



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



Revolutionary Technology

- 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 CPU, RAID) 4.5 hrs, 30.0+ GB
 - Dwarf 65 min 2.4 GB (Single CPU Pentium 4)
- Density 40 (496M))
 - Dwarf: 10.3 hrs 8.2 GB (Single CPU Pentium 4)

Never done before

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



Real Data

● Real data set (13,449,327):

- **Dimensions:** 8
- **Views:** 11,200
(6+1)(4+1)(4+1)(3+1)(1+1)(1+1)(1+1)(1+1)
- **Creation time:** 100 min
- **Size:** 6.7 GB
- **1000 Queries*:** 15.8 sec

Dimension	Level Cardinalities
A	7458 → 2265 → 737 → 188 → 32 → 11
B	2765 → 91 → 31 → 8
C	3857 → 841 → 111 → 16
D	213 → 68 → 8
E	3247
F	660
G	4
H	4

Table 4: Real Dataset Hierarchies

● 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. $v1 / v2 / \dots / v10$) and “all” for a 4th dimension- 10*10*10 point query



Dream DataCube

- 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

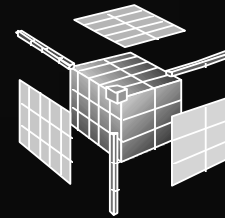
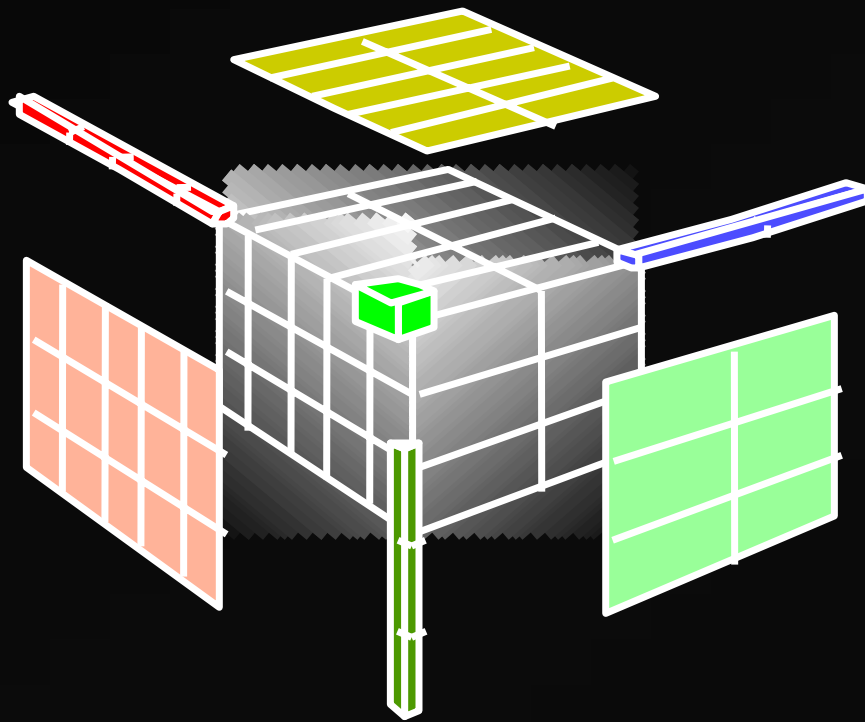
Never done before

- Challenge by XYZ

- This cube can never be built!



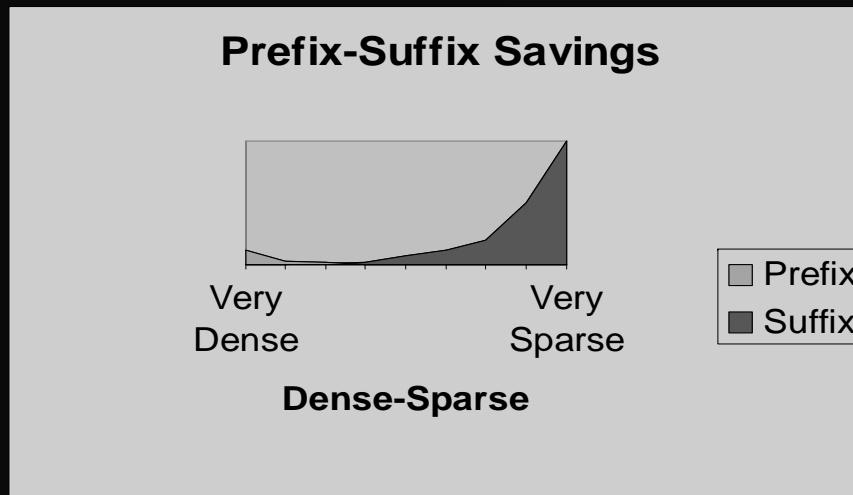
Dwarf Demo





What Makes Dwarf Tick

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



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

Patented



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

G_{min}	Space(MB)	Construction(sec)	Queries(sec)
0	490	202	154
100	400	74	110
1000	312	59	317
5000	166	29	408
20,000	151	25	476

Patented

● “The Knob”

Knob	Computation	Storage	Workloads	
			A	B
0	4860s	6.6GB	282s	340s
100	3388s	3.1GB	209s	249s
500	2038s	2.1GB	198s	238s
1,000	1794s	1.5GB	186s	222s
10,000	768s	806MB	191s	229s
Base Dwarf				
N/A	552s	764MB	1331s	1706s

Patented

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