Dwarf: A High Performance OLAP Engine

Nick Roussopoulos
ACT Inc.
Features

- Complete **OLAP** engine
  - Computes, indexes, and stores highly compressed data cubes
  - Queries, Incremental Updates

- Overcomes the “dimensionality-curse”
  - Independent of the number of dimensions and hierarchical levels within
  - Scalable
Highly compressed storage
- Full Cubes: ALL views answerable
- 100% Precision answers on all views including the fact table
- Stores a subset of the views in very tight space

Tremendous savings
- Storage
- Construction time

Efficient Query Retrieval
- Sub-second response
APB-1 Benchmark

- **Density 1 (1.3M)**
  - Dwarf (Thinkpad): 18 s 57 MB

- **Density 5 (65M)**
  - Oracle’s best benchmark: 4.5 hrs, 30.0+ GB (4 CPU, RAID)
  - Dwarf: 65 min 2.4 GB (Single CPU Pentium 4)

- **Density 40 (496M)**
  - Dwarf: 10.3 hrs 8.2 GB (Single CPU Pentium 4)

NOTE: fact table is 32GB in ASCII, 11.8GB in Binary
Real Data

- Real data set (13,449,327):
  - Dimensions: 8
  - Views: 11,200
  - Creation time: 100 min
  - Size: 6.7 GB
  - 1000 Queries*: 15.8 sec

- Challenge by XYZ
  - 48 hrs for a “wizard” to decide what to materialize
  - Several more hrs to create and index summary tables
  - Huge storage

* Each query asks for 10 different values for 3 randomly selected dimensions (e.g. $v1 | v2 | ... | v10$) and “all” for a 4th dimension- 10*10*10*10 point query
Fact table (5,000,000):
- Dimensions: 10 (3x9L, 4x4L, 3x2L)
- Views: 16,875,000
- Creation: 123 min
- Size: 6.3 GB
- 1000 Queries*: 325 sec

Challenge by XYZ
- This cube can never be built!
What Makes Dwarf Tick Tick

- Two breakthrough discoveries
  - Suffix redundancy
  - Fusion of prefix and suffix redundancy

- Identifies and factors out these redundancies before computing any aggregates for them
Dwarf Technology

- **Complete solution**
  - Extends to high dimensionality
  - Deep hierarchies
  - Queries the full cube - any dimension & level
  - Incremental updates
  - Indexing is inherent – all in one structure
  - Dwarf holds in the fact table too!

- **No gotchas**
  - No expensive preprocessing (just a single sort)
  - No TEMP space required for construction
  - No hidden post-construction costs
  - No information loss (100% precision)
Dwarf Software

- Lean optimized code
- Tools for discovery
  - Data correlation
  - Optimizing dwarfs
- A dozen of tuning knobs including
  - Gmin
  - The Knob
Data Driven Tuning

Gmin

<table>
<thead>
<tr>
<th>$G_{min}$</th>
<th>Space(MB)</th>
<th>Construction/sec</th>
<th>Queries/sec</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>490</td>
<td>202</td>
<td>154</td>
</tr>
<tr>
<td>100</td>
<td>400</td>
<td>74</td>
<td>110</td>
</tr>
<tr>
<td>1000</td>
<td>312</td>
<td>59</td>
<td>317</td>
</tr>
<tr>
<td>5000</td>
<td>166</td>
<td>29</td>
<td>408</td>
</tr>
<tr>
<td>20,000</td>
<td>151</td>
<td>25</td>
<td>476</td>
</tr>
</tbody>
</table>

“The Knob”

<table>
<thead>
<tr>
<th>Knob</th>
<th>Computation</th>
<th>Storage</th>
<th>Workloads A</th>
<th>Workloads B</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4860s</td>
<td>6.6GB</td>
<td>282s</td>
<td>340s</td>
</tr>
<tr>
<td>100</td>
<td>3388s</td>
<td>3.1GB</td>
<td>209s</td>
<td>249s</td>
</tr>
<tr>
<td>500</td>
<td>2038s</td>
<td>2.1GB</td>
<td>198s</td>
<td>238s</td>
</tr>
<tr>
<td>1,000</td>
<td>1794s</td>
<td>1.5GB</td>
<td>186s</td>
<td>222s</td>
</tr>
<tr>
<td>10,000</td>
<td>768s</td>
<td>806MB</td>
<td>191s</td>
<td>229s</td>
</tr>
</tbody>
</table>

Base Dwarf

<table>
<thead>
<tr>
<th>Knob</th>
<th>Computation</th>
<th>Storage</th>
<th>Workloads A</th>
<th>Workloads B</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>552s</td>
<td>764MB</td>
<td>1331s</td>
<td>1706s</td>
</tr>
</tbody>
</table>

Table 9: Knob Evaluation with 13,5 million tuples
Target Markets:
High-Dimensional Data

- Business Intelligence
- Security
- Telecom
- Scientific and sensor data
- Weather data
- Bioinformatics
- Web data (click statistics)
Dwarf’s Value

- Puts any OLAP engine on “steroids” and Delivers substantial performance improvement

- Dwarf is a fast and effective substitute of indexing for ROLAP products (supports SQL API)
Summary of Dwarf

- Practical all in one structure
- Remarkable Full Cube Size Reduction
- Unprecedented performance (construction and query retrieval)
- Scalable (number of dimensions, hierarchy depth, data size)
Dwarf Technology

- Math behind the scene
  - Exploit data dependencies & correlations
  - Probabilistic counting

- Dimension scalability
  - Savings/performance increases exponentially with sparseness (and dimensions)
  - Independence of # of dimensions
Product Status

- US Patent 7,133,876
- Metadata management
  - Mapping between external values and internal binaries
  - Can deal with partial cubes
- Implementation
  - Cross platform (Unix, MS)
  - Connects with all RDBMs
  - Dwarf Browser
ACT’s Experience

- UMD Group established materialized views and incremental access methods (over 50 publications since 1982)
- Data warehouse Cubetree Storage Organization started in 1997 (over 12 publications, ACM Best paper Award)
- Dwarf in 2001-2006