Ghislain Fourny

Big Data

13. Data Warehousing
The road to analytics
Another history of data management (T. Hofmann)

1970s – 2000s  Age of Transactions

1995 -  Age of Business Intelligence

2000s -  Age of Big Data
Paradigms

OLTP vs. OLAP
OnLine Transaction Processing

Consistent and Reliable Record-Keeping
OnLine Transaction Processing

Transactions and results on small portions of data
OnLine Transaction Processing

Lots of transactions on small portions of data
OnLine Transaction Processing

Normalized Data
OnLine Analytical Processing

Data-based Decision Support
OLAP is Big

Possibly many joins

Large portions of the data

Few long heavy queries
OLAP Examples

Web analytics

Sales analytics

Management support

Statistical analysis (census)

Scientific databases (e.g., bio-informatics)
OLTP vs. OLAP

Detailed Individual Records

Historical Summarized Consolidated Data

OLTP vs. OLAP

OLTP

OLAP
OLTP vs. OLAP

Lots of writes

OLTP

VS.

Lots of reads

OLAP
OLTP vs. OLAP

Small sets of records vs. Analysis over big chunks
OLTP vs. OLAP

fully interactive (< 1s)

VS.

Slow interactive

OLTP

OLAP
OLTP vs. OLAP

Consistency

Redundancy

Redundancy

Redundancy

Redundancy
OLAP
A data warehouse ... is a subject-oriented integrated time-variant nonvolatile collection of data in support of management's decision-making process.
Subject-oriented

customers  sales

products  events
Integrated
Time-variant

Time in data warehouses is paramount

(not so in OLTP systems)
Time-variant

Often past 5-10 years
Non-volatile


no updates
Architecture

- ERP
- CRM
- OLTP
- Files

ETL

- Analyze
- Report
- Mine
OLAP: Redundancy

Materialized views (denormalized)
1st Normal Form (tabular) – The Key

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- [ ]
- [x]
### 2nd Normal Form (not joined) – The Whole Key

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The Whole Key is correctly identified in the first table.
- The Whole Key is not correctly identified in the second table.
3rd Normal Form – Nothing But The Key
Why materialize?

Operational data sources are too heterogeneous
OLAP: Special-purpose indices
OLAP: Derived data
Querying OLAP

Slow interactive

Category 1 Category 2 Category 3 Category 4

Series 1  Series 2  Series 3

1 - 10s

Continuous monitoring/tracking

hours
## Summary of differences

<table>
<thead>
<tr>
<th></th>
<th>OLTP</th>
<th>OLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Source</strong></td>
<td>Original (operational)</td>
<td>Derived (consolidated)</td>
</tr>
<tr>
<td><strong>Purpose</strong></td>
<td>Business tasks</td>
<td>Decision support</td>
</tr>
<tr>
<td><strong>Interface</strong></td>
<td>Snapshot</td>
<td>Multidimensional views</td>
</tr>
<tr>
<td><strong>Writing</strong></td>
<td>short and fast, by end user</td>
<td>period refreshes, by batch jobs</td>
</tr>
<tr>
<td><strong>Queries</strong></td>
<td>Simple, small results</td>
<td>Complex and aggregating</td>
</tr>
<tr>
<td><strong>Design</strong></td>
<td>Many normalized tables</td>
<td>Few denormalized cubes</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>ACID</td>
<td>Sampling, confidence intervals</td>
</tr>
<tr>
<td><strong>Freshness</strong></td>
<td>Serializability</td>
<td>Reproducibility</td>
</tr>
<tr>
<td><strong>Speed</strong></td>
<td>Very fast</td>
<td>Often slow</td>
</tr>
<tr>
<td><strong>Optimization</strong></td>
<td>Inter-query</td>
<td>Intra-query</td>
</tr>
<tr>
<td><strong>Space</strong></td>
<td>Small, archiving old data</td>
<td>Large, less space efficient</td>
</tr>
<tr>
<td><strong>Backup</strong></td>
<td>Very important</td>
<td>Re-ETL</td>
</tr>
</tbody>
</table>
Data Model
Data Cubes

Data is stored in multidimensional hypercubes
Data Cubes

Year
Data Cubes
Data Cubes
Fact
Dimensions

Where

Who?

Dimensions

Which currency?

What?

When?

etc.
## Fact table

<table>
<thead>
<tr>
<th>Who</th>
<th>Year</th>
<th>Name</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>2016</td>
<td>Peter</td>
<td>1,000$</td>
</tr>
<tr>
<td>Germany</td>
<td>2015</td>
<td>Mary</td>
<td>15,000$</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2016</td>
<td>Mary</td>
<td>1,500$</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2015</td>
<td>Peter</td>
<td>3,000$</td>
</tr>
<tr>
<td>Australia</td>
<td>2015</td>
<td>Peter</td>
<td>6,000$</td>
</tr>
<tr>
<td>China</td>
<td>2015</td>
<td>Mary</td>
<td>1,000$</td>
</tr>
</tbody>
</table>
## Aggregation

<table>
<thead>
<tr>
<th>Where</th>
<th>Year</th>
<th>Who</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>2016</td>
<td>Peter</td>
<td>1,000$</td>
</tr>
<tr>
<td>Germany</td>
<td>2015</td>
<td>Mary</td>
<td>15,000$</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2016</td>
<td>Mary</td>
<td>1,500$</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2015</td>
<td>Peter</td>
<td>3,000$</td>
</tr>
<tr>
<td>Australia</td>
<td>2015</td>
<td>Peter</td>
<td>6,000$</td>
</tr>
<tr>
<td>China</td>
<td>2015</td>
<td>Mary</td>
<td>1,000$</td>
</tr>
</tbody>
</table>
Aggregation
## Aggregation

<table>
<thead>
<tr>
<th>Who</th>
<th>Year</th>
<th>Name</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Germany</td>
<td>2016</td>
<td>Peter</td>
<td>1,000$</td>
</tr>
<tr>
<td>Germany</td>
<td>2015</td>
<td>Mary</td>
<td>15,000$</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2016</td>
<td>Mary</td>
<td>1,500$</td>
</tr>
<tr>
<td>Switzerland</td>
<td>2015</td>
<td>Peter</td>
<td>3,000$</td>
</tr>
<tr>
<td>Australia</td>
<td>2015</td>
<td>Peter</td>
<td>6,000$</td>
</tr>
<tr>
<td>China</td>
<td>2015</td>
<td>Mary</td>
<td>1,000$</td>
</tr>
</tbody>
</table>
## Aggregation

<table>
<thead>
<tr>
<th>Year</th>
<th>Name</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>2016</td>
<td>Peter</td>
<td>$1,000</td>
</tr>
<tr>
<td>2015</td>
<td>Mary</td>
<td>$16,000</td>
</tr>
<tr>
<td>2016</td>
<td>Mary</td>
<td>$1,500</td>
</tr>
<tr>
<td>2015</td>
<td>Peter</td>
<td>$9,000</td>
</tr>
</tbody>
</table>
Slicing
Slicers and Dicers
Slicers and Dicers

Usually between 1 and 3 dicers,

often 2

Slicers

Dicers
### Slicers and Dicers

#### Slicers

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Servers</td>
<td></td>
</tr>
<tr>
<td>World</td>
<td></td>
</tr>
<tr>
<td>USD</td>
<td></td>
</tr>
</tbody>
</table>
## Slicers and Dicers

### Slicers

<table>
<thead>
<tr>
<th></th>
<th>2014</th>
<th>2015</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>$1,000,000$</td>
<td>$1,500,000$</td>
<td>$1,400,000$</td>
</tr>
<tr>
<td>Mary</td>
<td>$2,000,000$</td>
<td>$2,300,000$</td>
<td>$2,200,000$</td>
</tr>
</tbody>
</table>
Products: the big three

Oracle 12c

Essbase

IBM DB2

Analysis Services

Cognos

SQL Server

SAP HANA
ETLing
OLAP: Derived data
OLAP: Derived data
ETL

Extract
Transform
Load
## Extract

**Triggers**

<table>
<thead>
<tr>
<th>Incremental updates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Gateways**

**Log extraction**

![Hadoop Logo](https://www.hadoop.org/assets/images/hadoop-logo-md.png)
Transform

Derivation

Value transformation

Filter, split, merge, join
Load

Integrity constraints

Build indices

Sorting

Partition
Considerations

Granularity

When?

Infrastructure
Implementation
Two flavors of OLAP

ROLAP  MOLAP
## Fact table (ROLAP)

<table>
<thead>
<tr>
<th>Dim1</th>
<th>Dim2</th>
<th>Dim3</th>
<th>Dim4</th>
<th>Dim5</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Star Schema

<table>
<thead>
<tr>
<th></th>
<th>Dim1</th>
<th>Dim2</th>
<th>Dim3</th>
<th>Dim4</th>
<th>Dim5</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>63</td>
</tr>
</tbody>
</table>
Snow-flake schema
Querying
Querying cubes

Tables: SQL

Cubes: MDX
MDX stands for...

Multi-Dimensional eXpressions
Measures

Amount of licenses

Revenues

Taxes paid

...
Dimensions

Quarter
Salesperson
Product
Country
In short...

A cube is a list of dimensions indexing a list of measures
Hierarchies

Dimension values are organized in hierarchies.

i.e., slice and aggregate by geographic region, etc

i.e., slice and aggregate by economic partnership, etc
Members

Members correspond to levels in a hierarchy.

[Geo]

[Europe]
  [Switzerland]
    [ZH]
    [BE]
  [Germany]
...

[Asia]
  [China]
    [India]
    ...

[Africa]
  [Canada]
    [USA]
    [Brazil]
    ...

[Oceania]
Identifying a member

[Location].[Geo].[Europe].[Switzerland].[ZH].[Zurich]
Tuples

A list of members

([Location].[Geo].[Europe].[Switzerland].[ZH].[Zurich], [Salesmen].[People].[John], [Time].[Year].[2016].[Q4])

Associated with a dimensionality (list of hierarchies)

([Location].[Geo], [Salesmen].[People] [Time].[Year])
Sets

A set of tuples with same dimensionality

\{
  ([Location].[Geo].[Europe].[Switzerland].[ZH].[Zurich],
   [Salesmen].[People].[John],
   [Time].[Year].[2016].[Q4]),

  ([Location].[Geo].[Europe].[Switzerland].[BE].[Bärn],
   [Salesmen].[People].[Mary],
   [Time].[Year].[2016].[Q4]),

  ([Location].[Geo].[Europe].[Germany].[Berlin],
   [Salesmen].[People].[John],
   [Time].[Year].[2016].[Q3])
\}
MDX statements: dicing

SELECT
  [Measures].Members ON COLUMNS,
  [Location].[Geo].Members ON ROWS
FROM [Sales]
MDX statements: slicing

SELECT
  [Measures].Members ON COLUMNS,
  [Location].[Geo].Members ON ROWS
FROM [Sales]
WHERE [Products].[Line].[Laptops].[MBP]
Syntax
XBRL Architecture
Technologies

XML
XML Names
XML Schema
XML Link
### Fact

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>What?</td>
<td>Assets</td>
</tr>
<tr>
<td>Who?</td>
<td>Coca Cola</td>
</tr>
<tr>
<td>When?</td>
<td>Dec 31, 2011</td>
</tr>
<tr>
<td>Of what?</td>
<td>USD</td>
</tr>
</tbody>
</table>

```xml
<us-gaap:Assets
  contextRef="FI2012Q4"
  decimals="-6"
  id="Fact-600212FD4D06E63B4F8F6874C6E5BE74"
  unitRef="usd">
  86174000000
</us-gaap:Assets>
```
Context

<xbrli:context id="FI2011Q4">
  <xbrli:entity>
    <xbrli:identifier scheme="http://www.sec.gov/CIK">
      0000021344
    </xbrli:identifier>
  </xbrli:entity>
  <xbrli:period>
    <xbrli:instant>2011-12-31</xbrli:instant>
  </xbrli:period>
</xbrli:context>
Unit

<xbrli:unit id="usd">
  <xbrli:measure>iso4217:USD</xbrli:measure>
</xbrli:unit>
Concept (XML Schema)

```xml
<xs:element id='us-gaap_Assets'
    name='Assets'
    nillable='true'
    substitutionGroup='xbrli:item'
    type='xbrli:monetaryItemType'
    xbrli:balance='debit'
    xbrli:periodType='instant' />
```
Graphs
DAGs
Trees
Node: locator

xlink:type="locator" />
NET CHANGE IN OPERATING ASSETS AND LIABILITIES DISCLOSURE [Abstract]
Edge

<presentationArc
    order="10"
    preferredLabel="http://www.xbrl.org/2003/role/totalLabel"
    xlink:from="loc_us-gaap_AssetsAbstract_2F55ECB2BF7C1A62009CDA6BBC757094"
    xlink:to="loc_us-gaap_Assets_102D7A4D204ED45AC0DEDA6BBC78F386"
    xlink:type="arc" />


Summary