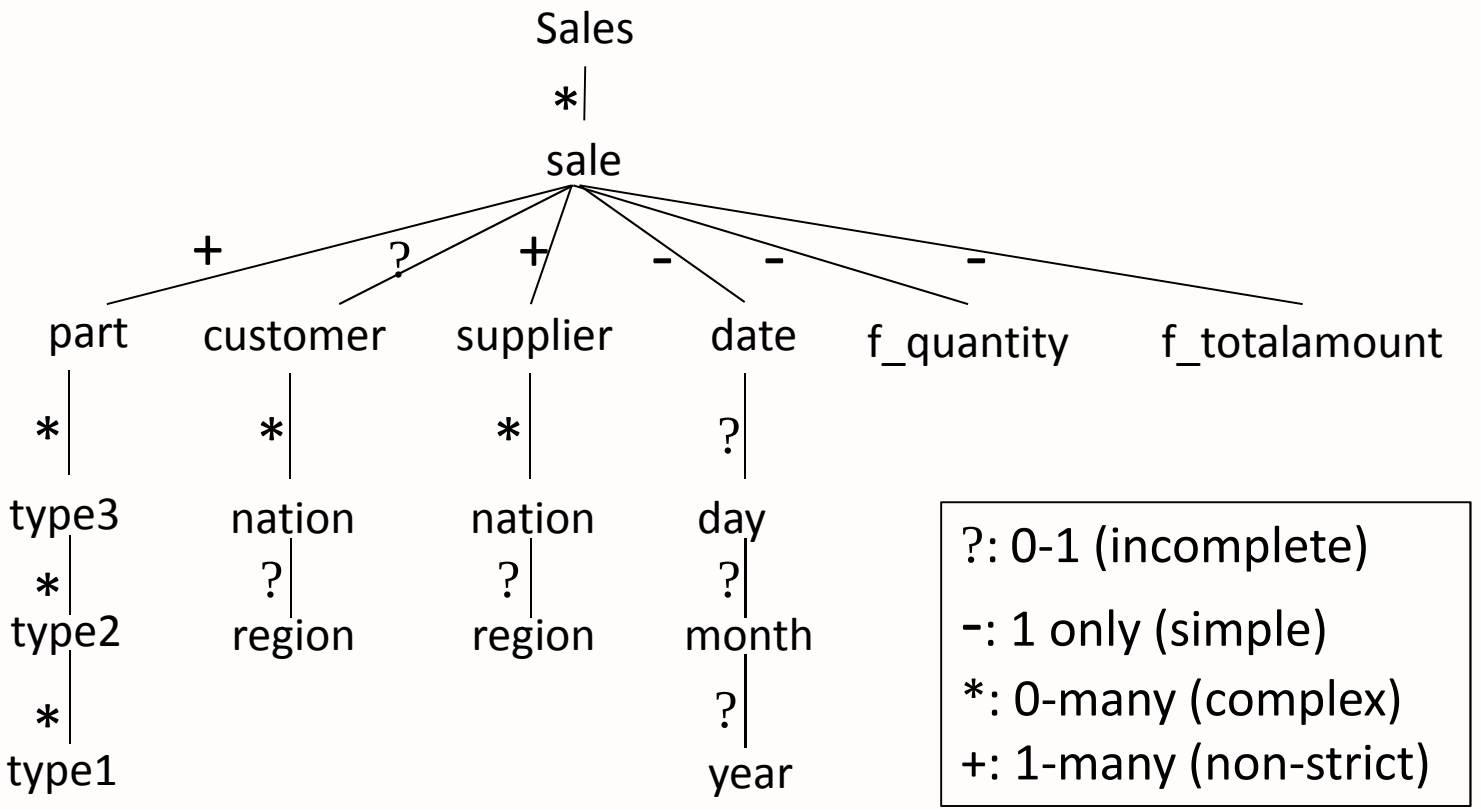
- Relational Decision Support Benchmark
  - TPC: TPC-H and TPC-DS [TPPC'12]
  - SSB [VLDB/TPCTC'09]
  - DWEB [IJBIDM'07]
- XML benchmarks: Michigan [VLDB'02], MemBer [SIGMOD'05], X-Mach, XMark [VLDB/EEXTT'02], XOO7[CIKM'01], and XBench [ICDE'04]
- XML decision support benchmarks: **XWeB** [VLDB/TPCTC'10]
  - Only one complex hierarchy workload
  - Complexity lies only on part-category dimension
  - Query on complex hierarchies is limited
  - Complex hierarchy is not scalable

Objective

Extending XWeB with:

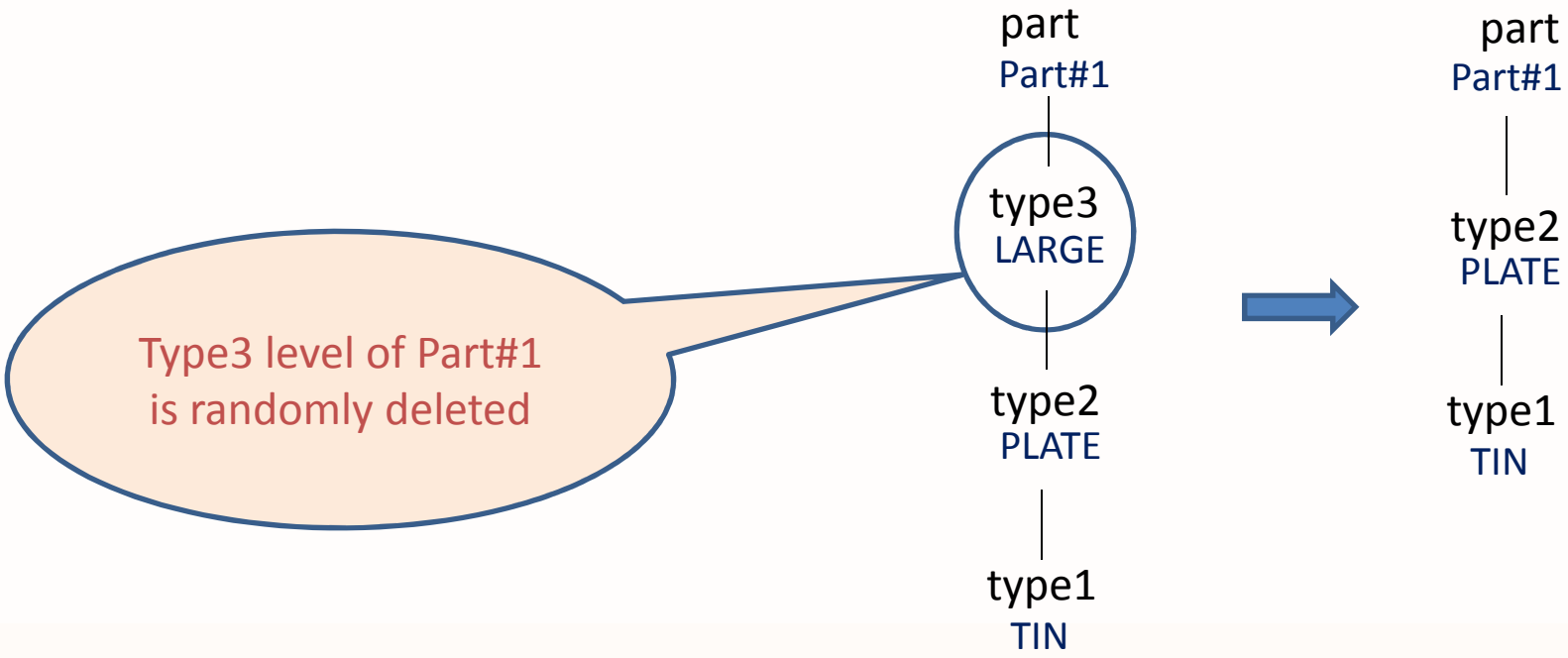
- Scalable complex hierarchies
- Summarizability processing

Data Model


Sales
*
sale
+
?
+
-
-
-
part
customer
supplier
date
f_quantity
f_totalamount
*
*
*
?
type3
nation
nation
day
*
?
?
?
type2
region
region
month
*
?
?
type1
year
?: 0-1 (incomplete)
-: 1 only (simple)
*: 0-many (complex)
+: 1-many (non-strict)

Generating Incomplete Hierarchies

- Randomly delete *ip* hierarchical levels
  - *ip:* incomplete percentage

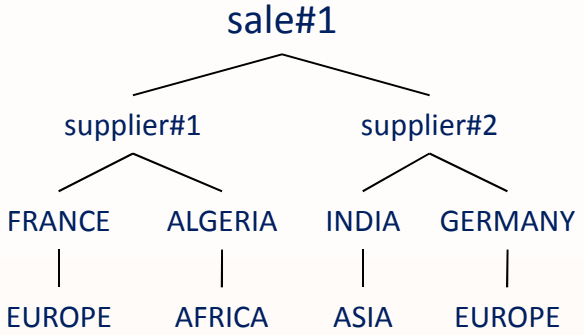


Generating Non-strict Hierarchies

- Randomly generate *np* non-strict hierarchies
  - *np*: non-strict percentage
  1. Randomly generate an array of *n* non-strict hierarchies
     - *n*: number of non-strict hierarchies. Ex. *n* = 4
  2. Convert the array into Hierarchical XML Data

4-non-strict-hierarchy array

Supplier#1	FRANCE	EUROPE
Supplier#2	INDIA	ASIA
Supplier#1	ALGERIA	AFRICA
Supplier#2	GERMANY	EUROPE



Generating Complex Hierarchies

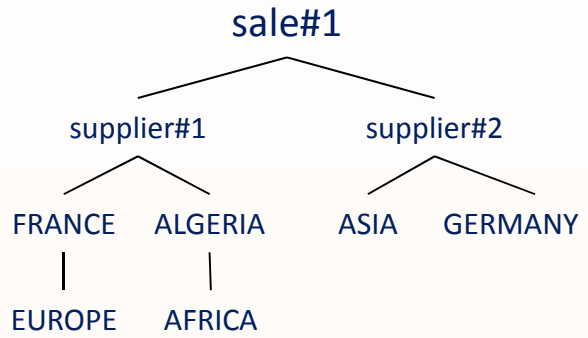
1. Generate *n*-non-strict array (as in slide #12)
2. Randomly delete some levels from non-strict array
3. Convert the array into Hierarchical XML Data

4-non-strict-hierarchy array

Supplier#1	FRANCE	EUROPE
Supplier#2	INDIA	ASIA
Supplier#1	ALGERIA	AFRICA
Supplier#2	GERMANY	EUROPE

complex-hierarchy array

Supplier#1	FRANCE	EUROPE
Supplier#2		ASIA
Supplier#1	ALGERIA	AFRICA
Supplier#2	GERMANY	



Query Workload

Q21	**Q23**
sum of *f_quantity*, *f_totalamount* from *part*, *customer*, *supplier*, *date* group by *part*, *customer*, *supplier*, *date*	max of *f_totalamount* from *date*, *part*, *supplier*, *customer* group by *month*, *type2*, *nation*, *region*
Q22	**Q24**
min of *f_quantity* from *customer*, *part*, *supplier*, *date* group by *nation*, *type3*, *nation*, *day*	average of *f_totalamount* from *supplier*, *part*, *customer*, *date* group by *region*, *type1*, *region*, *year*

Performance Metrics

- Quantitative metric: response time; the execution time of the query workload
- Qualitative metric: verifying the result whether the summarizability issues are correctly handled
  - Resulted groups are not duplicated
  - Total of aggregation values is equal to grand total
  - average value is the division of total and its number
  - Min is the least value
  - Max is the highest value

Experimental Study

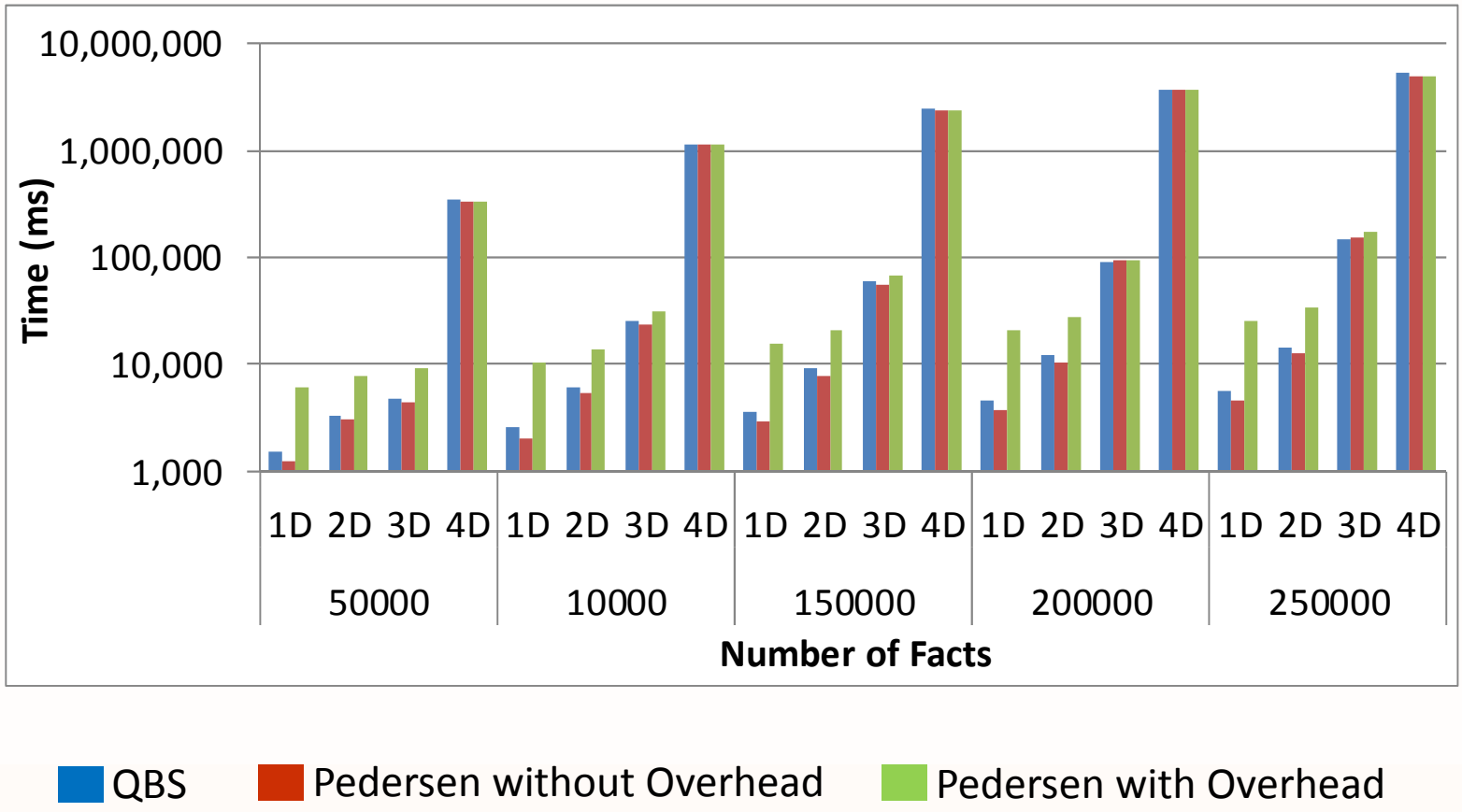
- Summarizability processing using:
  - Our proposed approach: Query Based Approach (QBS) [COMAD'12]
  - Previous approach: Pedersen's approach (Pedersen) [VLDB'99]

Experimental Study (Cont.)

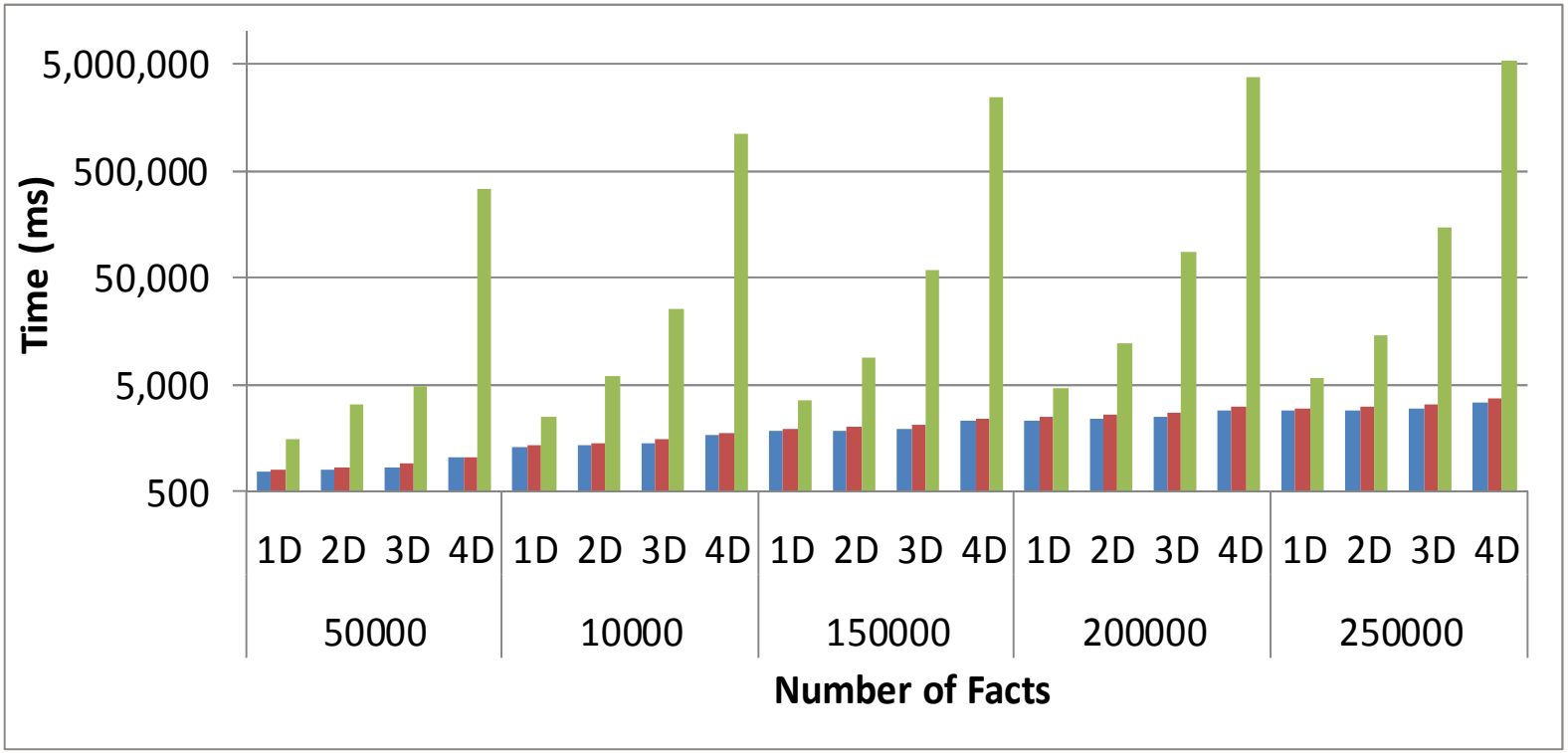
- Dataset size (KB)

No. Facts	50,000	100,000	150,000	200,000	250,000
Simple	27,700	55,390	82,800	110,577	138,015
Incomplete 5%	27,626	55,242	82,543	110,249	137,573
Non-strict 5%	28,669	57,328	85,671	114,422	142,786
Complex 5%	28,376	56,742	85,791	113,252	141,319
Incomplete 50%	25,020	50,030	74,769	99,842	124,601
Non-strict 50%	35,412	70,826	105,914	141,397	176,527
Complex 50%	32,522	65,031	97,263	129,839	162,088

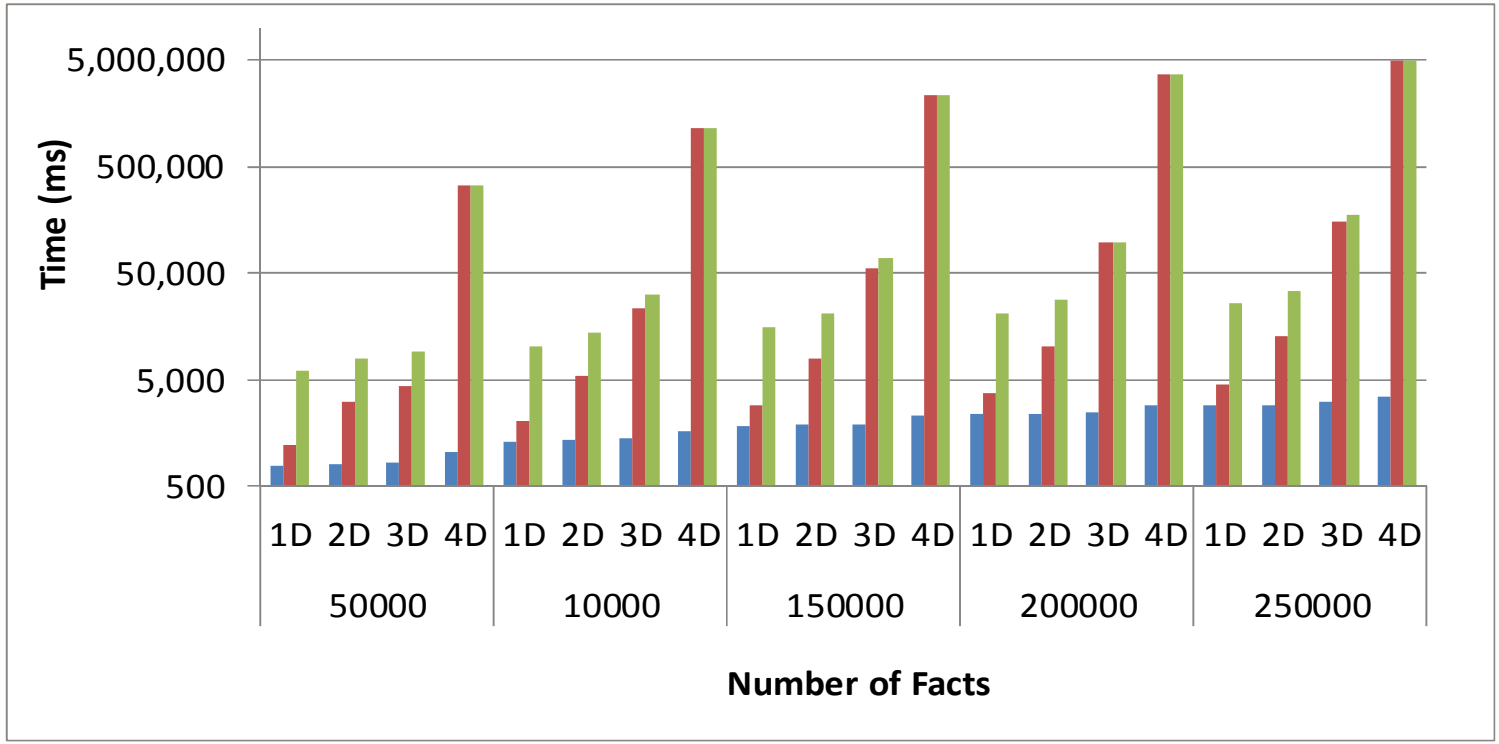
Exp. Results of Simple Hierarchy Grouping



Exp. Results of QBS Simple Hierarchy Group Matching



Exp. Results of Pedersen Simple Hierarchy Group Matching



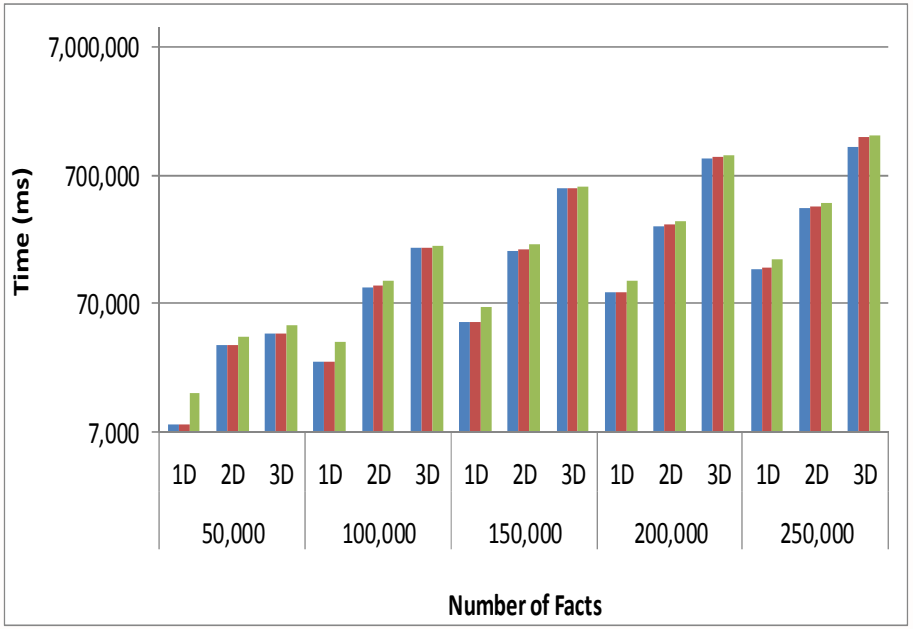
Pedersen without Overhead, without Group Matching

Pedersen without Overhead, with Group Matching

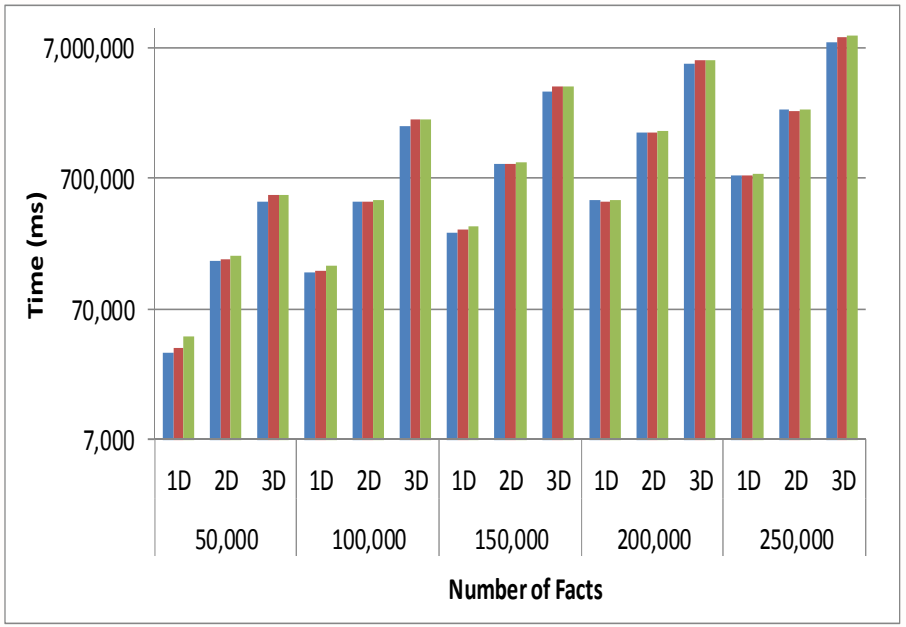
Pedersen with Overhead, with Group Matching

Exp. Results of Complex Hierarchy Grouping

5%

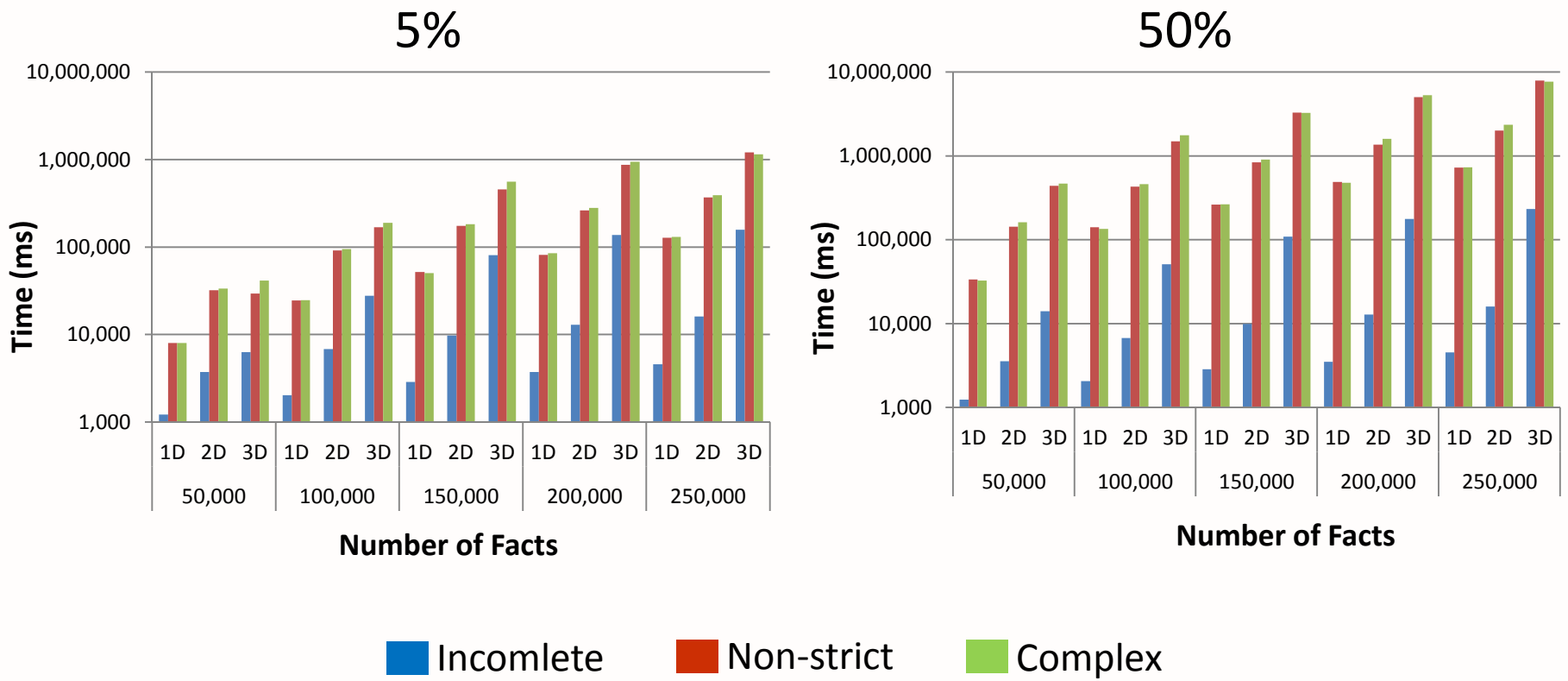


50%



QBS Pedersen without Overhead Pedersen with Overhead

Exp. Results of QBS Complex Hierarchy Grouping



Conclusion

- First XML data warehouse benchmark with complex hierarchies
- Conform to Gray's criteria: relevance, portability, scalability, and simplicity
- Experimentation addressing summariability processing:
  - Run-time summarizability management is feasible
  - Run-time of group matching process is still costly
- **Future work:**
  - Improve group matching process
  - Integrate with previous XML benchmarks: XWeB

QUESTIONS?

chantola.kit@univ-lyon2.fr

marouane.hachicha@univ-lyon2.fr

jerome.darmont@univ-lyon2.fr

Benchmark preliminary version:

http://eric.uni-lyon2.fr/~ckit/DOLAP12.zip