
ORACLE®

ORACLE®

Customer Experiences with Oracle XML DB

Aris Prassinos – *MorphoTrak, SAFRAN Group*
Asha Tarachandani & Thomas Baby – *Oracle XML DB Development*

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions.
The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

MorphoTrak
SAFRAN Group

- US subsidiary of Sagem Sécurité, SAFRAN Group
- Leading innovators in multi-modal Biometric Identification and Verification
  - Fingerprint, palmprint, iris, facial
- Government and Commercial customers
  - Law enforcement, border management, civil identification
  - Secure travel documents, e-passports, drivers' licenses, smart cards
  - Facility / IT access control
- Chosen as Biometric Provider for FBI Next Generation Identification Program
  http://www.sagem-securite.com/eng/site.php?spage=04010847

Printrak BIS

MorphoTrak
SAFRAN Group

- Printrak Biometrics Identification Solution
- Over 100 turnkey production installations worldwide
- Java-based application using Service Oriented Architecture
- Oracle Database 11g
  - Active Data Guard, RAC, XML DB, SecureFiles, ASM

Printrak BIS
Data

- Between data centric and document centric
  - Biometric data
  - Related scanned documents
  - Descriptive data serving as metadata for biometrics and documents

- Each customer requires a custom schema for its descriptive data
  - From basic demographics
    to comprehensive datasets such as NIEM http://www.niem.gov
  - Schema may evolve over time for a particular customer
  - Schema customization should not result in extensive application changes

Printrak BIS
Database Access

- Read intensive, frequent inserts, occasional updates / deletes
- Parts of descriptive data are searchable
  - Searchable fields vary widely among customers
    - May also change over time for a particular customer
- Need to query descriptive data at near-relational speeds
  - Range queries, arithmetic, group by, wildcards
  - Case insensitive, fuzzy searching for selected fields e.g. Name

Printrak BIS
Database Design

MorphoTrak
SAFRAN Group

- Folder / File / Link / Attributes metaphor for organizing content
  - Using a homegrown repository
- Descriptive data stored as XML documents in the repository
  - Documents of different types stored as rows in the same table
  - Schema-less XML used for maximum flexibility
  - Documents may contain collection (repeating) elements
  - Indexing selected XML fields
- Links and Attributes used to extend the model

Printrak BIS
Example system

- 75 million searchable XML documents
- Size of XML documents ranges from 100 bytes to 10 Kbytes
- 10 distinct XML schemas across all XML documents
- Number of tags per XML document ranges from 10 to 100
- 20 searchable tags across all XML documents

Printrak BIS
Indexing Challenges

MorphoTrak
SAFRAN Group

- Indexing technologies not specific to XML Domain
  - Functional indexes
    - Heavy impact on write performance as indexed nodes increase
    - Cannot index collection (repeating) elements
  - Oracle Text Indexes
    - No support for arithmetic queries
    - Poor performance for group by queries
- Indexing technologies specific to XML Domain
  - XMLIndex
- Cannot transparently change from one indexing method to the other without modifying application
  - query syntax different for each method

ORACLE OPEN WORLD

XMLIndex: An XML-Aware Index to Fit Your Use Case

ORACLE

XMLIndex
Overview

- Domain index that speeds up XML operators
- Organizes the information in the XML in relational tables beneath index
- XML operators can be in any part of a query
  - SELECT, FROM, WHERE, ORDER BY, GROUP BY
- XML operators in query rewritten to access the relational tables beneath index
- Two components to fit two different use cases!
  - Unstructured Component (Path-Based)
  - Structured Component

XMLIndex
Path-Based Component

- Available since 11gR1
- Single relational table called path table
- Each row in path table stores
  - Path of node
  - Dewey decimal key
  - Locator to node in base XML
  - Value of node
- Dewey decimal key used for ancestor-descendant checks
- Locator used for fragment extraction
- Ideal when xpath to be queried not known apriori

XMLIndex Path-Based
Path Table Layout

XML Data

```
<Address>
  <city>Fremont</city>
  <state>CA</state>
</Address>….

-------------------------------------------------
<Address>
  <bold> <city>Melbourne</city></bold>
  <font size="21">
    <state><!– state is not in   US --></state>
 </font>
  <country>Australia</country>
</Address>
```

Path Table Layout

RID	PathID	Dewey Key	Locator	Value
101	P1	01	XXX1	
101	P2	0101	XXX2	Fremont
101	P3	0102	XXX3	CA
102	P1	01	XXX4	
102	P4	0101	XXX5	
102	P5	0102	XXX6	Melbourne
…	…	…	…	…

XMLIndex Path-Based
DML Performance

- Allows easy indexing of interesting sub-trees
  - Include or exclude sub-trees
- Allows asynchronous maintenance
  - Manually triggered sync
  - Scheduler-job based sync
  - On-commit sync
- Updates to document result in piece-wise index updates
  - Binary XML with securefile

XMLIndex Path-based
11gR2 Enhancements

- Partitioning by range and by list
- Parallel index creation and parallel query supported
- Physical rewrite for path subsets
- Queries improve by 5x on average
  - XMark (10M)
  - 5 XMark queries improve by 20x
- Asynchronous DML performance improves 2.5x

XMLIndex
Structured Component

- New in 11gR2
- Usecase
  - Typical XML Queries based on structured attributes within XML
  - Example 1:
    Document centric content frequently queried on metadata attributes.
    Publications with Title, Author, Date, ..
  - Example 2:
    Relational Views over XML content

- Example Query:

```
SELECT * FROM DOCUMENT_TAB  doc
WHERE XMLEXISTS(
'$doc//Document [  title = "Indexing XML Techniques" and
                   pubdate > xs:date("2007-03-01") and
                   pubdate < xs:date("2007-12-31") and
                   affiliation = "Oracle" ]'
PASSING VALUE(doc) AS "doc")
```

XMLIndex Structured

Index Structured Metadata in XML Content

- Project out commonly searched structured data
  - All structured leaf data in the same group (having the same parent node) are stored in one row
- Physical rewrite using XQuery/XPath expression matching
  - All xpath matching is avoided at run time
  - All joins to ensure the structured leaf data from the same parent node is avoided
- Secondary Indexes can be created on Structured XMLIndex table
  - Relational indexes on projected scalar attributes
  - Text index on projected text attributes
  - Domain specific index on domain attributes, e.g. image

XMLIndex Structured
Table Layout

Table document_tab

```
<Document>
  <title>Indexing XML Techniques</title>
  <affiliation>Oracle<affiliation>
  <pubdate>2007-04-10</pubdate>
….
</Document>

<Document>
  <title>Object Relational Storage</title>
  <affiliation>Oracle<affiliation>
  <pubdate>2003-03-15</pubdate>
  …
</Document>
```

Structured XMLIndex Table doc_info_tab

Rid	Title	Affiliation	PubDate
ROWID	VARCHAR2	VARCHAR2	DATE
10	Indexing XML Techniques	Oracle	2007-04-10
20	Object Relational Storage	Oracle	2003-03-15

XMLIndex Structured
Index Creation DDL Statement

Table document_tab

```
<Document>
  <title>Indexing XML Techniques</title>
  <affiliation>Oracle<affiliation>
  <pubdate>2007-04-10</pubdate>
….
</Document>

<Document>
  <title>Object Relational Storage</title>
  <affiliation>Oracle<affiliation>
  <pubdate>2003-03-15</pubdate>
  …
</Document>
```

```
SELECT  doc.Title,  doc.Affiliation, doc.PubDate
FROM     document_tab,
         XMLTable('/Document'
         PASSING OBJECT_VALUE
         COLUMNS
      Title        VARCHAR2(30) PATH 'title',
      Affiliation VARCHAR2(30) PATH 'affiliation',
      PubDate    DATE              PATH 'pubdate') doc;
```

```
CREATE INDEX doc_xmlindex ON document_tab
(OBJECT_VALUE)
INDEXTYPE IS XDB.XMLIndex
PARAMETERS (
'XMLTable doc_info_tab "/Document"
    COLUMNS
      Title        VARCHAR2(30) PATH "title",
      Affiliation VARCHAR2(30) PATH "affiliation",
      PubDate    DATE              PATH "pubdate"');
```

XMLIndex Structured
Rewritten Query Is Purely Relational

- Query After Rewrite:

```
SELECT * FROM document_tab  doc
WHERE EXISTS(
        SELECT NULL FROM doc_info_tab
        WHERE doc.rowid=doc_info_tab.rid AND
        Title = 'Indexing XML Techniques' AND
        PubDate > TO_DATE('2007-03-01') AND
        PubDate < TO_DATE('2007-12-31') AND
        Affiliation = 'Oracle')
```

- Original Query:

```
SELECT * FROM document_tab doc
WHERE XMLEXISTS(
'$doc//Document [title = "Indexing XML Techniques" and
pubdate > xs:date("2007-03-01") and
pubdate < xs:date("2007-12-31") and
affiliation = "Oracle"]' PASSING VALUE(doc) AS "doc")
```

XML DB Use Cases: Criteria to Consider when Storing & Indexing XML

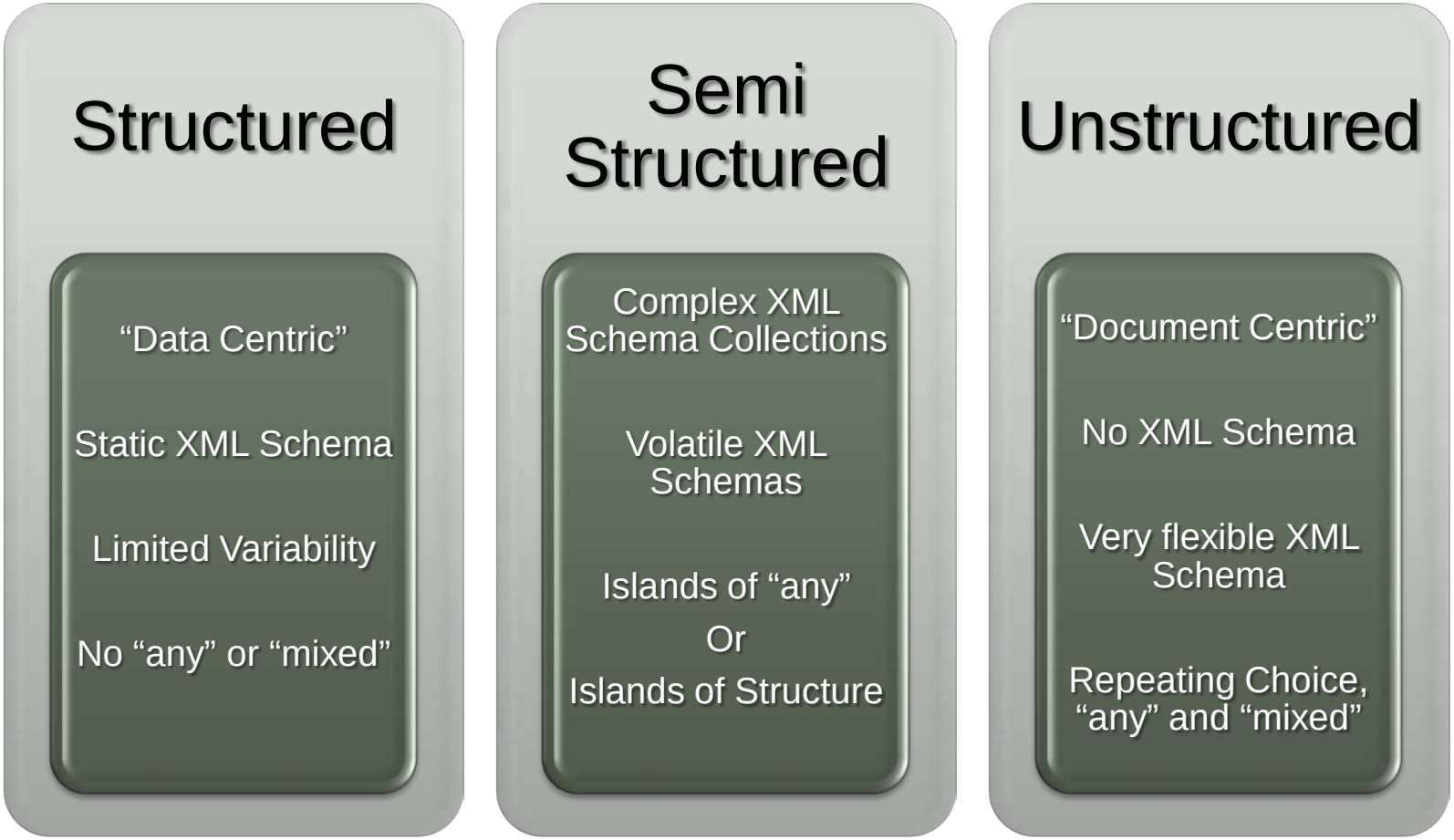
ORACLE

XMLDB Use Cases
Design Considerations

- XML Data characteristics
- Query criteria
- Application requirements

XMLDB Use Cases

XMLData Characteristics



XMLDB Use Cases
Query Criteria

- Ad-hoc queries or known XPaths
- Known structured or unstructured islands of data
- XPaths with wildcards or not
- Schema aware searches e.g. default values, data types
- Search across a mixed set of XML documents or one set at a time

XMLDB Use Cases
Application Requirements

- SQL-based or XML-based application
  - Relational or XQuery
- XML Usage
  - XML Content - specifications, reports (e.g. Office 2007 documents)
  - Content Management using XML
  - Both
- Other requirements – document ingestion rate, storage space, manageability

Printrak BIS & XML DB

MorphoTrak
SAFRAN Group

- Path-subsetted index helped achieve a high ingestion rate and optimal space usage
- High textual, numeric and aggregate query performance with the use of XMLIndex Structured component
- New elements can be easily added to the index online
- Additional advanced text query capabilities by creating Oracle Text indexes on top of XMLIndex

For more information:

http://www.oracle.com/technology/tech/xml/xmldb/

ORACLE IS THE INFORMATION COMPANY