

Dispelling Myths and

Creating LEGENDS

for your E-Biz Intelligence Warehouse

--- AOTC Conference 12/9/2005 ---

Jeffrey Bertman (jefflit@dbigusa.com)

Chief Engineer

DataBase Intelligence Group (DBIG)

- IT Consulting • Strategic Planning • On-Call Support •
- Advanced Technology Implementation and Troubleshooting •

Reston Town Center
11921 Freedom Drive, Suite 550
Reston, VA 20190

WDC/VA/MD Region
(703) 405-5432
www.dbigusa.com

Objectives

- Provide a Brief Background and Refresher:
Define the Data Warehouse (DW) and its Purpose
- Show **WHAT** we are Building
- Show **HOW** we Build It “If you Build it, They will Come”
➔ More Details in Separate Paper: “A REAL DWhs Project Plan...”
- Identify some Tips, Traps, and Best Practices
Road Trip from...
MYTH MEDICINE to LABOR LEGEND

Background

- Somewhat new term for old idea
- Use of the term dates back to early 1990s
- Formerly known as a Decision Support System or Executive Information System
- William Inmon (father) and Ralph Kimball are central innovators

What is a Data Warehouse (DW) ?

- An organized system of enterprise data derived from multiple data sources designed primarily for decision making in the organization.
- According to Bill Inmon, father of DW:
 - Subject Oriented
 - Time Variant (Historical)
 - Integrated
(standardized from multiple sources)
 - Nonvolatile (Query-only)

Purpose of the DW

- Provide a foundation for decision making
- Make information accessible
- Make information consistent
- Adapt to continuous change in information
- Protect information from abuses

How We View Data — Slice & Dice

Dimensions:
HOW we
Analyze

Facts:
WHAT we Analyze and Measure

\$Dollars	Quantity	Events	Facts/Info
-----------	----------	--------	------------

Region

*

*

*

*

Cost/Profit
Center

*

*

*

*

Campaign
/Promotion

*

*

*

*

Demographics

*

*

*

*

TIME

*

*

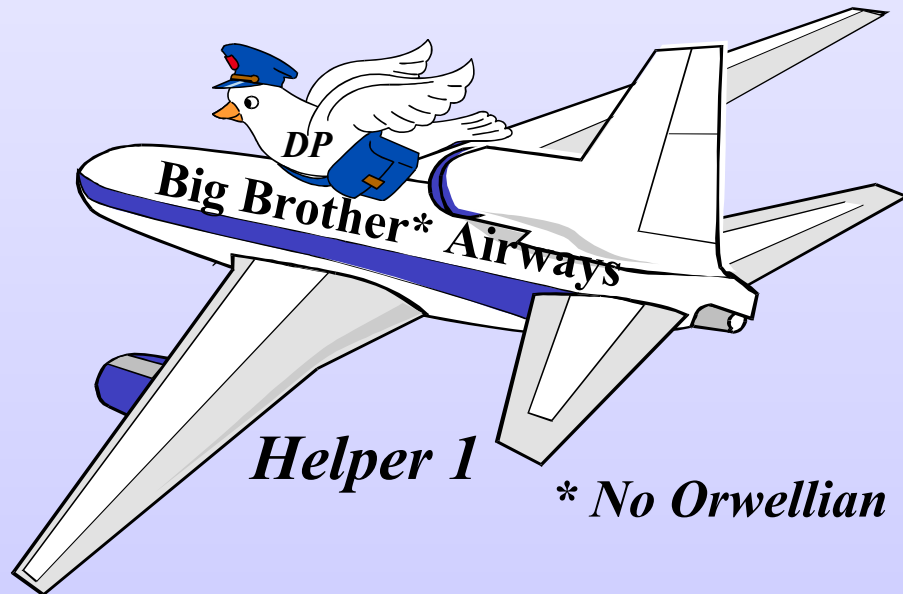
*

*

From Here to There

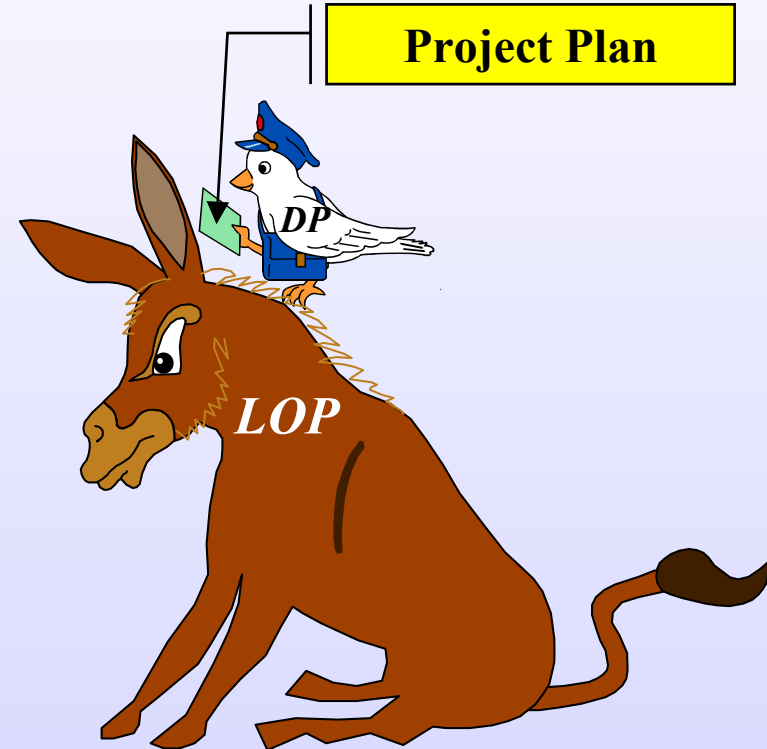


Helper 2



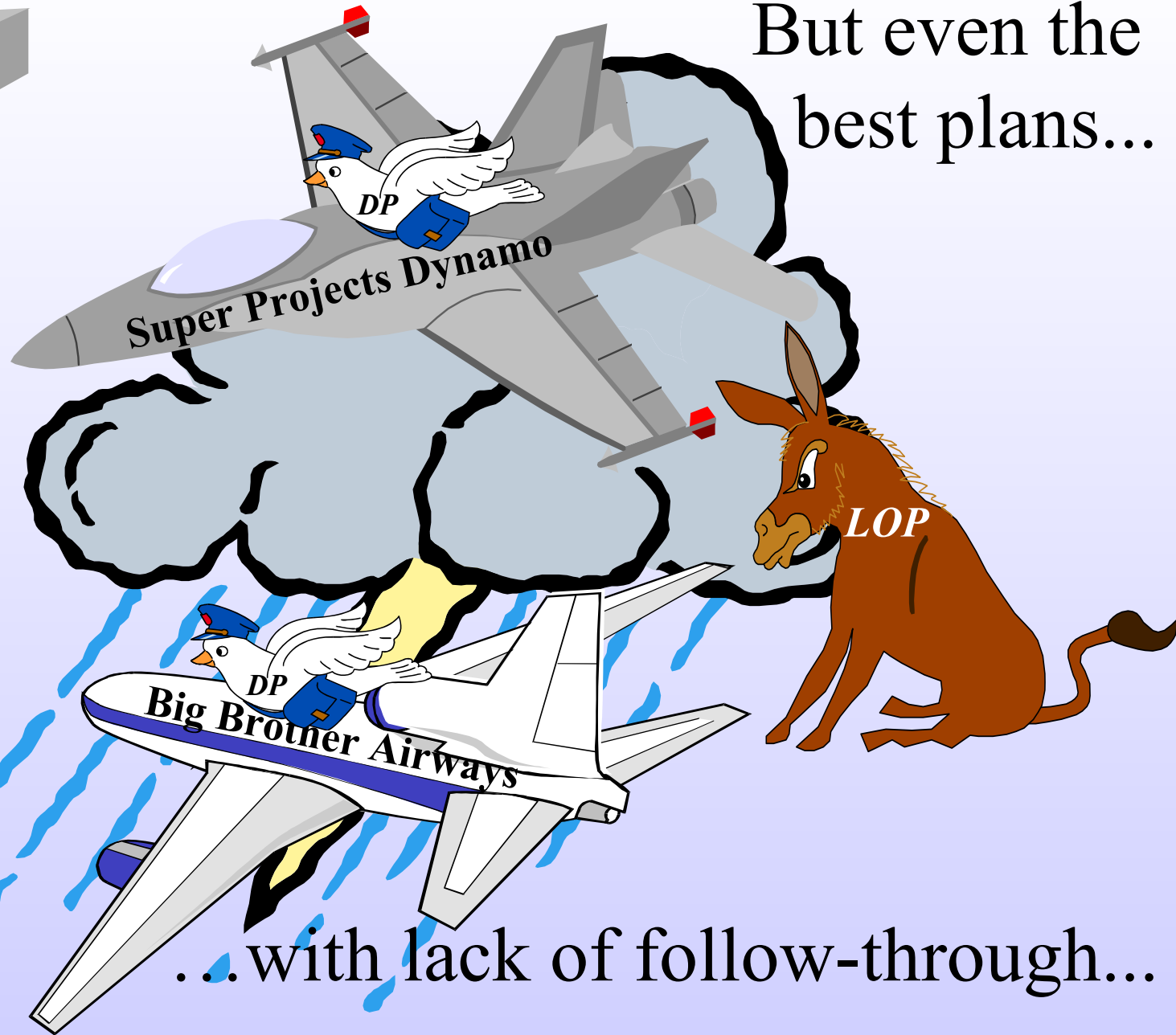
Helper 1

** No Orwellian Connotation*



*Challenge
Numero
Uno*

But even the
best plans...



...with lack of follow-through...

...can lead to...





Opening the door to success requires both
planning and follow-through...



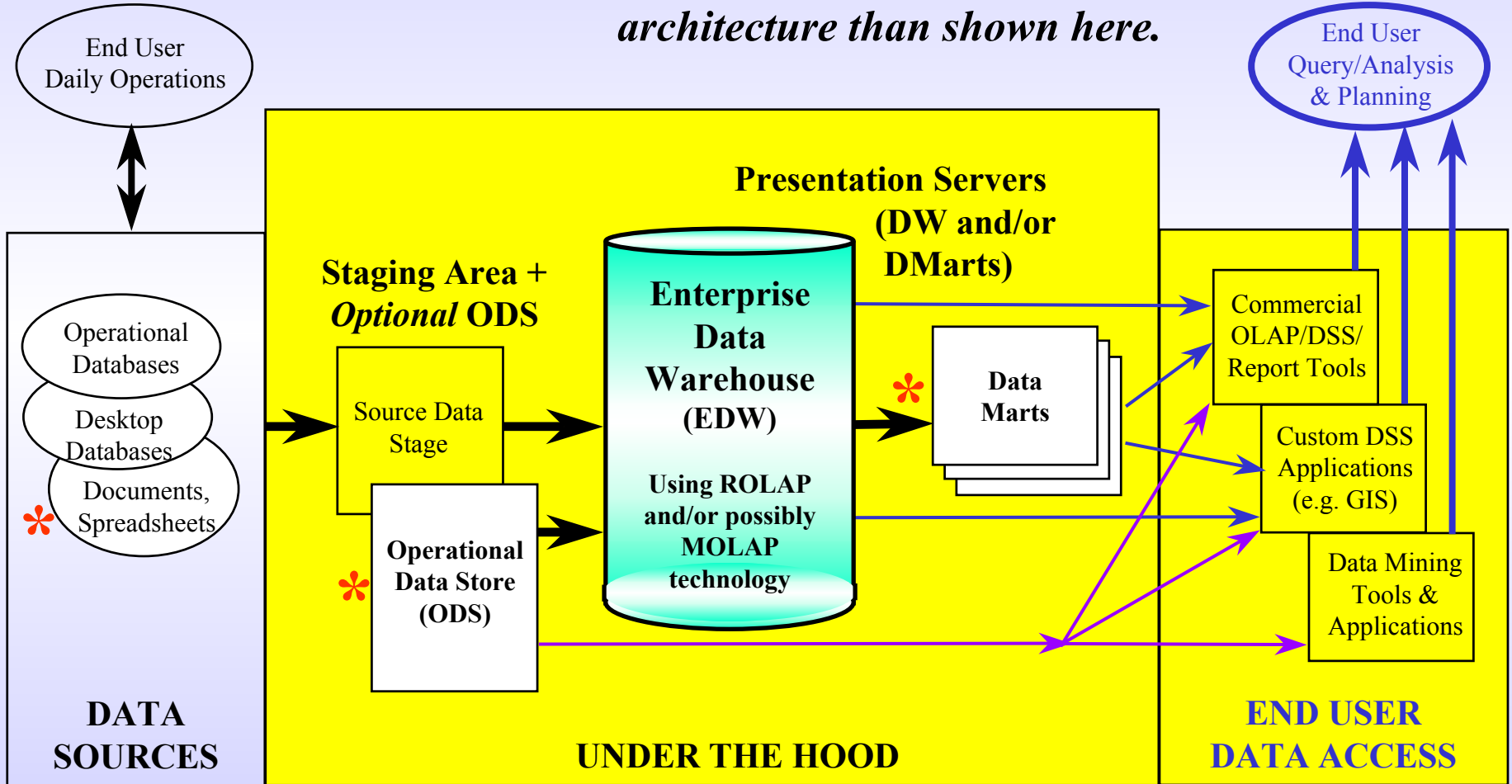
...to tame LOP the mule into LOP2 !

What are We Building? One Classic View...

Follow the Yellow Brick Road :-)

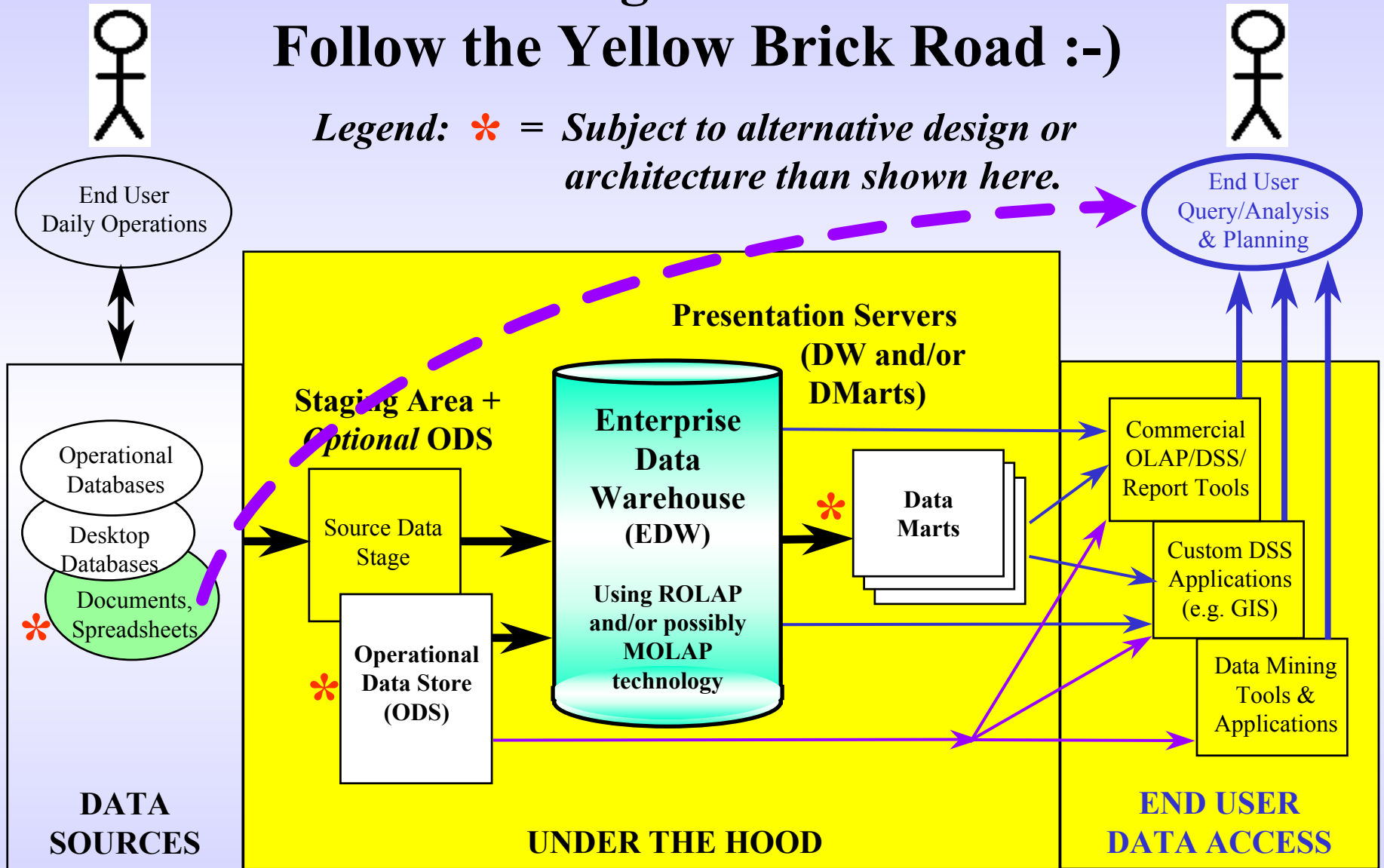


*Legend: * = Subject to alternative design or architecture than shown here.*



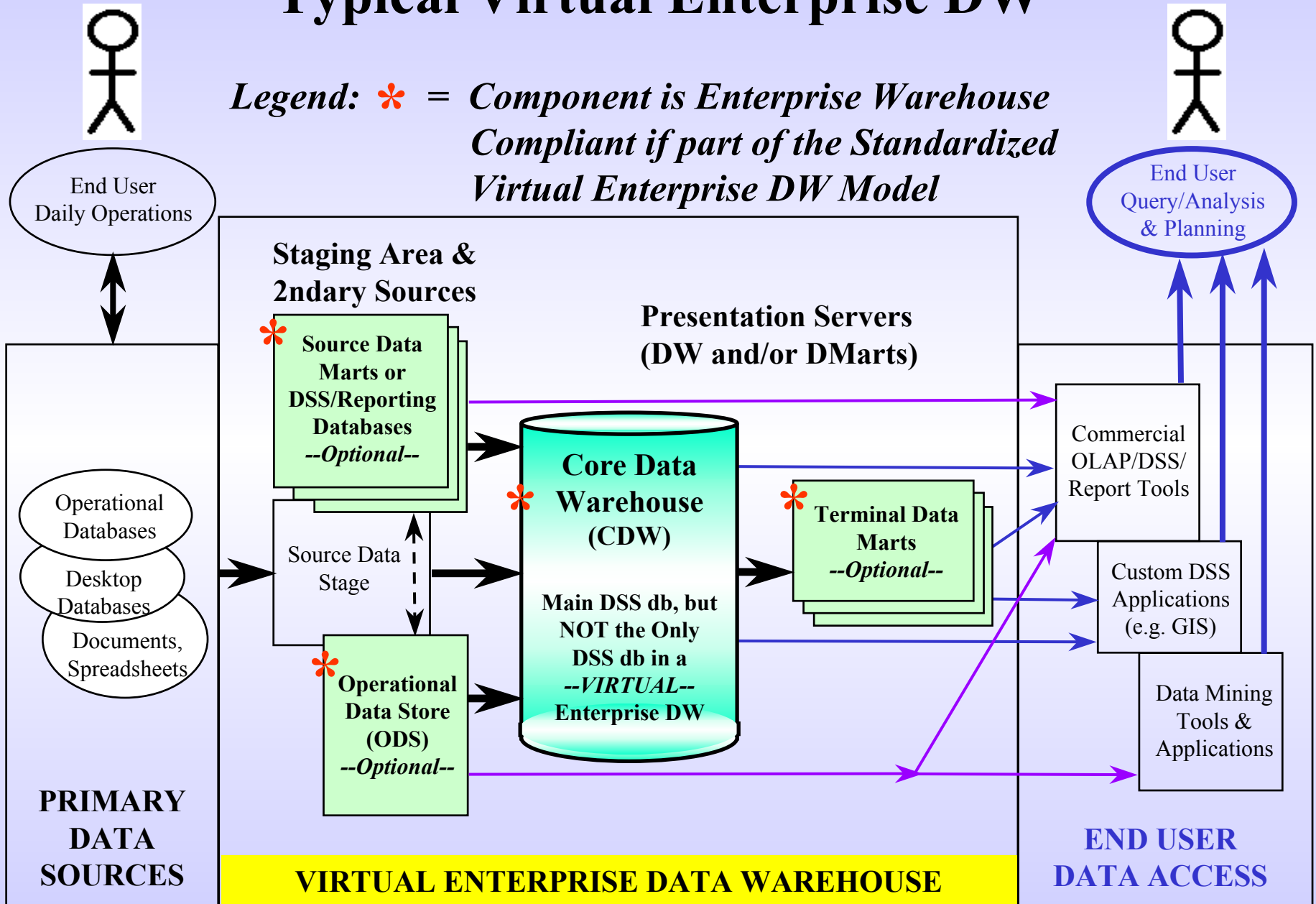
What are We Building? Another Classic View... Follow the Yellow Brick Road :-)

*Legend: * = Subject to alternative design or architecture than shown here.*



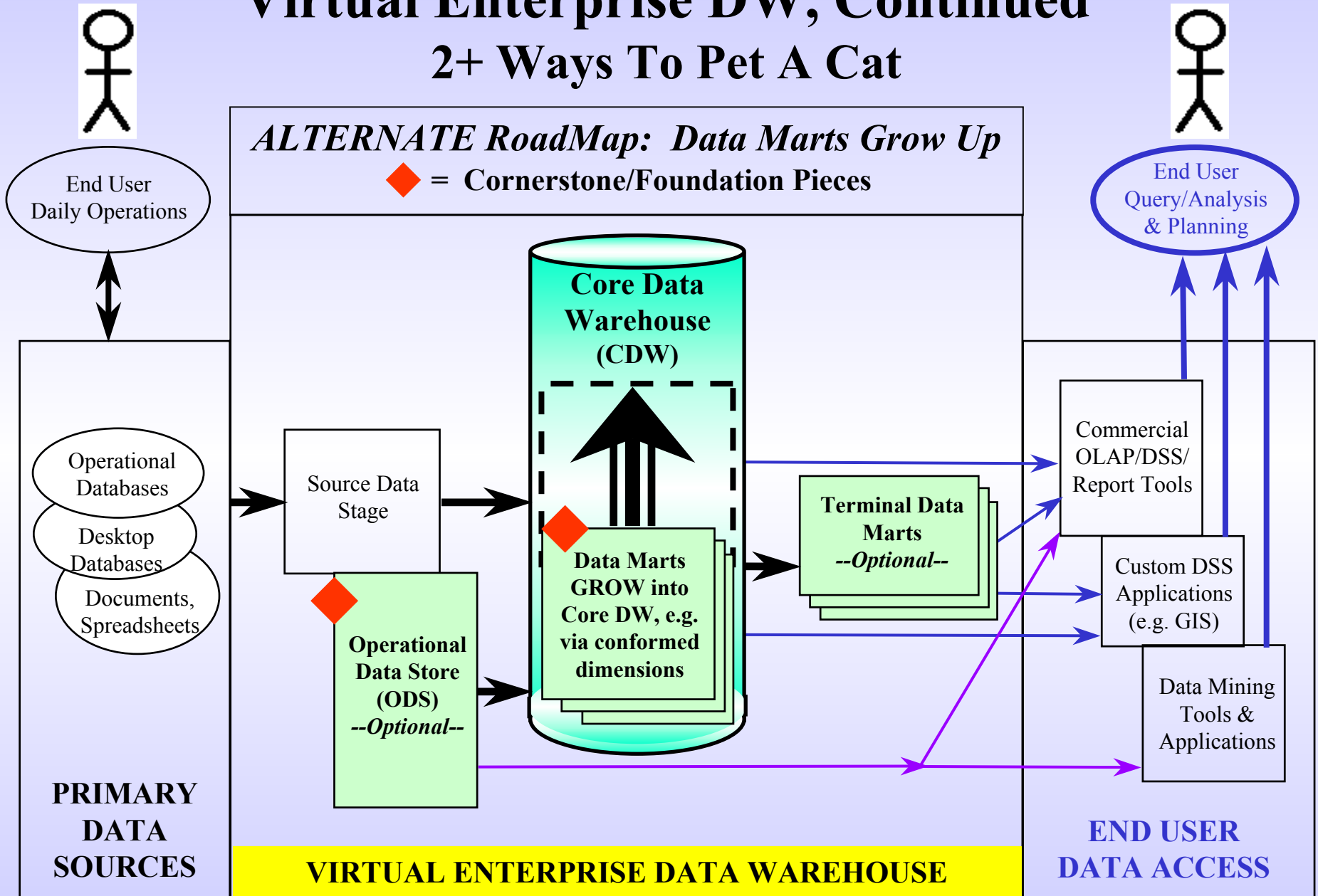
Typical Virtual Enterprise DW

*Legend: * = Component is Enterprise Warehouse Compliant if part of the Standardized Virtual Enterprise DW Model*



Virtual Enterprise DW, Continued

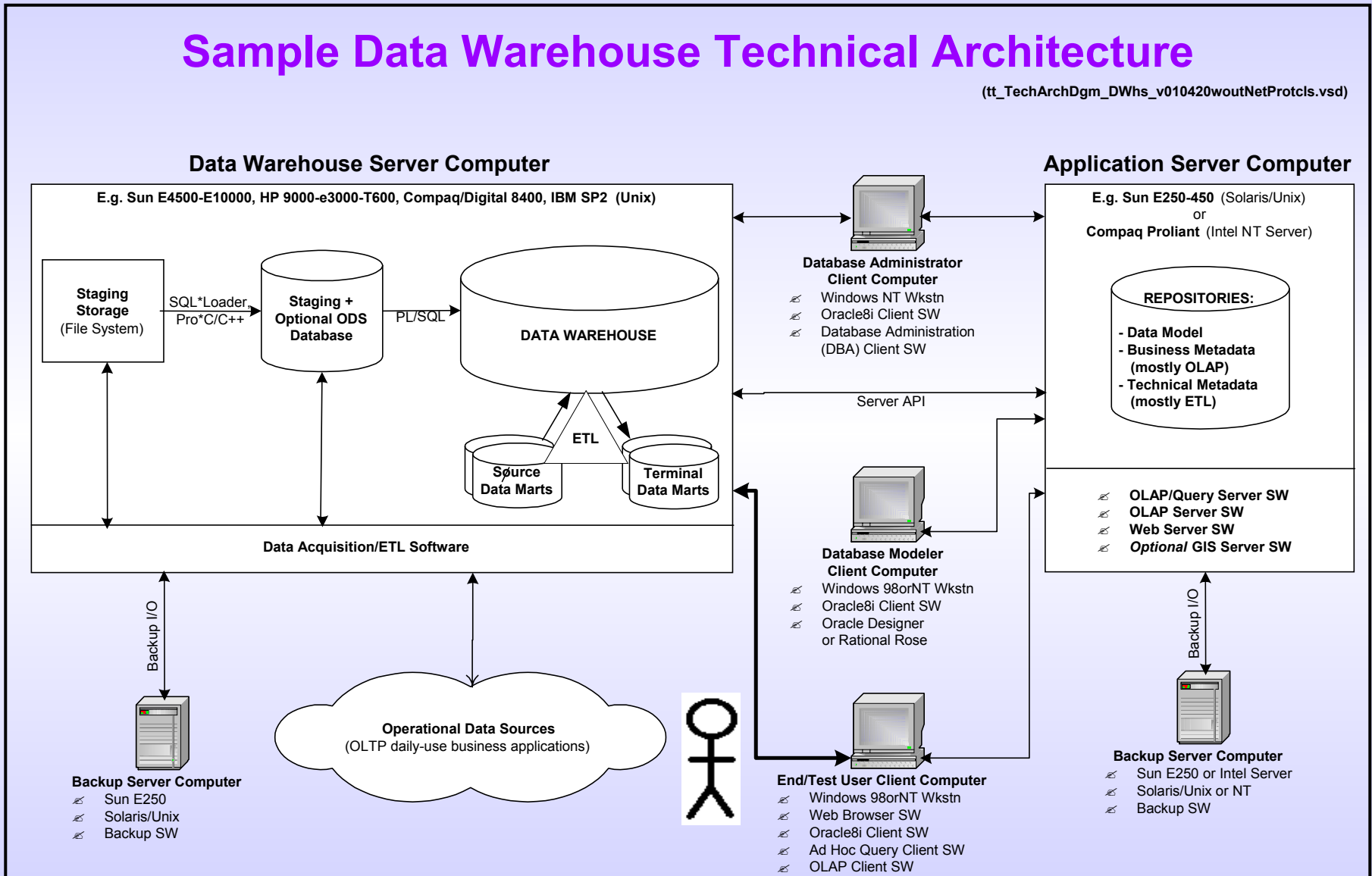
2+ Ways To Pet A Cat



What are We Building? One Tech-head Picture...

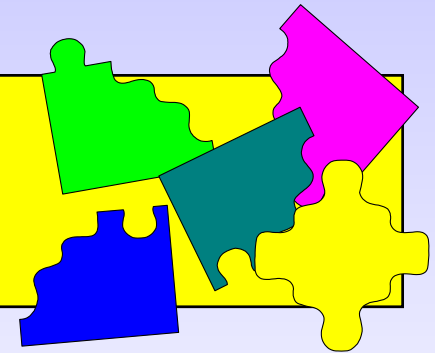
Sample Data Warehouse Technical Architecture

(tt_TechArchDgm_DWhs_v010420woutNetProtcls.vsd)



What are We Building? ...

1. Data Extraction from Data Sources



- Operational databases where the source data are stored -- typically used for daily business
- Diverse and uncoordinated
 - Different platforms in many locations
 - Multiple file formats and data types
 - Gaps and overlaps
 - Minimal control over [global] content and format

What are We Building? ...

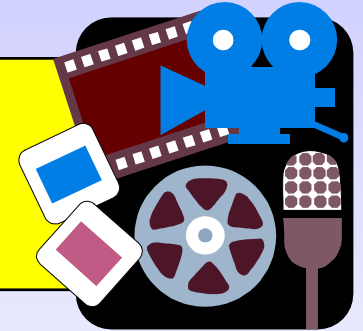
2. Staging Area for Data Cleansing/Transformation



- Interim location for storing data extracted from Data Sources
- A place for “cleansing” the data prior to storage in the Presentation Servers
 - Clean, prune, combine, reconcile duplicates
 - Translate, standardize
- A back room operation
 - NOT a place for “public” accessibility

What are We Building? ...

3. Presentation Platform (Production Environment)



- Cleansed / “Presentation” Data are moved from the Data Staging Area to one or more servers designed for access by decision makers and others
- Presentation servers are:
 - Subject oriented
 - User Community driven
 - For Data Marts: Locally implemented

3. Presentation Servers

a. Data Mart

- An organized subset of subject oriented data within the Presentation Server
 - Typically centered around one (or a few) business processes within a specific user group
(e.g., agricultural events, research findings, project expenses, etc.)
- Contain complete “pie-wedges” of the *DW*
 - In REAL WORLD *sometimes* separate little pies, but all “conform”
- Data Marts are organized, and consequently integrated, through entity-relation (ER) or dimensional modeling techniques

3. Presentation Servers

b. Data Organization

- The Overall DW and Data Marts are organized and integrated through modeling techniques:
 - Entity-Relation (ER) Modeling
 - Dimensional Modeling
- Even External Textual Data is Indexed via Data Model / DBMS

3. Presentation Servers

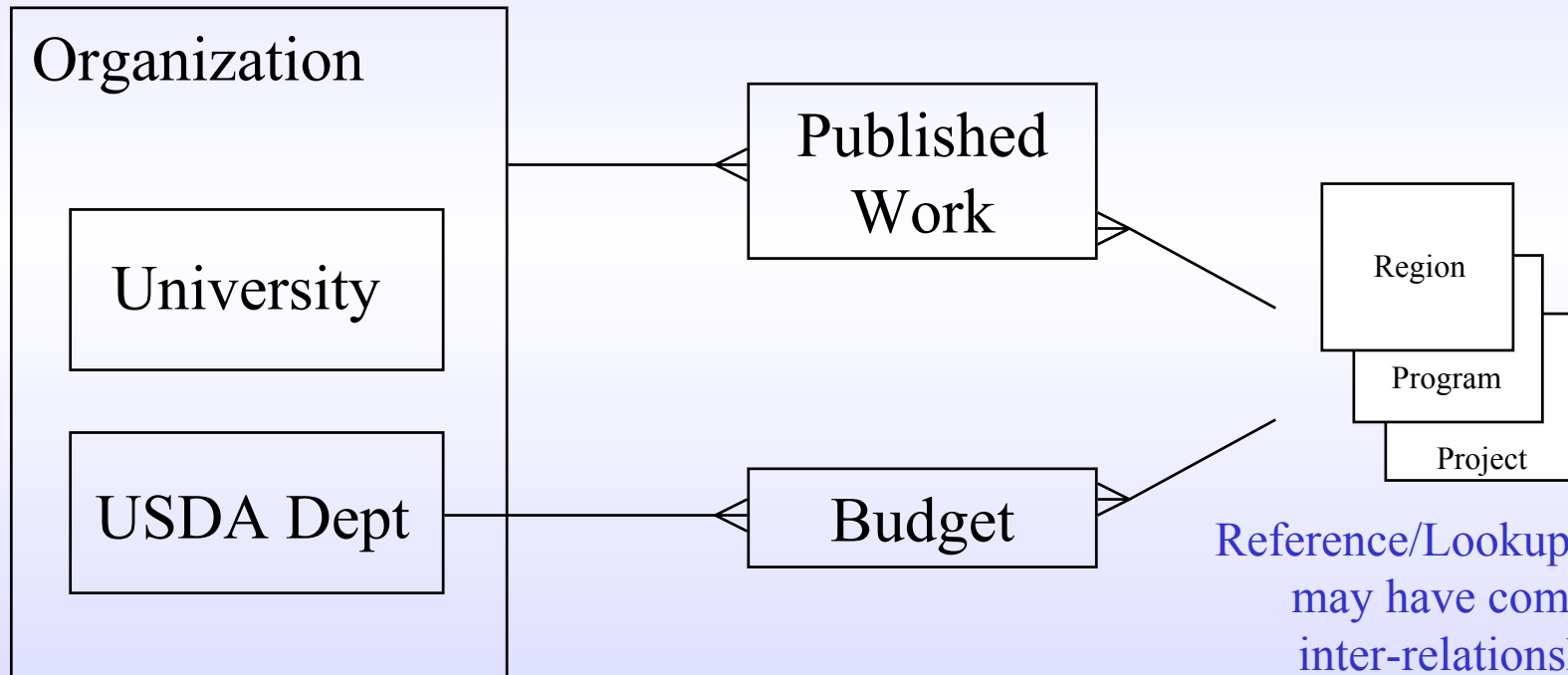
c. ER Data Modeling

- Primary Characteristics
 - Handle Daily Activity / Operations
(record every transaction)
 - Reduce Redundancy -- “Normalize”
(but not necessarily good for performance)
 - Enforce Data “Integrity”
(but doesn’t always happen, especially in non-RDBMSs)
 - Interact with Other Operational Systems
 - Provide Audit Trail

Optional
Detail

3. Presentation Servers

d. ER Diagram Example



Core Entities are often
“subtyped” to reduce redundancy

Reference/Lookup Entities
may have complex
inter-relationships
(not shown here)

3. Presentation Servers

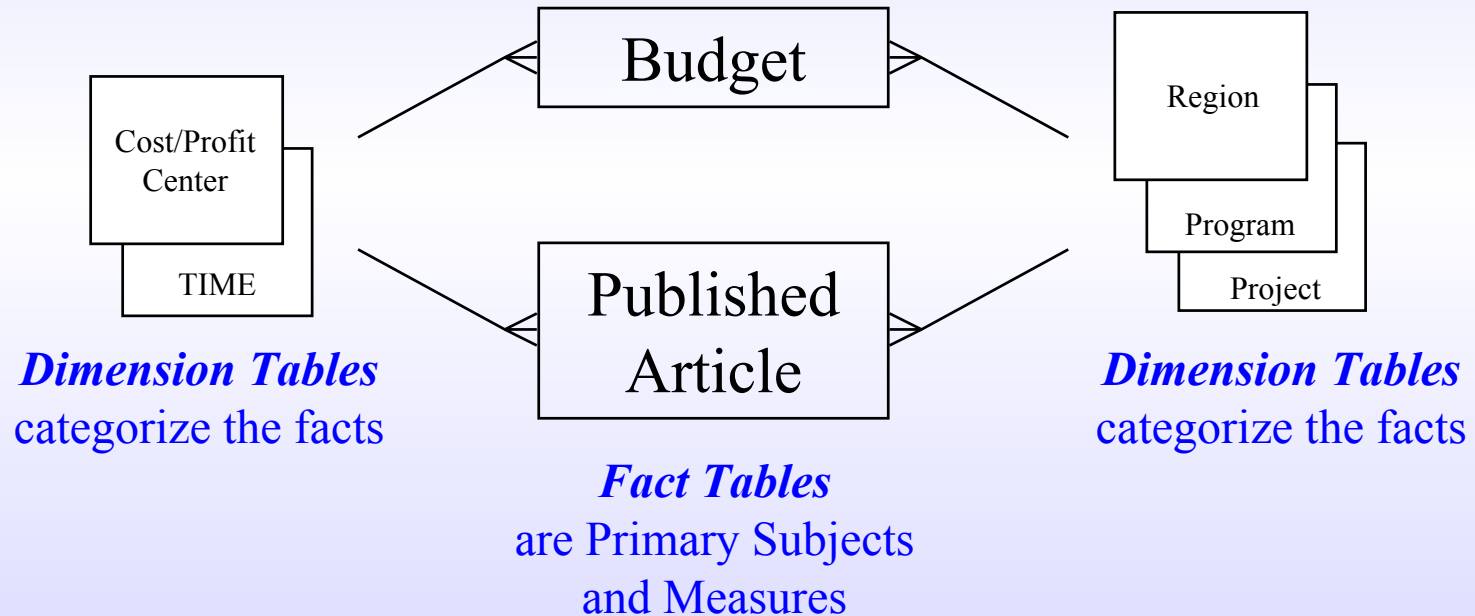
e. Dimensional Data Modeling

- Contains same information as ER model but organized “symmetrically” for
 - User understandability
 - Query performance (must handle millions of records)
 - Resilience to change
 - Simplicity
- Main components of a dimensional model are *Fact Tables* and *Dimension Tables*

**Optional
Detail**

3. Presentation Servers

b. Dimensional Modeling Example



This example contains TWO “STARS”

3. Presentation Servers

g. Fact and Dimension Tables

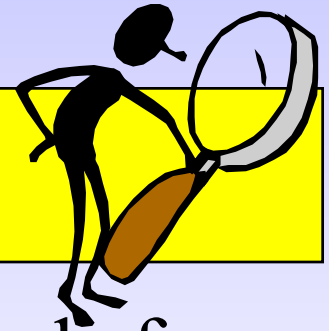
- A ***Fact Table*** contains prime organizational measures, are primarily numeric and additive, and are tied to ***Dimension Tables*** via keys
- A ***Dimension Table*** is a collection of text-like attributes about the ***Facts*** (e.g., geographic region, cost/profit center, demographics, marketing campaign/promotion, TIME)

3. Presentation Servers

h. Conformed Dimensions

- A key concept of a successful dimensional data model.
- Enables individual Data Marts to be built without requiring the entire set of all data to pre-exist in the DW.
- Meaning: The dimensions of the data mean the same thing across all Data Marts.

What are We Building? ...



4. End User Data Access

- Provide users with a diverse set of tools for accessing, manipulating, and reporting on data in the DW and Data Marts.
- Tools range from simple reporting and ad hoc query tools to sophisticated data analysis, mining, forecasting, and behavior modeling/scoring tools.
- Funny Marketing Example (For Parents)
- Data Discrepancy Analysis / Auditing

4. End User Data Access

a. Ad Hoc Query Tools

- Provide users with a powerful means of querying data in very flexible ways to meet specific information needs.
- Under-the-Hood query language is usually SQL.
- Can be effectively used by only 10% to 20% of end users (according to Kimball).
- Require techies to create End User “Layer” to facilitate query formulation.

4. End User Data Access

b. Behavior Modeling Applications

- Provides users with sophisticated analytical tools
 - Forecasting models
 - Behavior scoring models
 - Cost Allocation models
 - Data Mining tools
 - Homegrown models
- Have the power to transform the DW data

4. End User Data Access

c. Metadata

- Any data that are NOT the DATA ITSELF (effectively “DATA about DATA”).
- Technical / Control Metadata:
 - Descriptions of the data sources (e.g., as in a Database Catalog)
 - Information pertaining to the transfer of data from the source databases to the data staging area
- User / Business Metadata:
 - End User “Layer” to facilitate querying
 - Data Element Descriptions, Synonyms, Aliases
 - “Pre-Joins,” Summaries, Aggregations, Version Info

4. End User Data Access

d. OLAP, MOLAP, ROLAP, or Roloids?

- **OLAP -- Online Analytical Processing:**

- “The general activity of querying and presenting text and number data from data warehouses, as well as a specifically dimensional style of querying and presenting that is exemplified by a number of ‘OLAP vendors.’”
- Generally based on Multidimensional Cube of data, often involving MDDBs.

- **ROLAP -- Relational OLAP:**

- “A set of user interfaces and applications that give a relational database a dimensional flavor.”

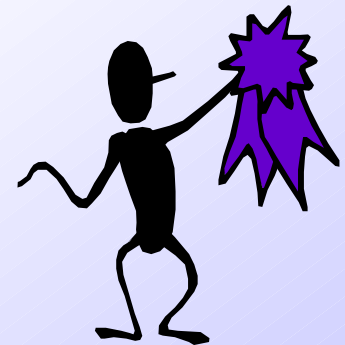
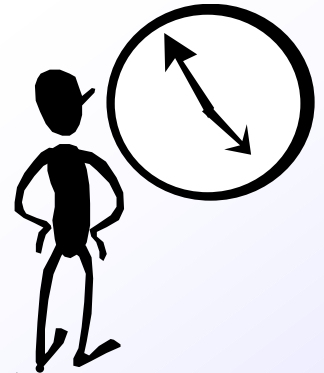
- **MOLAP -- Multidimensional OLAP:**

- “A set of user interfaces, applications, and proprietary database technologies that have a strongly dimensional flavor.”

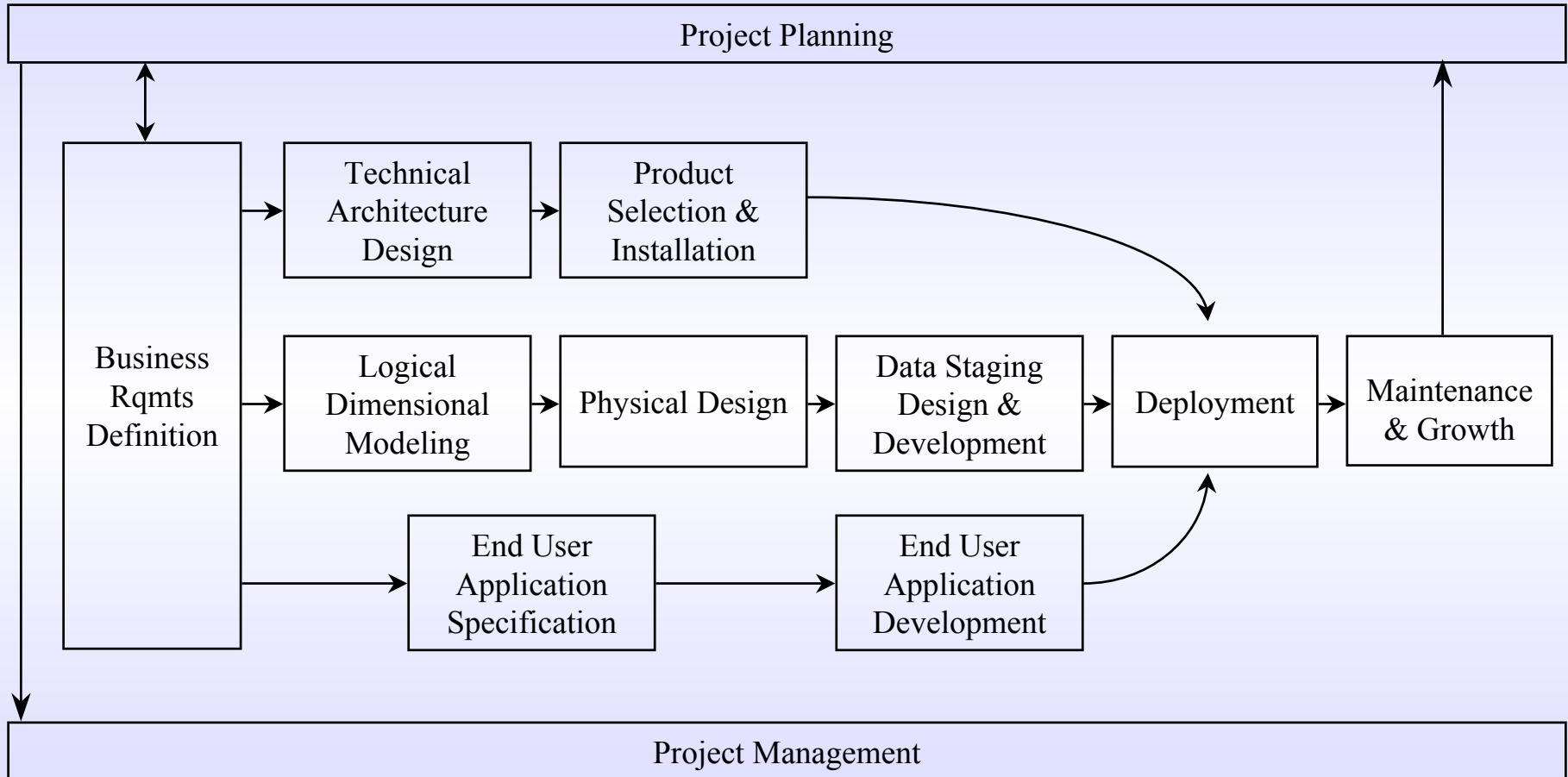
- (Quotes from *The Data Warehouse Lifecycle Toolkit*, Ralph Kimball)

5. DSS-Specific Project Lifecycle

- DW Project Lifecycle is an iterative, continuous process
- Orderly set of steps based on cumulative experience of those who have built hundreds of DWs
- Identifies roles and responsibilities throughout the lifecycle



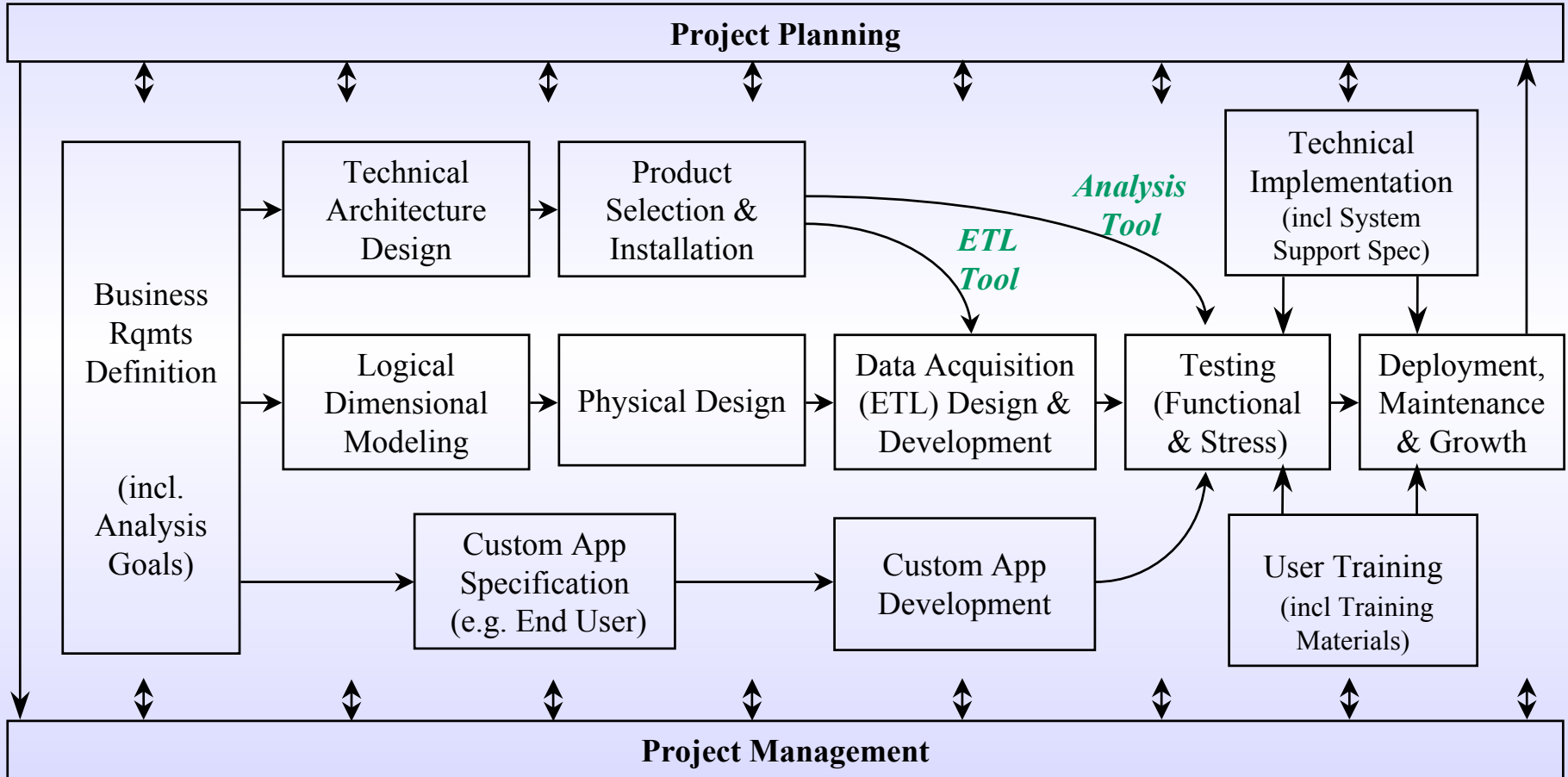
Business Dimensional Project Lifecycle



SLIGHTLY modified excerpt from *The Data Warehouse Lifecycle Toolkit*, by Ralph Kimball
(format changes only -- content/flow not modified)

Business Dimensional Project Lifecycle

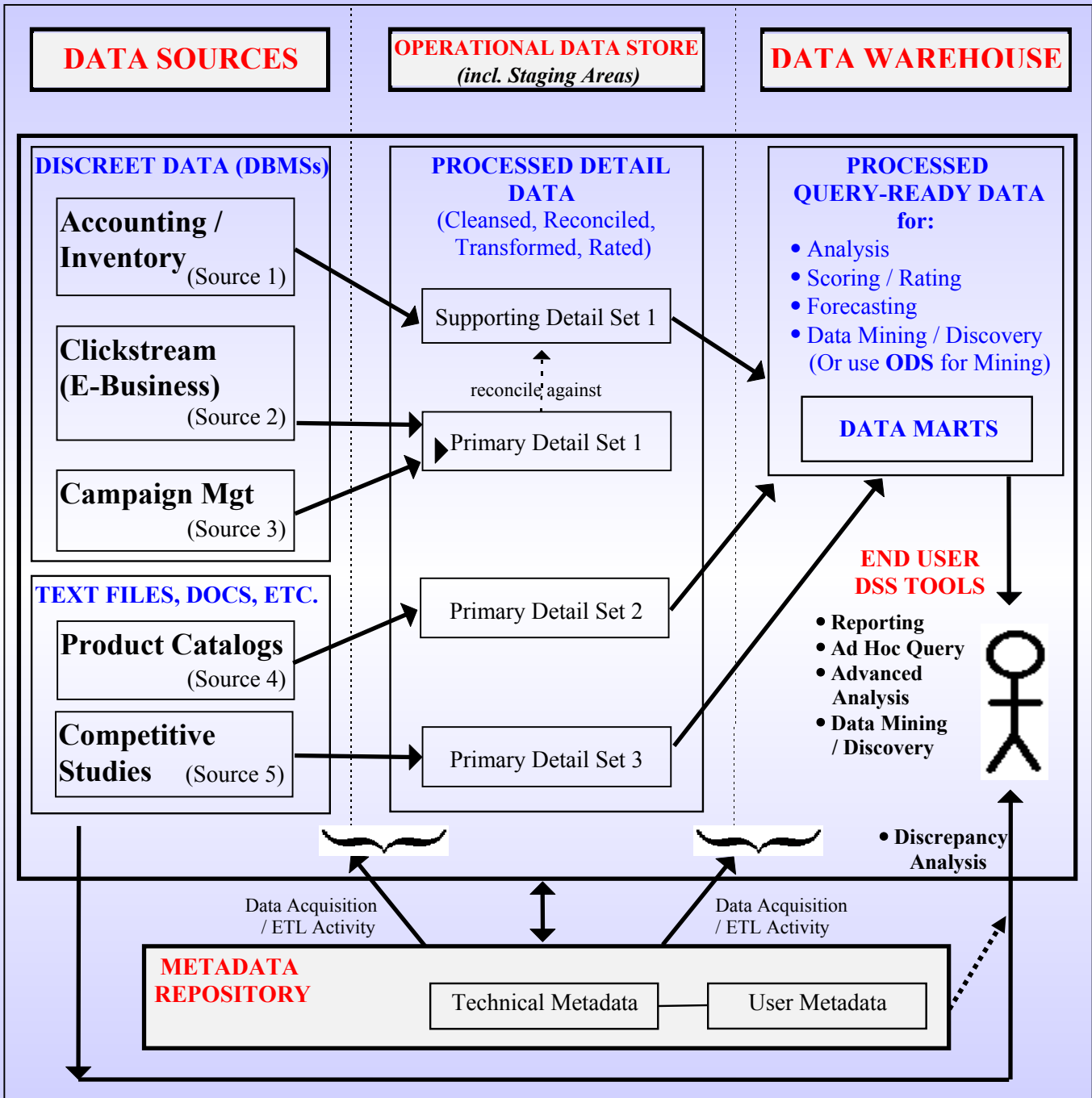
-- with some enhancements / extra detail --



Customized excerpt from *The Data Warehouse Lifecycle Toolkit*, by Ralph Kimball

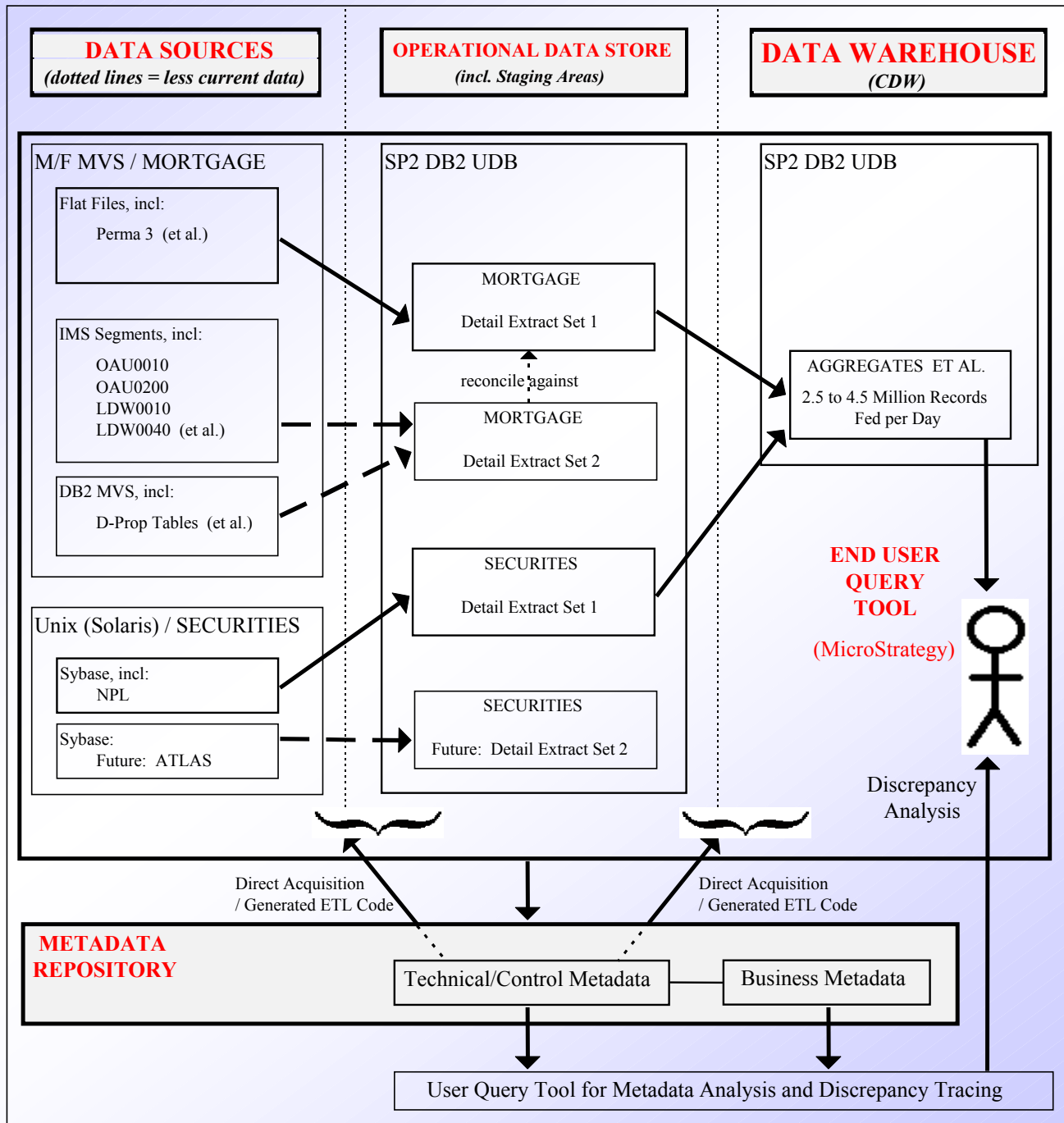
POPULATING THE DATA WAREHOUSE:

60 to 85% OF THE WORK!

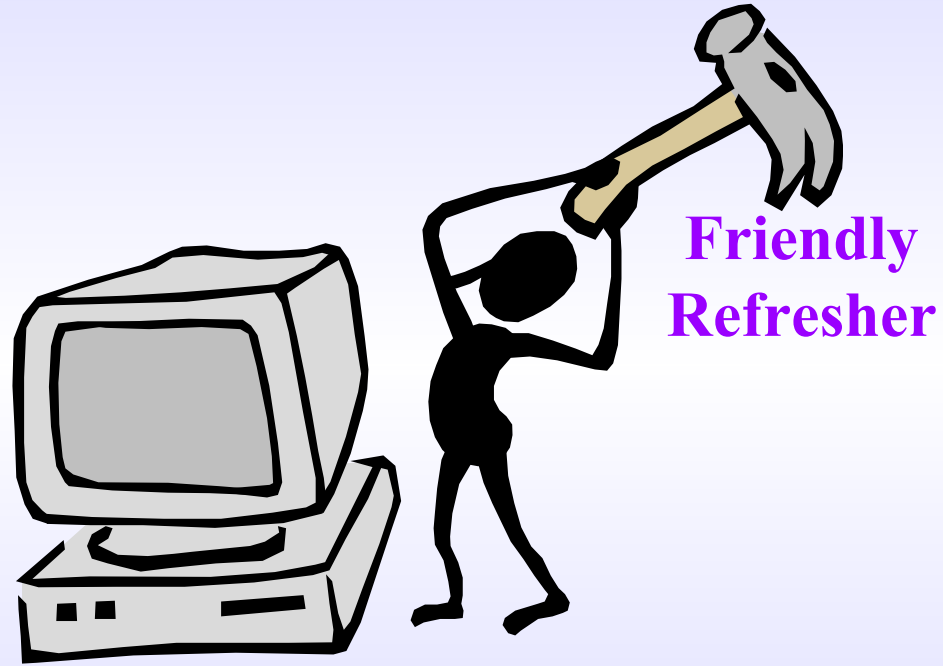


POPULATING THE DATA WAREHOUSE:

REAL WORLD EXAMPLE



WAKE UP !!!



It's STORY Time ...

Data Warehousing vs Decision Support/DSS vs Business Intelligence

- **DSS** is BROAD — Decision Enablement...
 - At Any Level
 - Via Any IT Means
- **Data Warehousing** implies more than its Formal Definition (sometimes synonymous with DSS)
- **Business Intelligence**
 - Focus on User Level Analysis, Discovery/Mining vs other framework, e.g. ETL
 - Primary Measures: Revenue, Profit, Market Share, Retention/Loyalty
- **Additional DW Goals**
 - 2ndary Measures: Quality, Fraud Detection, Etc...

e-Business Twists (or Twisty Bread?)

- Leverage e-Commerce, e-Marketing, e-Business
- First the **Conventional BI** Recipe:
 - DATA SOURCES: Subscription Data, Consumer Demographics, Special Interests, Spending History
 - PROFILING: Define *Typical* Attributes/Patterns (Typical Buyer, Loyal Customer, Churn Characteristics)
- Then Sprinkle in some **B2C Style *e***:
 - Fill the Abandoned Shopping Cart
 - Clickstream Analysis
 - Personalization



More *e*-Business Twists

- And a Dash of **B2B** Style *e*:
 - Foster vendor partnerships, not just competition: eRFP
 - Supply Chain Intelligence
SCI = Supply Chain Mgt (SCM) + Business Intelligence (BI)
 - ➔ Lower Procurement Costs
 - ➔ Improved Coverage (products, geography)
 - ➔ Less Faulty Inventory (higher quality)
 - ➔ Faster Lead Times
(Inventory Optimization, Less \$ Investment)

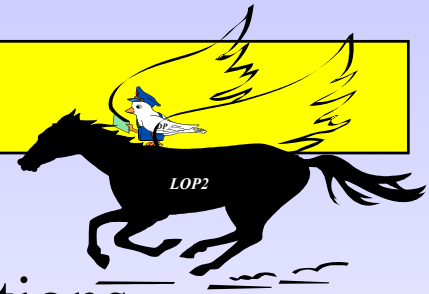


Chicken or Egg Syndrome: Which comes 1st — The Data Warehouse or Data Mart?

- Utopia
- Walk Before you Run
- Cost Benefit – Timing is Everything :-)
- Executive Support and User Confidence
- The Key is in the **PLAN** (our friend LOP2)...
Conforming Dimensions and Other Standards

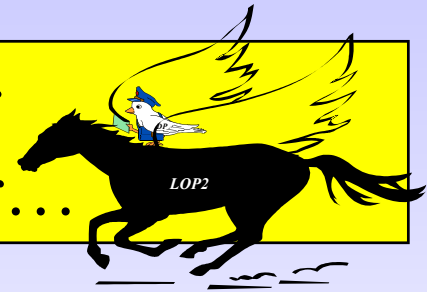


Scope or Listerine?



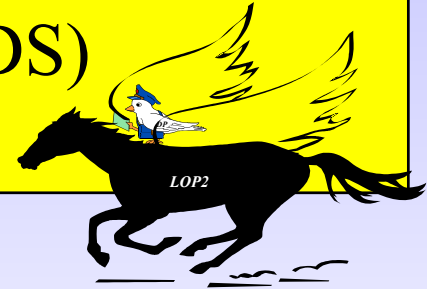
- DW/DMs have **Short Lifecycle** and Many Iterations
- Choose the **Grain** and an **Initial/Pilot Subject Area**
- Define Specific **VALUABLE** Analysis Goals (Prioritize)
- Validate the Goals with **Real Sample Queries**
- REMIND People the Initial Subject Area/Goals are Valuable
- **Prototype** the Sample Queries — Everyone's Baby
- **Leverage the ODS** for Analysis Requirements involving
 - Extra Dimensions which would make Summary Facts too Large
 - Data Mining

Objects In Mirror May Be Closer (and **BIGGER**) Than They Appear...



- Summary Facts (Aggregation Tables) do **NOT** Mean Less Rows!!!
- 1 Row * Dim1(# PossibleValues) * Dim2 (# PossibleValues) Etc...
 - 12 Months * 20,000 Customers * 50 Mktg Campaigns
* 10 Product Categories * 5000 Cities * 30 Salespeople
 - Total Rows/Year = 18,000,000,000,000 (**18 Trillion**)
- B2C Generally Larger than B2B
 - #Customers and ZIP5, e.g. if need to Generate **Mail Lists**
- Use **Categories/Classes** Whenever Possible or
Consider Splitting Into **Separate Stars** with Less Dimensions

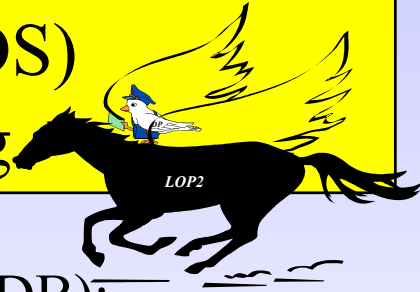
DW and Operational Data Store (ODS) Back to The Chicken or Egg



- If Build ODS **BEFORE** DW:
 - Leverage it for **Consolidated/Cleansed** Data Staging
 - ➔ Simplifies/Facilitates ETL for DW and all DMs
 - ➔ Can Optionally Centralize and Bring Source of Record Closer
 - Leverage it as **Alternative to Large (Uncategorized) Dimensions** in DW with Aggregated Granularity
 - ➔ Faster Performance for MOST DW Queries
 - Leverage it to **KEEP the DW GRAIN Higher**
 - ➔ Faster Performance for MOST DW Queries
 - Access **Transactional Data Quickly — Immediate Value**
 - ➔ Offload OLTP Envmt of Reporting Stress & Limited Batch Windows
 - ➔ Can Optionally Deliver **DATA MINING** Support Sooner
- If Build ODS **AFTER** DW (Discussion???):
 - Deliver **AGGREGATE** Results Sooner
(although if Build ODS BEFORE DW, can focus on DM before DW)

DW and Operational Data Store (ODS)

Tag Team — In or Out of the Ring



- If ODS & DW **CO-LOCATED** (on Same Machine/DB):
 - Easier for OLAP Tools to **Drill Down** to Transactional Detail
 - ➔ Improved User Functionality
 - ➔ Note Ralph Kimball’s Term “Operational Data *Warehouse*”
 - **Less Network Traffic** for Drill Down from DW/DMs to ODS
 - ➔ Faster User Performance
 - **Less Network Traffic** for DW Data Loads
 - ➔ Better Performance for MOST Queries
- If ODS & DW Reside on **SEPARATE Machines/DBs**:
 - Easier to Manage Capacity, Stress, and to **Tune DIFFERENTLY**
 - Discussion (???)

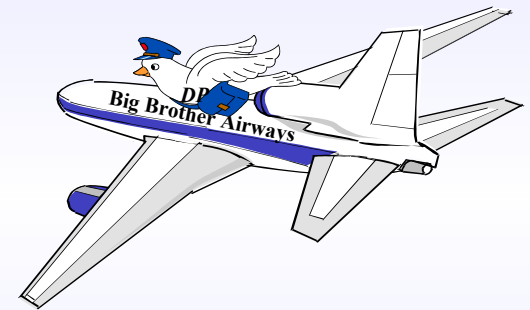
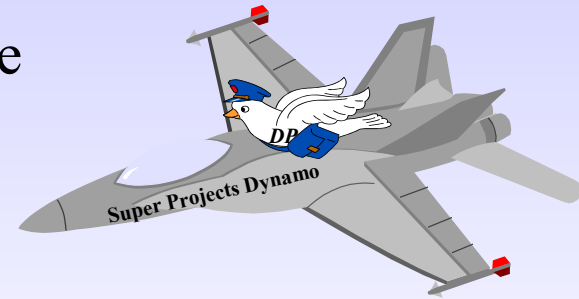
Reading, Writing, and Arithmetic



- Read-Only, Write — Right?????????
- **Conventional Batch ETL**
 - Off-Peak (night time)
 - Mostly Inserts + Some Updates for Dimensions and Retro-Aggregates
- **Recent Trend is ONGOING ETL** — Keeping Up with the Times
 - 7x24 Trickle or Real Time
 - Change Data Capture (CDC)
 - Message Queuing, Workflow or Async Replication (AVOID Synch/2PC Triggers)
- **Additional Kinds of Ongoing Writes**
 - Decision Support Workflow — Follow-up to Information PUSH (INSERTS/UPDATES to DSS Workflow Logs — possibly in ODS for analysis purposes)
 - Mass Mailings (INSERTS/UPDATES to Campaign Logs — typically in ODS)

Tools of the Trade: Can we Measure in \$ per Mile?

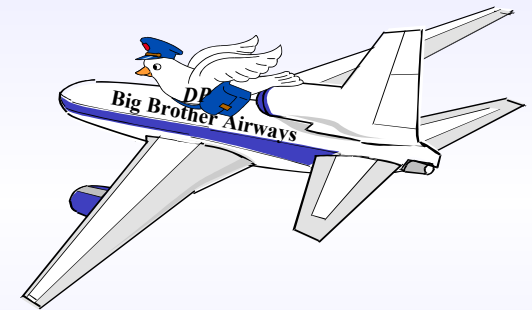
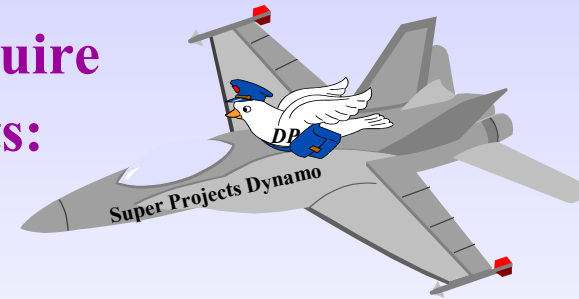
- **OLAP/Query/Reporting** Tools are NOT the Same
 - Some Differences, Should Match To Requirements
 - Usually a Must-Have
- **Data Acquisition/ETL** Tools are NOT the Same
 - **HUGE** Functional and \$ Differences
- **Data Quality Analysis and Data Generation** Tools
 - Significant Functional and \$ Differences
 - May Not Need (vs Generic Tools and Custom Extraction or Propagation)
- **DBA and CASE/Database Design** Tools
 - For Capacity Planning/Sizing/Forecasting, Space Mgt, Performance Tuning (OS, DB, App Servers, and App SQL), Diagnostics/Monitoring, Modeling, DDL Generation
 - Some Differences, Should Match To Requirements
 - Prefer to Have



Other Tools of the Trade

→ Following (optional) Tools Still USUALLY Require More Formal Evaluations against Requirements:

- E-Commerce
 - E.g. BroadVision, Blue Martini, et al
 - Significant Functional and \$ Differences
- Business Rules Engine
 - E.g. Blaze, JRules, et al
 - Significant Functional and \$ Differences
- Data Mining
 - E.g. Darwin plus products Bundled with E-Commerce
 - Significant Functional and \$ Differences



MetaData — Data about WHAT for WHAT?



- Often Easy to Create but **NOT** Maintain
- Many Products Advertise Ability to Analyze MetaData but Extremely **Awkward** (typically involving proprietary export)
- Many Products Support **Only 1-Direction** MetaData Interchange — THEIR Way (e.g. import only, no export)
- Standards Compliance is Underway, Should **Improve** in ~1 Year
 - XML: for ETL/EDI
 - XMI: one more step
 - Common Warehouse MetaData Interchange (CWMI):
 - ➔ Combination of Java, XML, and UML
 - OMG's Meta Object Facility (MOF):
 - ➔ Focus on INTERNAL metadata representation vs import/export format

GOTO PROJECT PLAN

DO NOT PASS GO.

DO NOT COLLECT \$2000.

1

→ Primary Focus of **Different Paper**, Thrown In Here for **Good MEASURE :-)** ...

See Project Plan embedded in Proceedings Manual article or e-mail jefflit@dbigusa.com for latest copy.

6. So NOW WHAT ?

The BIG QUESTION...



Will We SUCCEED?



Summary

*Questions...
Comments?*

Thank You!

-
- Provide a Brief DW Background/Refresher
 - Show **WHAT** we are Building
 - Show **HOW** we Build It “If you Build it, They will Come”
➔ More Details in Separate Paper: “A REAL DWhs Project Plan...”
 - Identify some Tips, Traps, and Best Practices
**Road Trip from...
MYTH MEDICINE to LABOR LEGEND**

Dispelling Myths and
Creating LEGENDS
for your E-Biz Intelligence Warehouse

--- AOTC Conference 12/9/2005 ---

Jeffrey Bertman (jefflit@dbigusa.com)
Chief Engineer

DataBase Intelligence Group (DBIG)

- IT Consulting • Strategic Planning • On-Call Support •
- Advanced Technology Implementation and Troubleshooting •

Reston Town Center
11921 Freedom Drive, Suite 550
Reston, VA 20190

WDC/VA/MD Region
(703) 405-5432
www.dbigusa.com

Bibliography

- v Ralph Kimball, *Data Warehouse Lifecycle Toolkit*, 1998
- v Ralph Kimball, *Data Warehouse Toolkit*, 1996
- v William Inmon, “What Happens When You Build the Data Mart First”, DM Review, October 1997
- v Douglas Hackney, “Picking a Data Mart Tool”, DM Review, October 1997
- v Gilbert Prine, Unisys, “Coherent Data Warehouse Initiative”, Unisys Presentation May 1998