



**Introduction to Data Warehousing  
and Business Reporting**

University of Washington : May 9, 2006

Nicholas Goodman

Director of Business Intelligence Solutions


[ngoodman@pentaho.org](mailto:ngoodman@pentaho.org)

# Presentation

- Perspective: I do this day in / day out
  - `select question from uw_students where q_elapsed_seconds < 2;`
- Business Reporting Basics (10 min)
- Dimensional Modeling Basics (20 min)
- Data Transformation (10 min)
- Oracle Specifics (10 min)
- Discuss / Questions (10 min)

# Basics: Reporting

- Access and format data from disparate sources
  - Oracle but then...
  - DB2, CSV, XML, Salesforce.com
- Holistic view of business
  - A customer Order touches:
    - Billing
    - Website
    - Fulfillment
    - Warehousing
    - Call center
    - Etc
- Inherently Semantic
  - Customers, Lifetime Value, Marketing Categories, Products



Steel Wheels  
500 International Speedway, Daytona, FL 32114  
(123) 456-7890 http://www.steelwheels.com  
Run Date: 2/20/06 1:24 PM

Steel Wheels  
500 International Speedway, Daytona, FL 32114  
(123) 456-7890 http://www.steelwheels.com  
Run Date: 2/20/06 1:24 PM

**TO:** Reims Collectables  
59 rue de l'Abbaye, null  
Reims, null 51100 France

Attn: Paul Henriot  
Sales Rep: 1337  
Terms: Net 30 days

**INVOICE**

Invoice #: 10121  
Account Number: 353  
Date: May 07, 2003

SKU	Product Description	Price/Unit	Qty Ordered	Total Price
S50_4713	2002 Yamaha YZR M1	\$74.95	44	\$3,293.40
S24_2390	1982 Ducati 900 Monster	\$76.88	32	\$2,460.16
S32_4485	1974 Ducati 350 MK3 Desmo	\$86.74	25	\$2,168.50
S12_2823	2002 Suzuki XSR60	\$165.58	50	\$8,279.00
S10_1678	1999 Harley Davidson Ultimate Chopper	\$81.35	34	\$2,765.90
				\$18,971.96

**Send Payment and Remittance Slip to:**  
Steel Wheels  
500 International Speedway  
Daytona, FL 32114

Thank you for your

Account Number: 353

**REMITTANCE**

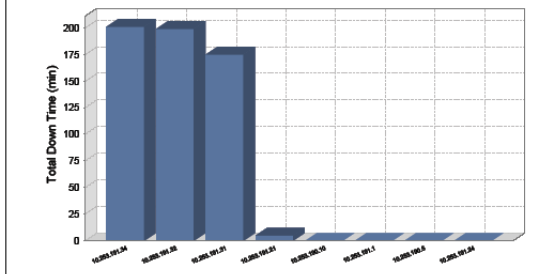
Reims Collectables  
59 rue de l'Abbaye, null  
Reims, null 51100 France

**TopN Least Available**  
TopN least available managed assets

**Summary** Report Period: Feb 01, 2006 12:00 AM to Feb 10, 2006 5:46 PM

Number of Elements Included: 10  
Availability Target: 99.000%  
Types of Outages Contributing to Downtime: Unplanned  
Business Day Policy: Full Day Including Weekends

**Availability Graph**



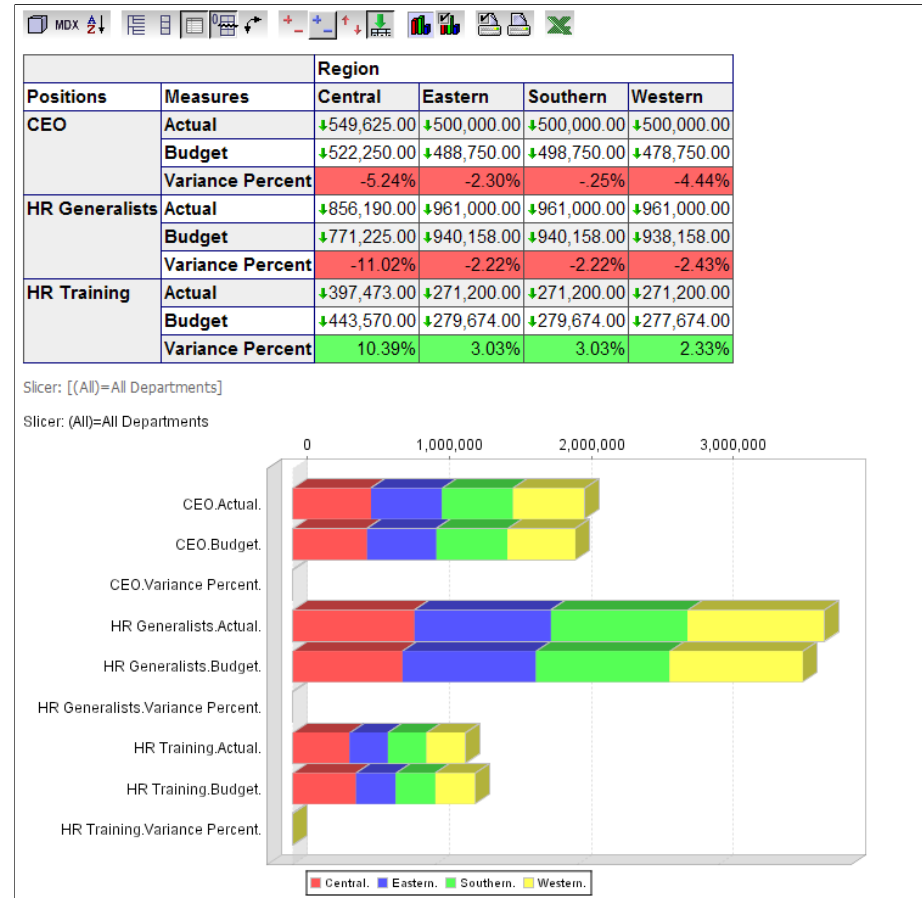
**Report Details**

Name	IP Address	Type	Total Downtime	No. Of Outages	Availability %
10.253.191.34	10.253.191.34	Windows Host	03h 20m 50s	11	98.568%
10.253.191.32	10.253.191.32	Windows Host	03h 18m 39s	12	98.581%
10.253.191.31	10.253.191.31	Windows Host	02h 55m 05s	15	98.752%
10.253.191.21	10.253.191.21	Windows Host	00h 05m 00s	1	99.964%
10.253.190.10	10.253.190.10	RS-3000	00h 00m 00s	0	100.000%
10.253.191.1	10.253.191.1	CiscoWSS302	00h 00m 00s	0	100.000%
10.253.190.5	10.253.190.5	RS-3000	00h 00m 00s	0	100.000%
10.253.191.24	10.253.191.24	Windows Host	00h 00m 00s	0	100.000%

Report Generated on Feb 12, 2006 10:54:53 AM Page 1 of 1

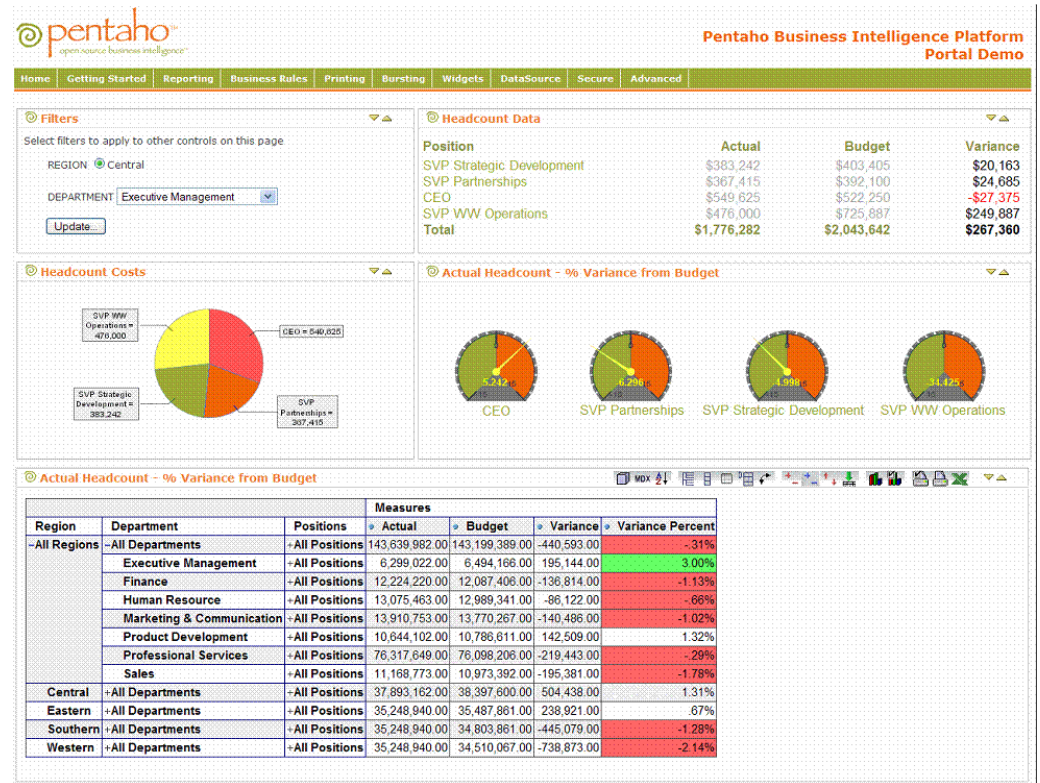
# Basics: Analysis

- View data “dimensionally”
  - i.e. Sales by region, by channel, by time period
- Navigate and explore
  - Ad Hoc analysis
  - “Drill-down” from year to quarter
  - Pivot
  - Select specific members for analysis
- Interact with high performance
  - Technology optimized for rapid interactive response



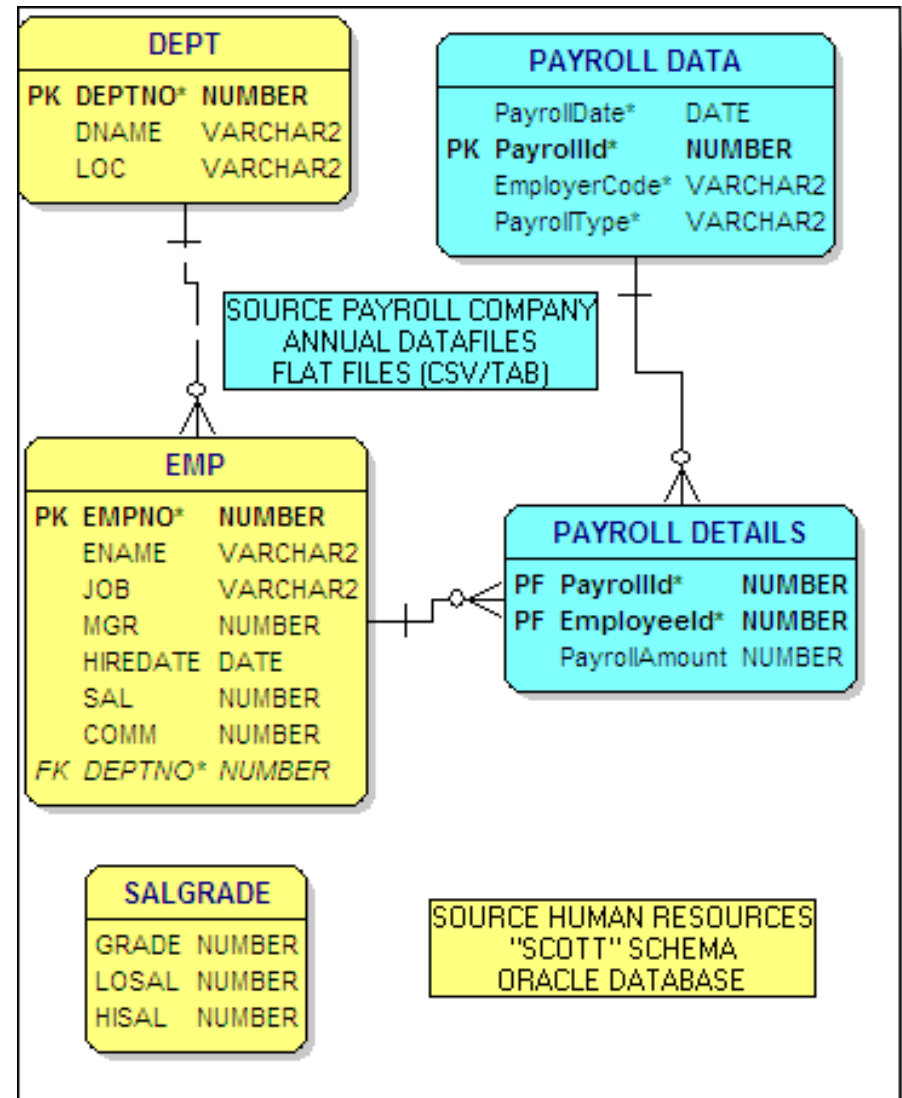
# Basics: Dashboards

- Monitor Key Performance Indicators (KPIs) / metrics
  - Investigate underlying details
    - Drill to supporting reports
- Track exceptions
  - Alert users based on business rules



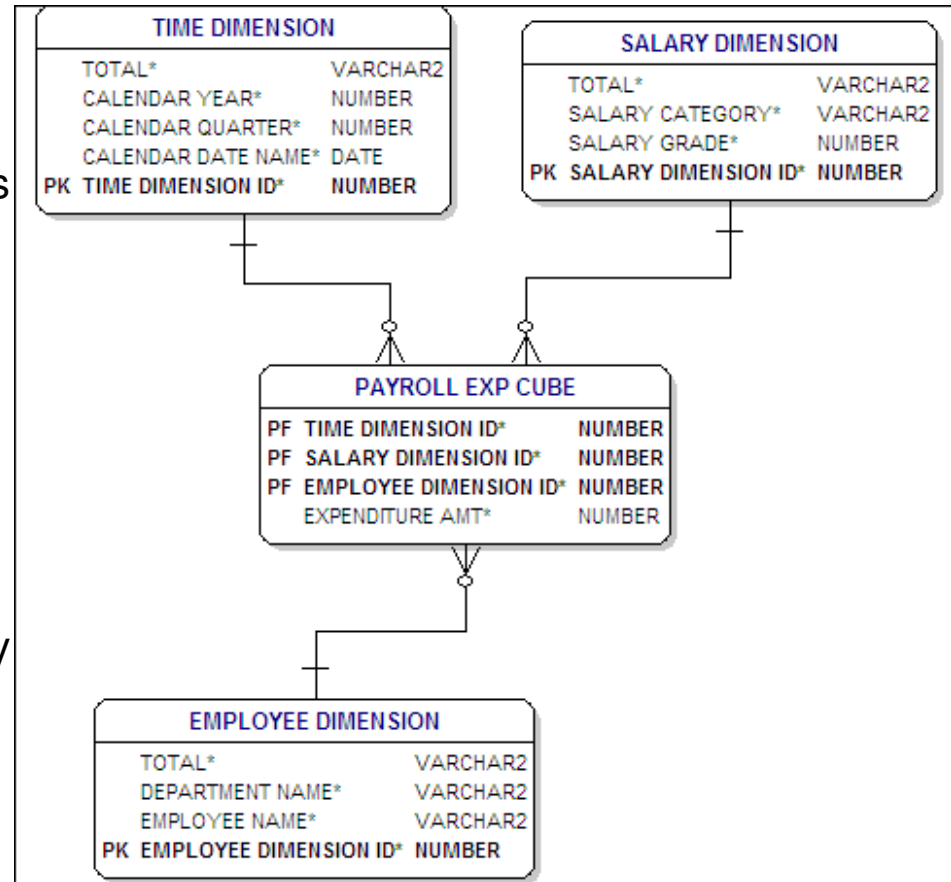
# Basics: Relational Rules, right?

- Most DATABASE Training:
  - Relational Databases
  - 3NF = IDEAL
  - Keys, Joins, Roles, Flexibility
- OLTP
  - OnLine Transaction Processing
  - Database to support your applications
  - IDEAL MODEL FOR:
    - Lots of Users, Small slices of Data
    - Ie, Debit account # 1002 \$40.00 from withdrawal at ATM #6551
  - BAD MODEL FOR:
    - Few Users, Large Slices of Data
    - Sums, Aggregations, Calculations

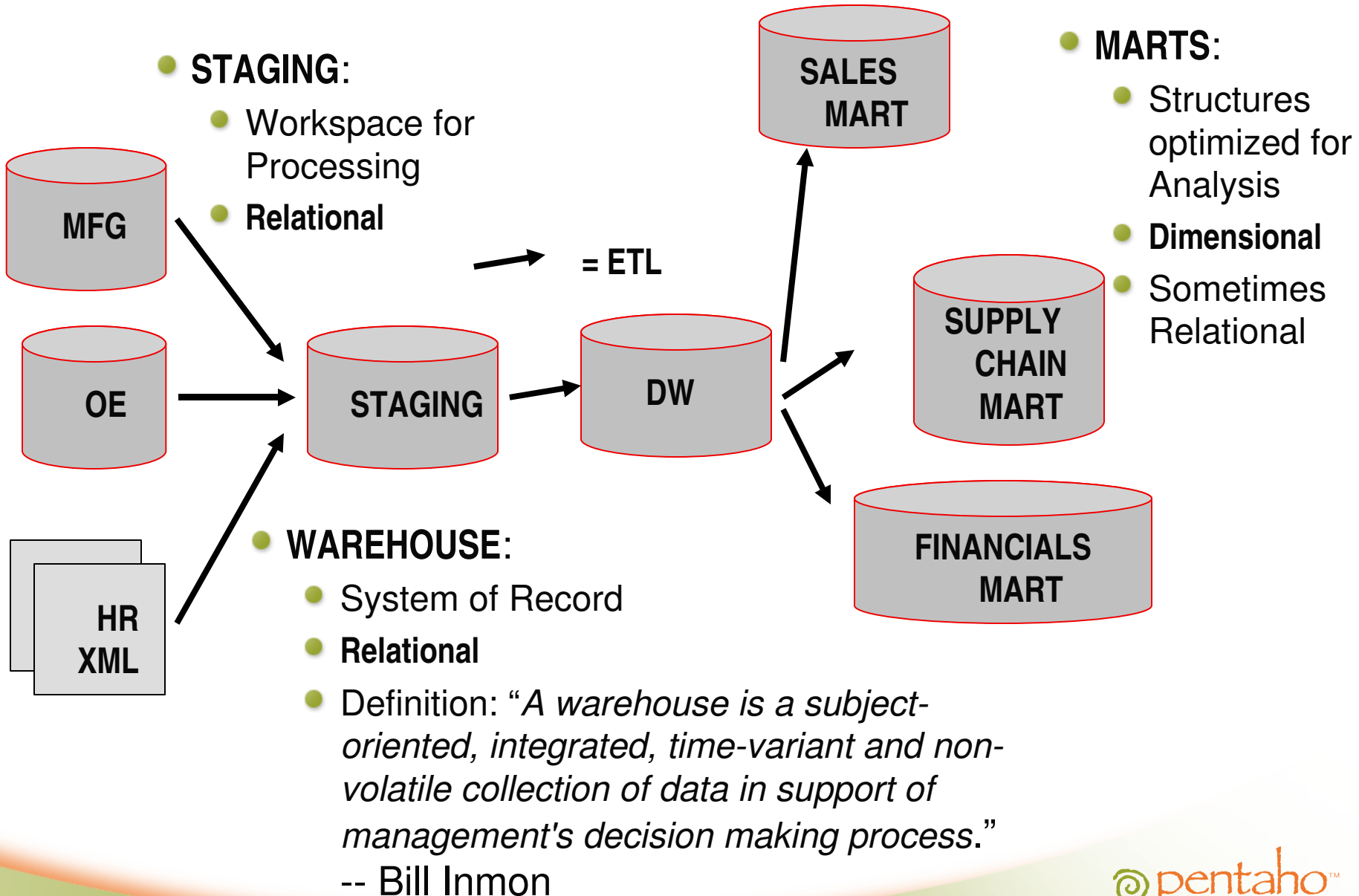


# Basics: Dimensional Models

- Reporting DATABASE Training:
  - Relational AND Dimensional Databases
  - Relational = ODS or Data Warehouse
  - Dimensional = Reporting Applications
- OLAP
  - OnLine Analytical Processing
  - IDEAL MODEL FOR:
    - Few Users, Huge Amounts of Data
    - Aggregates, slice and dice (sales by about 100 different qualifiers)
    - Ie, What is the proportion of ATM withdrawals that occur within 1 mile of the persons primary address?
  - BAD MODEL FOR:
    - Running your applications



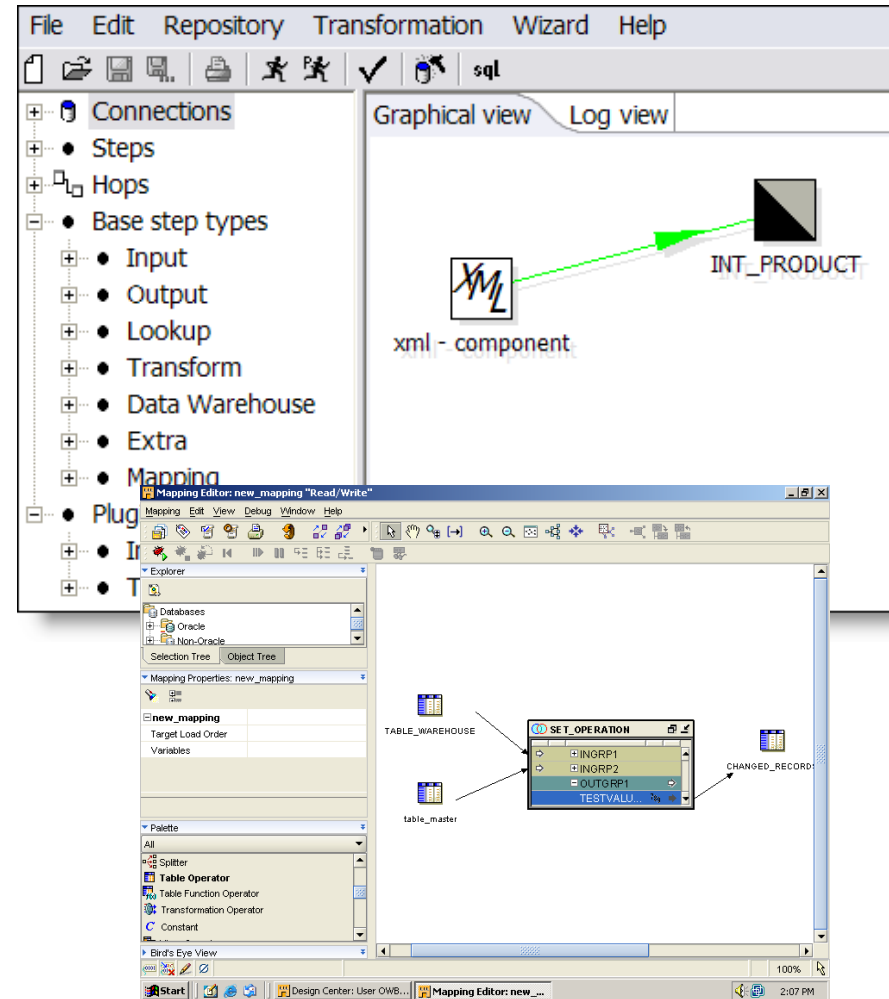
# Basics: Corporate Information Factory





# Basics: Extract Transform Load

- Data Processing
  - Pull data from X,Y,Z and insert or update in the Warehouse
- Logical Transformations
  - Sum, Join, Outer Join, Bucketize, calculate time variant items
  - *Everything you need to process flat files, XML, Tables into a set of tables that represent your reporting data.*
- Tools
  - Visual and include a Logical and Physical representation
  - Kettle and OWB
- SQL Scripts / Perl / Cron



# Dimensional Modeling: Star Schema

- **FACTS**

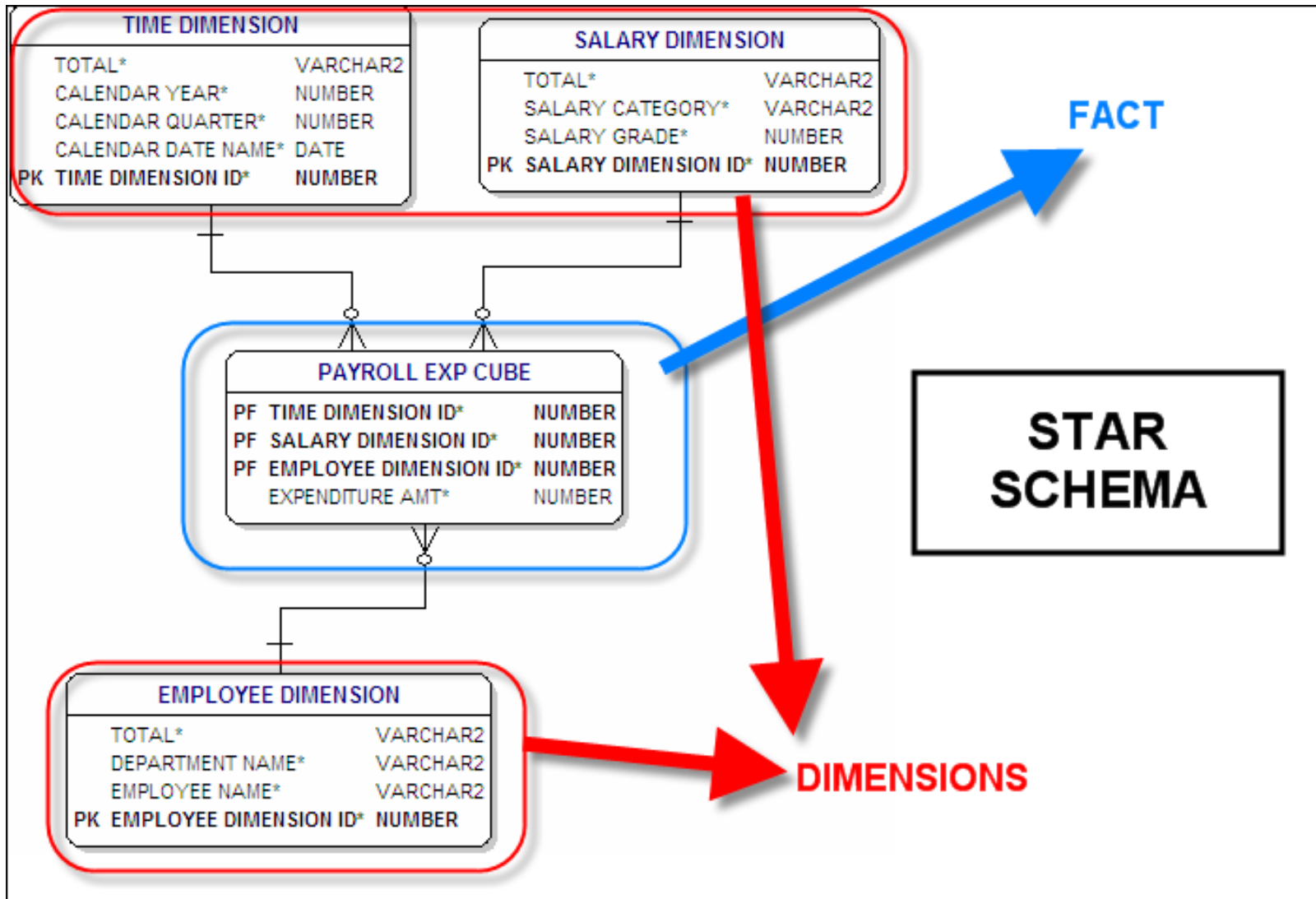
- Has what you are trying to **MEASURE** (Sales, Expenditures)
- Usually Numeric
- *Tough to model facts “correctly” when you’re learning*

- **DIMENSIONS**

- How you are trying to qualify measures. Products, Time, Department, etc.
- **DENORMALIZED**
- Usually Hierarchical (Year -> Qtr -> Mon -> Day)
- *Feels “weird” the first few times*

Positions	Measures	Region			
		Central	Eastern	Southern	Western
CEO	Actual	↓549,625.00	↓500,000.00	↓500,000.00	↓500,000.00
	Budget	↓522,250.00	↓488,750.00	↓498,750.00	↓478,750.00
	Variance Percent	-5.24%	-2.30%	-2.5%	-4.44%
HR Generalists	Actual	↓856,190.00	↓961,000.00	↓961,000.00	↓961,000.00
	Budget	↓771,225.00	↓940,158.00	↓940,158.00	↓938,158.00
	Variance Percent	-11.02%	-2.22%	-2.22%	-2.43%
HR Training	Actual	↓397,473.00	↓271,200.00	↓271,200.00	↓271,200.00
	Budget	↓443,570.00	↓279,674.00	↓279,674.00	↓277,674.00
	Variance Percent	10.39%	3.03%	3.03%	2.33%

# Dimensional Modeling: Star Schema



# Dimensional Modeling: Step 1

- REQUIREMENTS REQUIREMENTS REQUIREMENTS

- **Business Users Drive Process**

- Do NOT ask precisely what numbers do you want!
  - They ask for everything as a flat file or report so they can do their own analysis.
  - What they WANT and what they NEED are usually different.
- Have them express real English analytical 'wish list.'
- Examples:
  - I would like to know what is the proportion of Sales by my different product groups and customer types.
  - What is the proportion of revenue that comes from repeat versus first time customers.
  - What is the profile of customers (profile = Location, Income, and Gender) that make up 80% of my actual Profit as opposed to 80% of revenue.
- Have them show you their clandestine MS Access or Excel spreadsheet
  - Every numbers business group has "a guy" with "a spreadsheet" that consolidates, processes, and prepares the data like the business users desire.
  - Find this GUY and make him your best friend.

# Dimensional Modeling: Step 2

- FIND PATTERNS

- Begin to identify the WHAT's and the BY's
- WHAT = **FACT** (measures)
- BY = **DIMENSION**
  - Examples:
    - I would like to know what is the **proportion of Sales** by my different **product groups** and **customer types**.
    - What is the **proportion of revenue** that comes from **repeat versus first time customers**.
    - What is the profile of **customers** (profile = **Location, Income, and Gender**) that make up 80% of my **actual Profit** as opposed to 80% of **revenue**.
- Develop a rough dimensional model
  - Meant to help PROTOTYPE reports

# Dimensional Modeling: Step 3

- **PROTOTYPE A FEW “CROSS TAB”**

## REPORTS

- Use Excel because it's wicked easy
- Helps them “SEE” the dimensional model without abstract terms like “Dimensions” and “Facts”
- *Try before you buy*

Revenue	Profit	Product	State
15	3	Xbox	Washington
16	1	Playstation	Lousiana
14	2	Xbox	Wyoming
29	1	Xbox 360	Arizona
87	3	Xbox	Arizona
12	1	Playstation	Missouri
29	2	Xbox	Texas
101	1	Xbox 360	Washington
921	3	Xbox	Texas
83	1	Playstation	Oregon
18	2	Xbox	Oregon
291	1	Xbox 360	California

**PROTOTYPE  
SOURCE**

Sum of Revenue	Product			
State	Playstation	Xbox	Xbox 360	Grand Total
Arizona		87	29	116
California			291	291
Lousiana	16			16
Missouri	12			12
Oregon	83	18		101
Texas		950		950
Washington		15	101	116
Wyoming		14		14
Grand Total	111	1084	421	1616

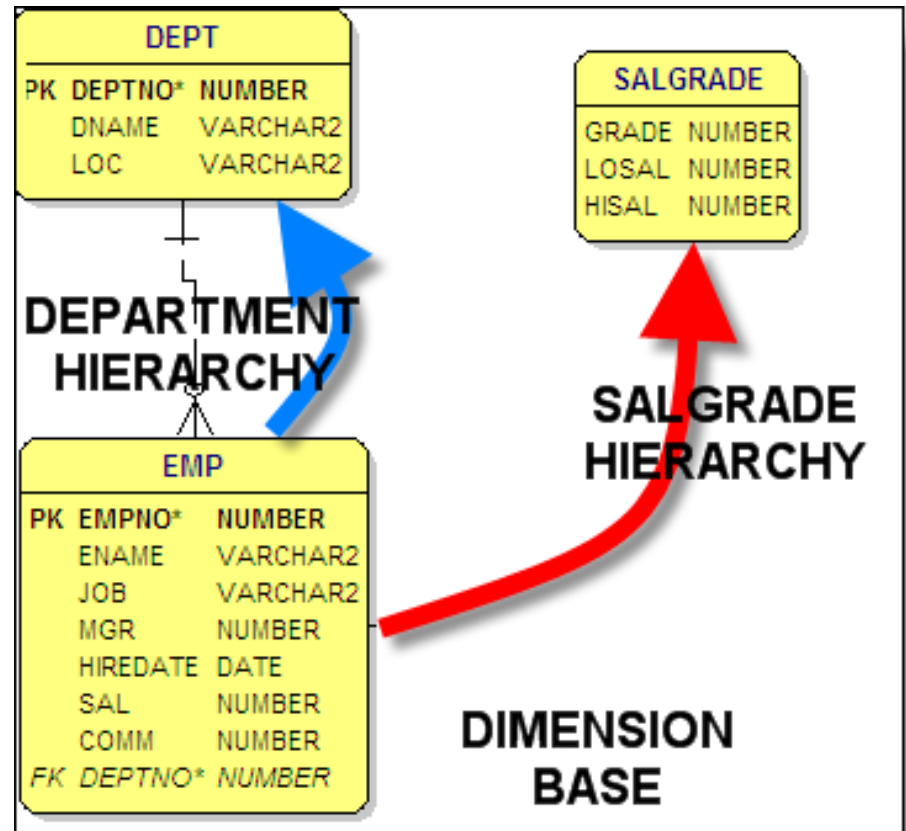
**CROSSTAB**

# Dimensional Modeling: Step 4

## REFINE MODEL

### AND IDENTIFY HIERARCHIES

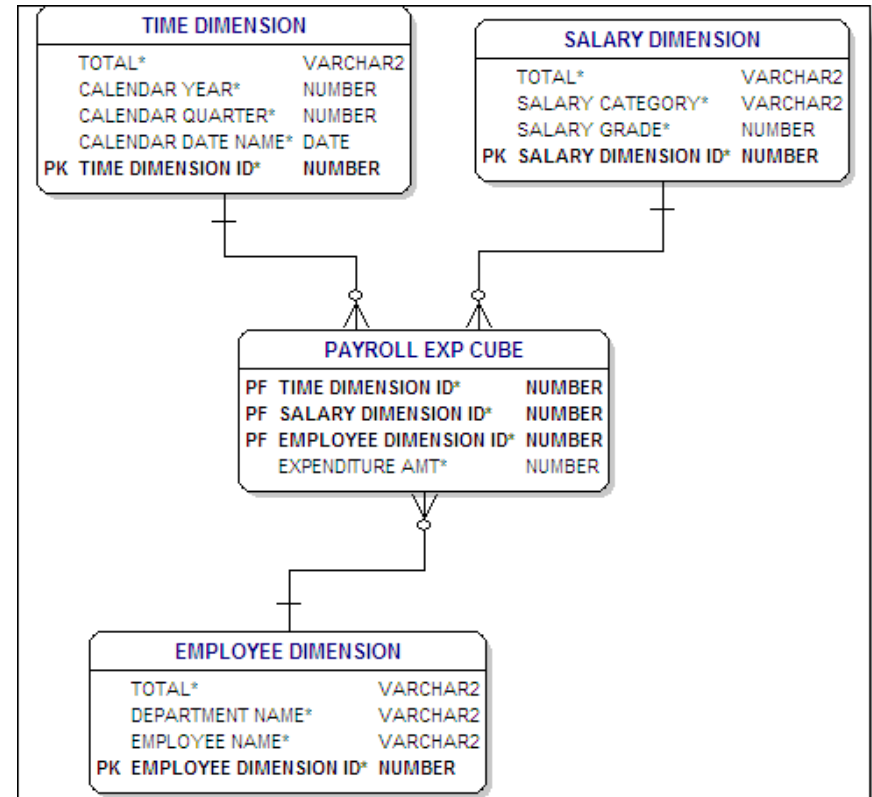
- Use feedback from the business users to further refine additional FACT measures (Revenue, Profit, Cost of Goods, etc).
- Grab other attributes close
  - Product Short Name → Product Long Name.
  - Country Name → Country ISO code.
- FIND HIERARCHIES
  - Requirements are good place to find them.
  - **SOURCE SYSTEM MASTER/DETAIL** is a good indicator of hierarchical data.



# Dimensional Modeling: Step 5

## ● FINISH MODEL AND SANITY CHECK

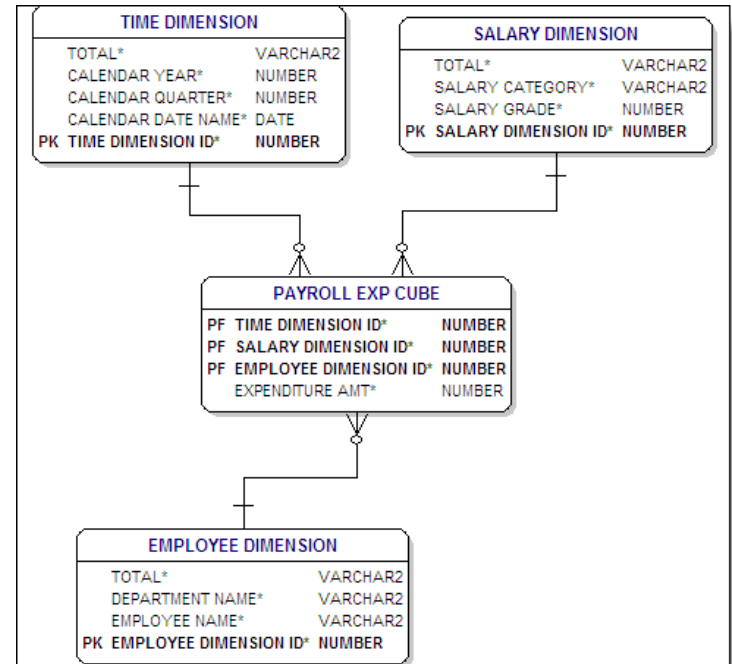
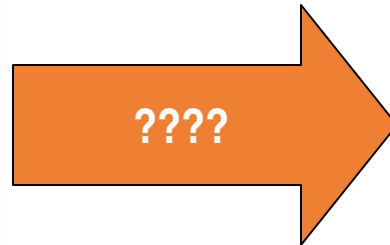
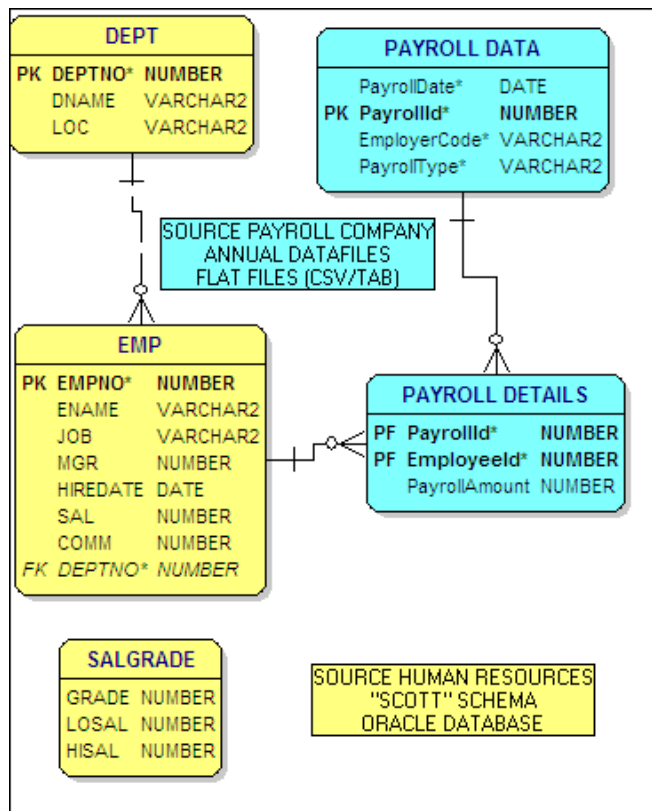
- Finish the STAR SCHEMA and build the DIMENSIONAL MODEL
- SANITY CHECK 1: Source Data?
  - Document the PSEUDO-ETL, a simple logical description of how you take your data in your source system and turn it into the dimension or fact.
  - Verifies that there's not a "missing" piece of data that makes the model useful.
- SANITY CHECK 2: Can you write SQL/MDX against your model?
  - Run through your mock up reports, and free text questions.
  - Mentally walk through your reports, and ensure you can answer your reports from this model





# Data Transformation: Problem to Solve

- Turn the OLTP data (source data) into our OLAP data (star schema)
- Known as Extract Transform and Load (ETL)



# Data Transformation: ETL

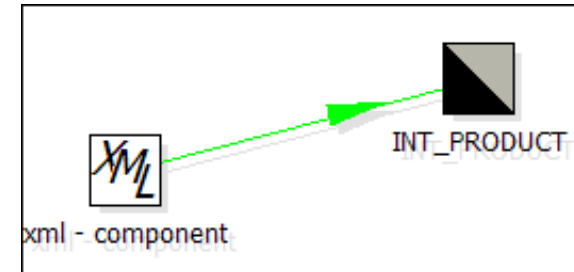
- Tools or Technologies that process source data and insert/update data in the warehouse based on the business rules defined.
- Example
  - We need to turn our source data into our warehouse data
  - Source System:
    - **ORDER\_LINE\_ITEM**: Quantity, Discount Amount, Actual Price
  - Data Warehouse:
    - **SALES\_FACT**: REVENUE, DISCOUNT PERCENT, etc.
- Technologies
  - SQL (if it's in the same database you can use SQL to do this)
  - Perl (original data warehouse toolkit, still in common use)
  - Commercial Tools (Oracle Warehouse Builder, Informatica)
  - Open Source Tools (Kettle, KETL)
- **BOTTOM LINE**: Get the job done!

# Data Transformation: ETL Topics

- Surrogate Keys
  - Protect yourself from source system changes.
  - Needed since Dimensions of TYPE II (see appendix) will have a different natural key.
  - Example:
    - Surrogate Id 1 / Customer Id : 100 / State: WA
    - Surrogate Id 2 / Customer Id : 100 / State: CA
- History
  - EFFECTIVE and EXPIRATION dates
  - Example:
    - Surrogate Id 1 / Customer Id: 100 / Eff 01-Jan-2006 / Exp 31-Mar-2006
    - Surrogate Id 2 / Customer Id: 100 / Eff 31-Mar-2006 / NULL
    - Accurate reports the sale two years ago and last week.

# Data Transformation: ETL Topics cont

- UPSERTS
  - INSERT/UPDATE is a common pattern in situations
- Process Everything or Just Changes
  - Deltas = Changes since last processing
  - Detected Deltas
    - Compare Yesterday's data to right now and build a list of changes
  - Application Managed Deltas
    - Corresponding SOURCE\_HISTORY table that has the data history
  - Database Managed Deltas
    - Streams / Triggers



# Oracle Specifics: OLAP Performance

- BITMAP Indexes
  - Ensures one pass through dimension tables (small, < 100k rows) and only ONE scan of FACT table (usually large, millions of rows)
- Parallel Query with Partitioned Fact tables
  - Allows for the “ONE scan” of the FACT table to be split across CPUs (nodes in RAC?) and I/O channels. I/O is MORE important than CPUs. Data Warehouse queries are almost ALWAYS waiting on disks.
- Materialized Views
  - Watch your reporting tool (Discoverer, Mondrian) and determine what SQL your “canned” reports are generating. Building a corresponding MView and refreshing after load will make these LIGHTNING quick!
- Oracle Tuning
  - Few Users, Lots of Sort Operations (group by)
  - Dedicated Connections

# Oracle Specifics: Misc

- ETL
  - Merge Statements
    - ROCK for doing UPSERTs in the database.
  - Sequences for Surrogate IDs.
- REDO
  - Lots of REDO during batch load.
  - Hardly ANY REDO during data access.
- Oracle Streams
  - Next generation message based Delta communication.
  - Log Miner ++.
  - Allows the warehouse to get a complete view of the Oracle source.
- Availability / Backup
  - Can USUALLY take cold backups (10pm at night).
  - Has less stringent availability then OLTP databases.

# Appendix: Where to go for more information

- Business Intelligence Tools
  - Oracle
    - Oracle Warehouse Builder, Oracle Discoverer, Oracle OLAP Option, Oracle Designer, Oracle BI Suite Enterprise Edition (2006)
  - Open Source
    - [www.pentaho.org](http://www.pentaho.org) (Reporting, OLAP, Data Integration, etc)
    - Free to use and prototype; use for your learning!
- Modeling and Data Warehousing
  - Ralph Kimball expert in Dimensional Modeling
    - Kimball University <http://www.kimballgroup.com/>
    - Data Warehouse Toolkit (Book, Kimball)
  - Data Warehousing
    - The Data Warehouse Institute Classes (<http://www.tdwi.org/>)
    - Corporate Information Factory (Book, Imhoff)

# Appendix : Slowly Changing Dimensions

- Type I
  - Corrections / Updates
  - There is no history kept in dimensions, changes in source are updated in warehouse.
- Type II
  - Historical
  - Multiple “versions” of the customer are kept in the warehouse.
  - Example:
    - Customer moves from WA → MA. Need to attribute one fact in WA and the other in MA but both from the same customer.
- Type III
  - Old and New Together
  - Typically used for change of classifications/rollups
  - Example:
    - Reorganization of Sales Organization.
    - New Sales Territory: Pacific, Old Sales Territory: West