Toward Variability Management to Tailor High Dimensional Index Implementations

An Analytical Model for Data Persistence in Business Data Warehouses

Veit Köppen, Thorsten Winsemann, Gunter Saake
Otto-von-Guericke University Magdeburg
May 14th, 2015

Presented by: Azeem Lodhi
Agenda

- Motivation
- Business Data Warehouse
- Decision Model
- Evaluation
- Conclusion
Motivation

- In-Memory Databases
  - Data Warehouses
  - Big Data Architectures
  - Data Lakes
- Where do we should store data?
  - Availability
  - Data Quality
  - Intermediate results
- Databases as primary storage
A BDW is a Data Warehouse to support decisions concerning the business on all organizational levels and it is, for instance, the basis for BI, planning, and CRM.
BDW: Layered Architecture

5 Layers

data value increase at every level

data can be used from every level

layer is not data storage

• Data Acquisition Layer
  „inbox“ of a BDW

• Quality & Harmonization Layer
  data integration

• Data Propagation Layer
  Single version of Truth without business Logic

• Business Transformation Layer
  transformation to business needs

• Reporting & Analysis Layer
  enhancing report performance
Reasons for Data Persistency

**Data acquisition**: decoupling, availability, data lineage, …

**Data modification**: changing transformation rules, addicted transformations, complexity

**Data management**: constant data basis, en-bloc data supply, authorization, single version of truth, corporate data memory

**Data availability**: information warranty, performance

**Laws and provisions**: corporate governance, laws and provisions

Winsemann & Köppen (2012)
Decision Model

(1) - (3): mandatorily stored data
(2) - (4): „easily“ answered
(5) - (7): fuzzy terms: complex, frequently

Focus on how to objectify
We use indicators & MCDA
Decision Model: Indicator system

Legend:
- Query (Q)
- Transformation ($\tau$)
- OLAP (O)
- Update (U)
- Modelling (Mo)
- Maintenance (Ma)
- Quality Assurance (QA)

Measured in time units from the BDW
Decision Model: Indicator system

\[
\text{Data Supply} = \sum_{i=1}^{Calls} F_{S_i} \cdot T_{S_i} = \sum_{i=1}^{Calls} F_{S_i} \cdot \left( T_{Q(S_i)} + T_{\tau(S_i)} + T_{O(S_i)} \right)
\]

\[
\text{Data Actualization} = \sum_{j=1}^{Loads} F_{A_j} \cdot T_{A_j} = \sum_{j=1}^{Loads} F_{A_j} \cdot \left( T_{Q(A_j)} + T_{\tau(A_j)} + T_{U(A_j)} \right)
\]

\[
\text{Data Reorganization} = \sum_{k=1}^{Runs} F_{R_k} \cdot T_{R_k}
\]

\[
\text{Cost} = C_P = C_{M_0} + C_{M_a} + C_{Q_A}
\]
Including User preferences

Multi-Criteria Decision Analysis (MCDA)

User preferences by pairwise comparison of criteria: $C_1$ to $C_2$

1 = equal
3 = moderate
5 = strong
7 = very strong or demonstrated
9 = extreme

Note, $C_2$ to $C_1$ is reciprocal

Result: weighted factors (WF) per criterion

\[
\begin{array}{|c|c|c|c|c|c|c|}
\hline
& DS & DA & DR & Cost & \sum c & WF_c \\
\hline
DS & 1 & 6 & 7 & 5 & 19.000 & 0.44225 \\
DA & 1/6 & 1 & 7 & 1/6 & 8.333 & 0.19397 \\
DR & 1/7 & 1/7 & 1 & 1/7 & 1.429 & 0.03325 \\
Cost & 1/5 & 6 & 7 & 1 & 14.200 & 0.33053 \\
\hline
\end{array}
\]

$\sum \sum c = 42.962$
Evaluation: Scenario

SAP NetWeaver Business Warehouse, Release 7.40, with SAP HANA (HDB, Release 1.0)
Evaluation: Data elements

Order Data Store (DSO1)
Invoice Data Store (DSO2)
Infocube 1 (IC1)
Infocube 2 (IC2)

M: Measurement (detailed)
F: Fact (aggregated)
Evaluation: Assumptions

Indicators are directly measured within the system

• Data Supply (6 different reports):
  
  20- 40 calls per day, 5 days

• Data Actualization (ETL processes):
  
  every 2 hours

• Data Reorganization (deletion from change logs + compression):
  
  Once a week

• Reports from Multiprovider → three alternatives
  
  DSOs, IC₁, IC₂
# Evaluation: Data Supply

<table>
<thead>
<tr>
<th>Report</th>
<th>Average response time (s)</th>
<th>Frequency</th>
<th>Overall time (s)</th>
<th>Time Data Manager (s)</th>
<th>Time OLAP (s)</th>
<th>Time Frontend (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Supply DSOs</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZTW_Q01A</td>
<td>4.780</td>
<td>100</td>
<td>478.077</td>
<td>26.504</td>
<td>447.678</td>
<td>3.893</td>
</tr>
<tr>
<td>ZTW_Q02A</td>
<td>2.887</td>
<td>100</td>
<td>288.668</td>
<td>17.088</td>
<td>268.557</td>
<td>3.024</td>
</tr>
<tr>
<td>ZTW_Q03A</td>
<td>2.812</td>
<td>100</td>
<td>281.231</td>
<td>19.236</td>
<td>258.701</td>
<td>3.293</td>
</tr>
<tr>
<td>ZTW_Q04A</td>
<td>4.020</td>
<td>100</td>
<td>402.016</td>
<td>19.612</td>
<td>378.97</td>
<td>3.436</td>
</tr>
<tr>
<td>ZTW_Q05A</td>
<td>1.794</td>
<td>200</td>
<td>358.738</td>
<td>18.439</td>
<td>333.891</td>
<td>6.418</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>2.915</td>
<td>800</td>
<td>2,331.58</td>
<td>119.036</td>
<td>2,186.022</td>
<td>26.528</td>
</tr>
<tr>
<td><strong>Data Supply IC1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZTW_Q01B</td>
<td>4.289</td>
<td>100</td>
<td>428.905</td>
<td>3.222</td>
<td>421.953</td>
<td>3.716</td>
</tr>
<tr>
<td>ZTW_Q02B</td>
<td>2.638</td>
<td>100</td>
<td>263.829</td>
<td>3.089</td>
<td>257.778</td>
<td>2.959</td>
</tr>
<tr>
<td>ZTW_Q03B</td>
<td>2.491</td>
<td>100</td>
<td>249.084</td>
<td>3.117</td>
<td>242.799</td>
<td>3.167</td>
</tr>
<tr>
<td>ZTW_Q04B</td>
<td>3.697</td>
<td>100</td>
<td>369.728</td>
<td>3.196</td>
<td>363.104</td>
<td>3.430</td>
</tr>
<tr>
<td>ZTW_Q05B</td>
<td>1.627</td>
<td>200</td>
<td>325.352</td>
<td>3.560</td>
<td>315.897</td>
<td>5.896</td>
</tr>
<tr>
<td>ZTW_Q06B</td>
<td>2.431</td>
<td>200</td>
<td>486.171</td>
<td>2.995</td>
<td>477.012</td>
<td>6.153</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>2.654</td>
<td>800</td>
<td>2,123.069</td>
<td>19.179</td>
<td>2,078.543</td>
<td>25.321</td>
</tr>
<tr>
<td><strong>Data Supply IC2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZTW_Q01C</td>
<td>4.248</td>
<td>100</td>
<td>424.771</td>
<td>3.147</td>
<td>417.939</td>
<td>3.681</td>
</tr>
<tr>
<td>ZTW_Q02C</td>
<td>2.636</td>
<td>100</td>
<td>263.619</td>
<td>2.994</td>
<td>257.709</td>
<td>2.919</td>
</tr>
<tr>
<td>ZTW_Q03C</td>
<td>2.465</td>
<td>100</td>
<td>246.497</td>
<td>3.076</td>
<td>240.245</td>
<td>3.172</td>
</tr>
<tr>
<td>ZTW_Q04C</td>
<td>3.721</td>
<td>100</td>
<td>372.125</td>
<td>3.189</td>
<td>365.511</td>
<td>3.428</td>
</tr>
<tr>
<td>ZTW_Q05C</td>
<td>1.620</td>
<td>200</td>
<td>324.063</td>
<td>4.131</td>
<td>314.060</td>
<td>5.867</td>
</tr>
<tr>
<td>ZTW_Q06C</td>
<td>2.437</td>
<td>200</td>
<td>487.301</td>
<td>2.878</td>
<td>478.217</td>
<td>6.197</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>2.648</td>
<td>800</td>
<td>2,118.376</td>
<td>19.415</td>
<td>2,073.681</td>
<td>25.264</td>
</tr>
</tbody>
</table>
### Actualization & Reorganization

<table>
<thead>
<tr>
<th>Data actualization</th>
<th>Process type</th>
<th>Average time (s)</th>
<th>Frequency</th>
<th>Overall time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSOs</td>
<td>Data transfer</td>
<td>7.972</td>
<td>120</td>
<td>956.633</td>
</tr>
<tr>
<td></td>
<td>Data activation</td>
<td>2.701</td>
<td>120</td>
<td>324.116</td>
</tr>
<tr>
<td></td>
<td>Overall</td>
<td>5.336</td>
<td>240</td>
<td>1280.749</td>
</tr>
<tr>
<td>IC1</td>
<td>Data transfer</td>
<td>7.981</td>
<td>120</td>
<td>957.727</td>
</tr>
<tr>
<td>IC2</td>
<td>Data transfer</td>
<td>7.199</td>
<td>60</td>
<td>431.916</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data reorganization</th>
<th>Process type</th>
<th>Average time (s)</th>
<th>Frequency</th>
<th>Overall time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSOs</td>
<td>CL deletion</td>
<td>22.082</td>
<td>1</td>
<td>22.082</td>
</tr>
<tr>
<td>IC1</td>
<td>Compression</td>
<td>26.717</td>
<td>1</td>
<td>26.717</td>
</tr>
<tr>
<td>IC2</td>
<td>Compression</td>
<td>24.138</td>
<td>1</td>
<td>24.138</td>
</tr>
</tbody>
</table>
Cost

Modelling: Two times a year
Including a new object at every element → 63 min
DSOs: 28 min, IC₁: 49 min, IC₂: 63 min

Maintenance: 10 min per week for all

Quality Assurance: after model change: 15 min
DSOs: 15 min, IC₁: 30 min, IC₂: 45 min
## MCDA Result

<table>
<thead>
<tr>
<th>Class</th>
<th>Alternative</th>
<th>IC2</th>
<th>IC1</th>
<th>DSOs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Supply</td>
<td>Overall in s</td>
<td>2118,380</td>
<td>2123,070</td>
<td>2331,580</td>
</tr>
<tr>
<td></td>
<td>Weighted utility</td>
<td>0.344</td>
<td>0.343</td>
<td>0.313</td>
</tr>
<tr>
<td>Data actualization</td>
<td>Overall in s</td>
<td>2670,420</td>
<td>2238,480</td>
<td>1280,760</td>
</tr>
<tr>
<td></td>
<td>Weighted utility</td>
<td>0.234</td>
<td>0.279</td>
<td>0.487</td>
</tr>
<tr>
<td>Data reorganization</td>
<td>Overall in s</td>
<td>72,937</td>
<td>48,799</td>
<td>22,082</td>
</tr>
<tr>
<td></td>
<td>Weighted utility</td>
<td>0.172</td>
<td>0.258</td>
<td>0.570</td>
</tr>
<tr>
<td>Cost</td>
<td>Overall in min per week</td>
<td>14,731</td>
<td>13,615</td>
<td>12,231</td>
</tr>
<tr>
<td></td>
<td>Weighted utility</td>
<td>0.304</td>
<td>0.329</td>
<td>0.366</td>
</tr>
</tbody>
</table>
Evaluation: User preferences

Fast data supply & low cost

<table>
<thead>
<tr>
<th></th>
<th>Data supply</th>
<th>Data actualization</th>
<th>Data reorganization</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data supply</td>
<td>1</td>
<td>7</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Data actualization</td>
<td>1 / 7</td>
<td>1</td>
<td>1</td>
<td>1 / 5</td>
</tr>
<tr>
<td>Data reorganization</td>
<td>1 / 9</td>
<td>1</td>
<td>1</td>
<td>1 / 7</td>
</tr>
<tr>
<td>Cost</td>
<td>1 / 5</td>
<td>5</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

\[w_{Data\ Supply} = 0.65, w_{Data\ Actualization} = 0.06\]

\[w_{Data\ Reorganization} = 0.05, w_{Cost} = 0.24.\]

DSOs \(\gg IC_1 \gg IC_2\)
Conclusion

- In-memory data management does not squeeze persistency out
- Decision model for data persistency
- Including user preferences with MCDA
- Can be easily adopted

Data Persistency decisions are necessary in Business Data Warehouses.
The End (?)

Thank you for your attention!

veit.koeppen@ovgu.de

thorsten.winseemann@t-online.de
References