NATIONAL ENGINEERING CENTER
University of the Philippines
Diliman, Quezon City

1.0 Introduction to Data Warehousing

Eugene Rex L. Jalao, Ph.D.

Associate Professor

Department Industrial Engineering and Operations Research

University of the Philippines Diliman

Module 2 of the Business Intelligence and Analytics Track of UP NEC and the UP Center of Business Intelligence

UP NEC BI Modules

- Analyst Level
  - Introduction to Business Intelligence and Data Mining
  - **Data Warehousing**
  - Data Mining
- Professional Level
  - Time Series Analysis and Forecasting
  - Optimization Analysis
  - R For Business Intelligence

Outline for This Training

1. Introduction to Data Warehousing
2. DW Lifecycle and Project Management
   - Case Study on DW PM
3. Dimensional Modeling
4. Designing Fact Tables
5. Designing Dimension Tables
   - Case Study on Dimension Modeling
6. Extraction Transformation and Loading
   - Case Study on ETL Planning
7. Transformation and Loading Methodologies
   - Case Study on ETL

Outline for This Session

- Introduction to Data Warehousing
- Data Warehousing and Data Mining
- Justifications for Data Warehousing
- OLAP and DW Compared
- Current Market Status of DW

Introduction to Data Warehousing

- Your Manager Wants to Know
  - Which are our lowest/highest margin customers?
  - Who are my customers and what products are they buying?
  - What is the most effective distribution channel?
  - What impact will new products/services have on revenue and margins?

Introduction to Data Warehousing

- Current Issues
  - Can't find the data
    - data is scattered over the network
    - many versions, subtle differences
  - Can't get the data
    - need an expert to get the data
  - Can't understand the data
    - available data poorly documented
  - Can't use the data
    - results are unexpected
    - data needs to be transformed from one form to other

Introduction to Data Warehousing

- End User's Requests
  - Data should be integrated across the enterprise
  - Summary data has a real value to the organization
  - Historical data holds the key to understanding data over time
  - What-if capabilities are required

Introduction to Data Warehousing

Definition 1.1 (Kimball Definition): Data Warehousing

- "The query-able source of data in the enterprise."

  - Ralph Kimball et. al. in The Data Warehouse Lifecycle Toolkit

Introduction to Data Warehousing

Definition 1.1 (CIO Definition): Data Warehousing

- "A data warehouse is the processes, tools, and facilities to manage and deliver complete, timely, accurate, and understandable business information to authorized individuals for effective decision making."

  - IBM Customer Council, 1990

Introduction to Data Warehousing

Definition 1.1 (Inmon Definition): Data Warehousing

- A data warehouse (DW) is a
  - subject-oriented
  - integrated
  - time-varying
  - non-volatile

  collection of data that is used primarily in organizational decision making.

  -- Bill Inmon, Building the Data Warehouse 1996

Introduction to Data Warehousing

Definition 1.2: Subject-Oriented DW

- Data is categorized and stored by business subject rather than by application

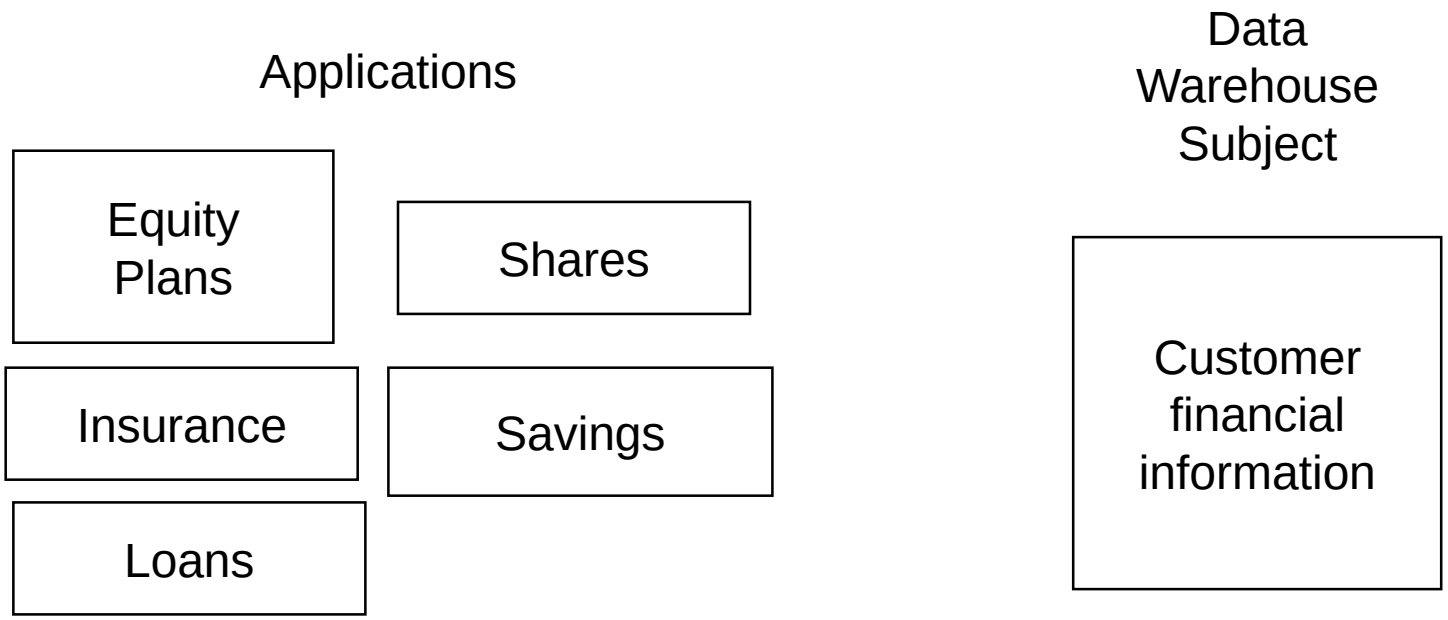


Figure 1.1: Applications vs Subjects

Introduction to Data Warehousing

Definition 1.3: Integrated DW

- Data on a given subject is defined and stored once.

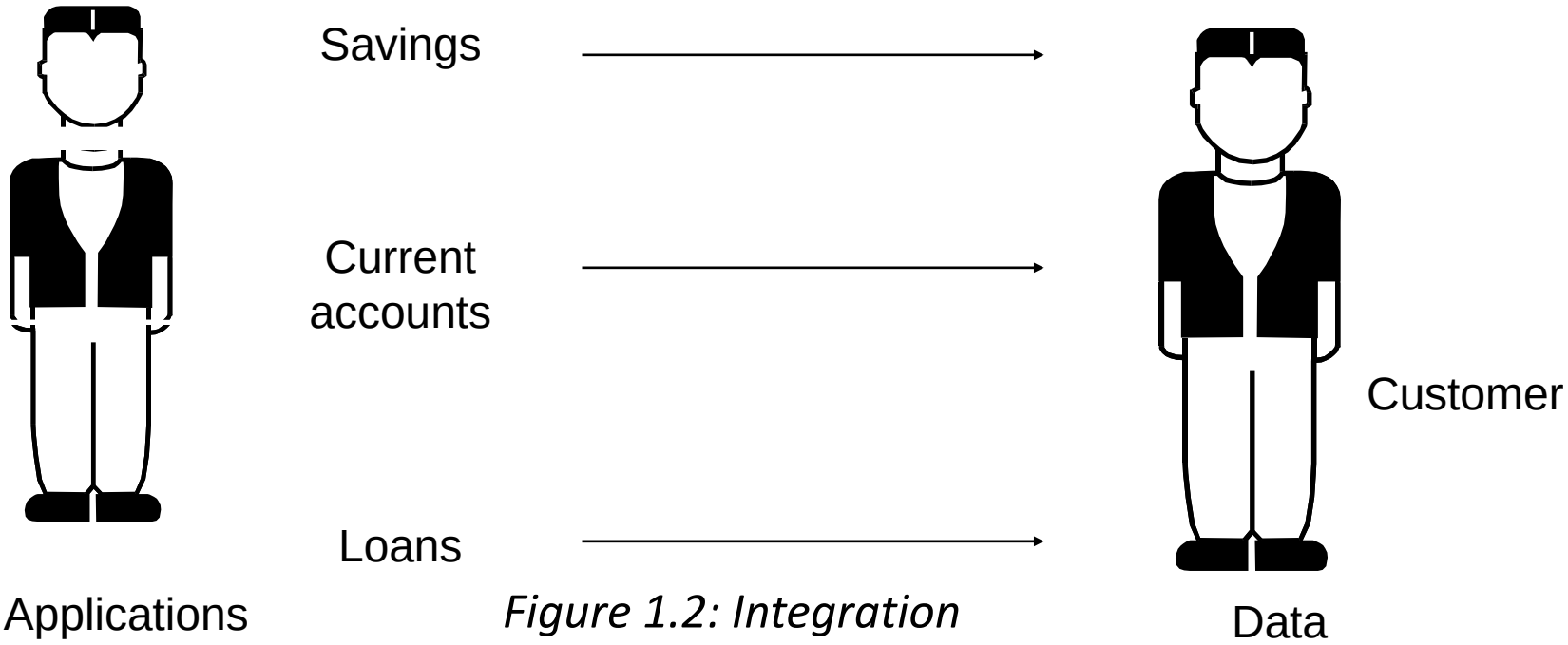


Figure 1.2: Integration

Introduction to Data Warehousing

Definition 1.4: Time-Variant DW

- Data is stored as a series of snapshots, each representing a period of time

Time	Data
Jan-97	January
Feb-97	February
Mar-97	March

Figure 1.3: Time Variant Example

Introduction to Data Warehousing

Definition 1.5: Non-Volatile DW

- Typically data in the data warehouse is not updated or deleted.

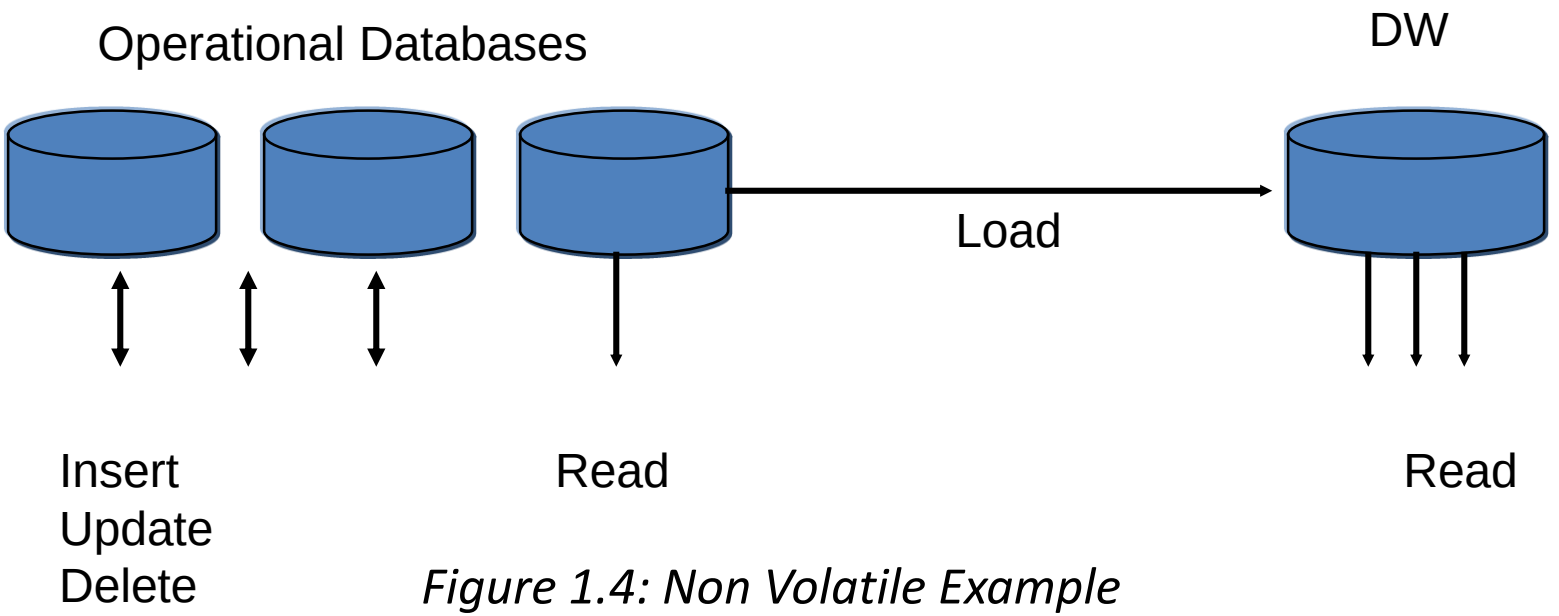


Figure 1.4: Non Volatile Example

Introduction to Data Warehousing

- Business Intelligence v. Data Warehouse
  - Data Warehouse is the Information Technology (IT) term
  - Users, especially senior management and users new to the concept of a data warehouse, identify more readily with Business Intelligence (BI)
  - We will use the term Data Warehouse most often in this module because much of our focus will be on how to build a DW system
  - Always keep in mind that the end purpose is Business Intelligence

Introduction to Data Warehousing

- Characteristics of Data Warehouse
  - Summarized Operational data are mapped into a decision-usable format
  - Large volume
    - Data sets are normally quite large.
  - Not normalized
    - DW data can be, and often are, redundant.
  - Metadata.
    - Data about data are stored.
  - Data sources.
    - Data come from internal and external unintegrated operational systems.

Introduction to Data Warehousing

- The "Who" of BI and DW
  - BI is the primary responsibility of business executives, managers, analysts
  - DW is the primary responsibility of information technology (IT) executives, managers, architects, and technicians
  - Each group has a collateral responsibility in most of the other groups' DW tasks
  - A strong mutual working relationship is essential

Introduction to Data Warehousing

- Why Build a BI Data Warehouse?
  Business View/Goals
  - Makes an organization's information accessible
  - Empowers end users and gives them control over their reporting needs
  - Makes an organization's information consistent
  - Is an adaptive and resilient source of information
  - Is a secure bastion that protects our information asset
  - Is the foundation for decision making
  - Yields excellent Return On Investment (ROI)

Introduction to Data Warehousing

- DW Mantra:

"A single version of the truth."

Introduction to Data Warehousing

- Quasi History of Data Warehousing
  - 1960's
    - Management Information Systems (MIS)
    - Database Management Systems (DBMS)
  - 1970
    - E. F. (Ted) Codd's ACM paper on relational database technology
  - 1970's
    - IBM's Business Systems Planning (BSP),
    - Enterprisewide information systems planning
  - 1980's
    - Executive Information Systems (EIS) and Decision Support Systems (DSS)
    - Pioneering DW work done by Ralph Kimball & colleagues

Introduction to Data Warehousing

- Quasi History of Data Warehousing
  - 1990's
    - Data Warehouse concept - W. H. (Bill) Inmon
    - Codd's Rules for On Line Analytical Processing (OLAP)
    - Proliferation of DW tools
    - Emergence of Very Large Data Bases (VLDBs)
    - Delivering DW data on the Web
  - 2000's
    - Role in e-commerce, Customer Relationship Management
    - (CRM), Supply Chain Management (SCM), Master Data
    - Management (MDM), Web Analytics, etc.

Outline for This Session

- Introduction to Data Warehousing
- **Data Warehousing and Data Mining**
- Justifications for Data Warehousing
- OLAP and DW Compared
- Current Market Status of DW

Data Warehousing and Data Mining

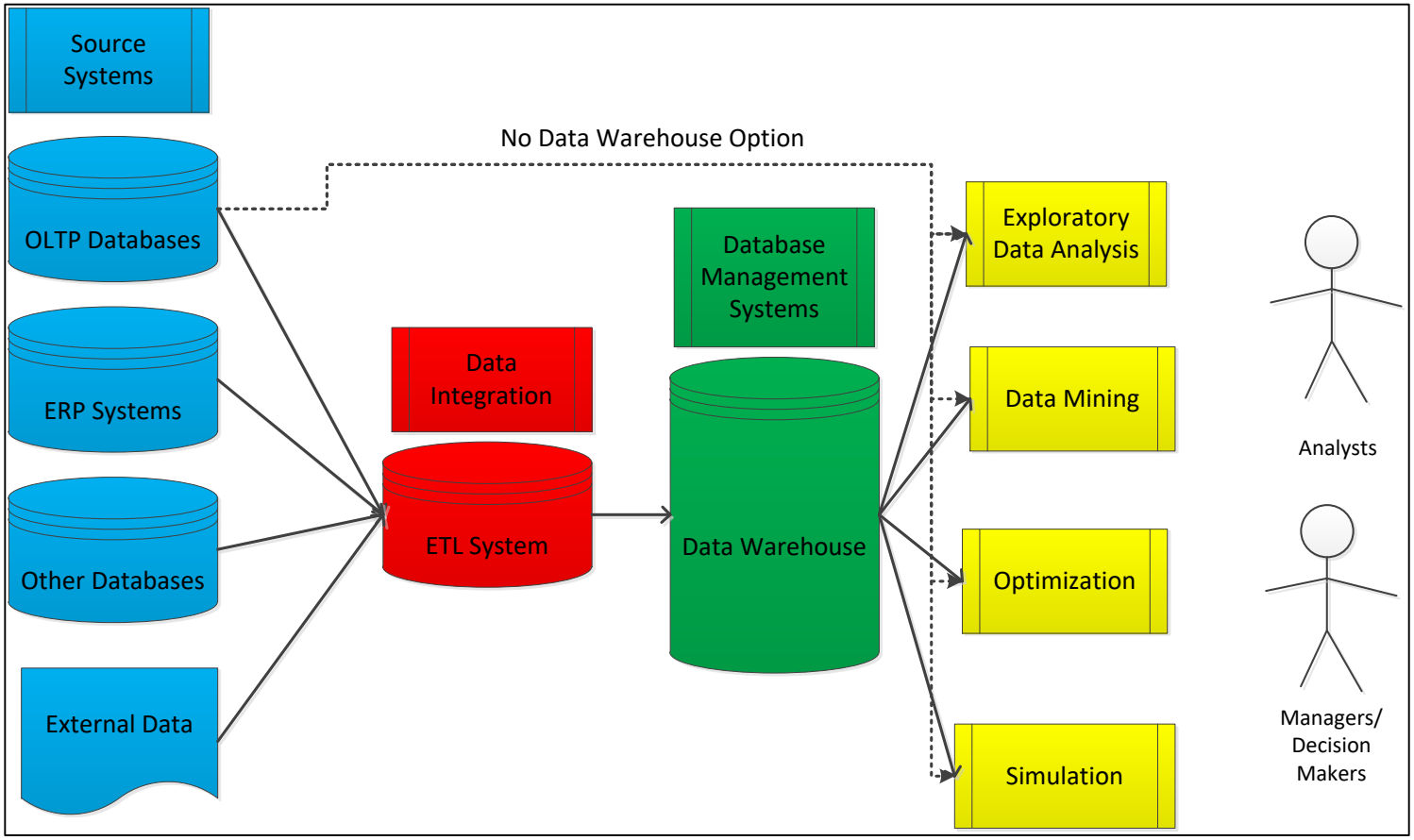


Figure 1.5: BA Framework

Data Warehousing and Data Mining

- Data Warehousing provides the Enterprise with a memory
- Data Mining provides the Enterprise with intelligence

Data Warehousing and Data Mining

- Profile of High End DW Organization
  - Leading edge DWs use dimensional models with conformed dimensions and facts
  - Environnent: Oracle, Microsoft SQL Server, or IBM DB2; on Unix or Windows
  - Informatica, Data Stage, SQL Server Integration Services or other ETL tool installed
  - Business Objects, Cognos, MicroStrategy, SQL Server Analysis Services or other end user BI tool installed
  - Web-enabled delivery to employees, customers suppliers
  - Seriously looking at data mining

Data Warehousing and Data Mining

- Advances in the following areas are making data mining deployable:
  - data warehousing
  - better and more data (i.e., operational, behavioral, and demographic)
  - the emergence of easily deployed data mining tools and
  - the advent of new data mining techniques.
    - -- Gartner Group

Data Warehousing and Data Mining

- Why Separate the Data Warehouse?
  - Performance
    - Operational Databases are designed & tuned for known transactions & workloads.
    - Complex queries would degrade performance
    - Special data organization, access & implementation methods needed for multidimensional views & queries.
  - Function
    - Missing data: Decision support requires historical data, which operational databases do not typically maintain.
    - Data consolidation: Decision support requires consolidation (aggregation, summarization) of data from many heterogeneous sources: operational databases, external sources.
    - Data quality: Different sources typically use inconsistent data representations, codes, and formats which have to be reconciled.

Data Warehousing and Data Mining

Table 1.1: Types of Analytics in Different Business Functions

FUNCTION	DESCRIPTION	EXEMPLARS
Supply chain	Simulate and optimize supply chain flows; reduce inventory and stock-outs.	Dell, Wal-Mart, Amazon
Customer selection, loyalty, and service	Identify customers with the greatest profit potential; increase likelihood that they will want the product or service offering; retain their loyalty.	Harrah's, Capital One, Barclays
Pricing	Identify the price that will maximize yield, or profit.	Progressive, Marriott
Human capital	Select the best employees for particular tasks or jobs, at particular compensation levels.	New England Patriots, Oakland A's, Boston Red Sox
Product and service quality	Detect quality problems early and minimize them.	Honda, Intel
Financial performance	Better understand the drivers of financial performance and the effects of nonfinancial factors.	MCI, Verizon
Research and development	Improve quality, efficacy, and, where applicable, safety of products and services.	Novartis, Amazon, Yahoo

Source: Davenport, Thomas H. "Competing on analytics." *Harvard Business Review* 84.1 (2006): 98.

Outline for This Session

- Introduction to Data Warehousing
- Data Warehousing and Data Mining
- **Justifications for Data Warehousing**
- OLAP and DW Compared
- Current Market Status of DW

Justifications of Data Warehousing

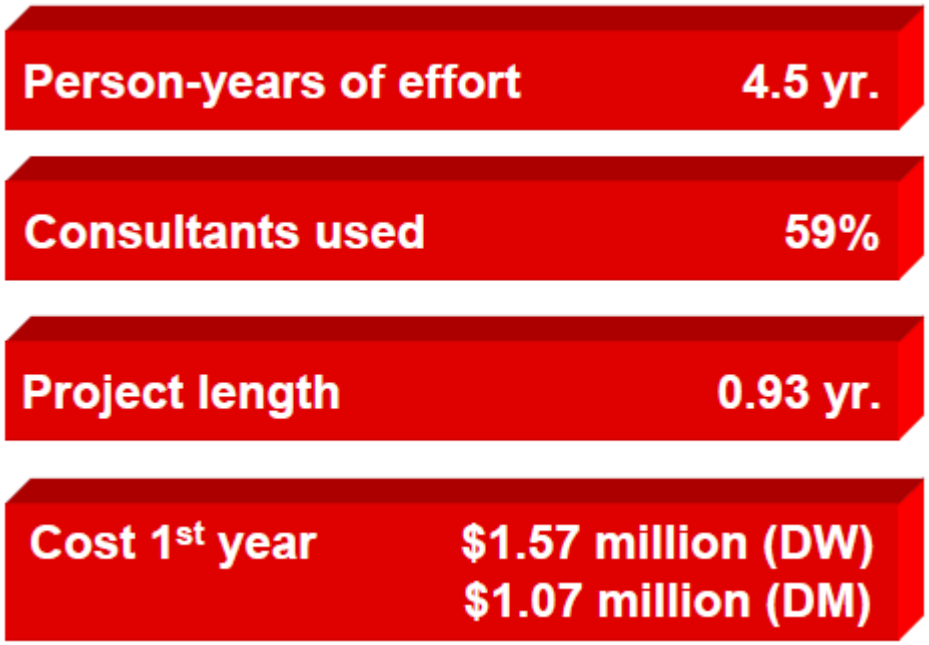
- Top 10 Reasons for Building a DW
  1. Get data out of OLTP environment
  2. Provide answers to business questions that thus far have been unanswered
  3. Provide a decentralized decision-making environment and not to rely on IT folk
  4. Integrate systems and applications
  5. Build the enterprise data models

Justifications of Data Warehousing

- Top 10 Reasons for Building a DW
  6. Improve quality of operational data without having to rebuild legacy systems
  7. Create a consistent view of the state of the business
  8. Provide answers to business users faster
  9. Offload reporting activities from the IT organization
  10. Provide good performance for both decision support and OLTP applications

Justifications of Data Warehousing

- How Much Do DWs Cost? (Medium to Large Enterprise)



Source: Gray, Paul & Hugh Watson, Decision Support in the Data Warehouse , Prentice Hall, 1998

Figure 1.6: Costs of Implementing a DW

Justifications of Data Warehousing

- DW justification: has shifted from cost justification to DW as a strategic business imperative
- Rapid increase in number of applications, users, and size of database
- Emphasis on business solutions: customer service, targeted marketing, supply chain logistics, quality management, and other
- Next high-ground:
  - Wide use of DW at medium and small size enterprises
  - DW in all functional areas
  - Dashboards/scorecards at all levels
  - Near-real-time DW/BI

Justifications of Data Warehousing

- IBM Data Warehousing Satisfaction Survey 2007
  - Some results and conclusions:
    - Surveyed 200 companies, 41 responded
    - 56% have DWs more than 6 years; 19% less than 3 years
    - Data warehousing is a mature approach to delivering business intelligence
  - DW is the foundation for step-by-step incremental progress in data management across applications
  - DW is a sustaining force in data management and applications
  - Enterprises should build for the long term since DWs tend to live long lives.
  - DW is not a short-term technology fad; it is a decades-long commitment.

Source: The DW Satisfaction Survey, Part 1,
DM Review BI Report, October 2007, dmreview.com

Justifications of Data Warehousing

- No. 1 Complaint: Lack of Data in Their Data Warehouse
- Much opportunity for improvement remains:
- 51% cited lack of data
- 41% cited insufficient/inadequate master data, especially customer and product data
- 24% cited not leveraging their info for impact on the business

Source: The DW Satisfaction Survey, Part 1,
DM Review BI Report, October 2007, dmreview.com

Justifications of Data Warehousing

- Do's
  - Heavy user involvement on front end
  - Identify champion(s)
  - Limited project scope
  - Realistic development schedule
  - Clean up production data
  - Present data in format users can grasp and use easily

Source: The Conference Board/PW Survey
ComputerWorld, March 23, 1998

Justifications of Data Warehousing

- Trends in DW
  - BI data warehouses finding broad applicability in virtually all organizations, all functional areas
  - Data warehouses now affordable by mid- and small-size enterprises
    - Microsoft SQL Server DW enhancements
  - The ETL tool "space" is changing
    - SQL Server bundled ETL tool
      - SQL Server 2000 Data Transformation Services (DTS)
      - SQL Server 2005 Integration Services (SSIS)
    - The industry is consolidating – big vendors are buying smaller vendors

Justifications of Data Warehousing

- Trends in DW
  - The BI tool space is changing
    - SQL Server Analysis Services - bundled BI tool
    - Cubes (MOLAP) gaining in practicality and popularity, especially in mid/small firms
    - Relational OLAP (ROLAP) is necessary and foundational; popular for large DWs and power users
    - BI tool vendor power bases keep changing
  - Packaged DWs having limited success
  - Increased career opportunities in DW

Justifications of Data Warehousing

- Trends in DW
  - DW increasingly recognized as needed in conjunction with other major programs:
    - Customer Relationship Management (CRM)
    - Supply Chain Management (SCM)
    - Master Data Management (MDM)
    - Web Analytics (WA)

Justifications of Data Warehousing

- Data Warehousing –A Career Opportunity
  - DW is now mainstream, not a fad
  - Many IT professionals now in DW
  - DW foundation for new trends – CRM,
  - eBusiness, Supply Chain Management, Master Data Management (MDM), Web Analytics
  - DW is "just now heating up"
  - "Getting data out is where all the action is"

Justifications of Data Warehousing

Industry	Drivers	Benefits
Health Care	Escalating Costs	Supplier Performance Cost Control
Financial Services	Retain (Profitable) Customers	Cross Product Profitability Credit Risk Management Trend Analysis
Telecom	Deregulation New Competition	Asset Utilization Fraud Detection Customer Service
Retail	Niche Players	Micro Marketing

Figure 1.7: Industry Drivers for Implementing a DW

Justifications of Data Warehousing

- DW Benefit Categories
  - Strategic value
  - Widespread demand
    - internal – enterprise-wide
    - external – customers, suppliers, alliances
  - Financial benefits (tangibles)
  - Business improvement (intangibles)

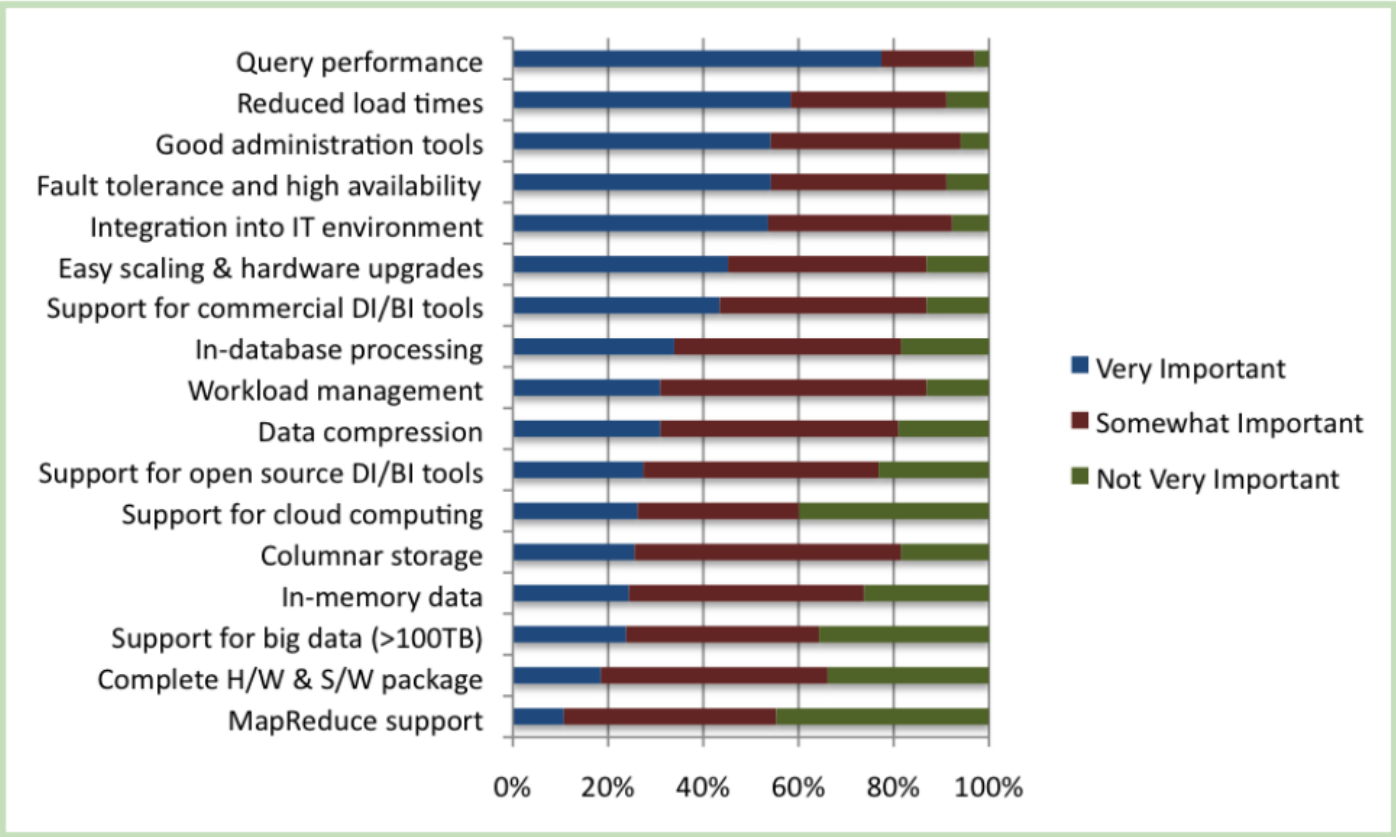
Justifications of Data Warehousing

- Strategic Value
  - Your "single version of the truth" – past, present, and future
  - Your decision-making tool of choice – at all levels of management and staff
  - Achieve more effective e-commerce (B2B, B2C)
  - Your newest, greatest legacy system – you can't buy it off the shelf, do it right the first time
  - Growth versus survival in an info-centric world

Justifications of Data Warehousing

- Widespread Demand
  - Internally: serves all levels of the organization, all departments and employees enterprise-wide with cross-functional data
    - Dashboards are popular for senior and middle management
    - See Planning and Control triangle (next slide)
  - Externally serves:
    - Customers
    - Suppliers
    - Channel partners
    - Alliance partners

Justifications of Data Warehousing



Source: http://vertica.com/wp-content/uploads/2010/12/beyond-traditional-data-warehouse.pdf

Figure 1.8: Important Features for BI Platforms

Justifications of Data Warehousing

- Why Now?
  - Data is being produced
  - ERP provides clean data
  - The computing power is available
  - The computing power is affordable
  - The competitive pressures are strong
  - Commercial products are available

Outline for This Session

- Introduction to Data Warehousing
- Data Warehousing and Data Mining
- Justifications for Data Warehousing
- **OLAP and DW Compared**
- Current Market Status of DW

OLAP and DW Compared

Definition 1.6: OLAP

- Online Analytical Processing - coined by EF Codd in 1994 paper contracted by Arbor Software
- Generally synonymous with earlier terms such as Decisions Support, Executive Information System
- OLAP = Multidimensional Database
- MOLAP: Multidimensional OLAP (Arbor Essbase, Oracle Express)
- ROLAP: Relational OLAP (Informix MetaCube, Microstrategy DSS Agent)

OLAP and DW Compared

Definition 1.6: Operational Systems

- They are OLTP systems
- Run mission critical applications
- Need to work with stringent performance requirements for routine tasks
- Used to run a business!

OLAP and DW Compared

- Operational Systems
  - Run the business in real time
  - Based on up-to-the-second data
  - Optimized to handle large numbers of simple read/write transactions
  - Optimized for fast response to predefined transactions
  - Used by people who deal with customers, products -- clerks, salespeople etc.
  - They are increasingly used by customers

OLAP and DW Compared

- RDBMSs have been used traditionally for OLTP
  - clerical data processing tasks
  - detailed, up to date data
  - structured repetitive tasks
  - read/update a few records
  - isolation, recovery and integrity are critical

OLAP and DW Compared

Table 1.2: Examples of Operational Data

Data	Industry	Usage	Technology	Volumes
Customer File	All	Track Customer Details	Legacy application, flat files, main frames	Small-medium
Account Balance	Finance	Control account activities	Legacy applications, hierarchical databases, mainframe	Large
Point-of-Sale data	Retail	Generate bills, manage stock	ERP, Client/Server, relational databases	Very Large
Call Record	Telecomm-unications	Billing	Legacy application, hierarchical database, mainframe	Very Large
Production Record	Manufact-uring	Control Production	ERP, relational databases, AS/400	Medium

OLAP and DW Compared

- OLAP systems are tuned for known transactions and workloads while workload is not known a priori in a data warehouse
- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)
  - e.g., average amount spent on phone calls between 9AM-5PM in Pune during the month of December

OLAP and DW Compared

- OLAP
  - Application Oriented
  - Used to run business
  - Detailed data
  - Current up to date
  - Isolated Data
  - Repetitive access
  - Clerical User
- Data Warehouse
  - Subject Oriented
  - Used to analyze business
  - Summarized and refined
  - Snapshot data
  - Integrated Data
  - Ad-hoc access
  - Knowledge User (Manager)

OLAP and DW Compared

- OLAP
  - Performance Sensitive
  - Few Records accessed at a time (tens)
  - Read/Update Access
  - No data redundancy
  - Database Size 100MB - 100 GB
- Data Warehouse
  - Performance relaxed
  - Large volumes accessed at a time(millions)
  - Mostly Read (Batch Update)
  - Redundancy present
  - Database Size 100 GB - few terabytes

OLAP and DW Compared

- OLAP
  - Transaction throughput is the performance metric
  - Thousands of users
  - Managed in entirety
- Data Warehouse
  - Query throughput is the performance metric
  - Hundreds of users
  - Managed by subsets

OLAP and DW Compared

OLAP Systems are used to *"run"* a business while the Data Warehouse helps to *"optimize"* the business.

OLAP and DW Compared

Example 1.1: Old Retail Paradigm

- WalMart
  - Inventory Management
  - Merchandise Accounts Payable
  - Purchasing
  - Supplier Promotions: National, Region, Store Level
- Suppliers
  - Accept Orders
  - Promote Products
  - Provide special Incentives
  - Monitor and Track The Incentives
  - Bill and Collect Receivables
  - Estimate Retailer Demands

OLAP and DW Compared

Example 1.2: New Just-In-Time Retail Paradigm

- No more deals
- Shelf-Pass Through (POS Application)
  - One Unit Price
    - Suppliers paid once a week on ACTUAL items sold
  - WallMart Manager
    - Daily Inventory Restock
    - Suppliers (sometimes SameDay) ship to WalMart
- Warehouse-Pass Through
  - Stock some Large Items
    - Delivery may come from supplier
  - Distribution Center
    - Supplier's merchandise unloaded directly onto WalMart Trucks

OLAP and DW Compared

Example 1.3: The WallMart DW System

- NCR 5100M 96 Nodes: *24 TB Raw Disk; 700 - 1000 Pentium CPUs*
- Number of Rows: *> 5 Billion*
- Historical Data: *65 weeks (5 Quarters)*
- New Daily Volume: *Current Apps: 75 Million, New Apps: 100 Million +*
- Number of Users: *Thousands*
- Number of Queries: *60,000 per week*

Outline for This Session

- Introduction to Data Warehousing
- Data Warehousing and Data Mining
- Justifications for Data Warehousing
- OLAP and DW Compared
- **Current Market Status of DW**

Current Market Status of DW

Table 1.3: Distribution of BI Reporting and DW

BI Reporting Solution Type	Data Warehouse Solution Type						Row Total*
	Homegrown	BI/Infrastructure Provider	Admin. Systems Provider	None	Same Solution as BI Reporting Solution	Other	
BI Provider (single)	14%	21%	12%	11%	5%	0%	**63%**
Homegrown	10%	1%	1%	3%	0%	0%	**15%**
BI Providers (multiple)	1%	4%	0%	0%	1%	0%	**7%**
Admin. Systems Provider	1%	1%	3%	1%	0%	0%	**6%**
None	3%	1%	1%	0%	0%	0%	**6%**
Admin. Provider with Third-Party BI Tools	0%	0%	1%	1%	0%	0%	**2%**
Other	0%	0%	0%	0%	0%	0%	**1%**
Column Total*	**29%**	**28%**	**19%**	**17%**	**6%**	**1%**	**100%**

** Certain totals differ from apparent sums due to rounding.*

Source: https://net.educause.edu/ir/library/pdf/ERB1403.pdf

Current Market Status of DW

Figure 1.9: Magic Quadrant for Business Intelligence and Analytics Platforms

Outline for This Session

- Introduction to Data Warehousing
- Data Warehousing and Data Mining
- Justifications for Data Warehousing
- OLAP and DW Compared
- Current Market Status of DW

References

- Simon, Alan. CIS 391 PPT Slides
- UCI Irvine Data Warehousing Notes
- S. Sudarshan, Krithi Ramamritham http://www.cse.iitb.ac.in/dbms/Data/Talks/krithi-talk-impact.ppt
- Gartner Research (2014): http://www.ioz.pwr.wroc.pl/Pracownicy/lubicz/SWD/KONSPEKTY/Magic-Quadrant-for-Business-Intelligence-and-Analytics-Platforms.pdf
- BI Reporting, Data Warehouse Systems, and Beyond (2013) https://net.educause.edu/ir/library/pdf/ERB1403.pdf