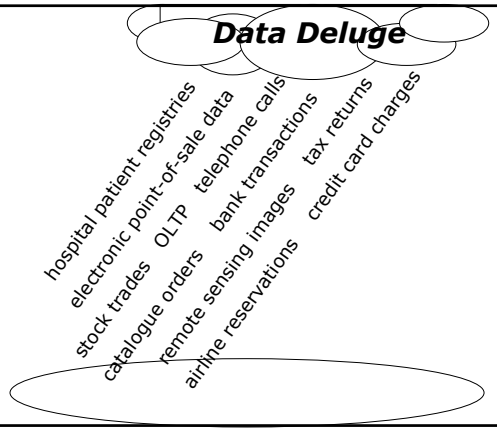
Data Warehouseing and Data Marts

Andrea Ahlemeyer-Stubbe
Vice President
European Center of Database Marketing
Member of Pro-ENBIS and ENBIS



"Only 40% of companies are 'very confident' with their own data quality."

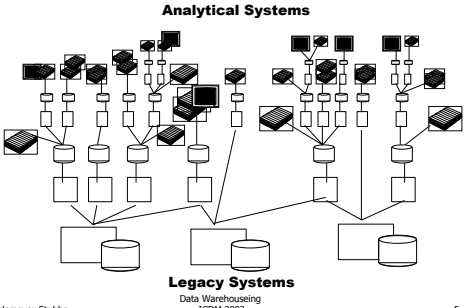
Pricewaterhouse Coopers
Data Management Survey, 2000

Customer focus impact

Customer focus results in:

- Data on more products and needs
- More history, more complex concepts of customer
- Data from more sources

Realistic IT-Horrorszenario



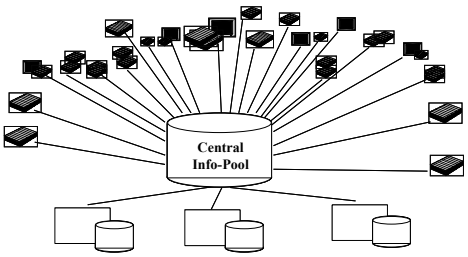
Solution(?): Data Warehouse

Data Warehouse (DW) =

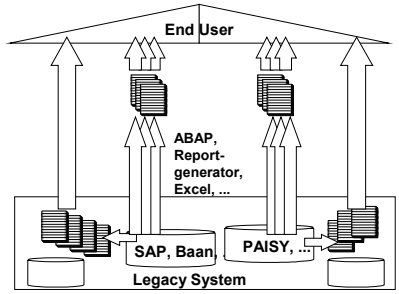
"A subject-oriented, integrated, non-volatile, time-variant collection of data organized to support management needs"

Inmon, Database Newsletter '92.

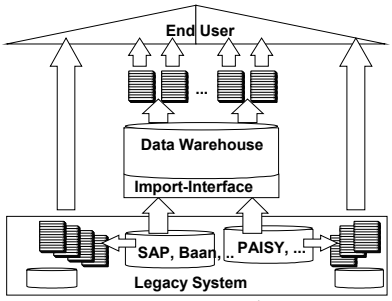
Ideal IT-World



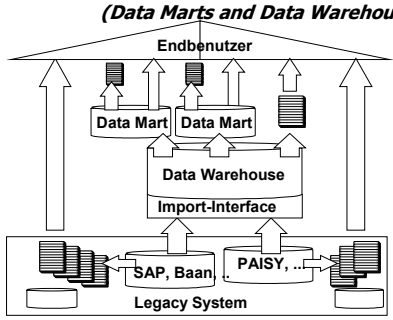
Architecturvariants (without Data Warehouse)



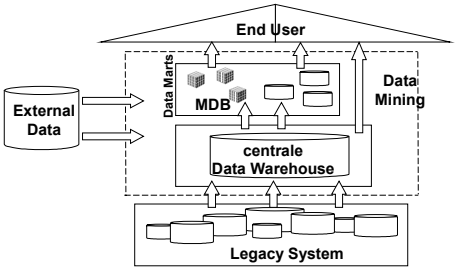
Architecturvariants (enterprisewide Data Warehouse)



Architecturvariantes (Data Marts and Data Warehouse)



Referencearchitectur



CRM and Data Quality

- The number one impediment to successful CRM programs is bad data.
- On average corporate databases are less than 75% accurate across the six key indices of data quality
  - Name Standardization
  - Address Hygiene
  - Demographic Accuracy
  - Completeness
  - Transaction Accuracy
  - Linkage (Household / Enterprise)

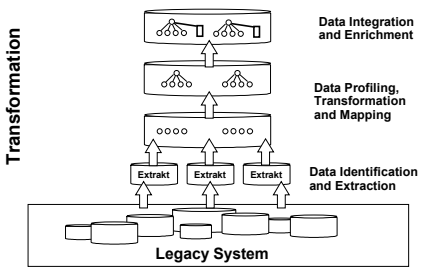
CRM and Data Quality

- CRM is by design a "entity based" process, therefore accurate, timely and complete entity identification is imperative
- Entity based data decays at a rate of 2 1/2 % to 3 1/2 % per month compounded

Objectives Objectives

- Review the most important element in a CRM system: DATA.
- Examine how external data combines with internal data to drive relationship applications.
- Understand the importance of Customer Data Integration (CDI) and how data quality and hygiene is critical.
- Explore applications which derive increased value through combining various data elements.
- Examine data in the context of both Business to Consumer and Business to Business Marketing.

ETL-Process



Database Requirements

Database Construction

Database should be able to handle and manage the complexities of:

- Transforming Data
- Cleansing, standardizing, and matching names & addresses
- Creating & maintaining constant relationship keys over time
- Support multiple, hierarchical definitions of the customer
- Identify significant changes in customer data from one cycle to the next

Refinery Requirements

- Data from internal & external sources should be collected, integrated and organized in a manner according to predefined business requirements stored as metadata in one central location

Refinery Requirements

Features & Functionality

- ETL - Extract, Transformation & Load
  - Data Identification and Extraction
  - Data Profiling, Transformation and Mapping
  - Data Integration
- Refinery MetaData Creation & Management
- Ease of Use and Development
- Scalability and Performance
- Operations Management, Target Loading Options

Refinery Requirements

Data Cleansing/Preparation (Name & Address Focused)
- Scrubbing
- Matching
- Householding (Consumer)/Siting (Commercial)

Process Flow Architecture
- Seamless
- Integrated
- Modular
- Batch and On-line/Real time

Refinery Requirements

Data Discovery
- Data Viability Analysis
- Data Element Population
- Data Hygiene Methodology
- Entity Identification
- Refinery Design

Refinery Requirements

Data Optimization
- Data Process Flow
- Edit and Validation Analysis
- ETL Optimization
- Meta Data Review
- Transformation Design
- Source Diagnostics

Unique Number Identification

- Refinery process drives unique number identification
- stable keys are established and matched against a reference file
- Customer specific unique numbering is employed where reference file can not maintain stable keys
- cross reference files are established where legacy systems can not support uni codes
- front end validations are much more effective than back end validations

Data is Energy

"Data is like kinetic energy. When it is stored in the database (I.e., the battery), it is only "potential" energy -- useless until accessed. Once data is connected into the application, the value comes from the kinetic flow of data into decision processes, marketing activities, and use."

Types of Data

Internal (Customer)
- Accounts Held
- Account Balances
- Transaction Data
- Promotion History
- Customer Service
- Demographics
- Purchase Behavior
- Address (street, telephone, email)
- Profitability/Lifetime Value

External
- Demographics
- Socioeconomic Data
- Lifestyles/ Psychographics
- Firmographics/Technographics
- Geography
- Address (street, telephone, email)

Value of Comprehensive Data

- Grow your customer base by efficient prospecting of qualified customers
- Keep customers by serving their needs through marketing appropriate products and services at the right time
- Up-sell and cross-sell to customers who need and expect the opportunity
- Increase share of wallet and life time value

The Fundamentals…

- Database design matters! No tool can overcome a poorly designed database
- A clear understanding of the business by the people designing the database is critical
- It takes more than loading lots of data into Oracle to make a marketing database

How to tell when your database is a failure:

- Users hate the database/won't use it
- It doesn't do what you want
- Only the IT guys know how it works
- Printout of ERD runs your printer out of toner
- Failure is not an option

How do you avoid failure?

- Blame it on somebody else
- Pay attention to the fundamentals
- Hire smart people and let them do their jobs
- Don't get too fancy
- Choose your supplier carefully
- Deliver a Functional specification

A few points about marketing databases

- They are different than a normal "relational" database
- Data warehouses don't make good marketing databases
  - Too detailed, too normalized
- No software product can overcome a poor database design
- Design of a database is predicated on business rules
- Not all processing has to be done as a function of the database

DO walk through critical reports and campaigns early in the design process

- Make sure the database design supports your business needs
- Step through sample queries and campaigns and make sure the data structure supports them
- The easier it is to perform the query, the faster the database will run

Make the database easy to use

- Data should be stored such that it's convenient to query/access, not so that it's easy to manage
- Segregate name/address data from query data
  - You don't use street names in cross tabs
- Don't use NULLs if you can avoid them
  - (This will cause most DBAs to go nuts)
  - Most marketers don't understand them anyway
  - Store empty strings as blank, numbers as zero

Use meaningful field names

- Get 'em right now 'cause you'll be looking at them for a LONG time
- Adopt standards
  - Dates end with "_DT", codes with "_CD", etc.
- Be Consistent
  - Don't say SEX in one place and GENDER in another
- Don't use dumb names like "CODE"
- Try to be consistent with your internal database to avoid confusion

Manage Expectations

- You WILL forget things
- You WILL make mistakes
- It's NOT the end of the world
- Deal with it
- Have contingency plans
- Have a disaster recovery plan
- Understand early on who is responsible to cover incurred costs

Processing efficiencies with CDI

Reduce redundant processing

- Once data has been assigned a unique ID, only updates need to be made to the database
- Based upon your industry, approximately 10% of records are updated on an annual basis.
- Traditional processing requires cycling through 90% of records unnecessarily
- Eliminate redundant data storage
- Duplicate records can be merged or eliminated to better manage large data volumes

Processing efficiencies with CDI

- Synchronize data across the organization
- Less physical movement of data is required to match records from data silos across your organization
- Once the data has been assigned a unique ID the ID can be used to link data regardless of the physical location
- Traditional processes require all data to be cycled through the matching process
- With large data volumes often data must be segmented for processing resulting in missed links
- Regional segmentation for matching misses links when the consumer moves outside of the region

Real-time access via CDI

- Reduction in cycle time required to load data to the data warehouse
- Instant database update
- Identification of consumers not currently resident within the data warehouse
- Assign unique IDs to new consumers
- Combine information from the current transaction with historic information to enable real-time scoring and continual learning

Large scale data warehousing trends increase focus on:

- Scalability: increased parallelism, better optimization throughout entire data warehousing process
- Efficiency in index structure and access technique
- Incremental refresh

Large scale data warehousing trends increase focus on:

- Near real time update
- Movement of OLAP & Data Mining into database
- Continuous availability
- Correct data due to "tactical" or "active" uses
- Self-managing systems

"75% of respondents reported significant problems as a result of defective data. The same proportion said they realized benefits from effective data management."

Solution

Data Assessment Study consists of three parts:

- Quantitative data analysis
- Domain studies to thoroughly examine each row and column of each data table
- Validity tests to assess the overall accuracy and deliverability of the data
- Qualitative data usage analysis
- User workshops to identify how data is being used and how it could be used better
- Qualitative future usage review
- Joint sessions to identify new opportunities to leverage data both internally and externally

Analyze available data

- Evaluate the data quality
- Identify needs for supplemental data
- Quantify data errors and omissions

Integration of data

- Identify duplicate records
- Assign unique consumer identifiers
- Identify records to purge

- **Conduct a data assessment**
  - Integrate persistent ID into database
  - Persistent ID's can be used to link data from disparate systems
- **Analyze available data**
- **Integration of data**
- **Implementation of persistent keys**

- Update records versus refreshing entire database
- Eliminate need for refresh cycles and establish a continual update process
- **Integration of data**
- **Implementation of persistent keys**
- **Data Management**

Enhancing data

- Enhance with data when it is needed
- Prioritize needs for supplemental data

- **Conduct a data assessment**
- **Analyze available data**
- **Integration of data**
- **Implementation of persistent keys**
- **Data management**
- **Enhancing data**

Meta Data Management Tool Should:

- Share meta data across business intelligence and data warehousing tools
- Apply data modeling to the translation & transformation process
- Provide a record of conversion & complex data changes
- Perform full meta data exchange of both form & content

Data Rationalization Process

- Data Conversion
- Analysis and Cleansing
- Integration Processes
- Transformation
- Investigations
- Data Append
- Data Updates

Data Conversion

- Convert records and data into a standardized format
- Understand the information needs
- Do a complete data inventory early in the process
- Identify common data elements
- Identify what is unique to departments or applications
- Understand how data differs

Data Conversion

- Use the results of the data inventory to establish the data format
- Allow for formats that accommodate current and future data needs
- Give special attention to the key demographic data needed for integration
  - Business name
  - Tradestyles
  - Mailing address
  - Physical address
  - Telephone
  - ...

Analysis and Cleansing

Analyze the data to identify necessary repairs, additions and corrections

Analysis and Cleansing

- Profile data limitations and defects
- Truncated data
- Incomplete data
- Outdated data
- Identify unnecessary data

Integration Processes

Consolidate all information about the business to provide a complete understanding of the relationship

Integration Processes

- Machine matching / aggregation
- Manual scanning / review capabilities
- Match against high quality
- standardized reference file

Integration Processes

- Many views of a business are possible
- Legal name versus tradestyle
- Physical address versus mailing address

Transformation

Establish business rules for resolving conflicting data

Investigations

- Some records cannot be resolved through machine and manual scanning
- Use telephone investigations to
- Add additional data
- Correct data

Data Append

To link customer aggregated data with external data to create greater intelligence and action

Data Append

- Expands information about customers
- Provides the ability to do advanced segmentation
- Provides the capability of comparing customers against the business environment
- Allows targeting of high potential accounts
- Becoming the base line in some industries

Data Update

- Data rationalization is a on-going process
- Introduction of new data into the system
- Update decaying data

Data Update - Decay

Data decays rapidly:

- Industry of business
- Size of business
- Age of business
- Type of data

- Address decay averages 2 to 3% per month across US business universe
- In one year 1 out of every three addresses would need to be updated

Value Proposition

- The true value of CRM applications
- can only be realized when high quality,
- well integrated, well managed,
- information is used

The CRM Solution

- **PREDICTIVE INFORMATION**
- **INTERNAL INFORMATION**
- **ACTIONABLE INFORMATION**
- **EXTERNAL INFORMATION**
- **BUSINESS RESULTS**

Cracking the Code

ID1	ID2	DATE	JOB	SEX	FIN	PRO3	CR_T	ERA
2612	624	941106	06	8	DEC	.	.	.
2613	625	940506	04	5	ETS	.	.	.
2614	626	940809	11	5	PBB	.	.	.
2615	627	941010	16	1	RVC	.	.	.
2616	628	940507	04	2	ETT	.	.	.
2617	629	940812	09	1	OFS	.	.	.
2618	630	950906	09	2	RFN	71	612	12
2618	631	951107	13	2	PBB	0	623	23
2619	632	950112	10	5	SLP	0	504	04
2620	633	950802	11	1	STL	34	611	11
2620	634	950908	06	0	DES	0	675	75
2620	635	950511	01	1	DLF	0	608	08

Errors, Outliers, and Missings

cking	#cking	ADB	NSF	dirdep	SVG	bal
Y	1	468.11	1	1876	Y	1208
Y	1	68.75	0	0	Y	0
Y	1	212.04	0	6		0
	.	.	0	0	Y	4301
y	2	585.05	0	7218	Y	234
Y	1	-47.69	2	1256		238
Y	1	4687.7	0	0		0
	.	.	1	0	Y	1208
Y	.	.	.	1598		0
	1	0.00	0	0		0
Y	3	89981.12	0	0	Y	45662
Y	2	585.05	0	7218	Y	234

Data Arrangement

Long-Narrow

Acct	type
2133	MTG
2133	SVG
2133	CK
2653	CK
2653	SVG
3544	MTG
3544	CK
3544	MMF
3544	CD
3544	LOC

Short-Wide

Acct	CK	SVG	MMF	CD	LOC	MTG
2133	1	1	0	0	0	1
2653	1	1	0	0	0	0
3544	1	0	1	1	1	1

Roll-Up

HH	Acct	Sales
4461	2133	160
4461	2244	42
4461	2773	212
4461	2653	250
4461	2801	122
4911	3544	786
5630	2496	458
5630	2635	328
6225	4244	27
6225	4165	759

HH	Acct	Sales
4461	2133	?
4911	3544	?
5630	2496	?
6225	4244	?

Derived Inputs

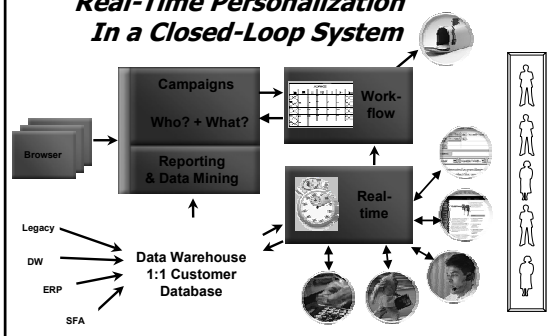
Claim Date	Accident Time
11nov96	102396/12:38
22dec95	012395/01:42
26apr95	042395/03:05
02jul94	070294/06:25
08mar96	123095/18:33
15dec96	061296/18:12
09nov94	110594/22:14

Delay	Season	Dark
19	fall	0
333	winter	1
3	spring	1
0	summer	0
69	winter	0
186	summer	0
4	fall	1

Massive

	Bytes	Paper
Kilobyte	2^{10}	1/2 sheet
Megabyte	2^{20}	1 ream
Gigabyte	2^{30}	167 feet
Terabyte	2^{40}	32 miles
Petabyte	2^{50}	32,000 miles

Real-Time Personalization In a Closed-Loop System



Thank You

Andrea Ahlemeyer-Stubbe
DataBase Management
Hauptstr.34
D-77723 Gengenbach
Tel: + 49 (0) 7803 92 93 59
Fax: + 49 (0) 7803 92 93 60
E-mail:ahlemeyer@ahlemeyer-stubbe.de

Member of
Pro – ENBIS www.enbis.org/pro-enbis
ENBIS www.enbis.org
Branta www.branat.de

ECDM EUROPEAN CENTRE OF DATABASE MARKETING