XWeB: the XML Warehouse Benchmark

Hadj Mahboubi and Jérôme Darmont

CEMAGREF Clermont-Ferrand -- Université de Lyon (ERIC Lyon 2)
hadj.mahboubi@cemagref.fr -- jerome.darmont@univ-lyon2.fr

September 17, 2010
New trends for business data warehousing and analysis

OLAP operation over irregular XML data

XML data warehouse

Performance is a crucial and critical issue

Performance assessment using benchmarks

XML database management system

Storage and querying

Extraction, Transformation and Loading (ETL)

XML data warehousing

Analysis

Context

Hadj Mahboubi and Jérôme Darmont

XWeB: the XML Warehouse Benchmark

September 17, 2010

2 / 23
Objective and contribution

- Existing XML benchmarks are not decision-oriented
  - Database schemas do not bear the multidimensional structure
  - Workload do not features typical OLAP-like queries

Objective

- Performance evaluation using a benchmark
  - A test XML data warehouse and its associated XQuery decision support workload

Contribution

- Complete and extend an early version of XWeB
  - Based on TPC-H
  - Complemented with XML irregular structures
  - Extended workload
Objective and contribution

- **Objective**
  - Performance evaluation using a benchmark
    - A test XML data warehouse and its associated XQuery decision support workload

- **Contribution**
  - Complete and extend an early version of XWeB
    - Based on TPC-H
    - Complemented with XML irregular structures
    - Extended workload
Objective and contribution

- Existing XML benchmarks are not decision-oriented
  - Database schemas do not bear the multidimensional structure
  - Workload do not feature typical OLAP-like queries

Objective

- Performance evaluation using a benchmark
  - A test XML data warehouse and its associated XQuery decision support workload

Contribution

- Complete and extend an early version of XWeB
  - Based on TPC-H
  - Complemented with XML irregular structures
  - Extended workload
Outline

1. Introduction

2. Related work

3. Reference XML Warehouse Model

4. XWeB Specifications

5. Sample Experiments

6. Conclusion and perspectives
# Relational Decision Support Benchmarks

| Benchmark Model | Description |
|-----------------|-------------|
| **OLAP Council – APB-1 Benchmark** (OLAP Council, 1998) | Data warehouse schema: four dimensions structured around Sale facts. Simple to understood and to use, but limited. |
| **Transaction Processing Performance Council – TPC standard benchmarks** (TPC, 2008) | TPC-H: classical *product-order-supplier* database model and 22 SQL-92 parameterized queries. TPC-DS: constellation schema, four classes of query templates. |
| **Star Schema Benchmark – SSB** (O’Neil et al., 2009) | A simpler alternative to TPC-DS, query workload with both functional and selectivity features. |
| **Data Warehouse Engineering Benchmark – DWEB** (Darmont et al., 2007) | Helps generate various ad-hoc synthetic data warehouses and typical OLAP query workloads. Conceived for testing the effect of design choices or optimization techniques. Extensive set of parameters. |
XML Benchmarks

XML micro-benchmarks
- Michigan Benchmark (Runapongsara et al., 2006) and MemBer (Afanasiev et al., 2005)
- Assess the individual performances of basic operation: projection, selection, join...
- Specialized and not adapted for decision support application evaluation

XML application benchmarks
- X-Mach1 (Böhme and Rahm, 2003), XMark Schmidt et al., 2003, XOO7 (Bressan et al., 2003) and XBench (Yao et al., 2004)
- Compare and evaluate the global performances of XML-native or compatible DBMSs
Outline

1. Introduction
2. Related work
3. Reference XML Warehouse Model
4. XWeB Specifications
5. Sample Experiments
6. Conclusion and perspectives
Reference XML Warehouse Model

| XML web warehouses | XML documents warehouses | XML data warehouses |
|--------------------|--------------------------|---------------------|
| Xyleme (2001)      | Baril & Bellahsène (2003)| Pokorný (2002)      |
| Golfarelli et al. (2001) | Nassis et al. (2005) | Hümmer et al. (2003) |
| Vrdoljak et al. (2003) | Rajugan et al. (2005) | Rusu et al. (2005)  |
|                     | Zhang et al. (2005)      | Park et al. (2005)  |
|                     |                          | Boussaïd et al. (2006) |

XML data warehouses

- Represent both facts and dimensions
- Converge toward a unified model
- Differ in the way dimensions are handled and in the number of XML documents used to store facts and dimensions

XML data warehouse reference model

- Performance evaluation *(Boukraa et al., 2006)*
- Represents facts in one single XML document and each dimension in one XML document
- Allows representing irregular XML data structures
Reference XML warehouse model

(a) facts_f.xml

(b) dimension_d.xml
Reference XML warehouse model

dw-model.xml
Outline

1 Introduction
2 Related work
3 Reference XML Warehouse Model
4 XWeB Specifications
5 Sample Experiments
6 Conclusion and perspectives
Principle

**Why deriving from TPC-H**

- To acknowledge the importance of TPC benchmarks’ standard status
- To fulfill Gray’s simplicity criterion for a good benchmark
- To benefit from TPC-H’s features, e.g., dbgen

**XWeB components**

- Database and workload models
- XWeB do not include ETL features
- The data Warehouse is a set of XML documents; loading can be timed
Parameterization

Size ($S$): helps control warehouse size

**Depends on**

- **Scale factor ($SF$):** inherited from TPC-H
- **Density ($D$):** helps control the overall size of facts independently from the size of dimensions
  
  $D=1$ $\longrightarrow$ all possible dimension references are present in the fact document

**Estimated as**

$$S = S_{\text{dimensions}} + S_{\text{facts}}$$

- $S_{\text{dimensions}} = \sum_{d \in D} |d|_SF \times \text{nodesize}(d)$, does not change where $SF$ is fixed
- $S_{\text{facts}} = \prod_{d \in D} |h^d_1|_SF \times D \times \text{fact\_size}$, depends on $D$

**Additional parameters (in fact instances)**

- Probability of missing values ($P_m$)
- Probability of element reordering ($P_0$)
### Schema Instantiation

#### Dimension data
1. Obtained from dbgen as flat files (size is tuned by SF)
2. Matched to `dw-model.xml` document $\rightarrow$ dimension$_d$.xml ($d \in D$) documents

#### Part category selection algorithm
- **Names** are taken from TPC-H and organized in three arbitrary hierarchy levels
- **Non-strict hierarchy**: names are interrelated thought rollup and drill-down relationships
- **Non-covering hierarchy**: randomly assign to each part element several categories at any level
Workload Model

Workload queries and parameterization

- Twenty typical aggregation queries for decision support
- Structured in increasing order of query complexity

- Subdivided into five categories: simple reporting queries, 1, 2 and 3-dimension cubes; and complex hierarchy cubes
- Boolean execution parameters: \( RE, 1D, 2D, 3D \) and \( CH \)
## Query workload

| Group                | Query | Specification                                      |
|----------------------|-------|---------------------------------------------------|
| Reporting            | Q01   | Min, Max, Sum, Avg of f_quantity and f_totalamount |
|                      | Q02   | f_quantity for each p_partkey                     |
|                      | Q03   | Sum of f_totalamount                              |
| 1D cube              | Q04   | Sum of f_quantity per p_partkey                   |
|                      | Q05   | Sum of f_quantity and f_total-amount per m_monthname |
|                      | Q06   | Sum of f_quantity and f_total-amount per d_dayname |
|                      | Q07   | Avg of f_quantity and f_total-amount per r_name   |
| 2D cube              | Q08   | Sum of f_quantity and f_total-amount per c_name and p_name |
|                      | Q09   | Sum of f_quantity and f_total-amount per n_name and p_name |
|                      | Q10   | Sum of f_quantity and f_total-amount per r_name and p_name |
|                      | Q11   | Max of f_quantity and f_total-amount per s_name and p_name |
| 3D cube              | Q12   | Sum of f_quantity and f_total-amount per c_name, p_name and y_yearkey |
|                      | Q13   | Sum of f_quantity and f_total-amount per c_name, p_name and y_yearkey |
|                      | Q14   | Sum of f_quantity and f_total-amount per c_name, p_name and y_yearkey |
| Complex hierarchy    | Q15   | Avg of f_quantity and f_total-amount per t_name   |
|                      | Q16   | Avg of f_quantity and f_total-amount per t_name   |
|                      | Q17   | Avg of f_quantity and f_total-amount per p_name   |
|                      | Q18   | Sum of f_quantity and f_total-amount per p_name   |
|                      | Q19   | Sum of f_quantity and f_total-amount per t_name   |
|                      | Q20   | Sum of f_quantity and f_total-amount per t_name   |
Execution protocol

1. **Load test:** load the XML warehouse into an XML DBMS;

2. **Performance test:**
   - *cold run* executed once (to fill in buffers), w.r.t. parameters $RE$, $1D$, $2D$, $3D$ and $CH$;
   - *warm run* executed $NRUN$ times, still w.r.t. workload parameters.

Performance metric: response time

- Load test, cold and warm runs are timed separately
- Global average, minimum and maximum execution times; and standard deviation
- Possibility to derive composite metrics
Experiments

Studied systems

- XML native systems: XQuery decision support query formulation facilities
- Five systems: BaseX, eXist, Sedna, X-Hive and xIndice

Highlight the performance differences among the studied systems

Parameters $p_m = p_0 = 0$

Total size of XML documents

| $SF$ | $D$               | Number of facts | Warehouse size (KB) |
|------|-------------------|-----------------|---------------------|
| 1    | $1/14 \times 10^{-7}$ | 500             | 1710                |
| 1    | $1/7 \times 10^{-7}$  | 1000            | 1865                |
| 1    | $2/7 \times 10^{-7}$  | 2000            | 2139                |
| 1    | $3/7 \times 10^{-7}$  | 3000            | 2340                |
| 1    | $4/7 \times 10^{-7}$  | 4000            | 2686                |
| 1    | $5/7 \times 10^{-7}$  | 5000            | 2942                |
| 1    | $6/7 \times 10^{-7}$  | 6000            | 3178                |
| 1    | $10^{-7}$           | 7000            | 3448                |
Load Test

**Fig.** Load test results

![Graph showing load test results for different XML databases: Sedna, Xindice, XHive, eXist, BaseX. The x-axis represents the number of facts, and the y-axis represents the loading time in milliseconds. The graph shows how the loading time increases with the number of facts for each database.](image-url)
Performance Test

**Fig.** RE performance test results

- X-Hive
- Sedna
- BaseX
- eXist
Performance Test

Fig. 1D performance test results
Performance Test

Fig. CH performance test results
1 Introduction

2 Related work

3 Reference XML Warehouse Model

4 XWeB Specifications

5 Sample Experiments

6 Conclusion and perspectives
Conclusion and perspectives

Conclusion

- XWeB: first XML decision support benchmark
- Gray’s criteria: Relevant, Portable, Scalable, Simple
- Experiments to illustrate XWeB’s relevance
- Also previously used to experimentally validate indexing and view materialization strategies

Perspectives

- Include update operations to improve workload relevance
- Filter factor and experimental feedbacks → Tune and broaden the benchmark scope and representativity
- Performance metrics: composite (as TPC benchmarks’) and qualitative metrics (query result correctness)
Conclusion and perspectives

**Conclusion**

- XWeB: first XML decision support benchmark
- Gray’s criteria: Relevant, Portable, Scalable, Simple
- Experiments to illustrate XWeB’s relevance
- Also previously used to experimentally validate indexing and view materialization strategies

**Perspectives**

- Include update operations to improve workload relevance
- Filter factor and experimental feedbacks → Tune and broaden the benchmark scope and representativity
- Performance metrics: composite (as TPC benchmarks’) and qualitative metrics (query result correctness)