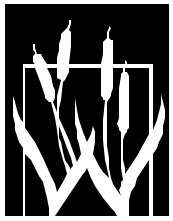


Whitemarsh
Information Systems Corporation

Achieving Data Standardization

Whitemarsh Information Systems Corporation
2008 Althea Lane
Bowie, Maryland 20716
Tele: 301-249-1142
Email: mmgorman@wiscorp.com
Web: www.wiscorp.com

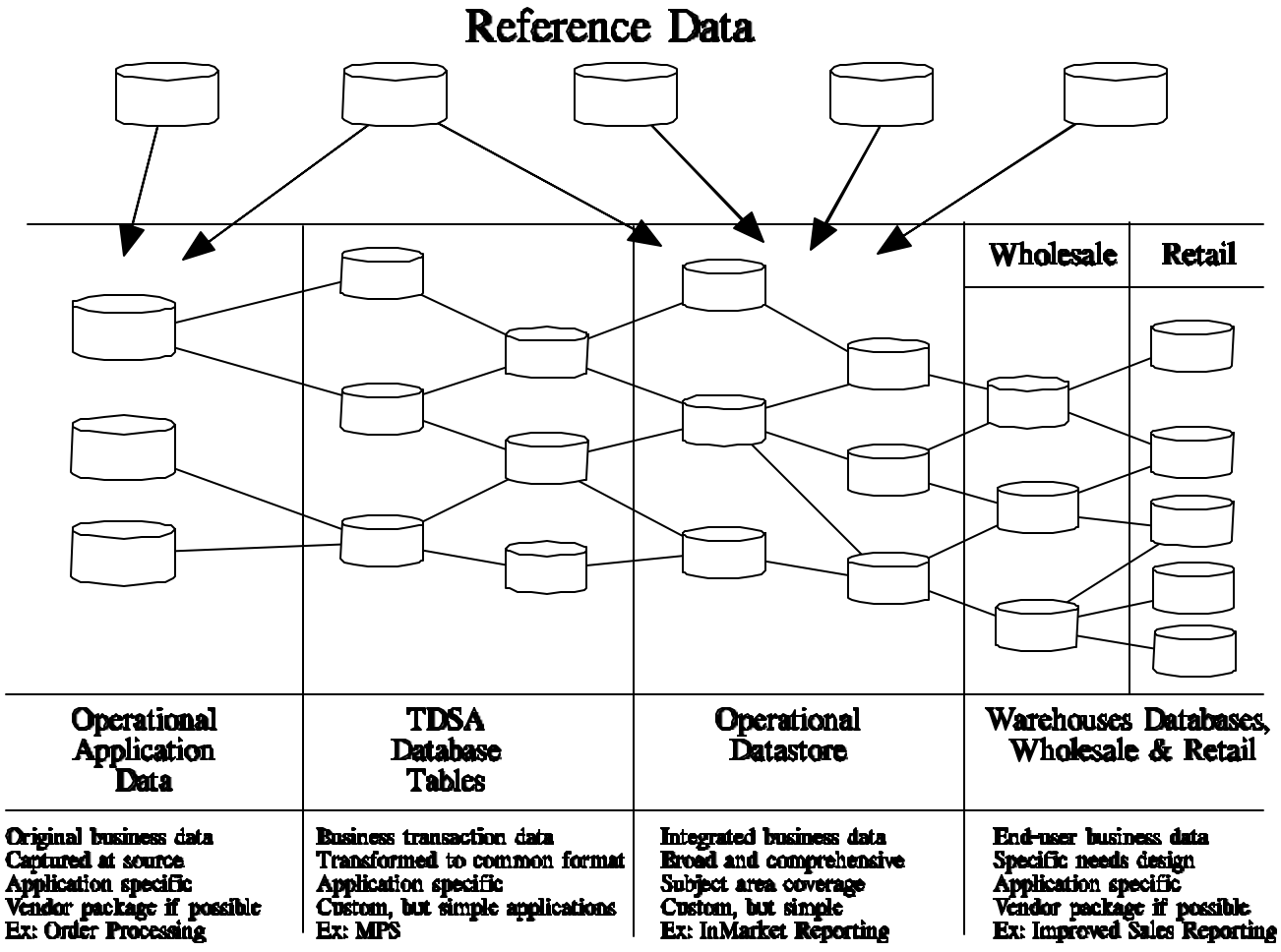
1.0 The Issue

If database objects are taken seriously and their implementation cannot succeed without:

- a standard data architecture,
- an approach to accommodate diverse data naming, and
- the ability to meld data across multiple database classes

Data is executed policy. Data definitions are the technical representations of policy specifications. Policy executions, that is data, are the medium of business communication.

2.0 Data Architecture Classes



2.5 Benefits from Distinct Data Architecture Classes

- People have a better understanding of business and its state because data is accepted and understood by wider audience.
- Accelerates incorporation of new data and new uses of old-data because data sources are able to be quickly understood and are seen through standardized value discriminators.
- Reduces size of data by eliminating redundant data or data that is different merely for reason of style
- Frees up staff time to work on real business problem areas rather that ferreting out the same data hidden under different names.
- Increases the quality of decision making because data is valid, reliable, and represents discriminating facts about business activities

3.0 Semantics

- Semantics: Rules for meaning and usage
- Data Semantics: semantics for persistent data acquisition, storage, manipulation, and reporting
- Process Semantics: semantics for data transformations

3.1 Data Semantics Problems

When the same concept is known under two different semantic representations

Data semantics irregularities are most commonly evidenced through differences in:

- data names
- data types
- data lengths/precisions
- data structures

Examples:

Name Problems: SSN vs Social Security Number

Type Problems: SSN INT(9) vs SSN Char (11)

Precision Problems: ANNUAL_OVERTIME_HOURS (Decimal)

Structure Problems:

Inventory Quantity January, ..., Inventory Quantity December
Vs
Inventory Year, Inventory Month, Inventory Quantity

3.2 Process Semantics

Process semantics are the rules that govern data transformation. Typical problems include:

- Improper specification
- Wrong placement leading to multiple, but different executions
- Improper maintenance and evolution

Examples:

Improper Specification: Computing an average but not considering null values to reduce count

Encoding different data quality process in different programs. Given: Sex Char(1),

Sex: Y|N
Sex: M|F
Sex: 1|2

Updating 98% of all the instances of someone's address. That is, not having "the golden source" and then reference data replication.

3.3 Testimonials????

- **50% of all software costs are attributable to error corrections–The U.S. DoD**
- **60% of all corporate IS budgets are devoted to correction of errors–Software Quality by Mordecai Ben-Menachem and Gary Marliss**
- **Software maintenance time: 47% analysis, 28% testing & debugging, 19% coding, and 6% documentation--errors–Software Quality by Mordecai Ben-Menachem and Gary Marliss**
- **700,000 Americans were shorted $850,000,000 in Social Security payments due to software error.–U.S. Social Security Administration**
- **3 times more errors are introduced during maintenance than in original development.**

Data Standardization is not just an abstract concept...

1. A single data conversion and/or reformatting program is about 20 pages @ 50 lines per page. At 10 days of staff time per program (they're pretty simple), the cost would be about $5,000 per program.

2. A single Government Agency currently spends $1,000,000,000 per year in such programs.

3. That's 2,000,000 programs per year!

4. For $1 Billion, you can:

    - Build 5,000 houses
    - Educate 100,000 high-school students
    - Build 1/5th of an aircraft carrier
    - Buy 20,000 Mercedes

4.0 Common Reasons for Data Standardization Failures

- A fundamentally flawed data standardization model
- No accommodation for enterprise wide data architectures
- Multiple implementation technologies
- Central standardization and maintenance authority