Even Straight Forward Data Warehouses are Complicated

Panel paper presented at

ACM Tenth International Workshop on Data Warehousing and OLAP (DOLAP 2007)

(in conjunction with ACM CIKM 2007)

Nov. 9, 2007, Lisbon, Portugal

Jens Otto Sørensen

Data Warehouse Architect

jsq@energinet.dk

What is Energinet.dk?

- Energinet.dk is the national Danish Transmission System Operator (TSO)

- As a TSO, Energinet.dk bears the overall responsibility for the electricity and natural gas systems in Denmark

- Only one entity can be responsible for the overall balancing of production and consumption of energy within a certain geographical area

- Denmark has, like a few other countries in EU, a deregulated energy market for electricity and natural gas

- In a deregulated market there has to be an exchange where energy can be traded. Energinet.dk is a co-owner of the Nordic power exchange Nord Pool

Main electricity and gas systems in Denmark

Legend

ELECTRICITY

— 400 kV overhead line, AC

········· 400 kV cable, AC

— Overhead line, DC

········· Cable, DC

◉ Substation

☒ Converter station

Primary power station

Offshore wind farm

GAS

— Gas transmission pipeline

— Gas transmission pipeline owned by others

—◇— Meter and regulator station

Gas storage facility

Business Intelligence at Energinet.dk

- We have a BI solution based on Oracle technology with
  - Oracle Warehouse Builder (OWB) as the development tool
    - We're using OWB version 10g Release 2 (a.k.a. *the version with the new dimensions*)
  - Oracle 10.2 *g* database as data management system for the staging areas and the data marts
  - Oracle Business Intelligence Enterprise Edition (OBIEE) (formerly Siebel) as the presentation tool
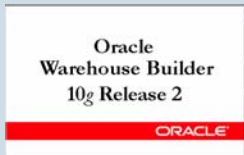
Sources for the Data Warehouse

- We rely on a number of external source for e.g. weather forecast, exchange rates, prices from Nord Pool and so on
- To simplify one could say that our two main sources are the Panda settlement system and the SAP ERP system
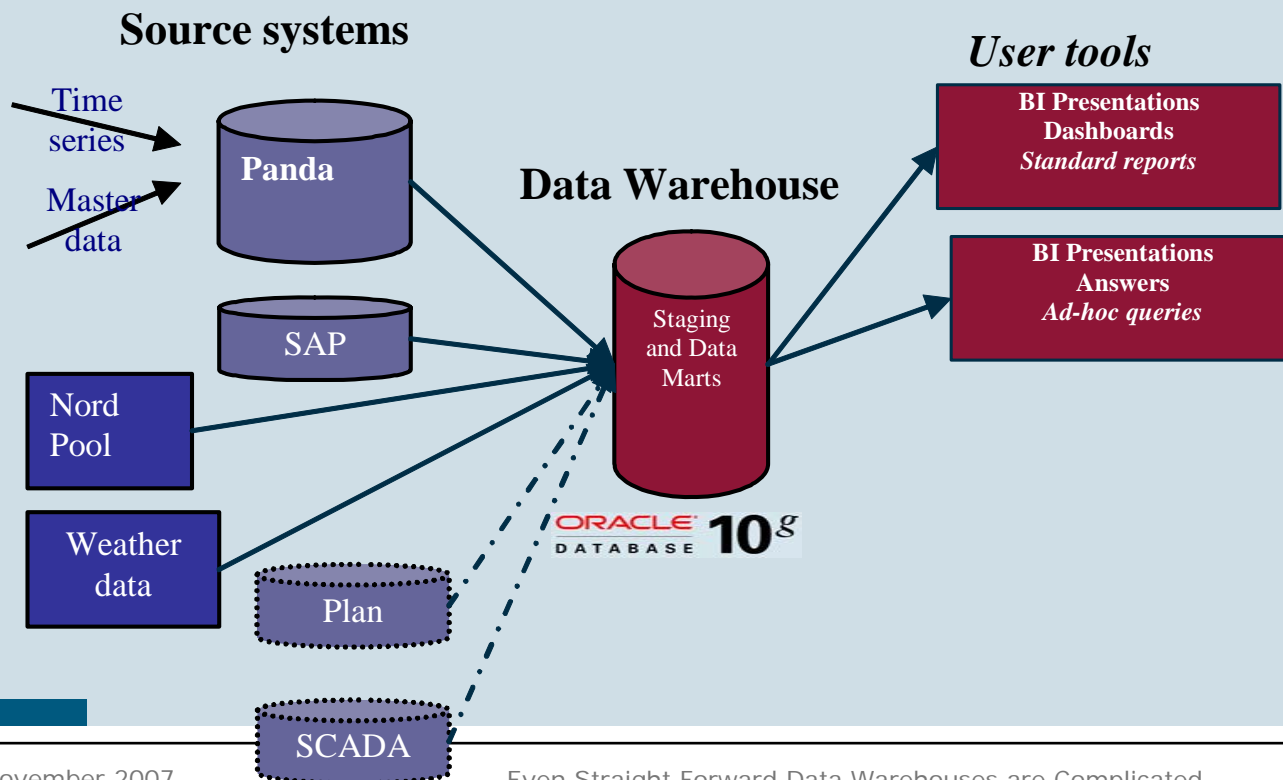
Issue no. 1: Conforming dimensions

- I would like to see a model and implementation details for dimensions that would take into consideration the 5 types of conforming dimensions that I see we could use even in our simple data warehouse:

- The 5 types are

1. Physical identical

2. Copies

3. The schema of one dimension is a subset of the schema of another dimension

4. The two dimensions describe the same but at different levels of granularity

5. The set of tuples in one dimension is a subset of the set of tuples another dimension

Issue no. 2: The time dimension and DLST

Let's assume that a year only has 5 weeks, each week has two days and each day has two hours

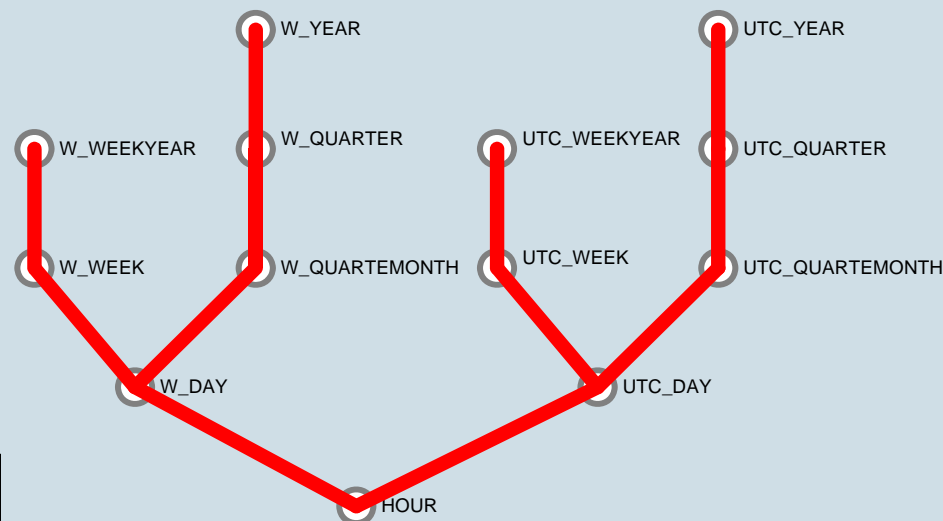A year *without* daylight savings (UTC)

	Month Summer		Month Winter		
	Week 1	Week 2	Week 3	Week 4	Week 5
Day 1	Hour 1	Hour 5	Hour 9	Hour 13	Hour 17
	Hour 2	Hour 6	Hour 10	Hour 14	Hour 18
Day 2	Hour 3	Hour 7	Hour 11	Hour 15	Hour 19
	Hour 4	Hour 8	Hour 12	Hour 16	Hour 20

A year *with* daylight savings (Wrist Watch time)

	Month Summer		Month Winter		
	Week 1	Week 2	Week 3	Week 4	Week 5
Day 1	Hour 1	Hour 5	Hour 8	Hour 12	Hour 17
	Hour 2	Hour 6	Hour 9	Hour 13	Hour 18
Day 2	Hour 3	Hour 7	Hour 10	Hour 14	Hour 19
	Hour 4		Hour 11	Hour 15	Hour 20
				Hour 16	

I.e. if you want the production for hour 8 you must chose

- the 2nd. hour in Day 2 in the summer month *if* you are using UTC but you must chose

- the 1st. hour in week 3 of the winter month *if* you are using Wrist Watch time

W_YEAR

UTC_YEAR

W_WEEKYEAR   W_QUARTER   UTC_WEEKYEAR   UTC_QUARTER

W_WEEK   W_QUARTEMONTH   UTC_WEEK   UTC_QUARTEMONTH

W_DAY   UTC_DAY

HOUR

Issue no. 3: Tools Integration or Lack Thereof

- No tools cover everything
- When retrieving source data, very often, from a number of heterogeneous sources what model should be used to signal e.g. failure to contact source? failure to load?

Issue no. 4: Continuous updates

- Much research has gone into doing efficient reads in cubes (querying the cubes)
- But can we do equally fast writes (updates) into cubes and dimensions without seriously impairing the response time that the user experience

Issue no. 5: A Query Langue for Multi Dimensional Cubes

- It would be beneficial to all involved if there could be a standard query language for multidimensional cubes

- We do have MDX and XMLA, but it is not a standard that all will adapt as e.g. SQL

Conclusion

- There seems to be a tendency within the academic field that moving into totally new areas carries more money and more prestige than finishing or adding consistency and rigorous models to an area or field that has been somewhat commercialized