

# Information Resource Management (IRM) Round Table

February 4, 2004

Hosted by:



# Agenda

- Introductions & Agenda 10 min
- Data Architecture & Data Quality 45 min
- Break 15 min
- Data Warehousing & ETL 45 min
- Metadata Management 35 min
- Q&A 15 min
- Wrap Up 5 min

# Data Architecture & Data Quality

Gregg Wyant

Chief Data Architect  
Intel

# Data Architecture

## Strategy / Approach:

- Enterprise Architecture based on Zachman Framework
- Centralized Enterprise Data Architecture org and establish Chief Data Architect role
- DA/DM Resources assigned to programs/projects
- Common Tools and Governance Process
- Enterprise Model

## Progress To Date:

- Data ADG running with Architecture Review and Variance Process; Sub-teams for Data and Process Model Reviews
- Data Architects/Modelers assigned to all major programs
- Data Analyst Competency Center

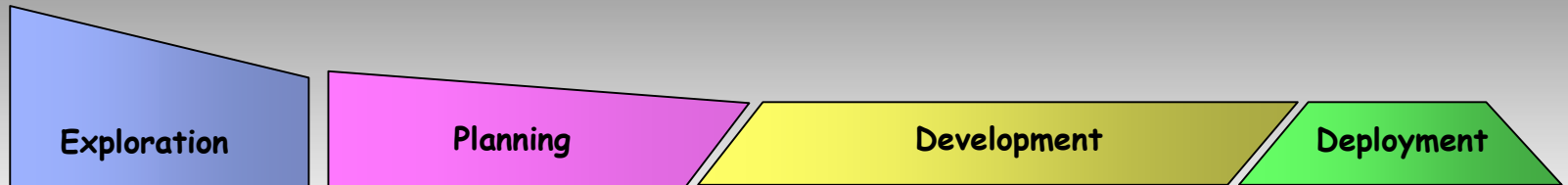
## Issues / Roadblocks:

- Demand outstripping Supply
- Measuring Architectural Impact (business value)

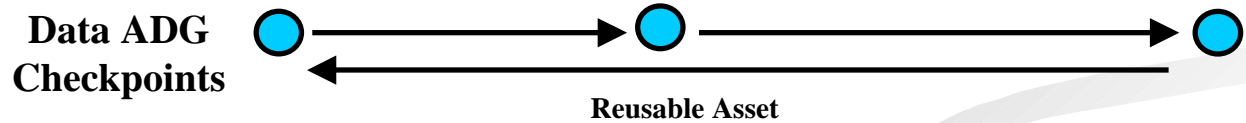
# Enterprise Data Architecture Principles

- **Single ROO and ROR** for each data element; **single business owner** for each of these data elements; **unnecessary copies of data removed** from environment
- **Common data models and data definitions** across organizations & business units
- **Single point of entry** for each data element; improve the quality of the data at point of entry; ensure its integrity throughout its lifecycle; data lifecycles documented
- **Defined ROT(s) (record of transport(s))** for each data element
- **Minimized data movement/transformation**; all data transformations are documented
- Necessary **tracking, managing and operational metadata captured** throughout data lifecycle
- All **reuse elements (code, messaging, reporting) built on compliant data**; reuse driven throughout the data lifecycle
- Data **security centrally administered**; data **access centrally controlled**
- All data (and its associated metadata) aligned to the **Enterprise Model**
- Drive to an **Enterprise Data Warehouse**; data defined consistently and accessible across organizations; **analytical data structures ready for x-org business consumption**

# Data ADG Checkpoints Example

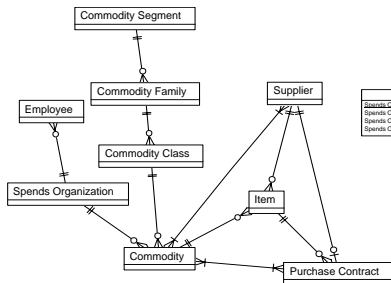


- **(Planning)**
  - Information Requirements
  - Conceptual Data Model
  - SIPOC Diagram
  - Approved Team Roster for Data Definition Acceptance
- **(Analysis)**
  - Data Requirements
  - Logical Data Model
- **(Design)**
  - Physical Models
  - Value Cost Chain Diagram
- **(Construction)**
  - Test plan/cases that meets knowledge workers data quality requirements.
- **(Test)**
  - Test plan/cases that meets knowledge workers data quality requirements.
  - Data Models have been placed into the Metadata Repository.
- **(Implementation)**
  - Data & Information training included into business process training.
  - Operational Model (Physical implementation information has been placed into the Metadata Repository.)

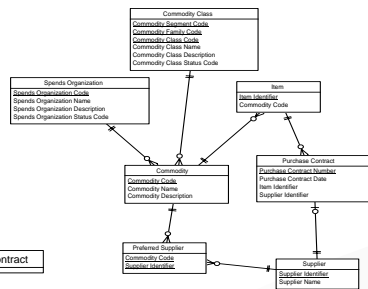


## Example Deliverables

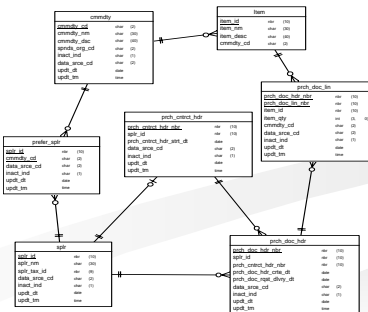
### Conceptual Data Model



### Logical Data Model



### Physical Data Model



**Critical DADG Questions Answered**

- Reuse of enterprise data
- Engineering to single ROO
- Improving data flow
- Reducing manual touchpoints
- Following data policies/procedures

# Data Quality

## Strategy / Approach:

- Utilizing Larry English TQdM process
- Data Czars, Data Pipeline Managers, TQdM Coaches

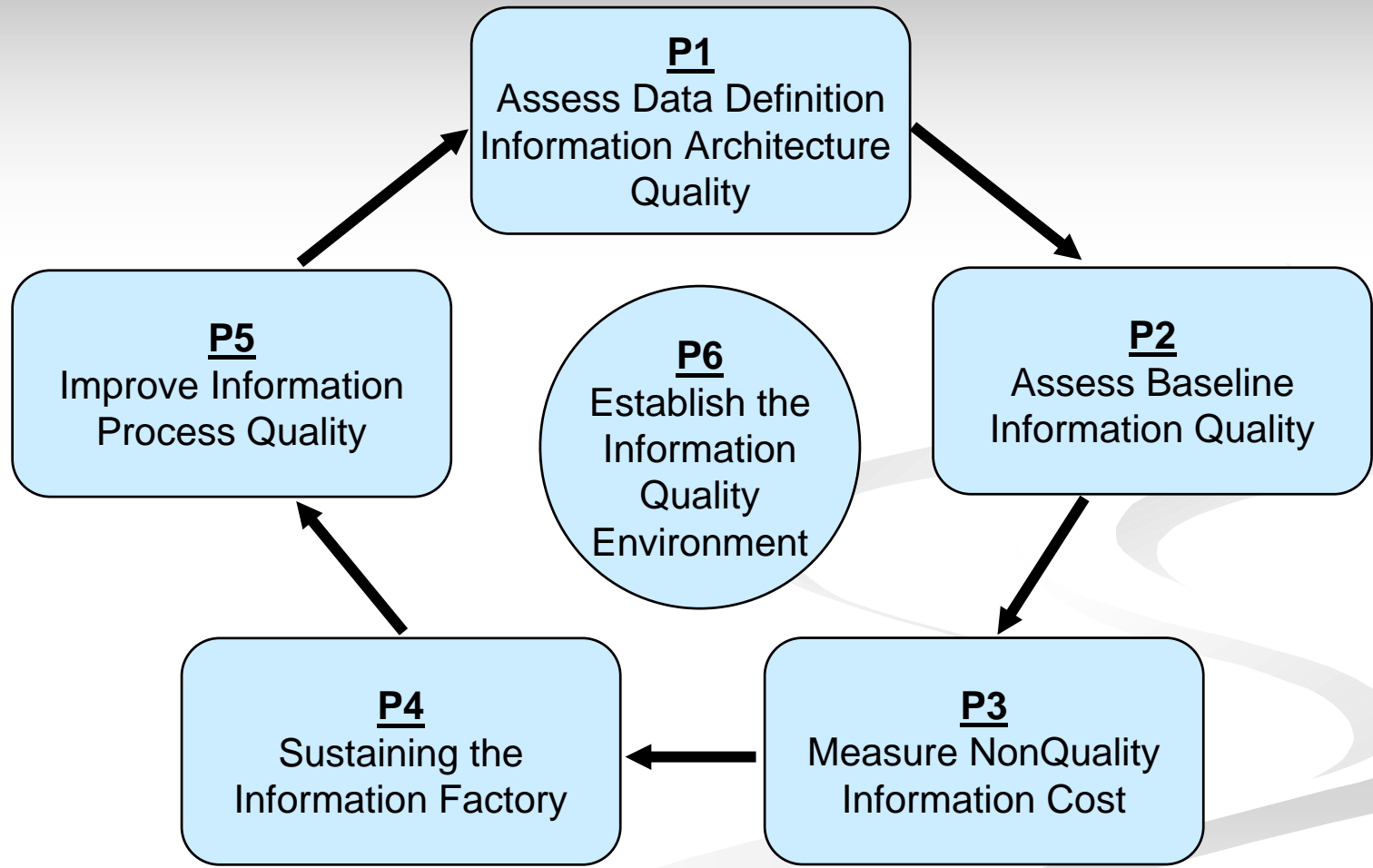
## Progress To Date:

- DQ incorporated into each major program
- Using re-engineering efforts to iteratively drive improvements
- 2002 Focus: P1-P3
- 2003 Focus: P4/P5 and monitor baselines
- 2004 Focus: Quality Monitoring in Place

## Issues / Roadblocks:

- Overhead of DQ in SDLC Processes (ExP)
- Management of “done” expectations

# TQdM Is an Iterative Process

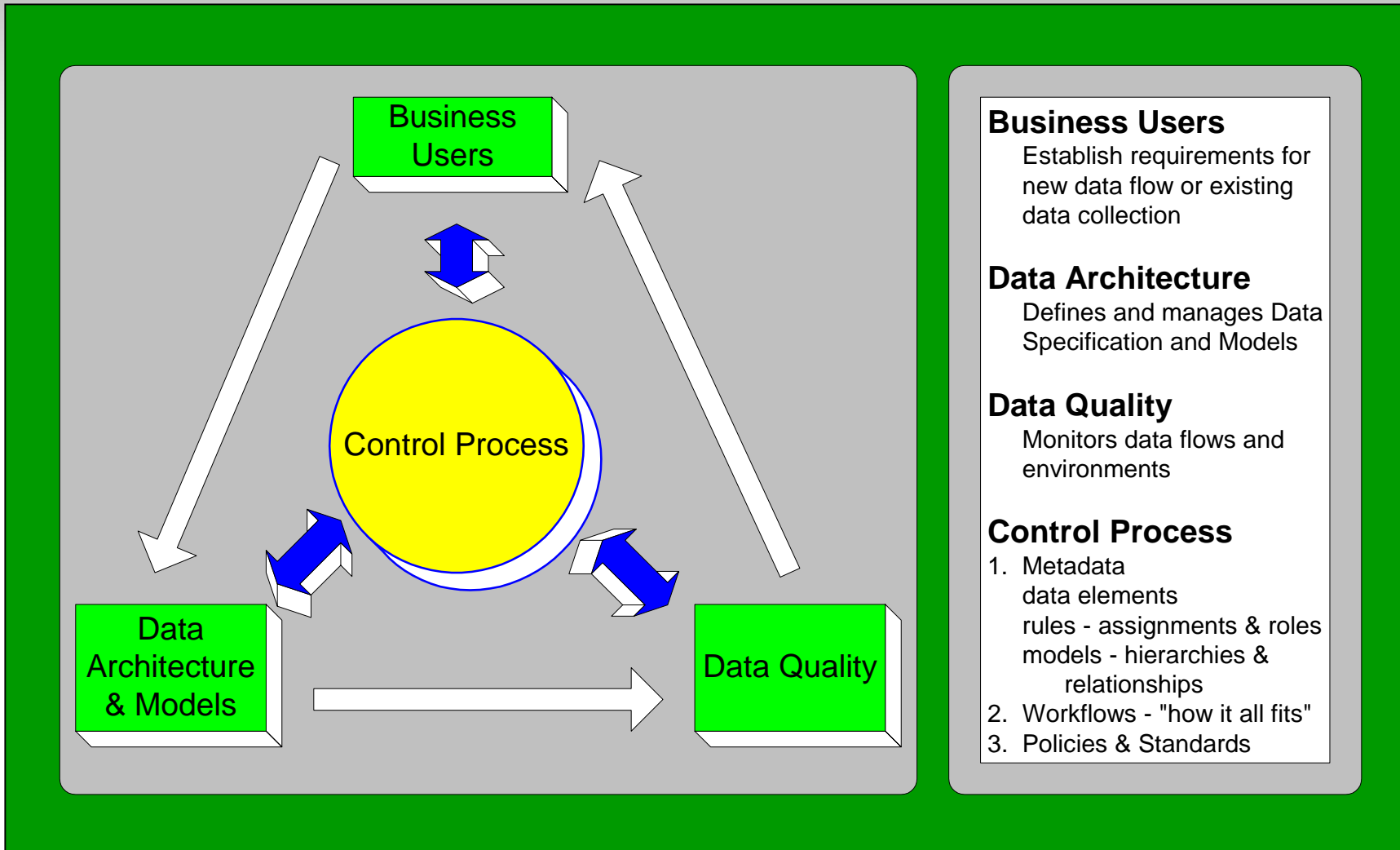




# Information Resource Round Table Data Architecture & Data Quality

Presented by: Juanita M. Mercado  
Lead Data Architect

# Data Architecture & Data Quality At Work



*Data Architecture and Data  
Quality within an Overall BI  
Architecture Framework*

Cass Squire  
Associate Partner  
IBM Business Consulting Services  
csquire@us.ibm.com  
(650) 520-7247

# Topics

- The Business Problem
- A Strategy & Approach for Solving it
- Progress To Date
- The Real World: Issues and Road Blocks

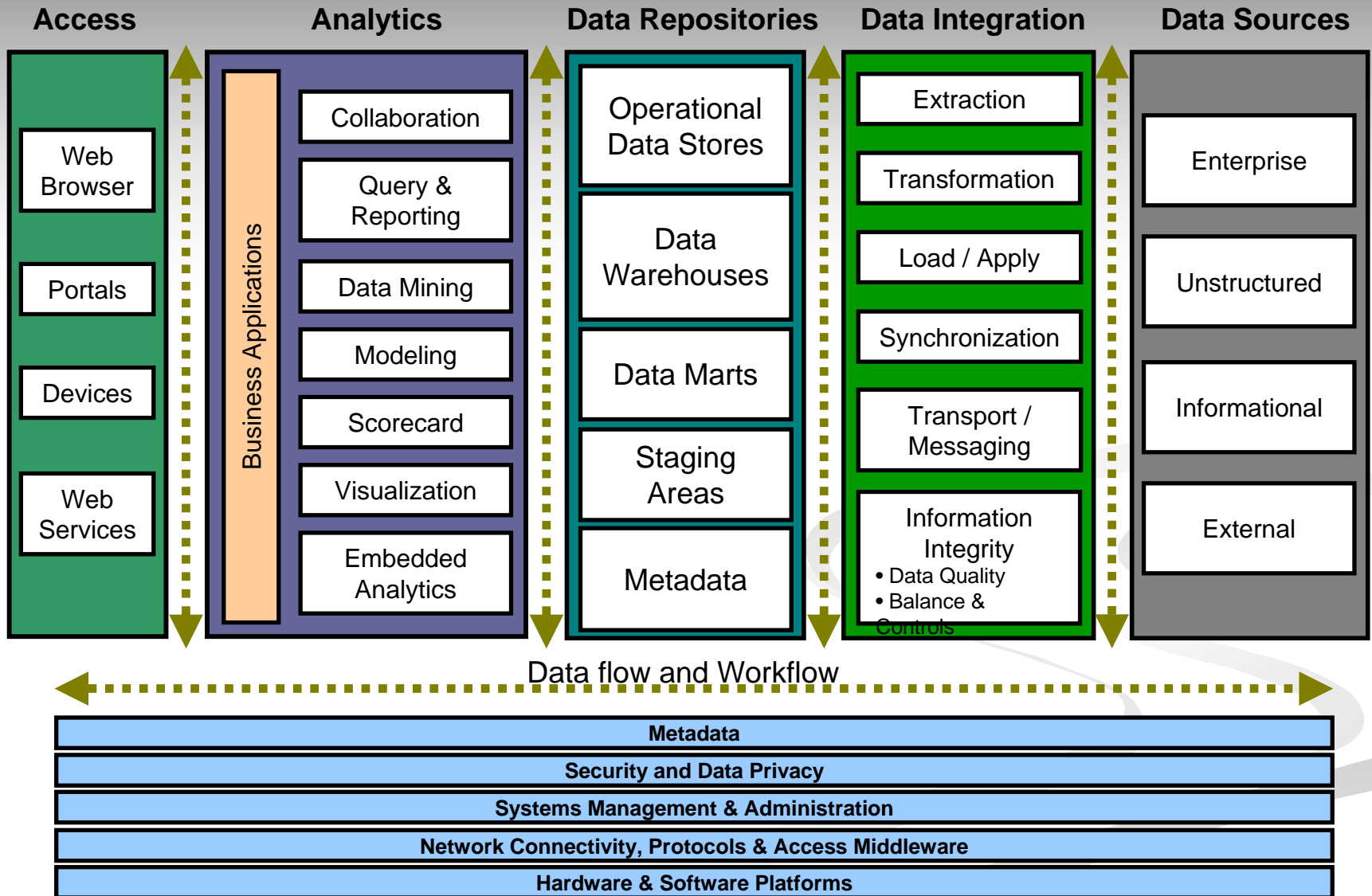
# The Business Environment

- For the past 10 years (life of the company to-date) the drive has been on acquiring new customers
- The company's direction is on being the premier provider rather than low cost
- The market (at least in the US) is pretty much saturated
- The market has changed from exclusively Dialup to increasingly higher Broadband access (DSL or Cable)
- The focus must now shift to Retention and Up sell
- All data, analytics, campaigns, etc. currently in place are geared towards acquisition
- Strategic shift to towards analytics to enable retention/upsell,... required
- This entails building a true EDW environment

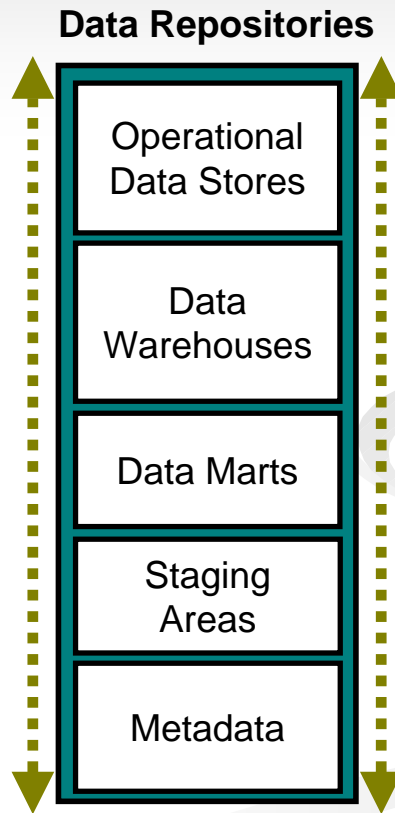
# Volumes: A Primary Frame of Reference For Why an Architecture Is Important

Amount	Statistic
20,000	Servers
3,500,000	Modems
28,000,000	Current number of customers
400,000,000	emails - <b>daily</b>
2,400,000,000	spams blocked - <b>daily</b>
4,500,000,000	Impressions served - <b>daily</b>
15,000,000,000	URLs downloaded - <b>daily</b>
3.5 Terabytes	<b>one daily</b> feed into the warehouse
35 Terabytes	Estimated EDW size

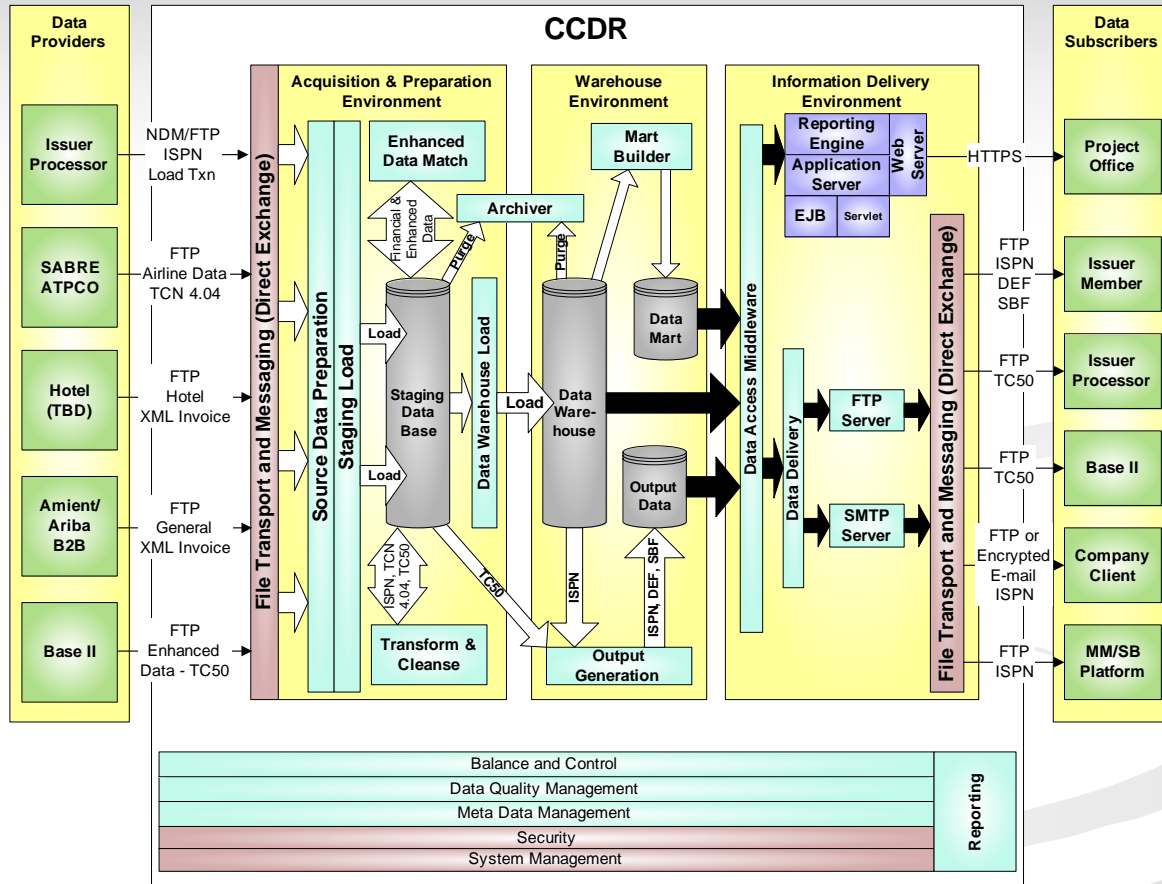
# Business Intelligence – Reference Architecture



# Picking The Right Platforms is Crucial



# Translating the Data Architecture into a Component Model



# Data Becomes Information If and Only If You:

■ **Have** the data,

and

■ **Know** you have it,

and can

■ **Access** the data,

and can

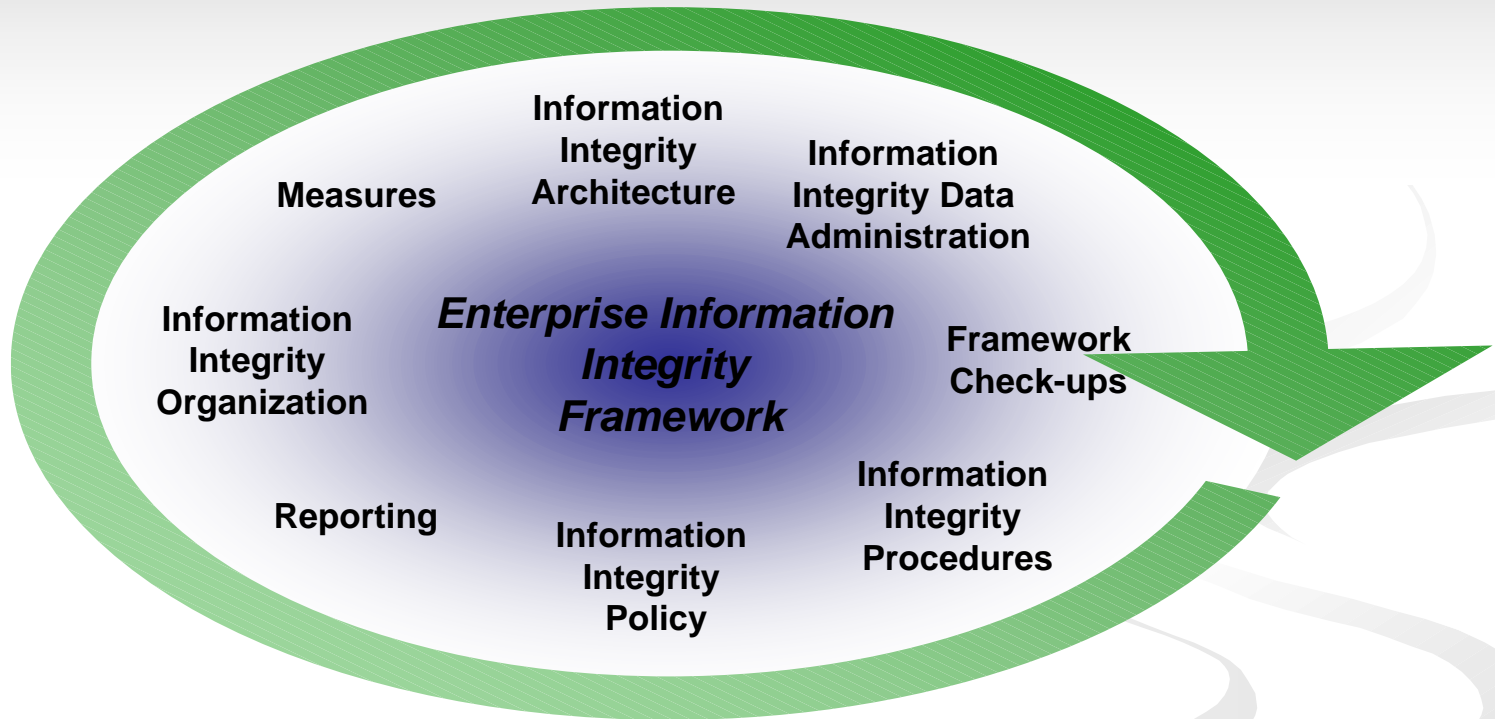
■ Can **use** the data,

and can

■ **Trust** the data

# Enterprise Information Integrity Framework

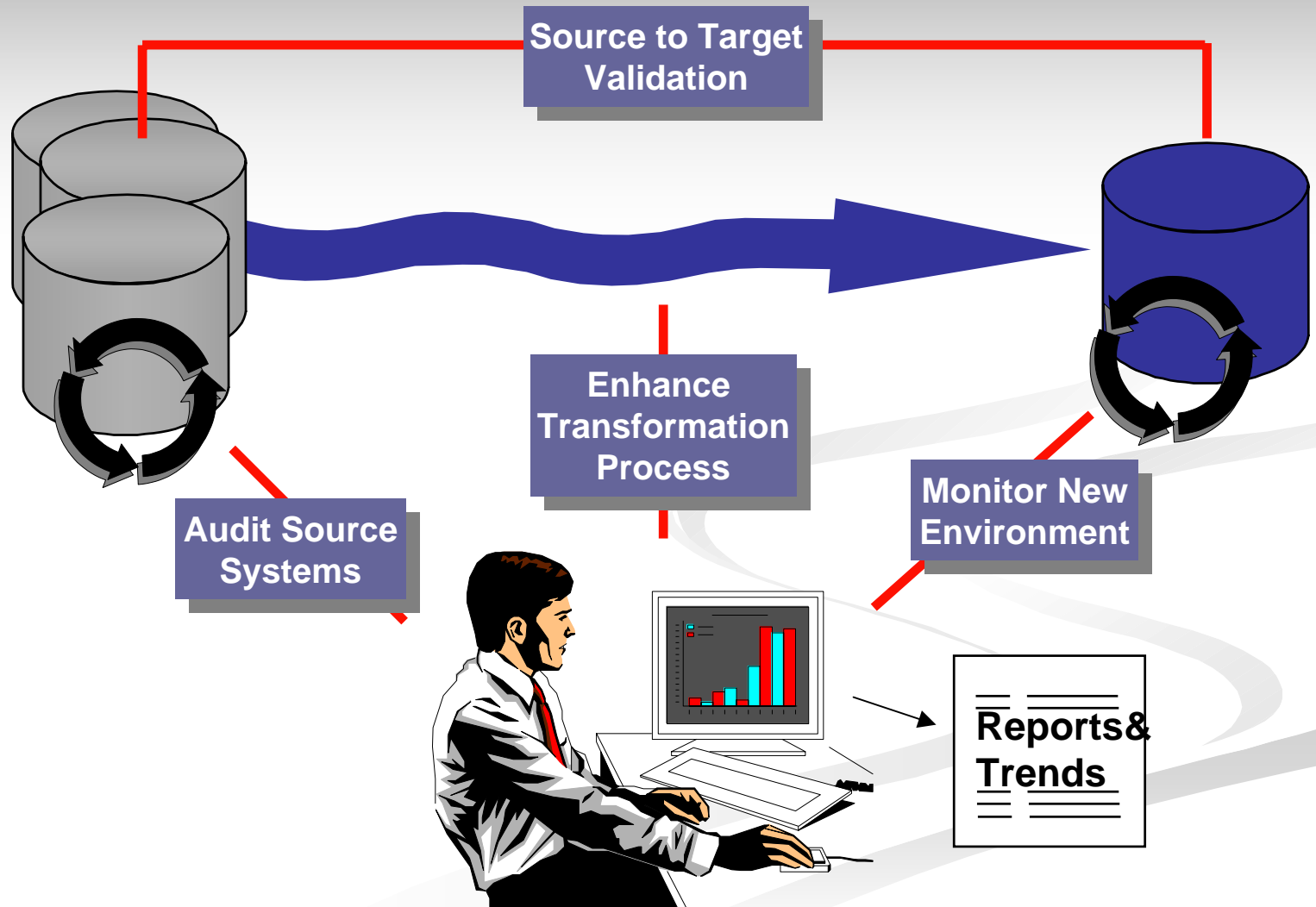
Information Integrity Enterprise



Enterprise Information Integrity Methodology



# Data Quality in the Transformation Process



# Basic Data Quality Analysis

## Consists of Multiple Levels

- **Level 1 Analysis – Domain Values** (basic content and distribution analysis)
- **Level 2 Analysis - Completeness and Correctness** (check for blank/null/missing and data type conformance)
- **Level 3 Analysis - Relational Integrity** (structural integrity, codes found in code/lookup table, etc.)
- **Level 4 – Business Rule Compliance** (e.g., ship date after order date)
- **Level 5 – Transformation Rule Compliance** (check data in warehouse/mart to ensure ETL transforms were applied properly)

# Progress to Date

- Information Governance Program Established
- Team Hired
- Data Stewards Identified in the Business Areas
- Data Quality Suite Purchased
- All Source Data Being Tested for Actual vs. Expected Values
- EDW Designed – conceptual Model Finished
- Logical model in Progress
- Guidelines for what data belongs where defined and published

# If You Do Nothing Else, do a Simple Content and Distribution analysis of Code Fields

Gender Code:

- M
- F
- U
- B

Then follow up with your business users on how to interpret and handle the unexpected.

# The Real World: Issues and Road Blocks

- Little Participation by Business
- “Build it and they will come” mentality
- Reporting Requirements Unknown
- Need to fix data at the source
  - Incentives of Call Center Employees directly cause data quality problems
- DBA’s used to RedBrick
  - Everything forced into star schema
  - Many Data Marts that can’t be “joined”
- GUI standards defined by developers – not targeted towards business users

# Data Becomes Information If and Only If You:

- **Have** the data, and
- **Know** you have it, and can
- **Access** the data, and can
- **Can use** the data, and can
- **Trust** the data

# Enterprise Architecture Process

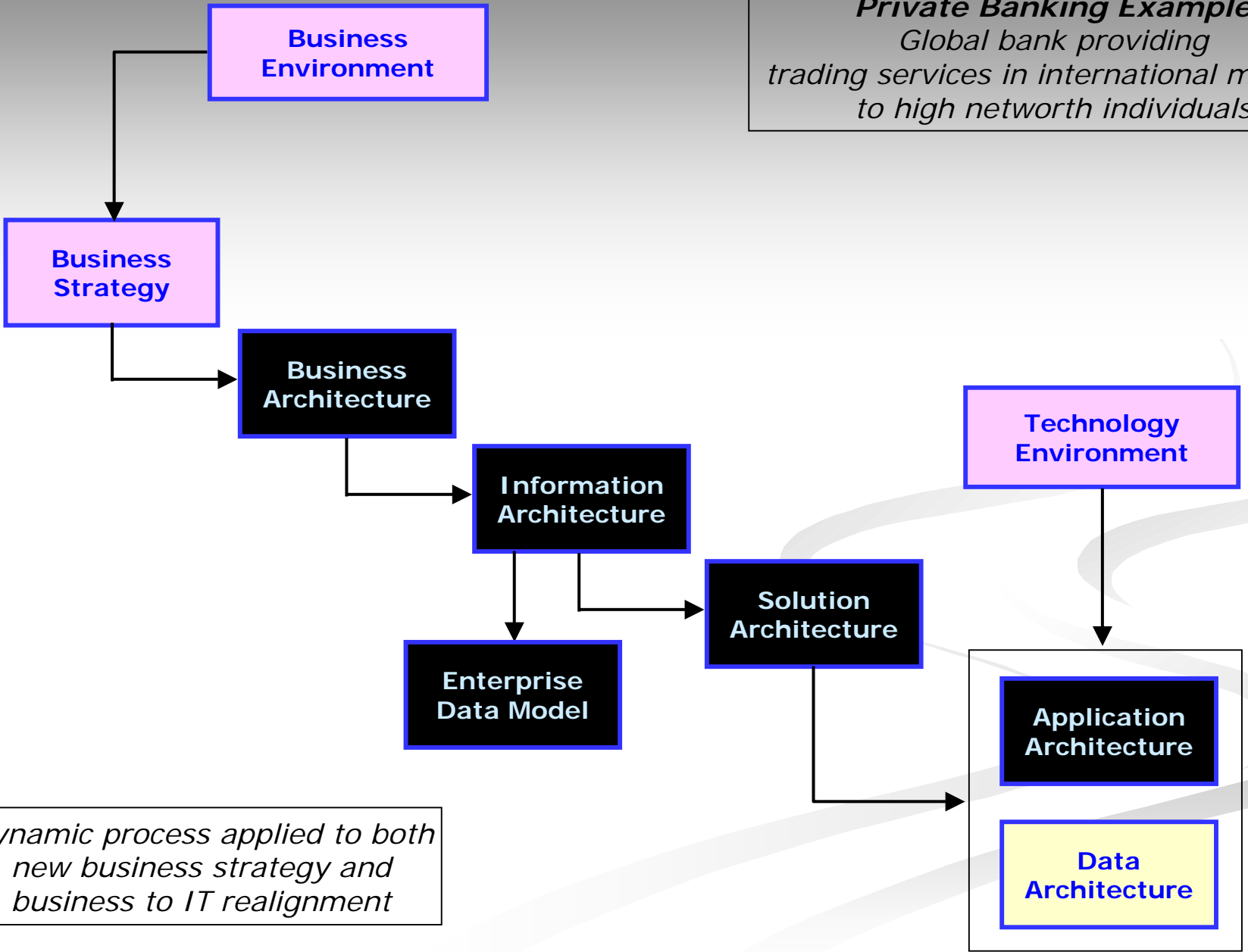
DAMA Roundtable  
February 4, 2004

Nicholas Khabbaz, Ph.D.  
e-Modelers, Inc.  
[www.emodelers.com](http://www.emodelers.com)



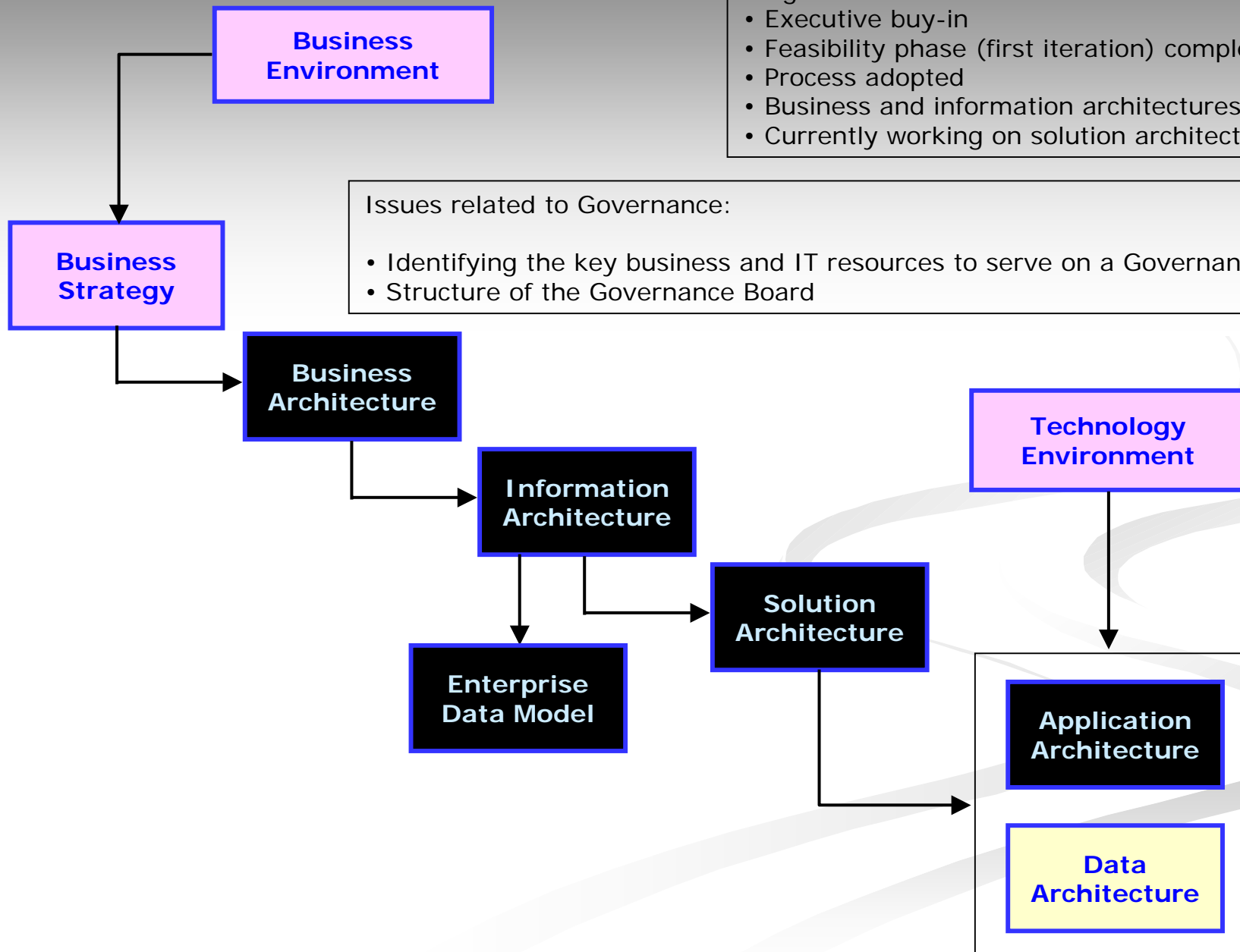
# Enterprise Architecture Process

*Private Banking Example*  
Global bank providing trading services in international markets to high networth individuals



*Dynamic process applied to both new business strategy and business to IT realignment*

# Enterprise Architecture Process



## Progress to Date:

- Executive buy-in
- Feasibility phase (first iteration) completed
- Process adopted
- Business and information architectures adopted
- Currently working on solution architecture

## Issues related to Governance:

- Identifying the key business and IT resources to serve on a Governance Board
- Structure of the Governance Board

# Business Environment & Strategy

## Business Environment

2003 Global Stock  
Market Returns



## Business Strategy

Provide high networth  
individuals with capability  
to trade directly in  
foreign markets



## Business Benefits

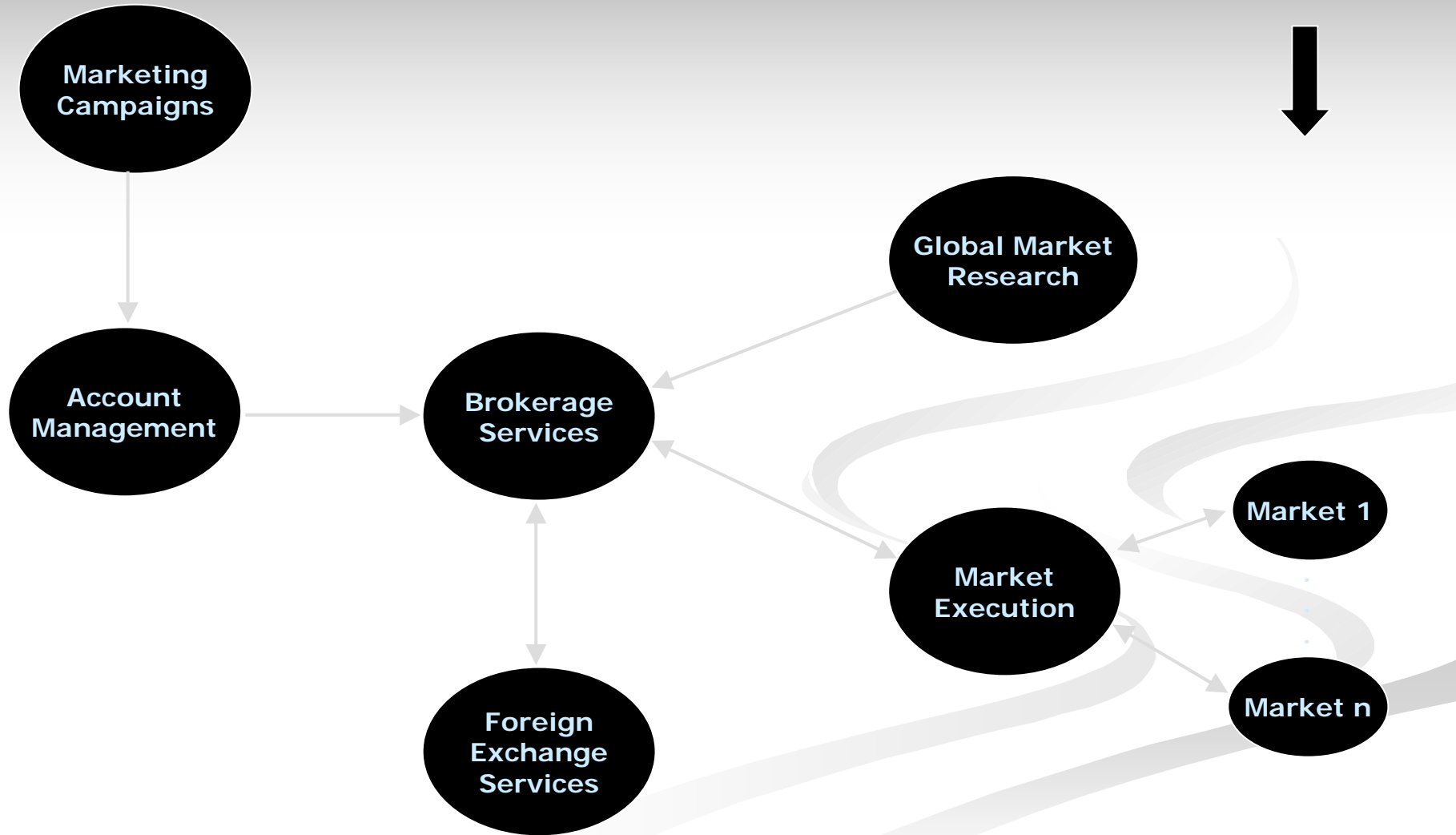
Client retention and  
increased brokerage fees

Region	Country	Index	2003 Close	2003 Return
Africa	South Africa	Johannesburg All-Share	10,387	12.0%
Asia	China	Shanghai A Shares	1,569	10.6%
	Hong Kong	Hang Seng	12,576	34.9%
	India	Bombay Sensex	5,839	72.9%
	Japan	Nikkei	10,677	24.5%
	Singapore	Straits Times	1,765	31.6%
	South Korea	Kospi	811	29.6%
	Taiwan	Weighted	5,891	32.3%
	Thailand	SET	772	116.6%
	Australia&NZ	Australia	All Ordinaries	3,306
New Zealand		Top 40	2,278	17.1%
N. America	Canada	Toronto 300 Composite	8,221	24.3%
	Mexico	IPC All-Share	8,795	43.6%
	U.S.	DJIA	10,453	25.3%
S. America	Argentina	Merval	1,072	104.2%
	Brazil	Sao Paulo Bovespa	22,236	97.3%
	Chile	Santiago IPSA	1,485	48.5%
	Venezuela	IBC General	22,203	177.0%
W. Europe	Austria	ATX	1,545	34.4%
	Denmark	KBX	214	28.5%
	France	Paris CAC 40	3,558	16.1%
	Germany	Frankfurt Xetra DAX	3,965	37.1%
	Italy	Milan MIBtel	19,922	13.9%
	Norway	All-Share	178	48.0%
	Spain	IBEX 35	7,737	28.2%
	Sweden	SX All-Share	194	29.8%
	Switzerland	Zurich Swiss Market	5,488	18.5%
	U.K.	London FTSE 100	4,477	13.6%

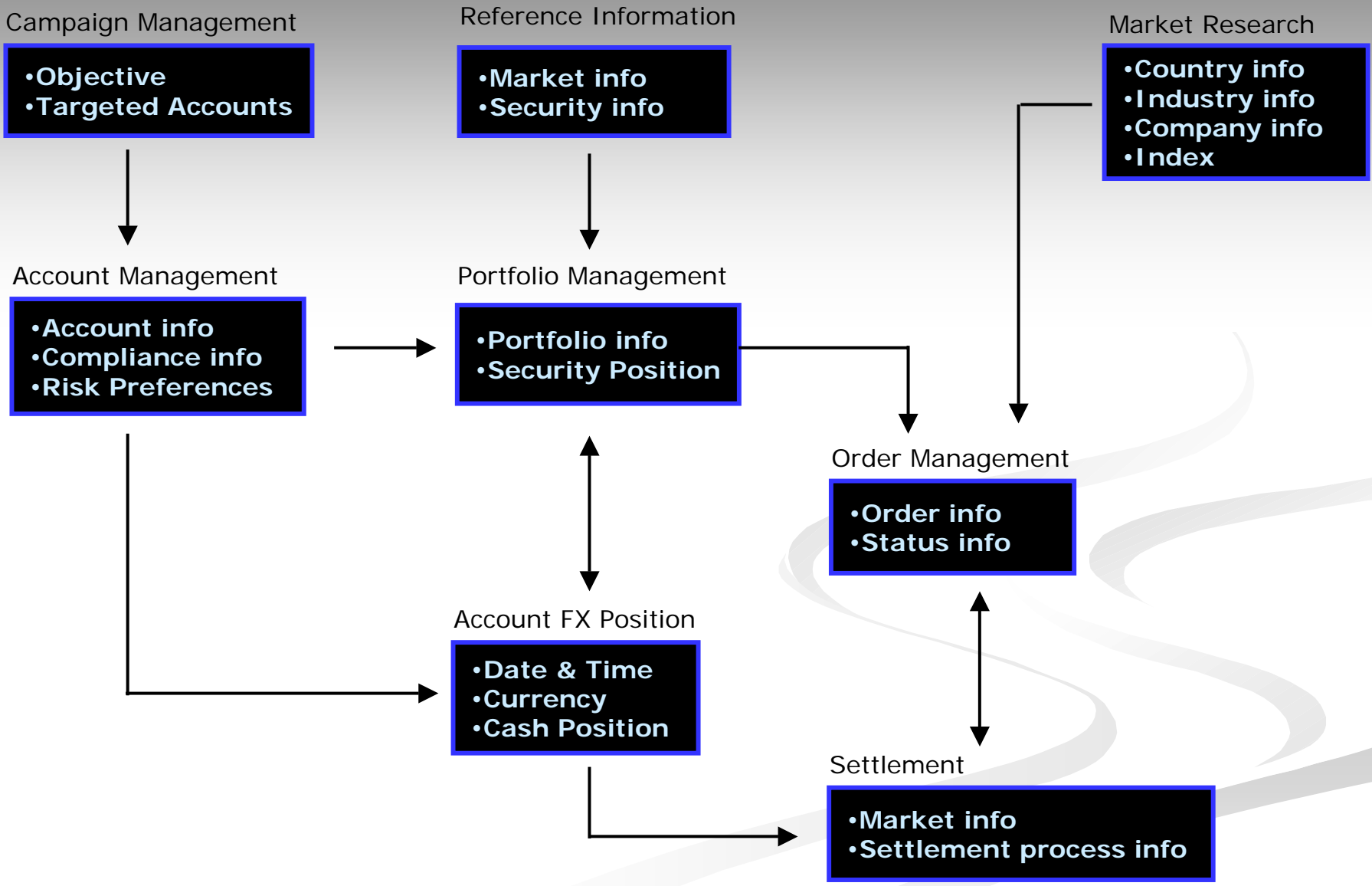
# Business Architecture

Issues/roadblocks related to trading in different markets:

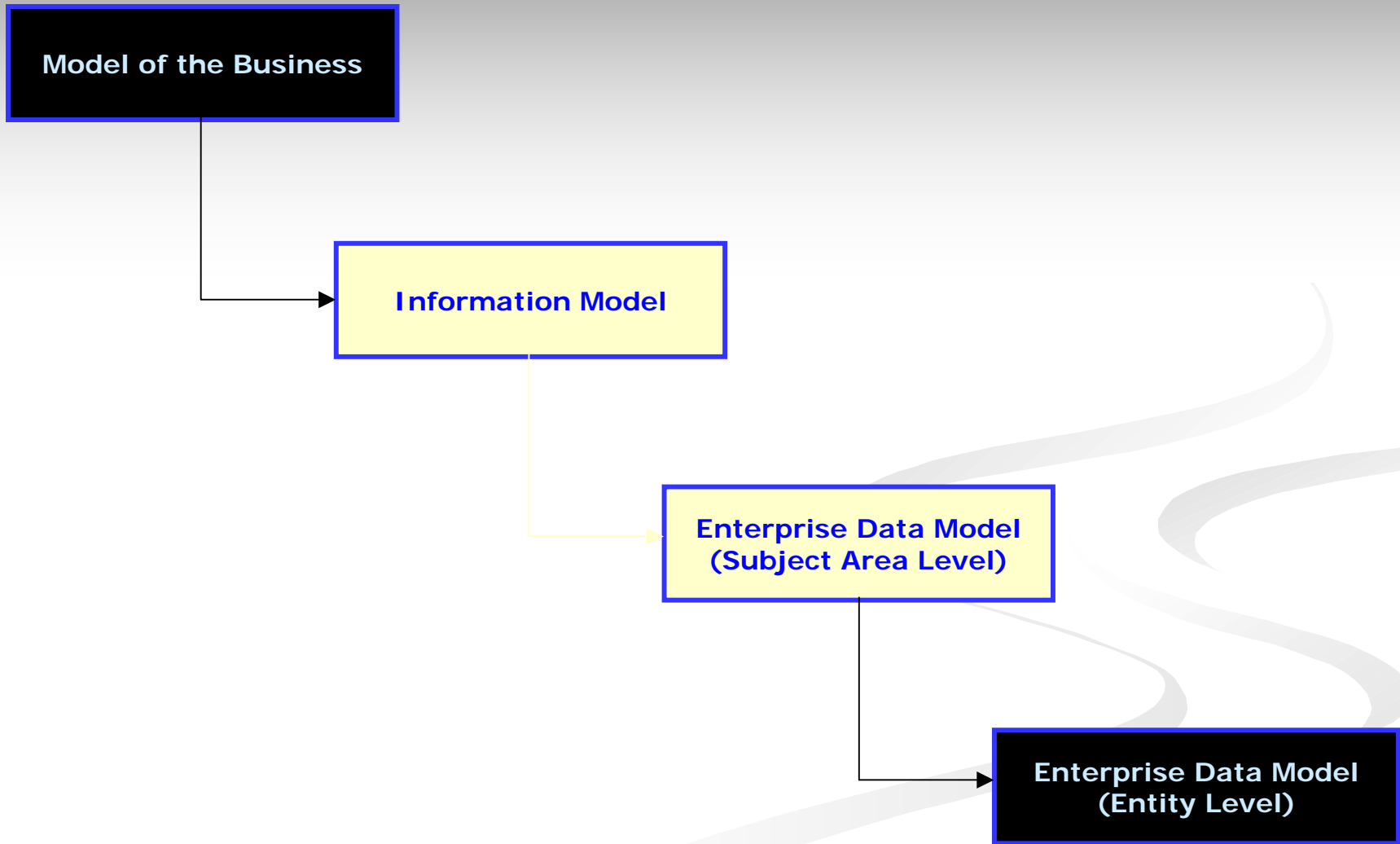
- Each market had its own timing and its own business processes
- There were no common data standards related to these markets



# Information Architecture



# Enterprise Data Model



# Solution Architecture

Issues/roadblocks:

- The extent to which existing solutions could or should be re-used

**Customer Relationship Management**

- Campaign Management
- Account Management
- Consolidated Reporting

Reference Information Management

**Front-Office Trading**

- Market Research
- Portfolio Management
- Real-time Compliance
- Decision Support
- Order Management

**Middle-Office**

- FX Management
- Settlement
- Exception Reporting

Market Interface

*Operational Systems*

**Risk Management**

- Analytics
- Exposure Reporting

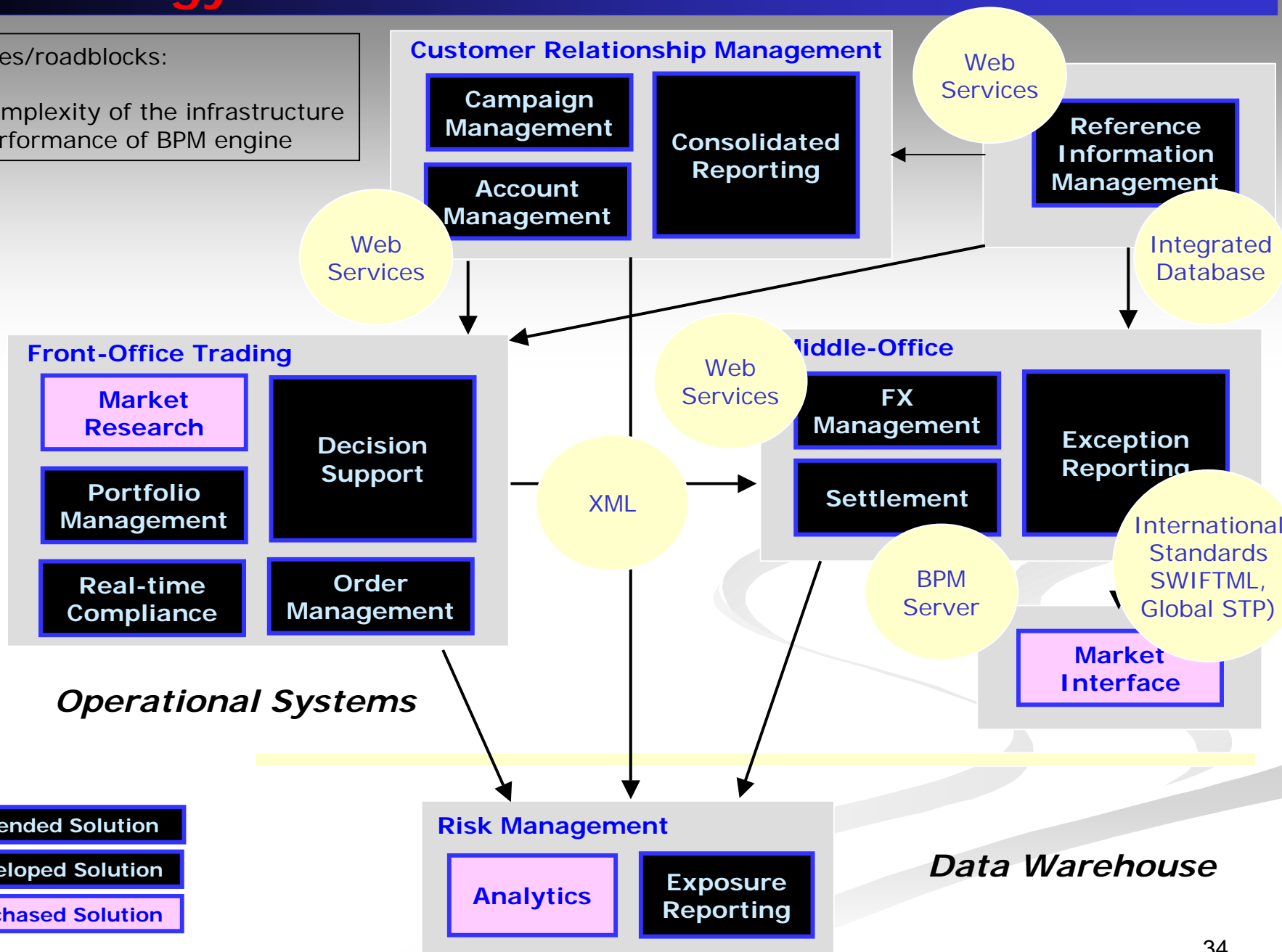
*Data Warehouse*

- Extended Solution
- Developed Solution
- Purchased Solution

# Technology Environment

Issues/roadblocks:

- Complexity of the infrastructure
- Performance of BPM engine



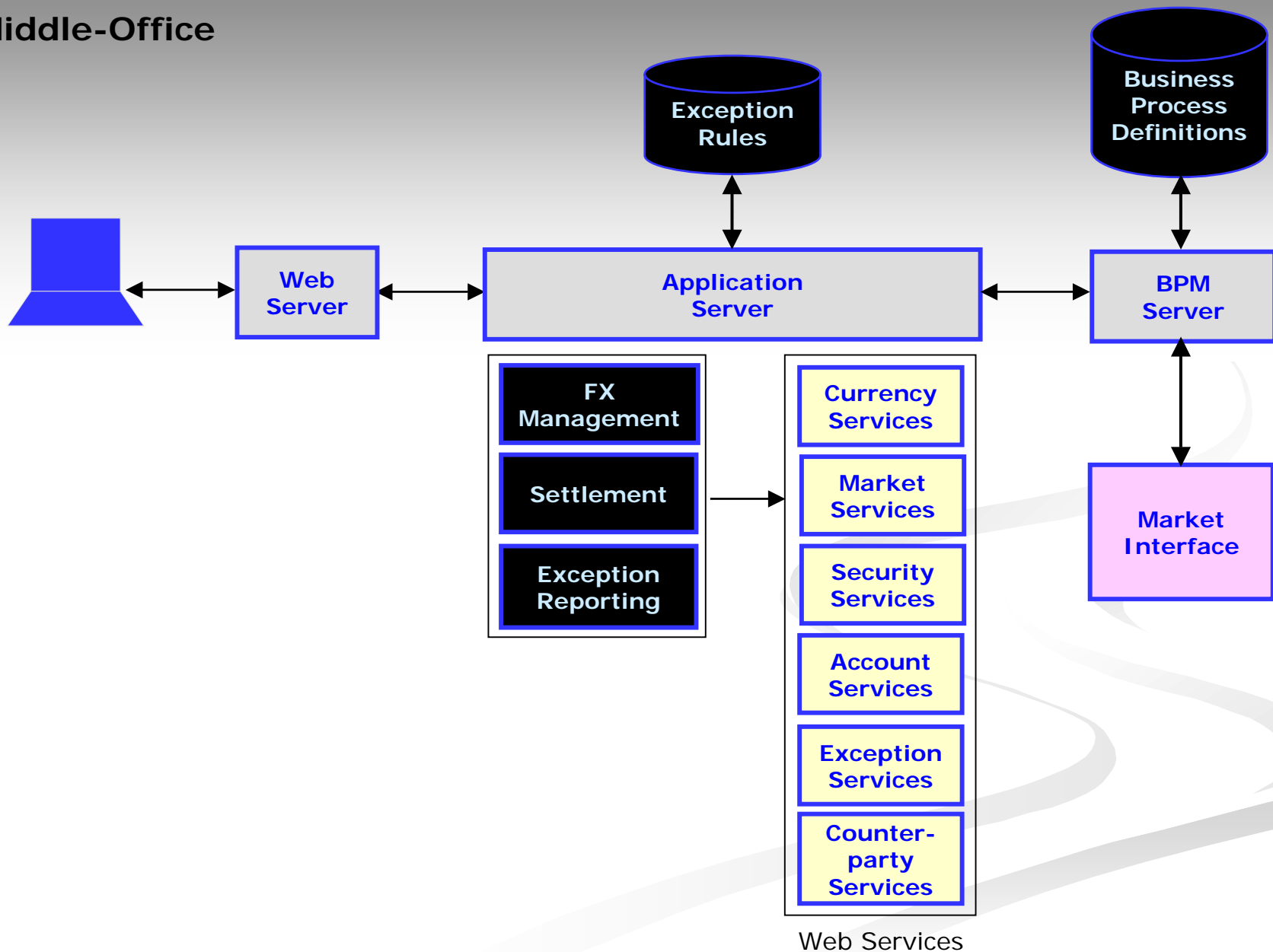
Extended Solution

Developed Solution

Purchased Solution

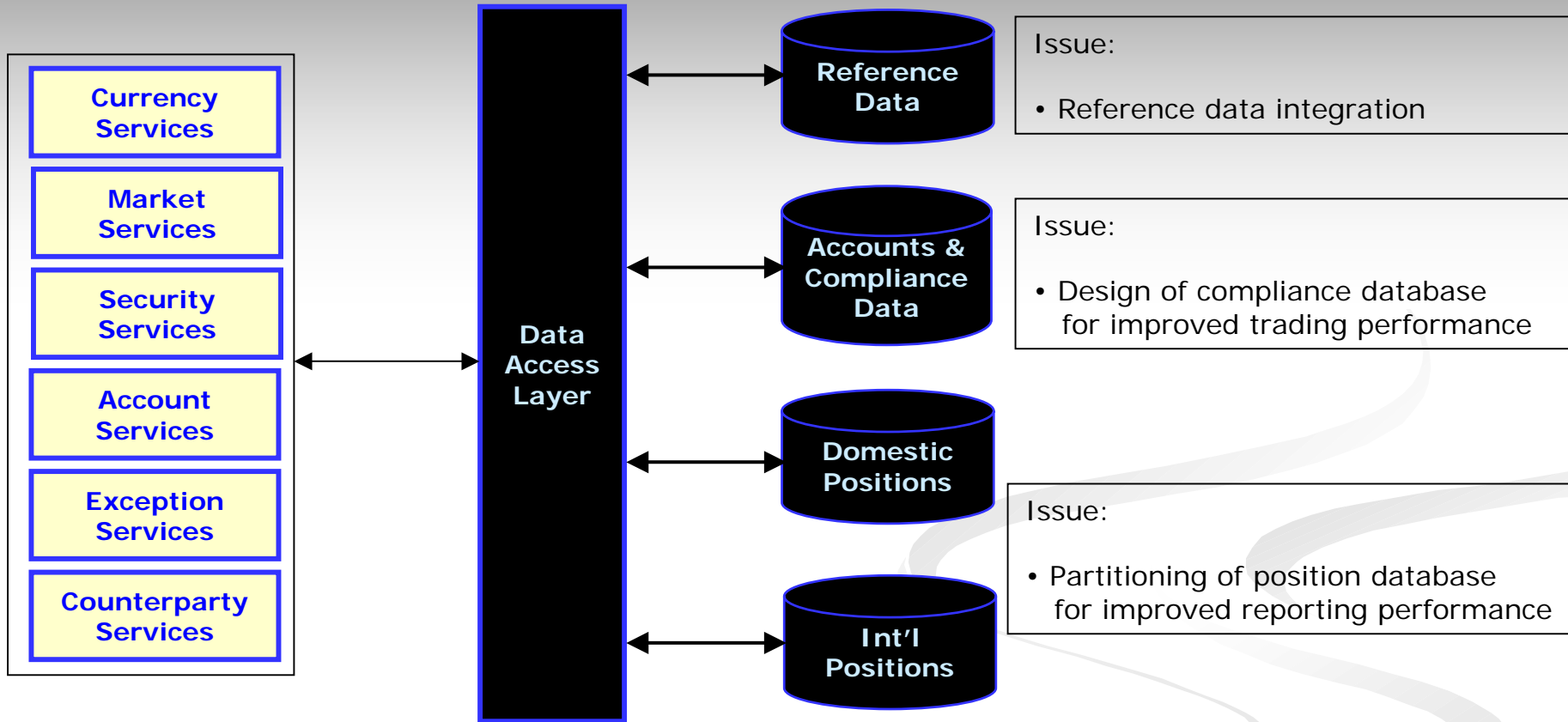
# Application Architecture

## Middle-Office



# Data Architecture

## Middle-Office



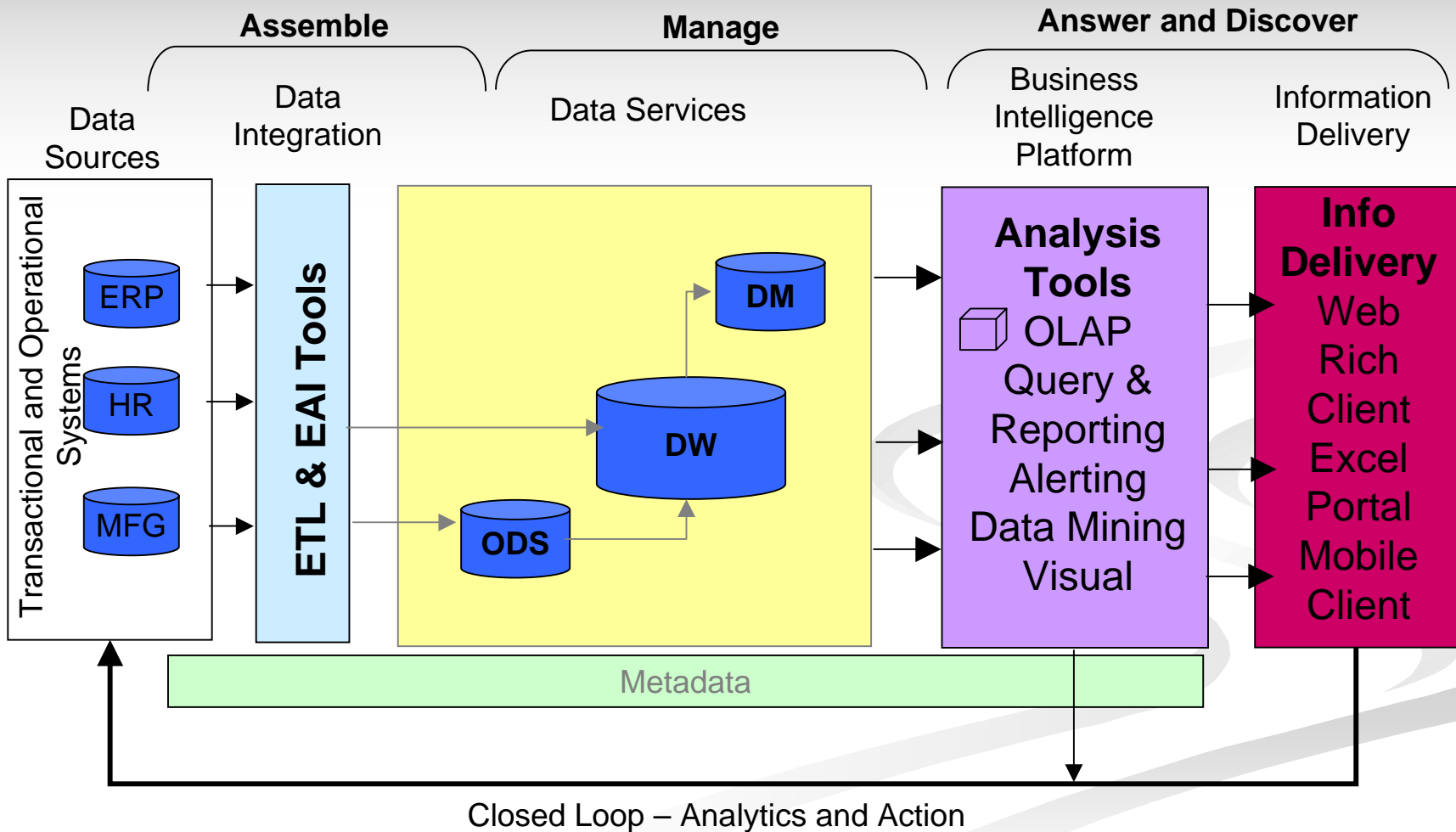
*Data Access Layer Services*  
Transaction management,  
access to multiple databases,  
security, performance, reliability,  
data availability, etc.

# Data Warehousing & ETL

Gregg Wyant

Chief Data Architect  
Intel

# DSS/Analytics conceptual view



# Data Warehousing

## Objective / Approach:

- Integrated View of Enterprise Data
- One Version of Truth
- Enterprise Datawarehouse (Centralized Infrastructure)
- Incremental Build

## Progress To Date:

- In Production since September 2002
- Refreshed every 4 – 8 hours
- ERP Data – GL, Revenue, Product, Channel

## Issues/Roadblocks:

- Balancing Quality and Speed of Delivery
- Experienced Resource Pool

# Extract Transform & Load (ETL)

## Objective / Approach:

- Increased Developer Productivity, Lower Maintenance Cost
- Scalable, Reliable
- Homegrown, ELT Vs. ETL
- Drive to an Enterprise ETL solution

## Progress To Date:

- Loading ~700 Tables/Day Every 4 – 8 Hours
- Support for Full and Net Change Data

## Issues/Roadblocks :

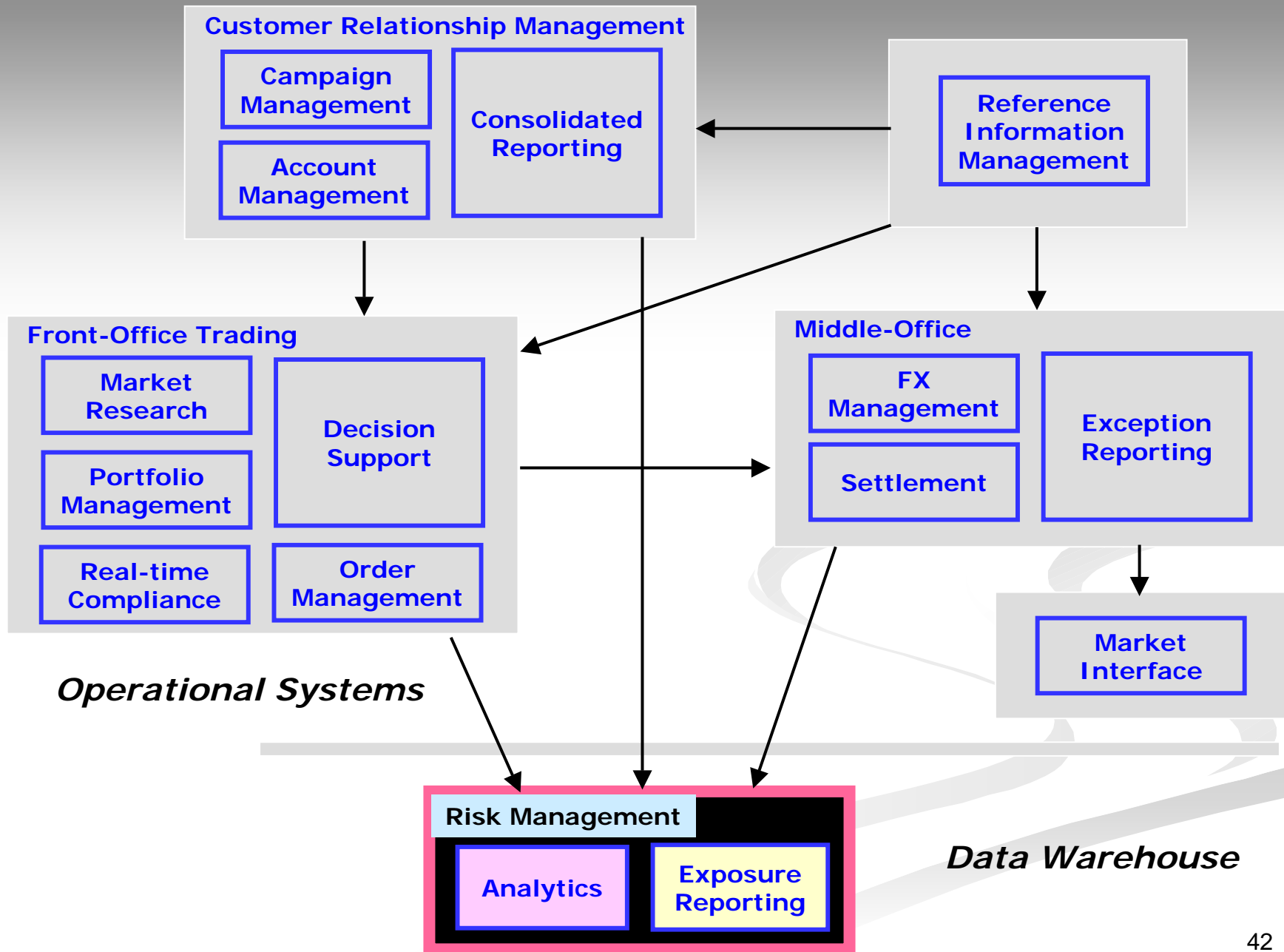
- Cost Effective Transformation Solution
  - Platform
  - Developer Skill set

# Data Warehouse Architecture Process

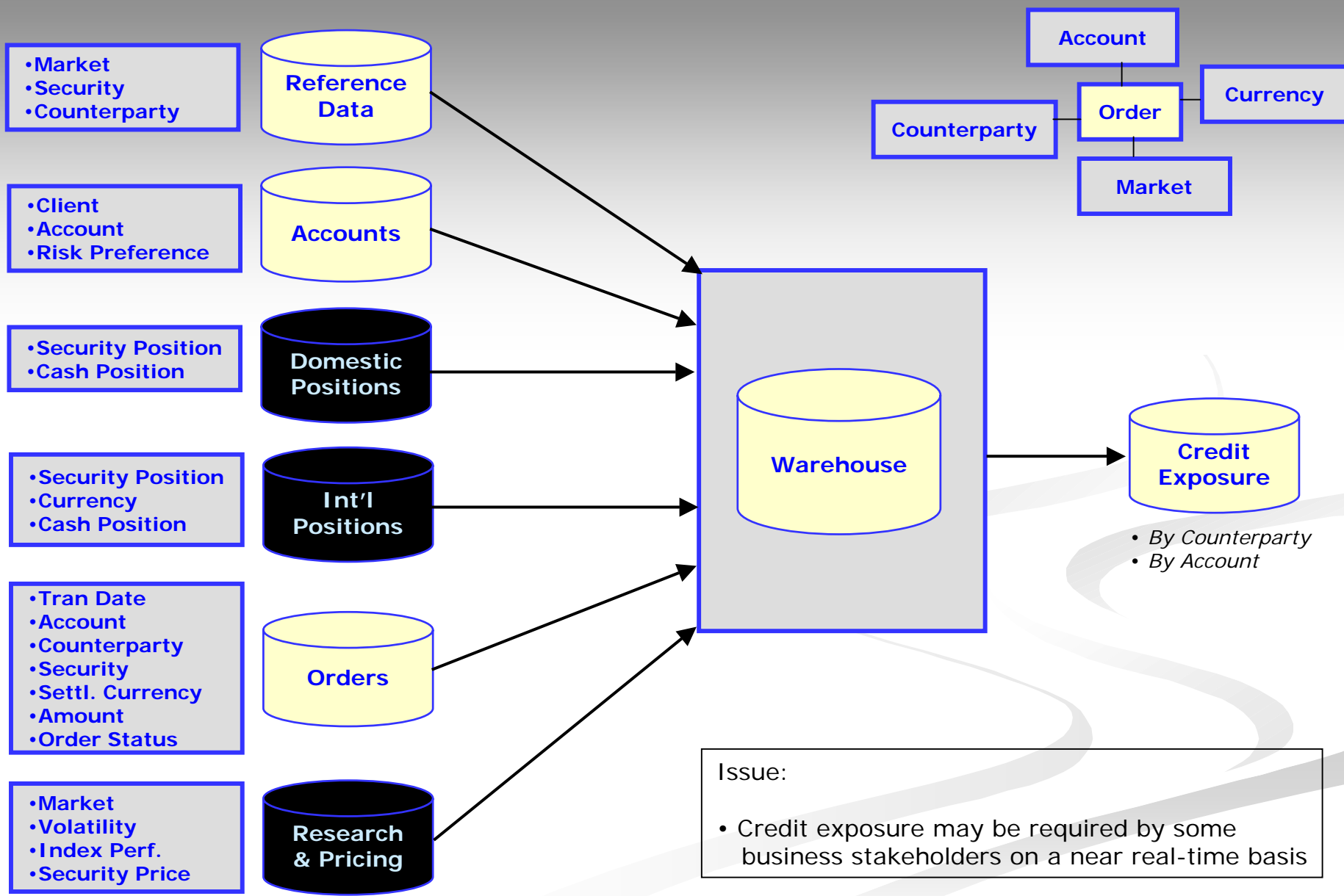
DAMA Roundtable  
February 4, 2004

Nicholas Khabbaz, Ph.D.  
e-Modelers, Inc.  
[www.emodelers.com](http://www.emodelers.com)

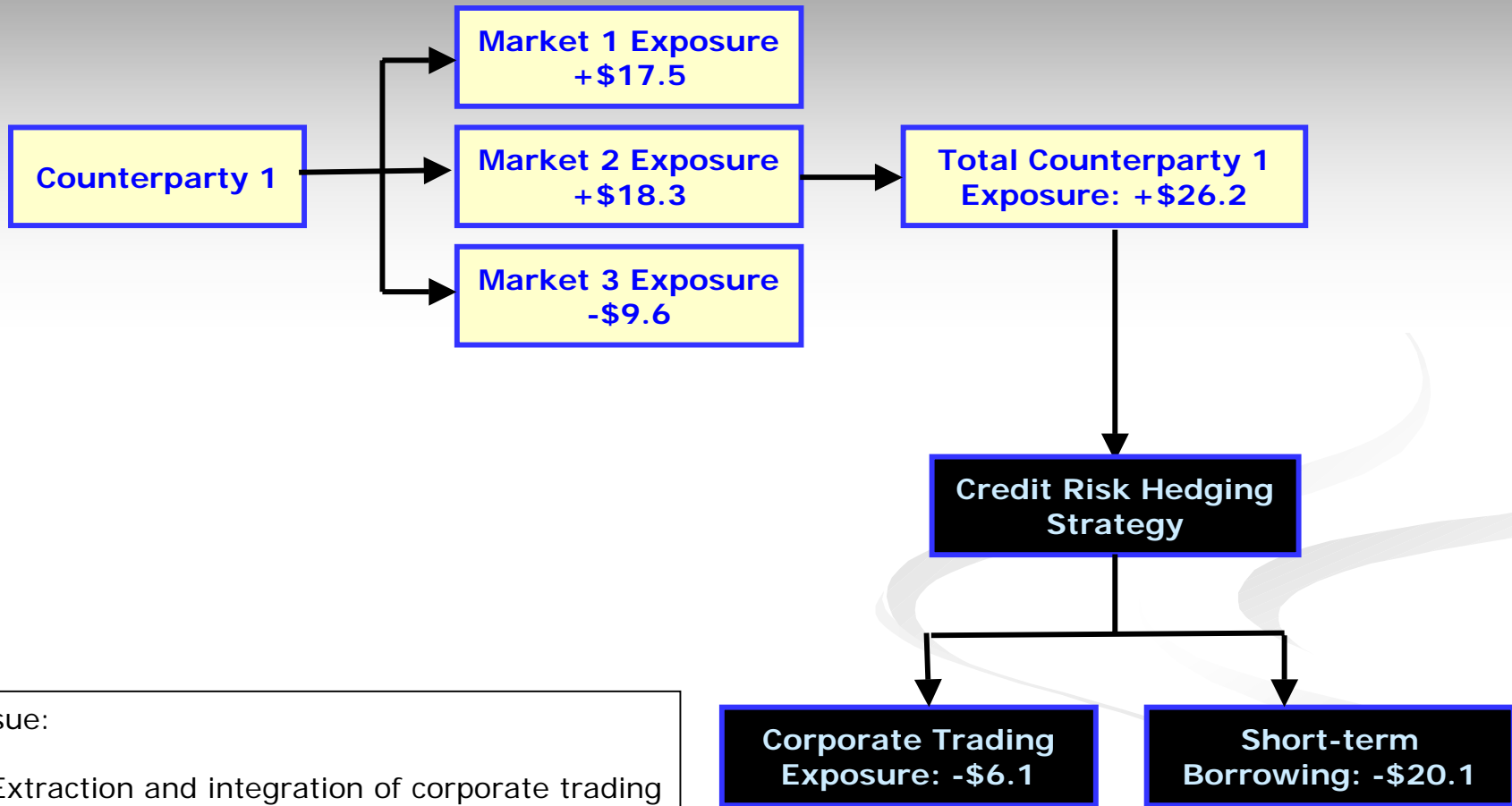
# Solution Architecture



# Credit Exposure & Hedging



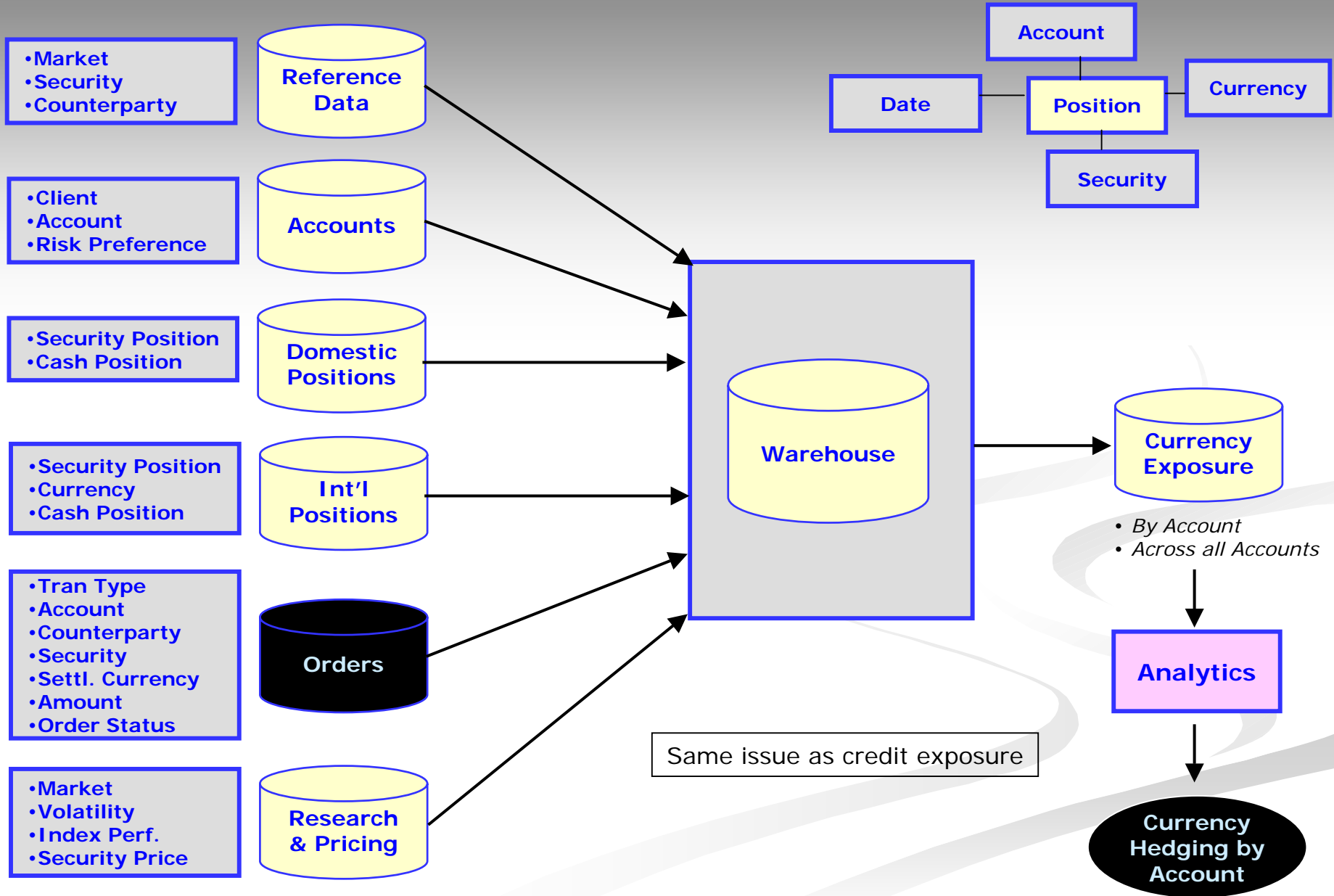
# Credit Exposure & Hedging



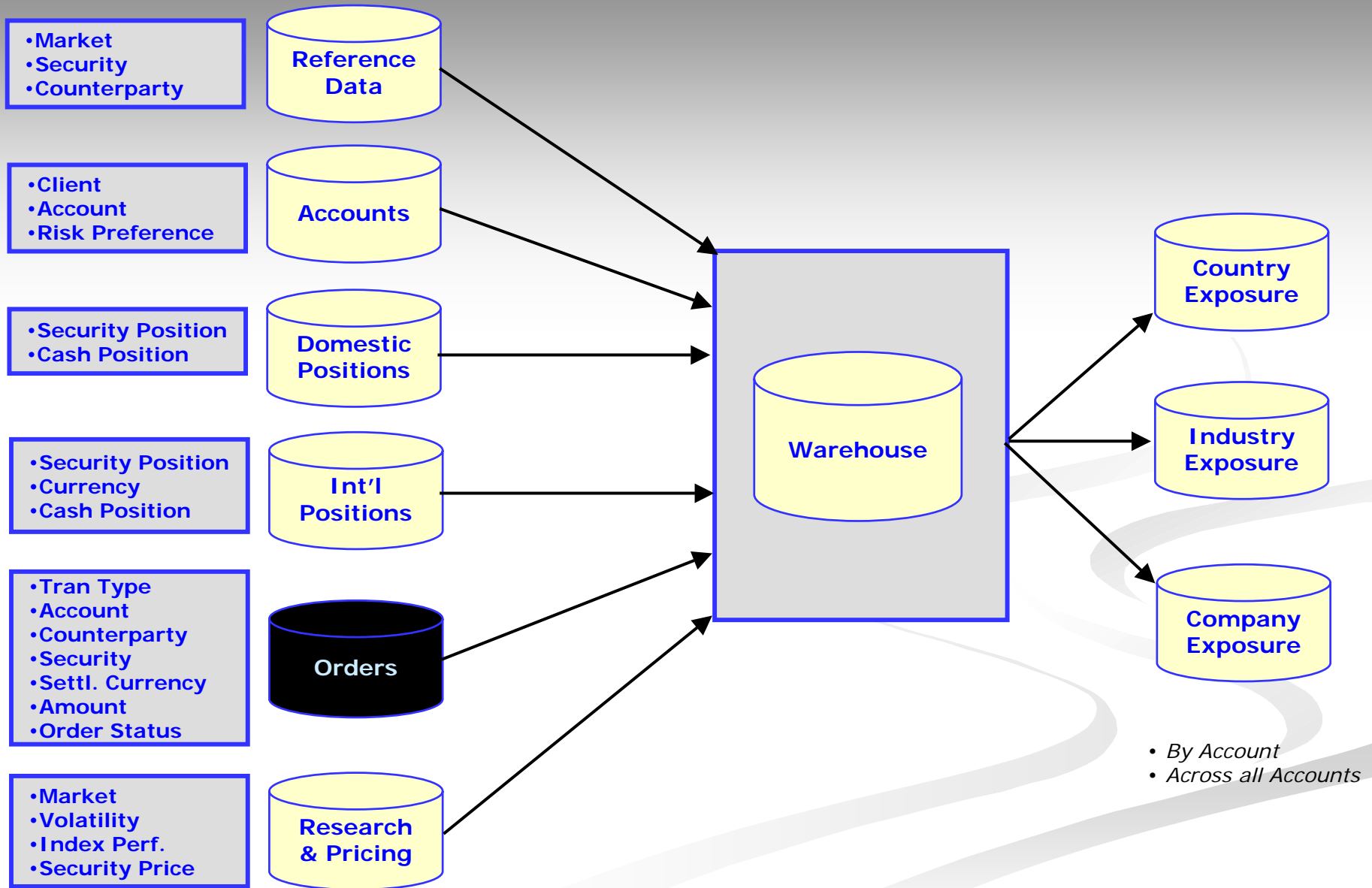
Issue:

- Extraction and integration of corporate trading exposure from legacy databases

# Currency Exposure & Hedging



# Country, Industry & Company Exposure

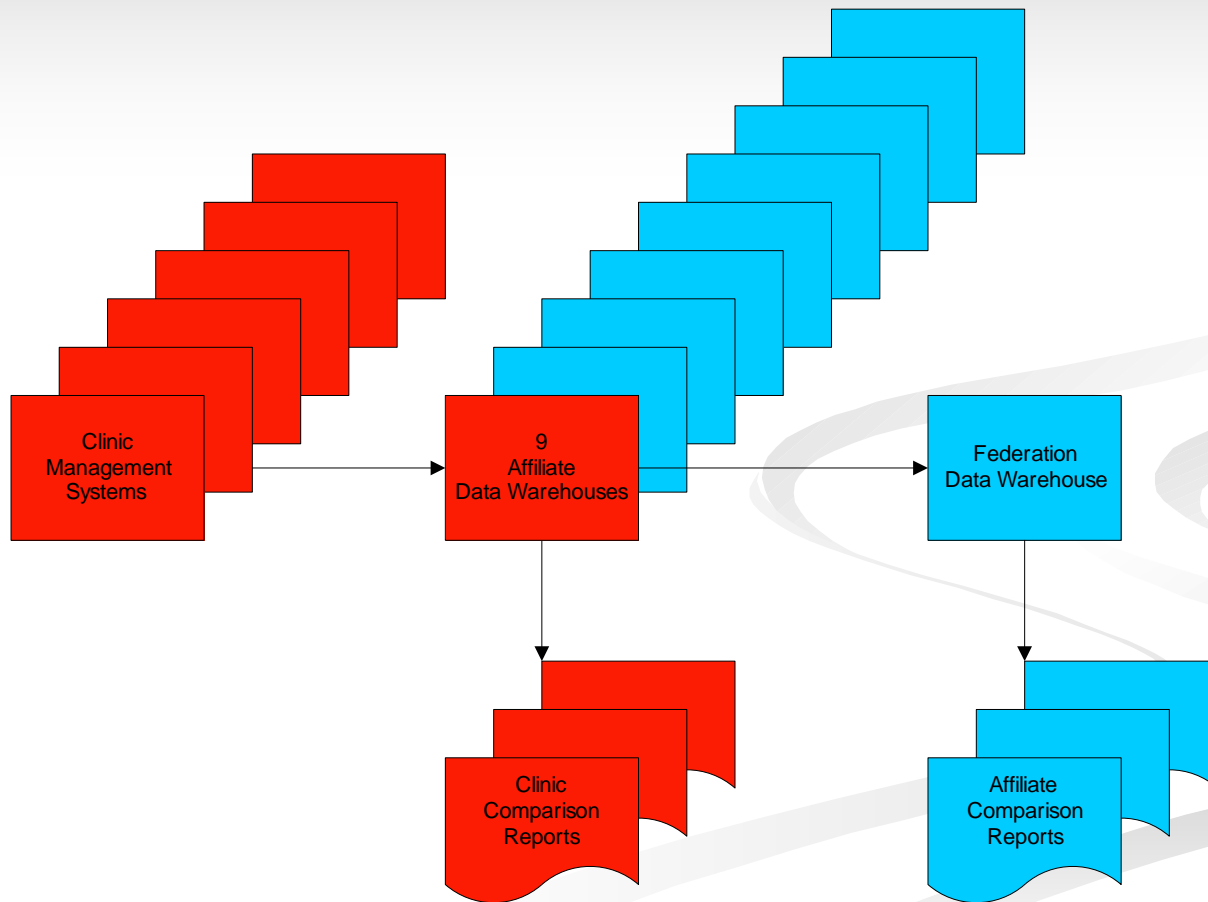


# Data Warehousing and ETL

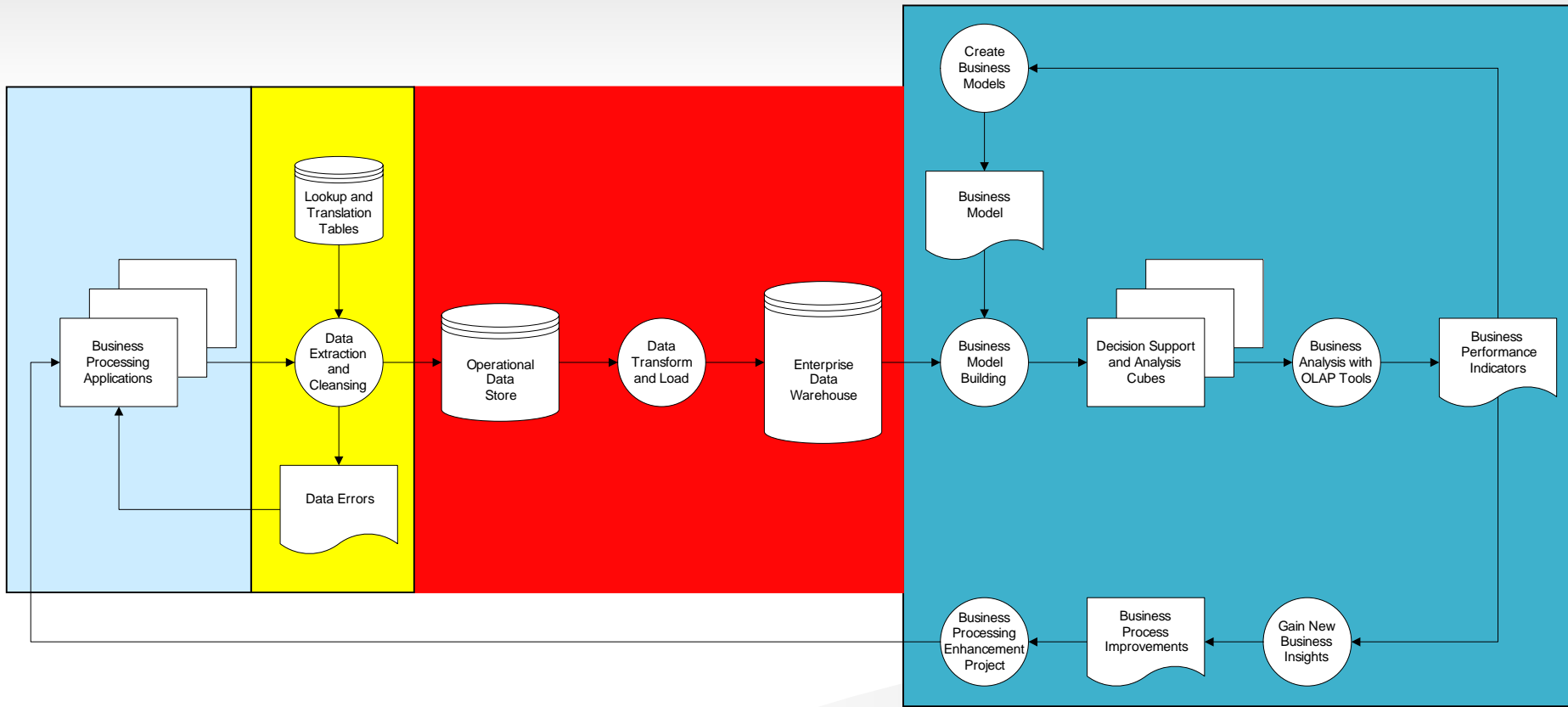
Created and Presented  
By  
Rainer Schoenrank  
The Data Organization

Copyright © 1998-2004  
Thursday, February 05, 2004

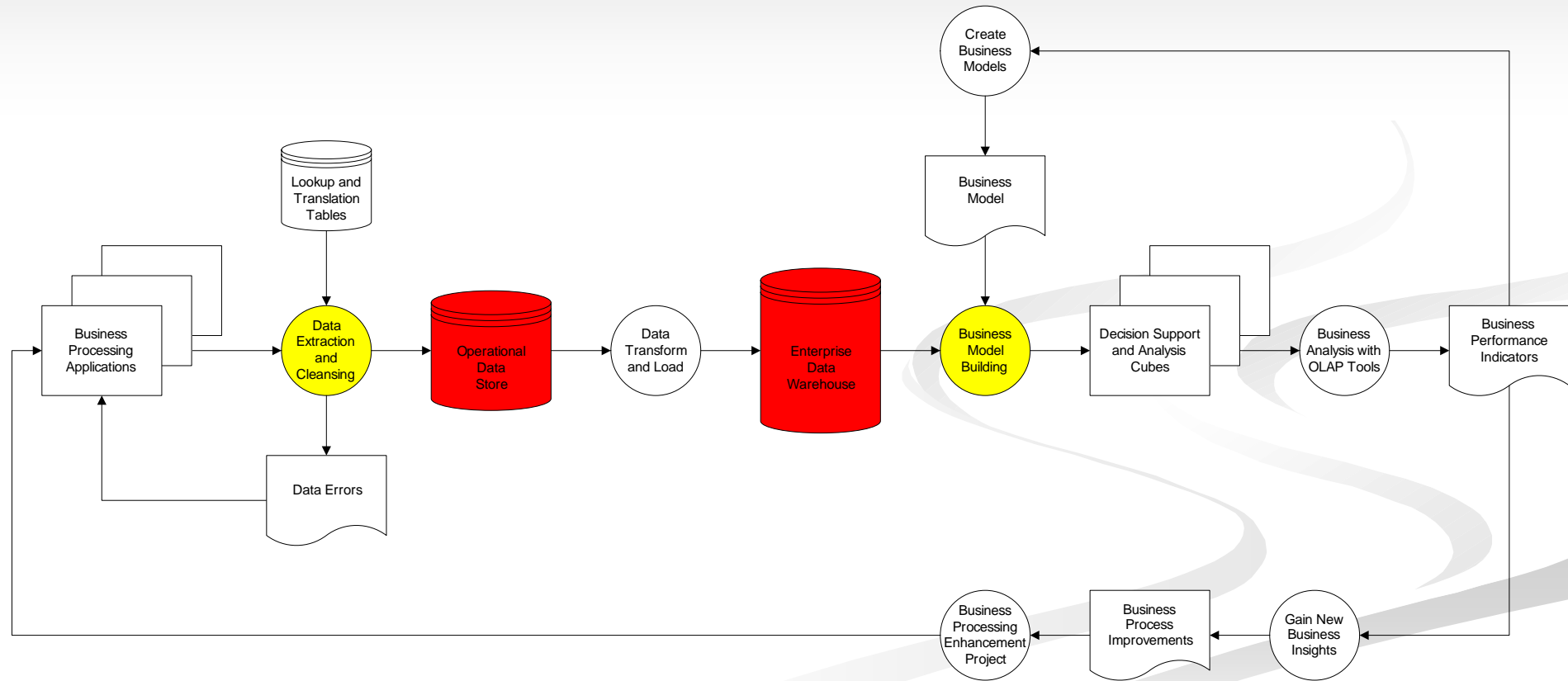
# Project Scope



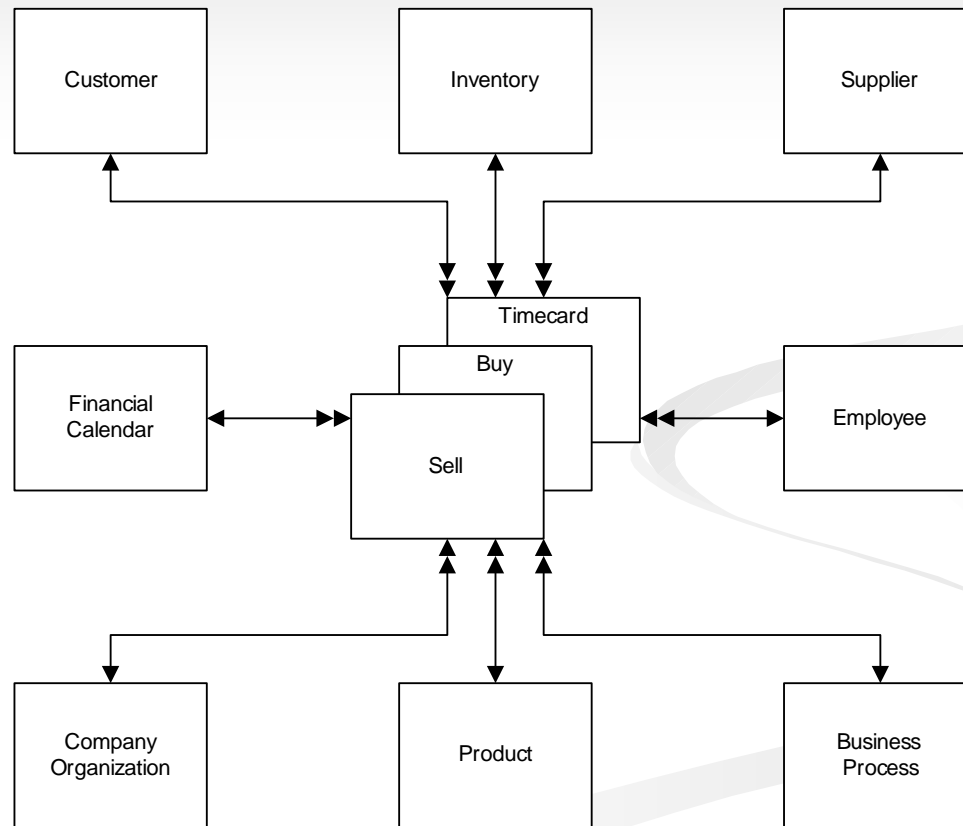
# Data Warehousing Process



# Data Warehouse Target



# Conceptual Data Warehouse



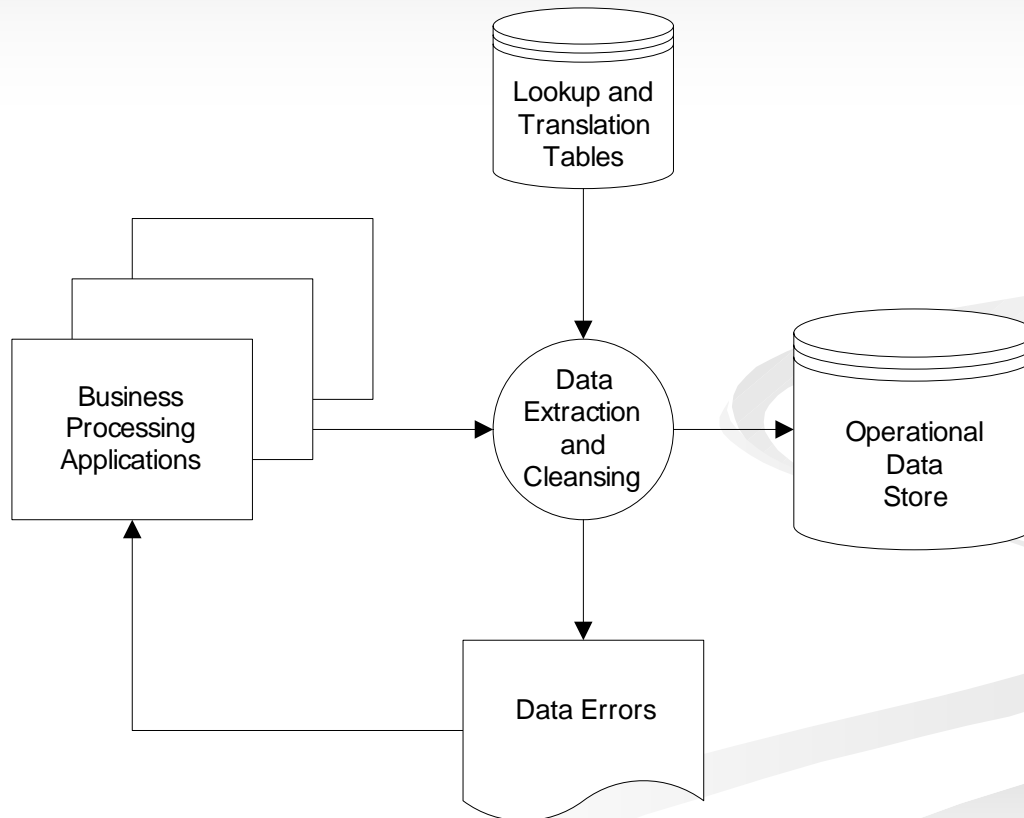
# Unified Business Model

Company Organization		Data Dimensions							Business Transactions		
		Operational Data				Organizational Data			Buy	Sell	Timecards
		Customers	Inventory	Suppliers	Employees	Business Processes	Products	Company Organization			
Management Processes	Understand Markets and Customers										
	Develop Vision and Strategy										
	Manage Improvement and Change										
	Manage External Relationships										
	Execute Environmental Management Program										
Operating Processes	Design Products and Services										
	Market and Sell										
	Produce and Deliver Product										
	Invoice and Service Customer										
Support Processes	Develop and Manage Human Resources										
	Manage Information Resources										
	Manage Financial and Physical Resources										

# Unified Business Model Validation

Company Organization		Data Dimensions						Business Transactions				
		Operational Data				Organizational Data		Buy	Sell	Timecards		
		Customers	Inventory	Suppliers	Employees	Business Processes	Products				Company Organization	Financial Calendar
Management Processes	Understand Markets and Customers	SFA	SFA		SFA		SFA					
	Develop Vision and Strategy	Decision Support Reporting										
	Manage Improvement and Change	Decision Support Reporting										
	Manage External Relationships			Capital Mngmt								
	Execute Environmental Management Program											
Operating Processes	Design Products and Services					PCM						
	Market and Sell	Order Entry								Order Entry		
	Produce and Deliver Product	Fulfillment	Receiving	Supplier Mngmt	Scheduling				Purchasing	Scheduling		
	Invoice and Service Customer	CRM								CRM		
Support Processes	Develop and Manage Human Resources				HR						Time Capture	
	Manage Information Resources		Fixed Assets									
	Manage Financial and Physical Resources		Fixed Assets				G/L					
		A/R		A/P	Payroll				A/P	A/R	Payroll	

# Data Quality Assurance Process (ETL)



# Progress to Date

- Project Status
  - 80% complete
  - Cost is about \$1M of a planned \$1.4M
- Project Time Line
  - Started Oct 95
  - Specification completed Jun 97
  - Proof of Concept completed Nov 00
  - QA Review and Evaluation completed Mar 01
  - Pilot completed Mar 02
  - Rollout to 10 installations is in progress

# Progress to Date

## ■ Installation Status

- One site is being used daily as part of operational auditing
- Two sites are being used monthly as part of enterprise reporting
- Four sites are attempting to integrate new processes into the operational culture
- Three sites are preparing for installation

# Issues / Roadblocks

- Cross-department cooperation for an infrastructure initiative
  - Support for changes to enterprise models including new policies, procedures, and organizing structures
- Resistance to change of existing processes
  - Changes to enterprise processes by collecting more complete, accurate, and timely data

# ETL Architecture & Related Tools

Cass Squire  
Associate Partner  
IBM Business Consulting Services  
[csquire@us.ibm.com](mailto:csquire@us.ibm.com)  
(650) 520-7247

# Topics

- The Environment
- A Strategy & Approach for Solving it
- Progress To Date
- The Real World: Issues and Road Blocks

# The Business Environment

- A major retailer – three distinct store brands and on-line sites
- 4,200 Stores world-wide
- 25 distribution centers
- Recently purchased the Retek RMS system to manage its inventory and shipping processes
- Assured delivery required
- Sarbanes Oxley (SOX) compliance a company emphasis

# The Technical Environment

- Retaining parts of its Legacy Systems for Allocations and other functions
- Both systems need virtually all data – whether they are system of record or not
- Much is batch oriented, but some is real-time or near real-time
- DB2/MVS and IMS on the legacy side
- Oracle 9i on the Retek side

# Strategy & Approach

- Build an Integration layer to keep the legacy and Retek systems synchronized
- Build a reporting ODS for analysis, determining synchronization issues, etc.
- (look in the future towards building a true ODS so that both sides share the same database and data need not be shuttled back and forth)
- This leads to complex Extract, Transform, and Load requirements
- Select tools to build this rather than custom, hand-written code

# Progress to Date

- Tool evaluation process conducted
- Ascential DataStage (Server and Parallel Canvases) selected for ETL
- IBM Websphere Business Integrator (WBI) selected for messaging/real-time/near real-time interfaces
- Crystal Reports selected for reporting off the ODS
- Common services for Auditing and Message-retry built
- Approximately 1/4 of the interfaces built and tested
- “Best Practices” guidelines for both ETL and EAI developed
- 2 other projects in the DW/ODS space now utilizing DataStage and the best practices and architecture established in this project

# Benchmark: ETL Versus Code and In-DataBase Processing

- **69,000,000 rows of Source Data**

- **3 Table Lookups**
- **2 Against 1,000,000 row Tables**

- **Sort and Aggregate Operations**

## The Data

- The data was representative of Store Inventory. It included a full download of the Legacy TSKU table.
- The Source data included 69,000,000 rows.

## The Rules

- The data was first analyzed for base quality (format of Style Code)
- The data was then validated against a 1,000,000 row table.
- The data was further validated against a 6,000 row table.
- The Data was then cross referenced against a compound key 1,000,000 row table
- Finally, the data was aggregated on the 1,000,000 keys of the final lookup, and summary values for each key were sent to 5 different output targets.

- **Data Available in both Flat File and Oracle Database**
- **Ultimate solution must support these, as well as MQ Series and JMS**

- **Target of both File and Oracle Database**
- **Oracle prototype only used Oracle DB as both source and target**

# Results – DataStage Server

- Typical 4200 transactions per second (tps) per CPU for File to File activity
- Using a Database as a source, the performance was degraded by roughly 10% to 3900 tps per CPU
- Using a Database as a target, the performance was degraded by roughly 30% to 2900 tps per CPU
- The effects were cumulative – Using a Database as both source and target degraded performance by roughly 40% to 2500 tps per CPU

**Baseline performance of 4200 tps per CPU**

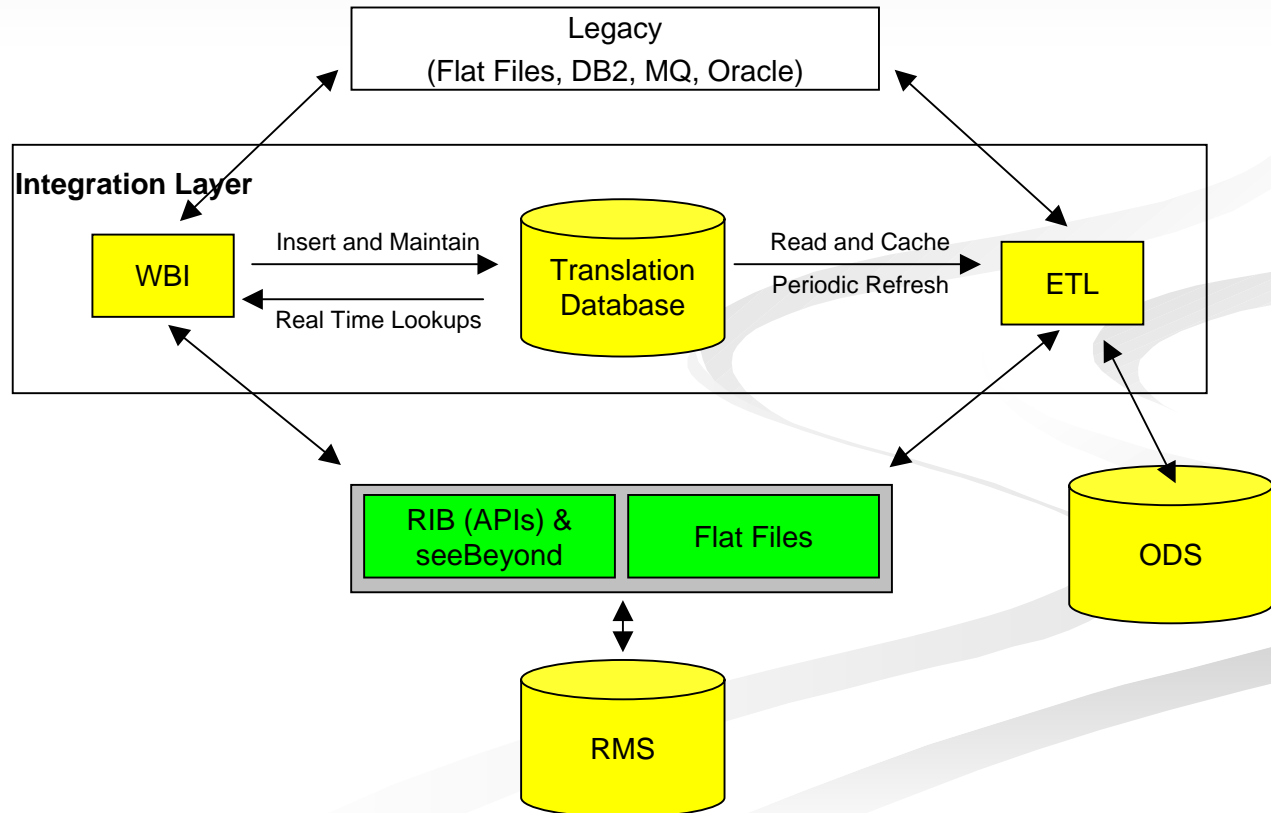
**Database Access degrades performance by up to  
40 % - reduce to 2500 tps per CPU**

# Results: Code & In-RDBMS

- Raw results were on the order of 200-700 Transactions per second per CPU.
- Database Tuning had little impact on the overall performance of the custom coding approach (120 tps/cpu – 130 tps/cpu)

# The Real-World Becomes Complicated

Combinations of tools come into play



# The Real World: Issues and Road Blocks

- This requires moving massive amounts of data (60-100 million rows a day) back and forth between the systems
- Oracle tables with 500,000,000 to 2,000,000,000 or more rows
- Some of these Retek tables have no timestamps
- Forcing the need to come up with some form of Change Data Capture (CDC) strategy – using replication, logs, etc.
- Rate of change on the Oracle tables results in hundreds of megabytes – or more - per second being written to the Oracle logs
- The good news is Ascential PX takes advantage of every resource on its box when running its massively parallel processes
- The bad news is Ascential PX takes every resource on its box when running its massively parallel processes

# Real-World Environment

- Real-time or Near Real-time
  - IBM Websphere Business Integrator (WBI)
  - SeeBeyond e\*Gate (Retek RIB)
  - Ascential RTI(?) DataStage TX(?)
- Batch
  - DataStage PX
  - DataStage Server
- Change Data Capture
  - Ascential CDC for Oracle?
  - Oracle Replication and LogMiner ruled out
  - Some brute-force file compares required. Using hash-totals to gain efficiencies

# Data Becomes Information If and Only If You:

- **Have** the data, and
- **Know** you have it, and can
- **Access** the data, and can
- **Can use** the data, and can
- **Trust** the data

# Metadata Management

Gregg Wyant

Chief Data Architect

Intel

# Metadata Management

## Strategy / Approach:

- Leverage progress of TQdM program
  - The TQdM program at Intel has made data issues visible
  - Metadata management needs to intersect the positive aspects of TQdM
- “Repeatable process, standard deliverables”
  - Metadata cannot be managed without consistent processes and fixed deliverables

## Progress To Date:

- Metadata program established and funded
- Focus on data-oriented metadata management
- First release of Enterprise Metadata Repository completed

## Issues / Roadblocks:

- Culture rewards solving each problem anew
  - This is slowly changing; pushing Reuse Awards to recognize desired behavior
  - Creating metrics which can be used as carrot and stick
- Consensus management approach
  - Obtaining commitment to a proposed change is extremely time-consuming

# Example Screenshot from Intel's Enterprise Metadata Repository

The screenshot shows the Intel Enterprise Metadata Repository website. The header features the site name and a search bar with a 'GO' button. Below the header, the user is logged in as 'SIMPSON JR, JOHN E' and the date is '10/7/2003 9:46:07 AM'. A left sidebar contains navigation links under 'What do you want to do?' and 'Administrative Only'. The main content area has a 'WELCOME' message and six featured sections, each with an icon and a brief description.

**enterprise metadata repository** search [advanced] GO

Metadata as of: 10/7/2003 9:46:07 AM SIMPSON JR, JOHN E logged in as Repository Admin

**What do you want to do?**  
[Home]  
Browse Object Hierarchy  
Request Training  
Run a Report  
Search For Objects  
View FAQ's/User Help  
View Metrics

**Administrative Only**  
Audit The Repository  
Check In  
Check Out  
Manage Code Tables  
Populate Seed Model  
Template File Load

**WELCOME** to the new Enterprise Metadata Repository  
Your search for high-quality, reusable data stops here!

**browse**  
**Browse Object Hierarchy**  
When you are unsure of the desired object to view, drill into our catalog of items in order to find objects that you are looking for. Browsing is an intelligent method for drilling into items based on categorization versus blind searching.

**request**  
**Request Training**  
Even the most intuitive system with highly skilled workers require a little assistance. See what training is available as well as what you can do to increase your knowledge in the area of Metadata at Intel.

**reports**  
**Run a Report**  
Data is useless until you place in in categories, analyze, and create information from it. Reports - yes, we have those too.

**search**  
**Search for Objects**  
Straight text search that looks in the description as well as the name of all objects selected.

**help**  
**Frequently Asked Questions**  
Help is supplied at your fingertips enabling to be more effective in the use of the repository.

**metrics**  
**Metrics**  
Remember those reports? Now think about taking data, and running analysis on those, crunching some numbers, such that some decisions can be made without being in the dark.

Copyright © Intel Corporation, 2003. All Rights Reserved.

# Information Resource Round Table Metadata Management

Presented by: Juanita M. Mercado  
Lead Data Architect  
2004-Feb-04

# Visa Global Business Elements (VGBE)

## *Definition Metadata*



### ■ What it is

- Formal specification about ways to accurately and unambiguously describe of information elements that are critical to VisaNet interoperability
- Standard definitions for meaning, acceptable content and relationships

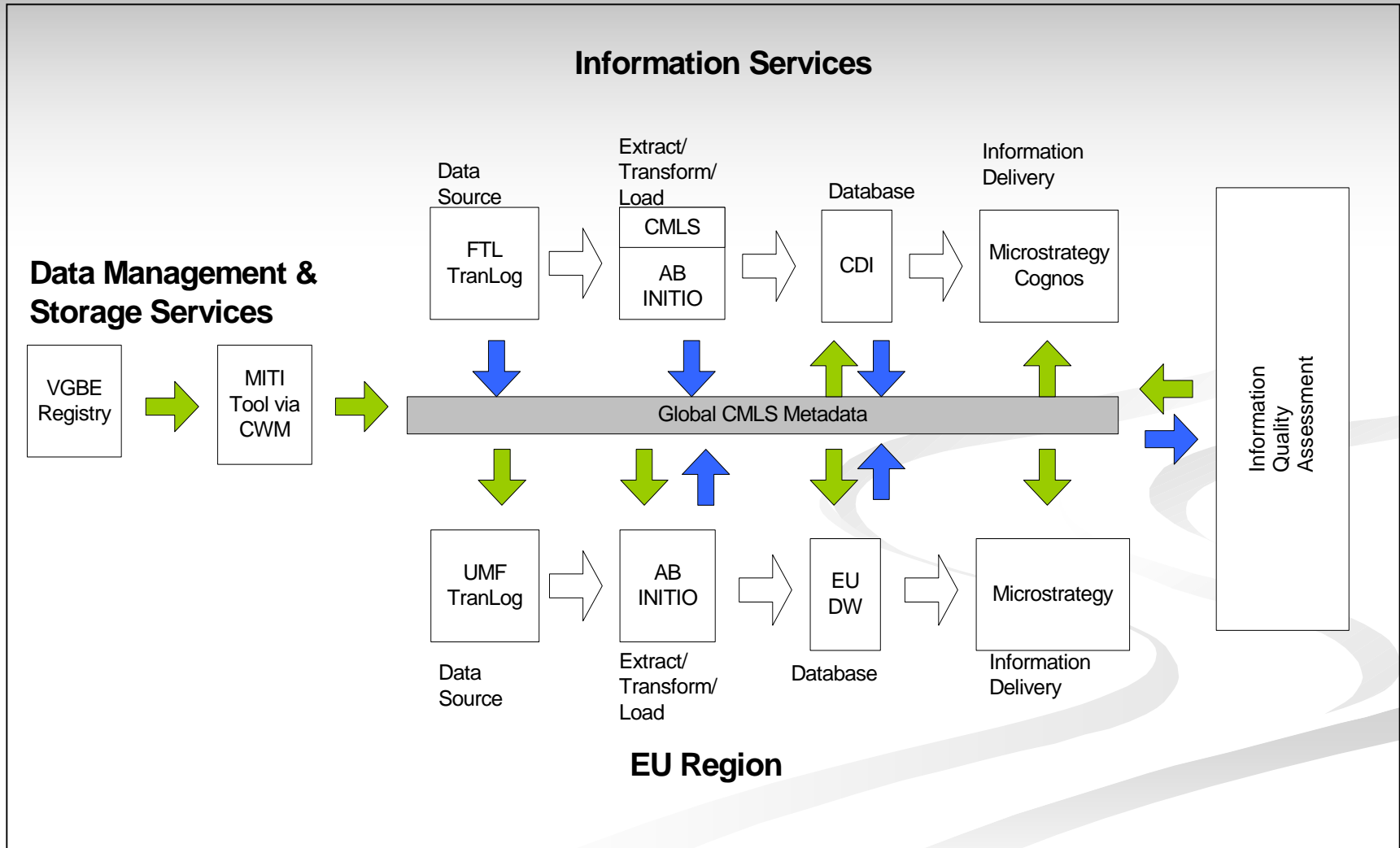
### ■ What it accomplishes

- Share metadata consistently across Visa and with IT partners
- Assures consistent characteristics and behavior for all implementations
- Extensible and flexible to support evolving business strategies

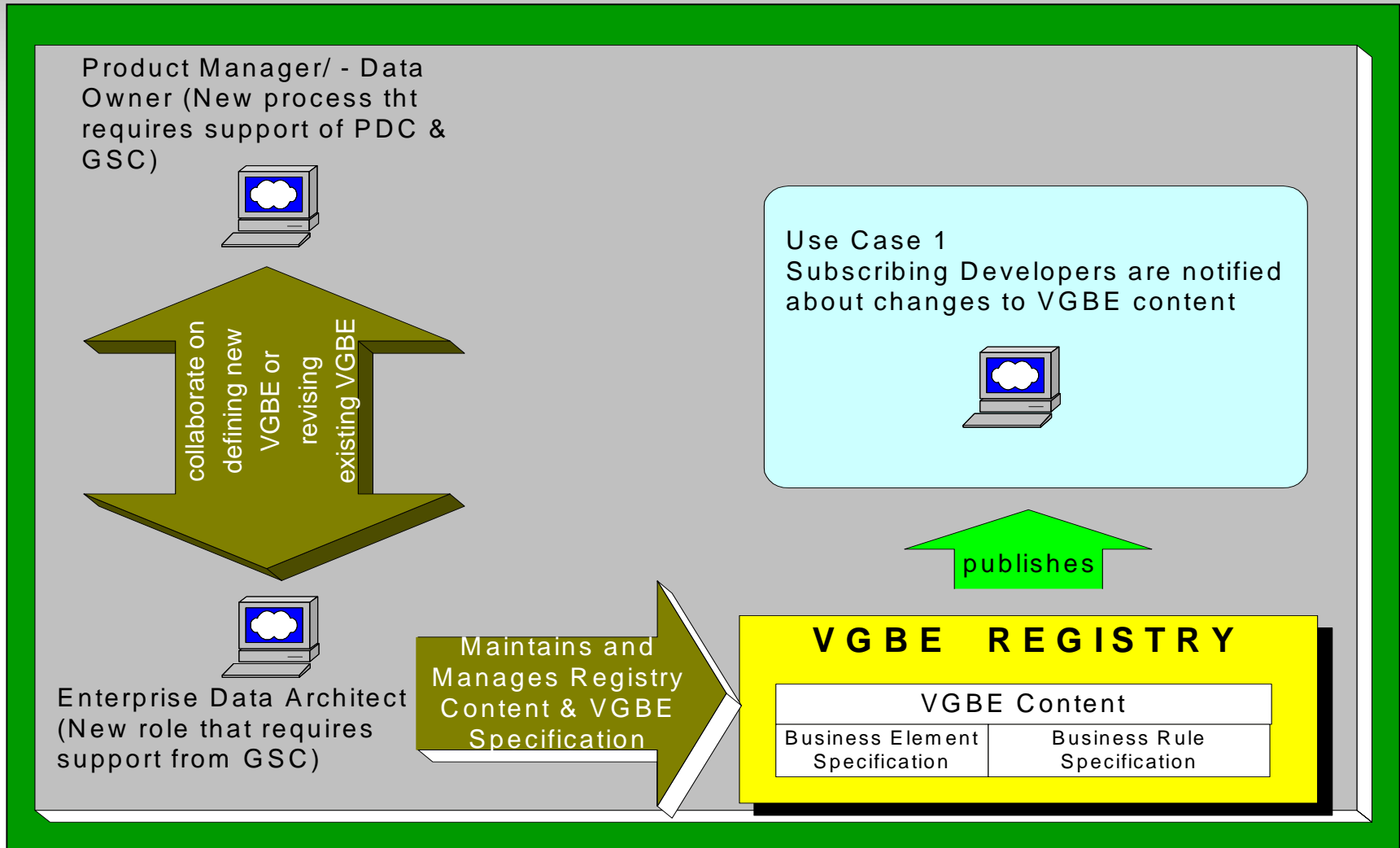
### ■ How it works

- A structurally stable yet dynamic document (UML) that integrates into development environments
- Automates inheritance of definitions, behaviors and relationships directly into application codes

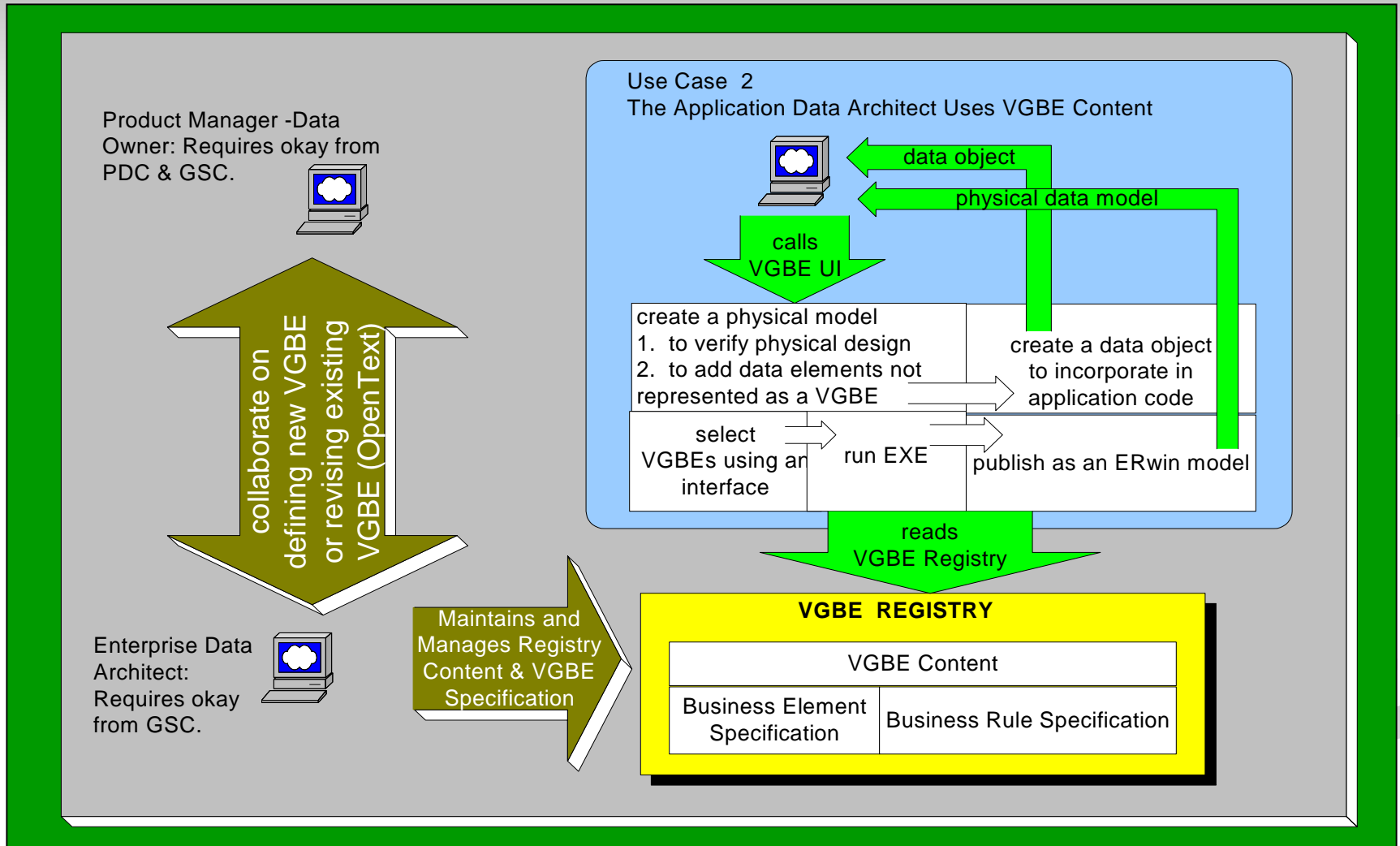
# High Level View Proposed Metadata Controls



# Proposal to Integrate Metadata Control Service: Use Case 1



# Proposal to Integrate Metadata Control Service: Use Case 2



# Metadata

## The Promise Versus The Reality

Cass Squire

Associate Partner

IBM Business Consulting Services

(650) 520-7247

[csquire@us.ibm.com](mailto:csquire@us.ibm.com)

# Topics

- The Need
- A Strategy & Approach for Solving it
- Progress To Date
- The Real World: Issues and Road Blocks

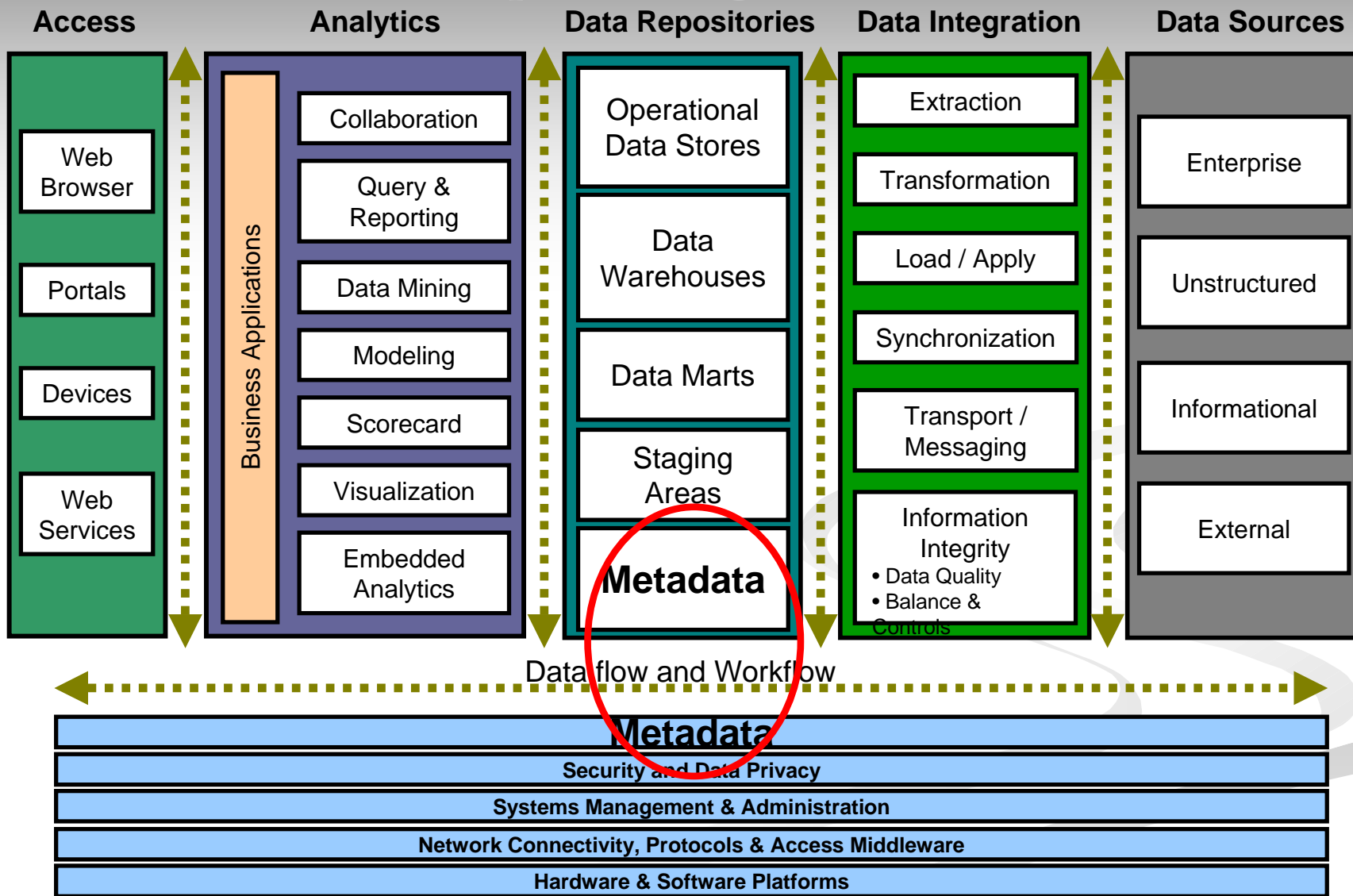
# The Need

- An entirely new set of data for a new kind of analytics is being rolled out
- The business community has not been properly engaged
- The business community needs to understand:
  - what data is available
  - What it means
  - What its source is
  - What its currency is
  - Who to go to to ask questions about the data and its meaning
- A new tool for querying (Business Objects) is being rolled out as well

# Strategy & Approach

- A clear need for a mechanism for capturing and sharing metadata surfaces as essential to the successful roll out of the new analytical environment
- AbInitio is the corporate ETL tool – leverage its metadata for technical metadata
- Determine the applicability of its repository – the Enterprise Metadata Environment (EME) for serving as the repository for all metadata
- ERwin contains Business and Technical names, definitions, and allowable values – use it as the source for this metadata
- Determine Data Stewards in the both the business and technical arenas to be the go-to people for questions
- Evaluate other tools for applicability
- Integrate Operational Metadata and SOX auditability
- KISS

# In IBM's Business Intelligence Reference Architecture, Metadata is one of the components that glues the whole process together.



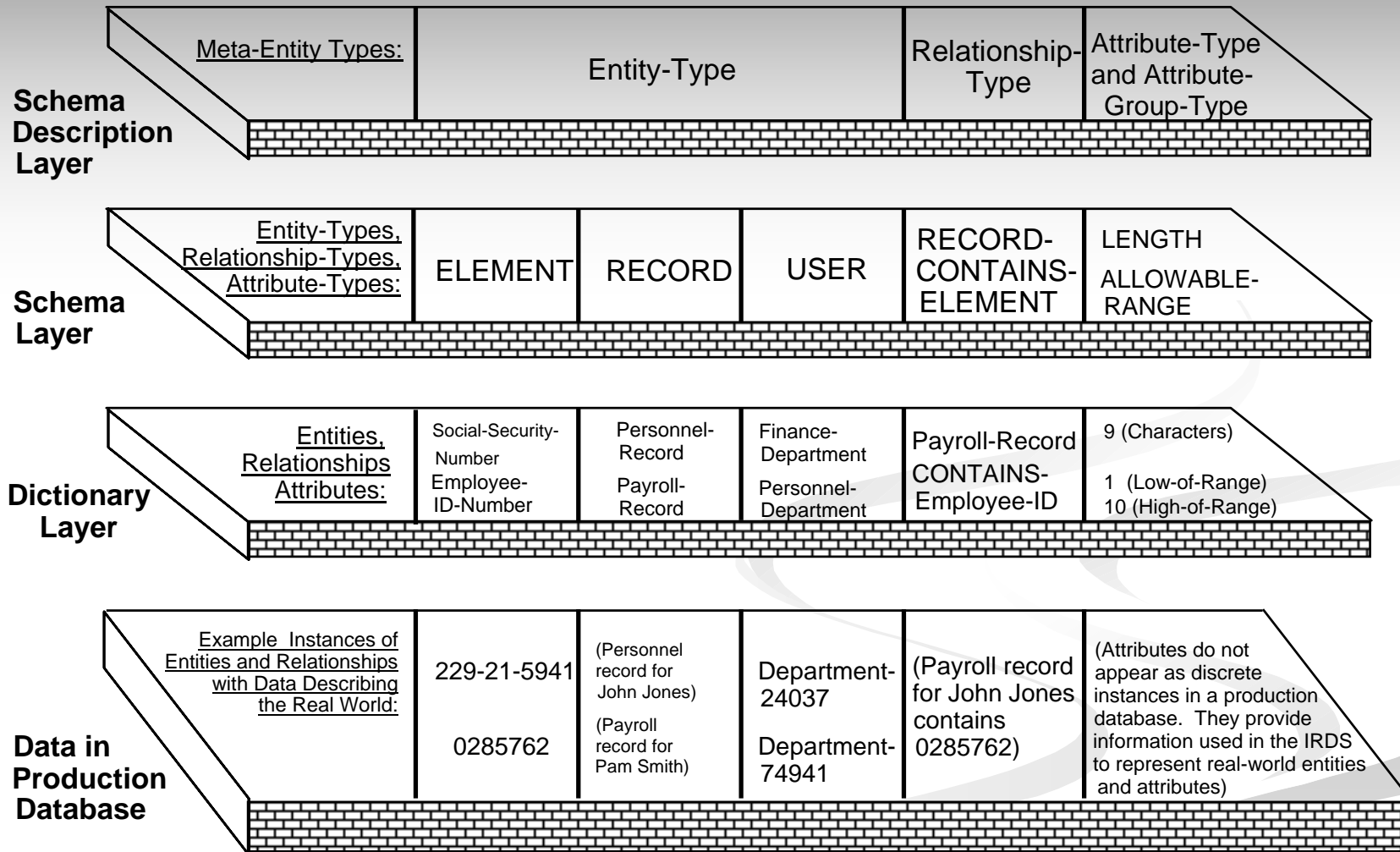
# Progress to Date

- AbInitio EME determined to be easily enough extensible to hold all metadata required – at least for initial phases
- Their web-based reporting is deemed acceptable for technical users but not for business users
- The ODBC API into their flat-file based repository is new and performance is “not ready for prime time” yet
- The decision was made to extract from the EME to relational tables (15 +/-)
- Business Objects (web version) as the user interface to the metadata since that’s what users will use to see the data – single universe and a dozen or so queries
- Extracts from the EME into the relational tables have been built.
- Data lineage simplified to ultimate source to ultimate target (intermediary ETL steps hidden) for business users
- Processes and extracts built for getting data from ERwin into the repository
- Processes for ensuring data analysts on all projects use the same tools/processes for capturing and publishing metadata for the repository
- Identified Data Stewards and created a cross reference between them and the entities
- Business users have applauded it as very useful in helping them understand and use the new data for analytics
- Common processes for capturing Operational Metadata (Statistics, Error Reporting, Auditing) built and used by every ETL process
- Unicorn evaluated as a possible tool – received high praises especially for its ability to speed up the mapping process - but the determination was made to postpone further testing of it until the above environment is in production for awhile

# The Real World: Issues and Road Blocks

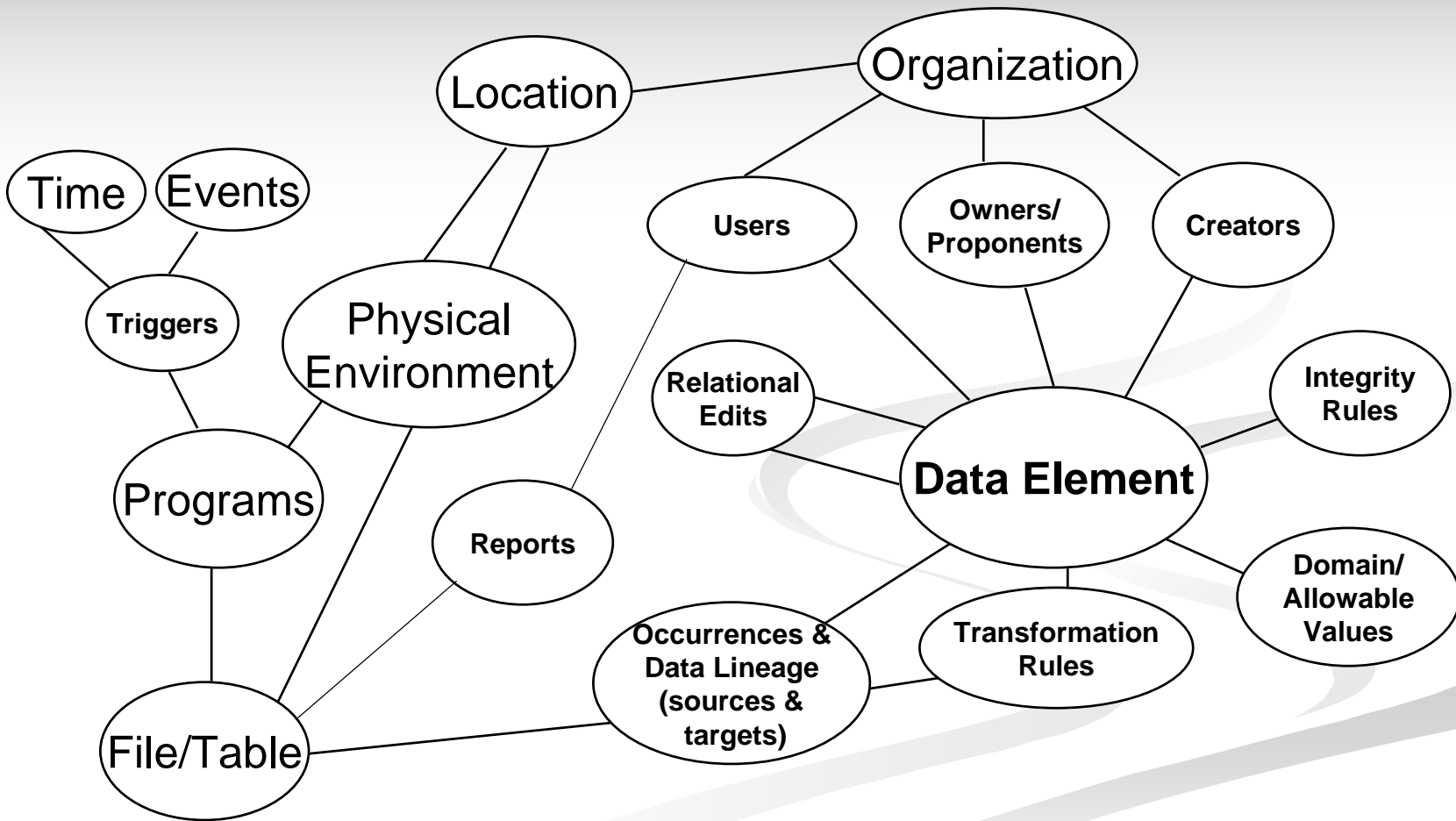
- Lots of hype about metadata – little in the way of tools to deliver
- There are lots and lots of type of metadata – picking the right subset to implement is key
- Ensuring automated maintenance is key
- New data unfamiliar to the users
- Demographics data – initial rollout a fiasco; queries produce wildly inaccurate numbers because the data is not well enough understood
- Lots of churning while technical team got enough understanding of the data to identify the problems
- User confidence in the data seriously hurt
- Education program in the data and what queries would generate correct results required

# Layers & Perspectives of Data & Metadata



The 1984/5 (+/-) Information Resources Dictionary Standard (IRDS) was an attempt to define a syntax for metadata exchange.

# Simple Metadata Model



# John Zachman's Enterprise Architecture Framework

also provides us a way to categorize the sources of metadata.

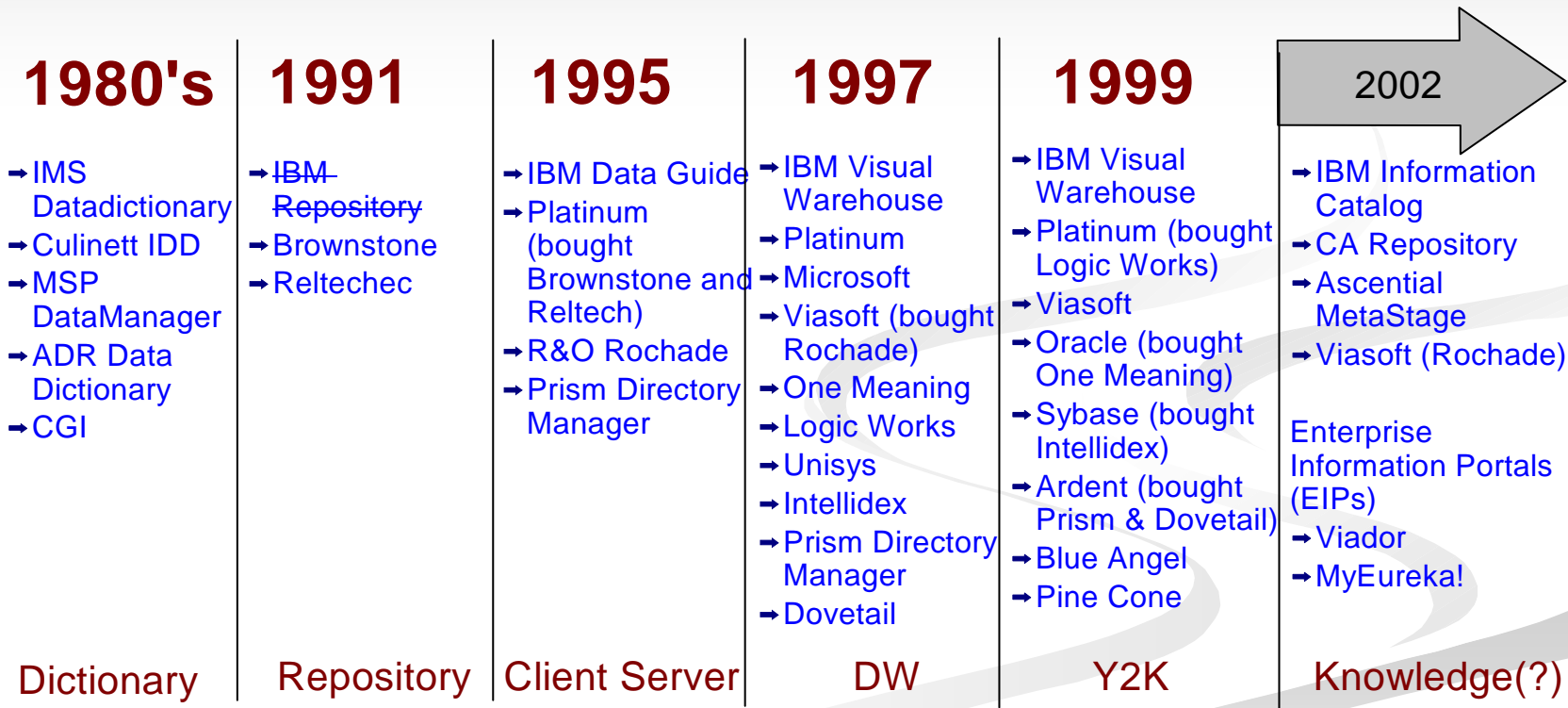
ENTERPRISE ARCHITECTURE - A FRAMEWORK <sup>TM</sup>							
	DATA <i>What</i>	FUNCTION <i>How</i>	NETWORK <i>Where</i>	PEOPLE <i>Who</i>	TIME <i>When</i>	MOTIVATION <i>Why</i>	
SCOPE (CONTEXTUAL)	List of Things Important to the Business 	List of Processes the Business Performs 	List of Locations in which the Business Operates 	List of Organizations Important to the Business 	List of Events Significant to the Business 	List of Business Goals/Strat. 	SCOPE (CONTEXTUAL)
<i>Planner</i>	ENTITY = Class of Business Thing	Function = Class of Business Process	Node = Major Business Location	People = Major Organizations	Time = Major Business Event	Ends/Means = Major Bus. Goal/Critical Success Factor	<i>Planner</i>
ENTERPRISE MODEL (CONCEPTUAL)	e.g. Semantic Model 	e.g. Business Process Model 	e.g. Logistics Network 	e.g. Work Flow Model 	e.g. Master Schedule 	e.g. Business Plan 	ENTERPRISE MODEL (CONCEPTUAL)
<i>Owner</i>	Ent = Business Entity Rel = Business Relationship	Proc = Business Process RO = Business Resources	Node = Business Location Link = Business Linkage	People = Organizational Unit Work = Work Product	Time = Business Event Cycle = Business Cycle	End = Business Objective Means = Business Strategy	<i>Owner</i>
SYSTEM MODEL (LOGICAL)	e.g. Logical Data Model 	e.g. Application Software 	e.g. Distributed System Architecture 	e.g. Process Architecture 	e.g. Programming Structure 	e.g. Business Rule Model 	SYSTEM MODEL (LOGICAL)
<i>Designer</i>	Ent = Data Entity Rel = Data Relationship	Proc = Application Function RO = User Views	Node = ER Function (Database Schema Unit) Link = Line Characteristics	People = Role Work = Deliverable	Time = System Event Cycle = Processing Cycle	End = Structural Assertion Means = Action Assertion	<i>Designer</i>
TECHNOLOGY MODEL (PHYSICAL)	e.g. Physical Data Model 	e.g. System Design 	e.g. System Architecture 	e.g. Presentation Architecture 	e.g. Control Structure 	e.g. Rule Design 	TECHNOLOGY MODEL (PHYSICAL)
<i>Builder</i>	Ent = Segment/Block Rel = Relationship	Proc = Computer Function RO = Screen/Device Formats	Node = Hardware/Software Link = Line Specifications	People = User Work = Screen Format	Time = Epoch Cycle = Component Cycle	End = Condition Means = Action	<i>Builder</i>
DETAILED REPRESENTATIONS (OUT-OF-CONTEXT)	e.g. Data Definition 	e.g. 'Pages' 	e.g. 'Network Architecture' 	e.g. Security Architecture 	e.g. Timing Definition 	e.g. Rule Specifications 	DETAILED REPRESENTATIONS (OUT-OF-CONTEXT)
<i>Sub-Contractor</i>	Ent = Field Rel = Address	Proc = Language Struct. RO = Control Block	Node = Address Link = Protocol	People = Manly Work = Job	Time = Interval Cycle = Machine Cycle	End = Subcondition Means = Step	<i>Sub-Contractor</i>
FUNCTIONING ENTERPRISE	e.g. DATA	e.g. FUNCTION	e.g. NETWORK	e.g. ORGANIZATION	e.g. SCHEDULE	e.g. STRATEGY	FUNCTIONING ENTERPRISE

# The Enterprise Architecture defines five views and six aspects of the enterprise.

View	Data	Function	Network	People	Time	Motivation
Planner	Subject Area List	Business Process List	Business Location	Organization List	Significant Events	Business Goals List
Owner	E-R Diagram	Functional Flow Diagram	Logistic Network	Organization Chart	Master Schedule	Business Plan
Designer	Data Model	Data Flow Diagram	Distribution System Architecture	Human Interface Architecture	Processing Structure	Knowledge Architecture
Builder	Data Design	Structure Chart	System Architecture	Human Technology Architecture	Control Structure	Knowledge Design
Sub-Contractor	Database Description	Language Statement	Network Architecture	Security	Timing Definition	Knowledge Definition
Enterprise	Data	Function	Communications	Organization	Schedule	Strategy

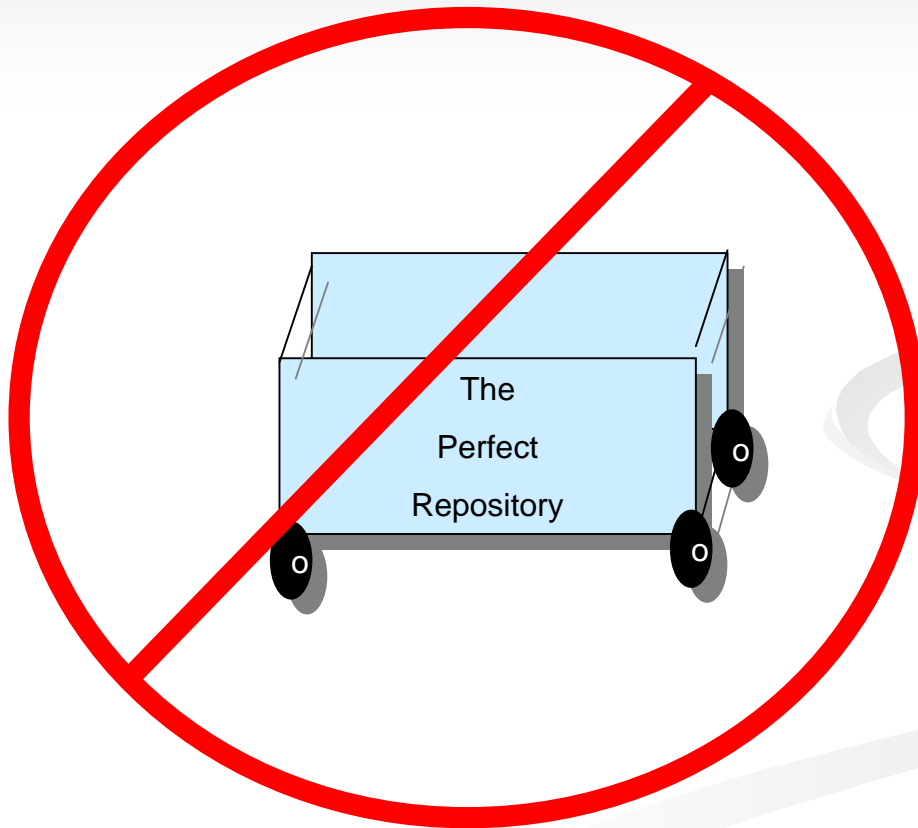
- John Zachman compares delivering technology to the enterprise to building an airplane. It is a complicated task involving several stages of design and many builders whose activities must be coordinated. He asks: Who would build an airplane without conceptual drawings? Without detailed sub-assembly charts? His premise is that the IT industry often has these design drawings and sub-assembly charts, but fails to file, cross-reference and maintain them.
- Not only do IT practitioners need to keep multiple views of the enterprise, the relationships between the cells in the framework must also be tracked. It is important to know which functions use which data elements.
- He says that as builders of data warehouses we have many of these “specification sheets” and he states that they contain metadata.

# Over time, repositories and metadata management tools have changed with, in spite of, or regardless of, the IT industry focus.



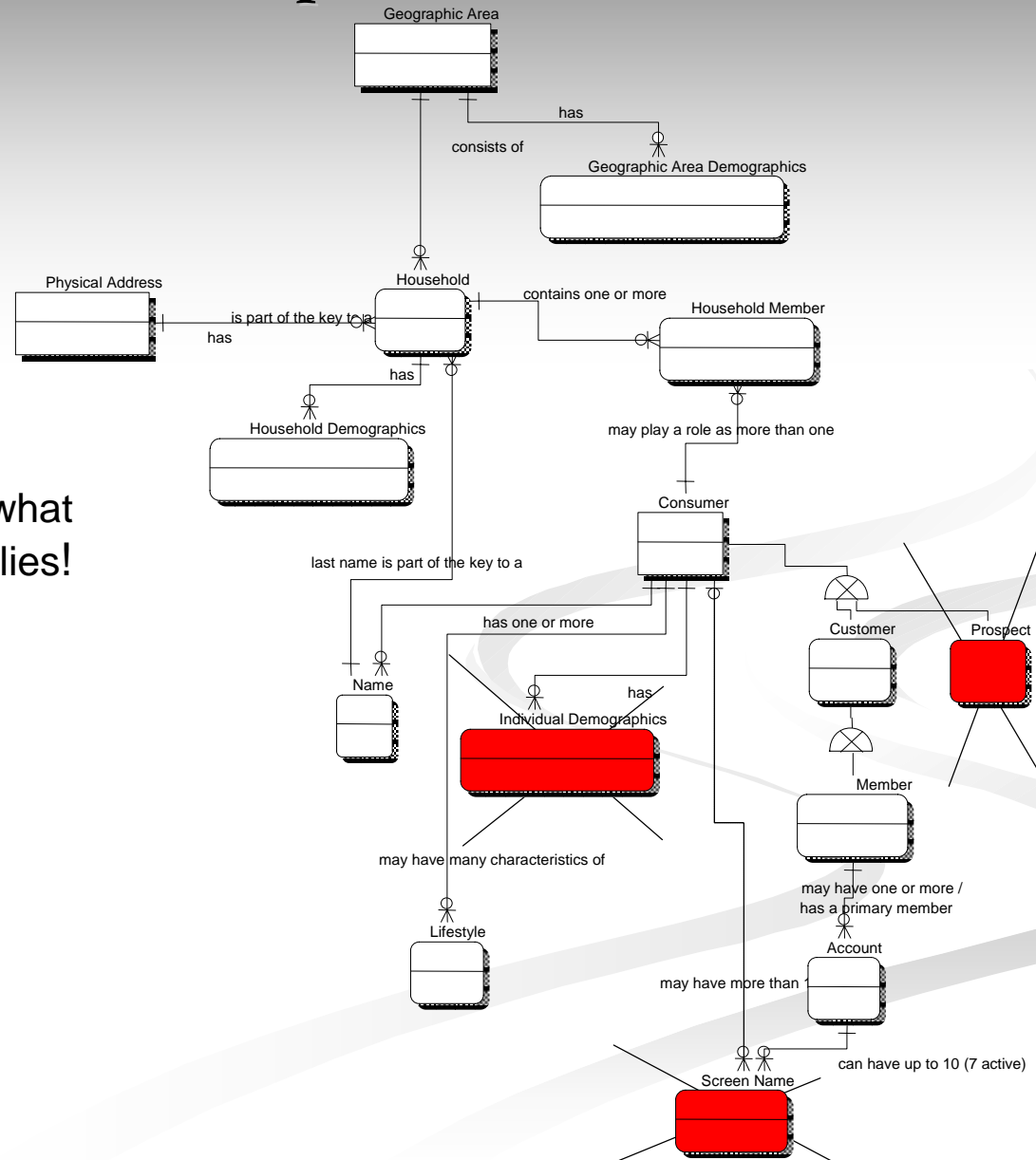
# The Need and Promise is Great.

The delivery isn't there yet.



- They often take more effort to feed than the benefit derived
- Many repository tools/vendors won't expose (share) their metadata
- They tend to be passive, and thus can get out of synch with the real world

# Demographic Data: Conceptual Model

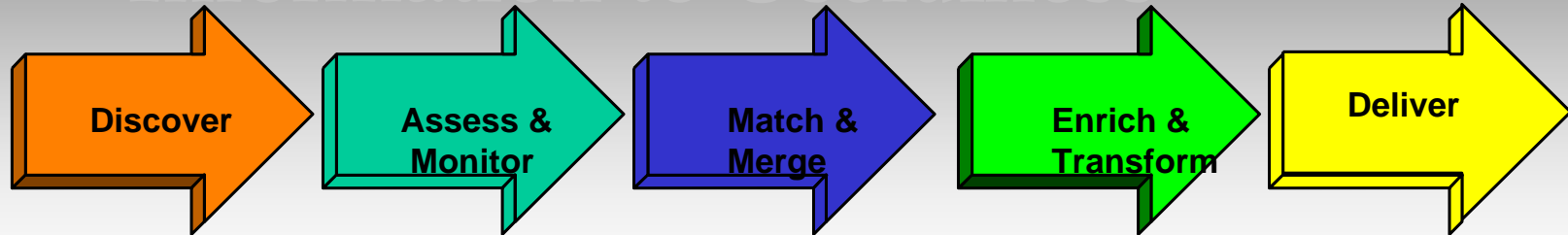


Make sure you know at what level the data applies!

# How to implement?

- “It’s like pinning jello to the wall”
- There are no “best practices”
- Are there analogies we can use?

# From Raw Data to Standardized Information to Usefulness



<p><b>The Problems:</b></p>	<p>What's in the source data? Does it mean what you think it should? How is it structured? How might it be structured?</p>	<p>Does it contain what you think it should? How complete is it? How clean is it? Does it follow the business rules? How is the quality changing over time?</p>	<p>Resolve duplicates. Standardize names. Assign unique Id's. Identify households.</p>	<p>Correct and improve it. Change to standard values. Transform codes to meaningful terms, Summarize it.</p>	<p>Deliver new sets of data on a periodic basis. Capture changes as required. Deliver updates/new transactions in as timely a fashion as required.</p>
-----------------------------	--	---	--	--	--

<p><b>Sample Tools:</b></p>	<p>ProfileStage (MetaRecon) Evoke Axio</p>	<p>AuditStage (Quality Manager) Ab Initio Data Profiler</p>	<p>QualityStage (Integrity) Trillium First Data Innovative Systems</p>	<p>DataStage Informatica Ab Initio Warehouse Manager</p>	<p>ETL tools, Propagation, Change Data Capture tools, MQ</p>
-----------------------------	--	---	--	--	--

# Data Becomes Information If and Only If You:



## The Problem

## Tools, Techniques, & Processes

1. Have the data and

1. Capture Process; Business  
Process Re-Engineering

2. Know you have it and

2. Metadata; Evangelism

3. Can access it and

3. BI Environment; Data  
Structured for Access;  
End-User Analysis tools

4. Can use it and

4. Business Metrics Captured

5. Can trust it!

5. Data Quality Process;  
Metadata