

Managing Data Resources


Presented by
Eldon Y. Li

*** All right reserved. Video or audio recording and distributing are prohibited without the author's consent. Reference to this document should be made as follows: Li, E.Y. "Managing Data Resources," unpublished lecture, National Chung Cheng University, 2020.

Copyright © 2020 E.Y. Li

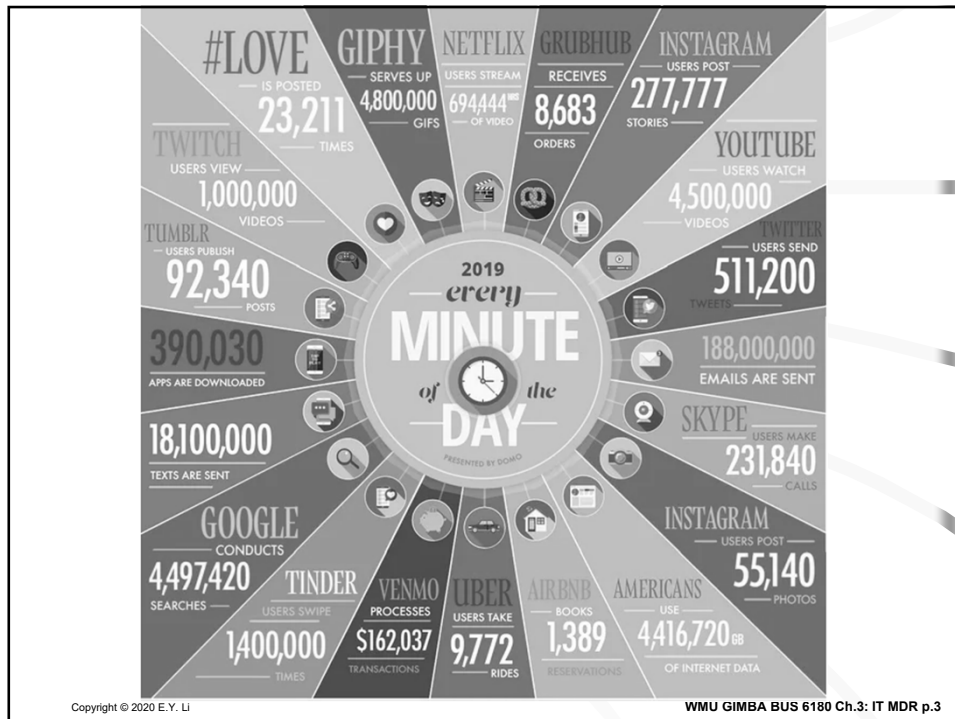
WMU GIMBA BUS 6180 Ch.3: IT MDR p.1

Sharing Moment

- Do you process data at home or at work?
What kinds of data do you process?
- What program do you use to process data?
- What is the difference between a file and a database?
- In the case study II-4, what are the choices available to make a decision? 

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.2









What is Big Data

- **Big Data** is a general term used to describe the massive amount of data available to today's managers. Majority (80%) of corporate data are messy and unstructured, and are too big and costly to easily work through use of conventional databases, but new tools (e.g., Hadoop) are making these massive datasets available for analysis and insight.
- Big Data has 6 V's characteristics:
 - Volume: the huge amount of data
 - Variety: the diversity of data types (structured, semi-structured, unstructured).
 - Velocity: the speed of data changes
 - Veracity: the quality/integrity/credibility/accuracy of the data.
 - Variability: the diversity of data formats and data sources (data fusion).
 - Value: the added value for companies.

The six Vs of big data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: *volume, variety and velocity*. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.
					

Turning Big Data into Value:



- The 'Datafication' of our World;
- Activities
 - Conversations
 - Words
 - Voice
 - Social Media
 - Browser logs
 - Photos
 - Videos
 - Sensors
 - Etc.

- Volume
- Velocity
- Variety
- Veracity

- Analysing Big Data:
- Text analytics
 - Sentiment analysis
 - Face recognition
 - Voice analytics
 - Movement analytics
 - Etc.



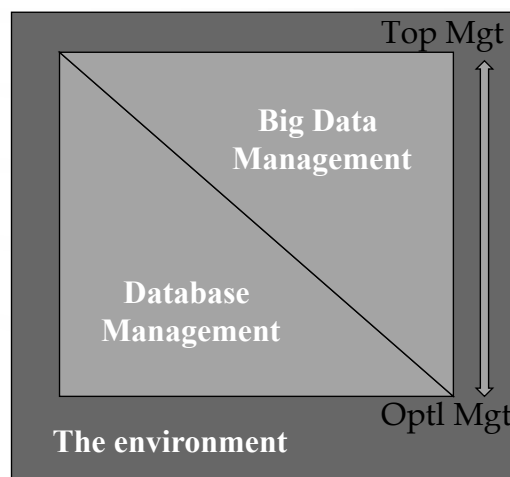
Data Structures in Big Data

- **Structured data** is highly organized data that is easy to search and is predictable. The data is usually in a fixed field or predetermined record and can be related to other data records within its structure. An example of structured data would be a relational database or a spreadsheet.
- **Semi-structured data** does not reside in fixed fields or records, but does contain elements that can separate the data into various hierarchies. Examples of semi structured data are tab delimited files or XML files.
- **Unstructured data** has no defined format. Examples may include books, journals, documents, metadata, health records, audio, video, analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document.

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.7

Two Milieus of Data Management



Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.8

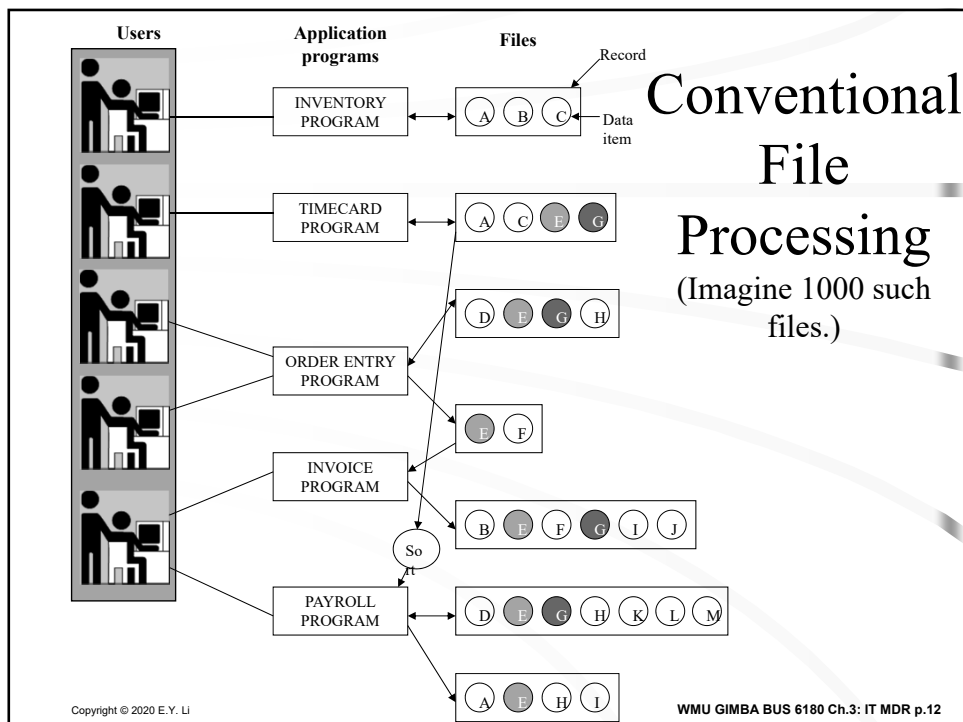
Database Management

Learning Objectives

- What are the problems in processing data files?
- How can database approach solve these problems?
- What are the available database structures?
- Which database structure is the best of all?

The Hierarchy of Data

- Bit
- Byte
- Field/Column
- Record/Row
- File/Table
- Database

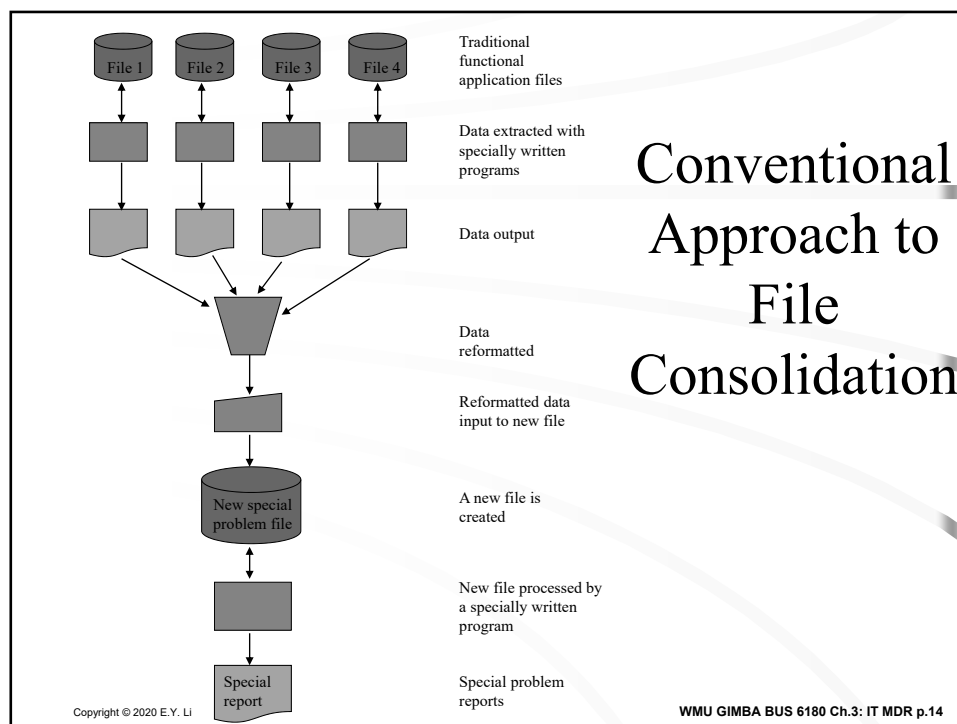


Problems in File Processing



- **uncontrolled redundancy:** duplication of data or files
- **poor standardization:** on data item names and formats
- **poor integrity:** non-synchronized file update
- **poor sharability:** data formats different in application programs
- **poor accessibility:** (see Fig.) difficult to query horizontally across the file boundaries for consolidation reporting purpose
- **poor flexibility:** difficult to accommodate changes of users' information needs, such as insert, delete, and change fields
- **poor maintainability:** difficult to modify programs and data files
- **poor programmer productivity:** need great effort to specify data definition, file access method, and read/write statements
- **poor security control:** no centralized security control

Copyright © 2020 E.Y. Li



WMU GIMBA BUS 6180 Ch.3: IT MDR p.13



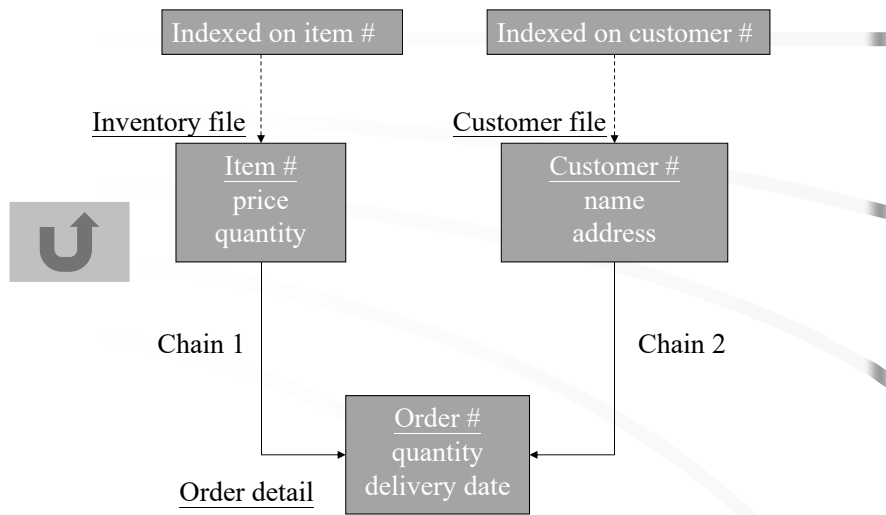
The Database Approach

- Create file linkage:
 - Linked by pointer fields
A pointer field is a physical field in a table created by the system to link the record of the current table to a record in another table. 
 - Linked by common fields
 - Linked by relationship files if no common fields exist
A relationship file is a table that uses the primary keys of two or more other tables as its composite primary key. This table serves as the liaison for the other tables to link one another. A relationship file is common called an "intersection table."
- Create data independence:
 - Three-level views of data (The origin of middleware)
 - Two logical views of the data 

The Database Approach

- Use schema to control data integrity and security:
 - Schema 
 - Subschema
- Implement new database structures:
 - Hierarchical 
 - Network
 - Relational

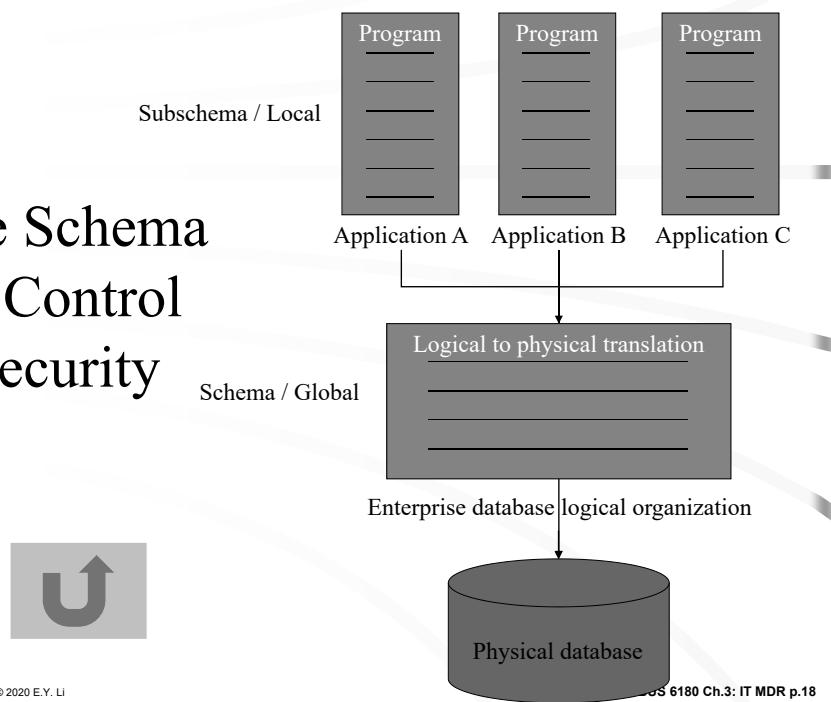
The File Linkage



Copyright © 2020 E.Y. Li

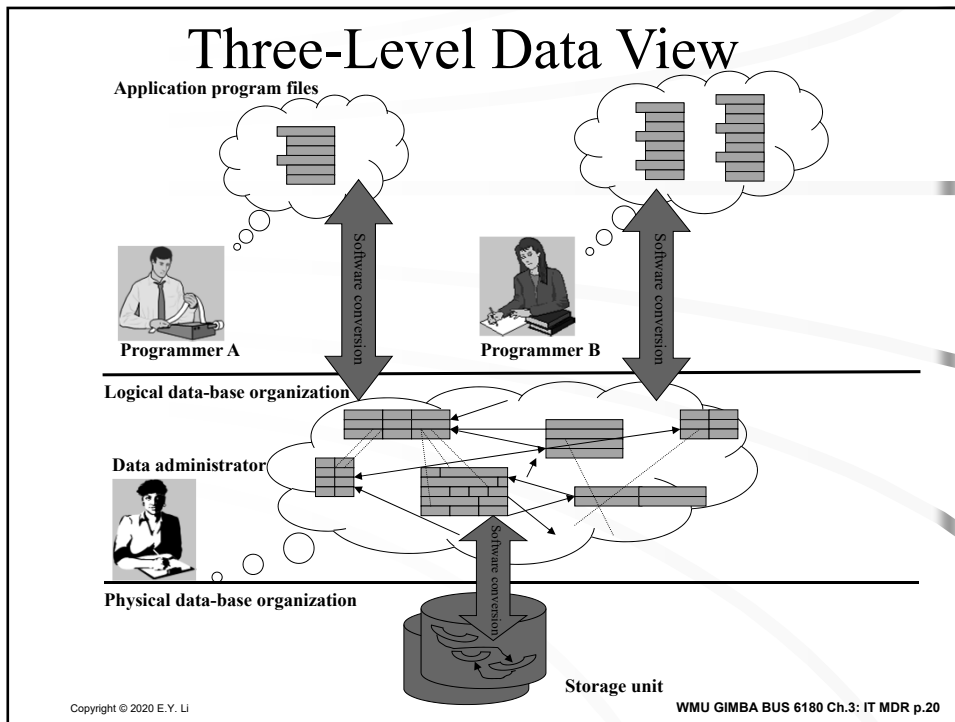
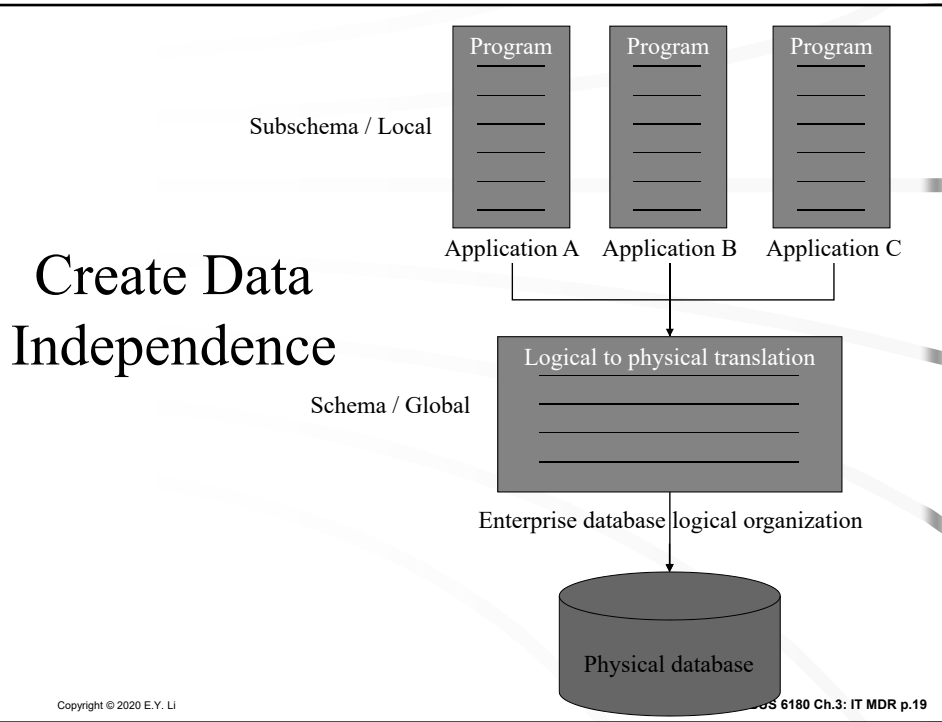
WMU GIMBA BUS 6180 Ch.3: IT MDR p.17

Use Schema to Control Security

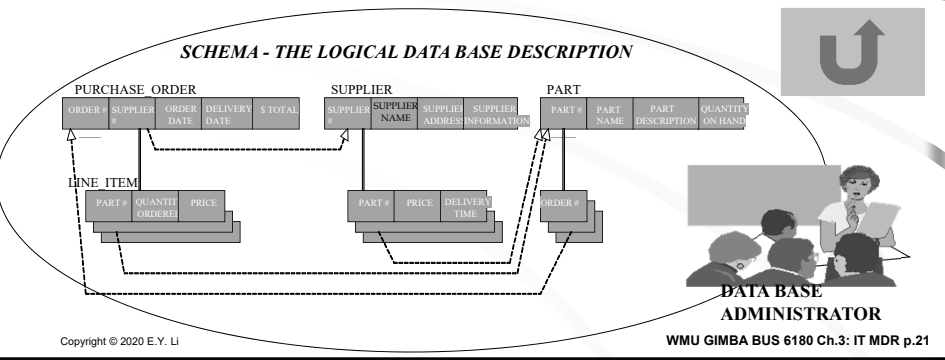
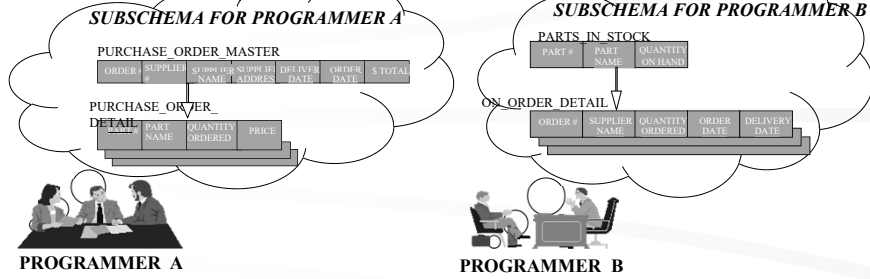


Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.18



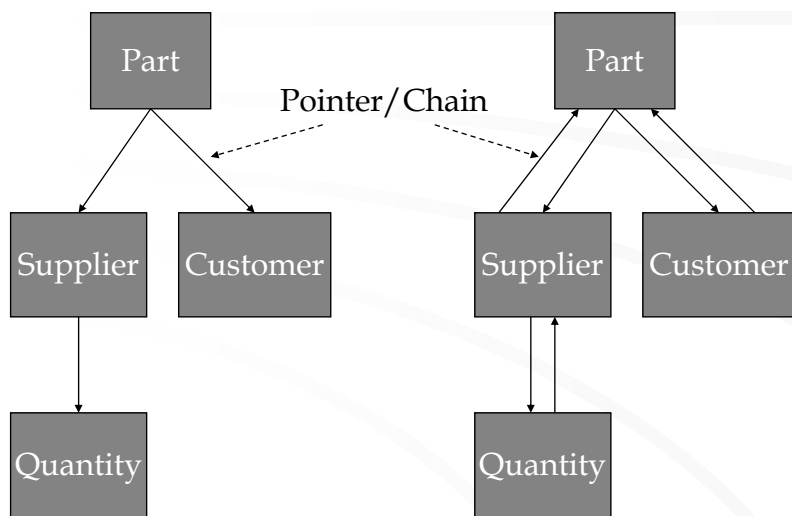
Two Logical View of Data



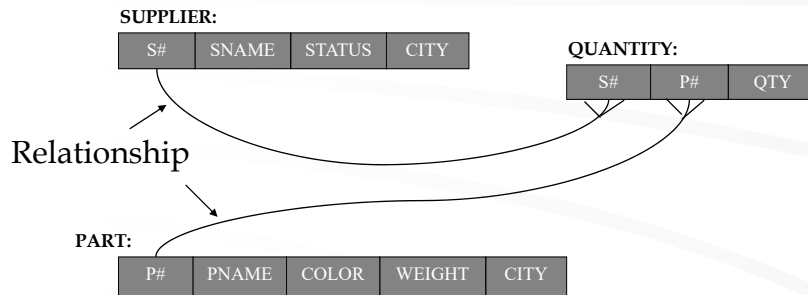
Non-Relational Database Models

Hierarchical Database

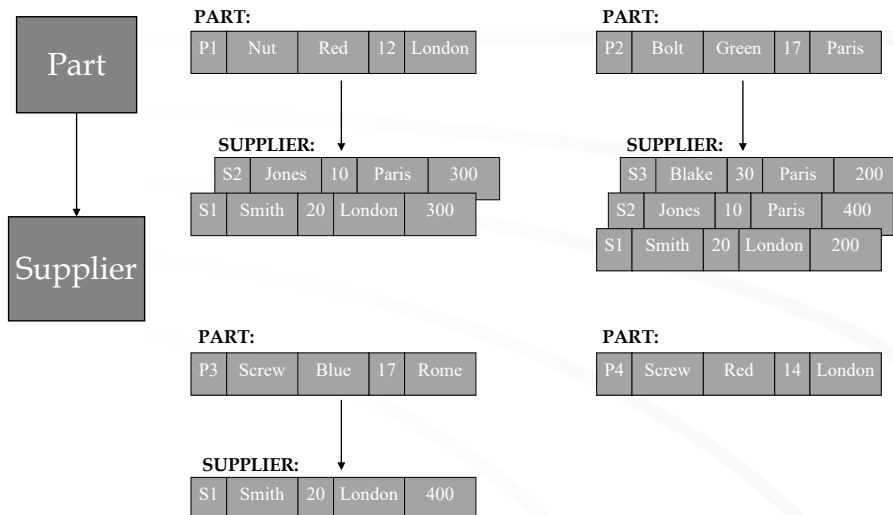
Network Database



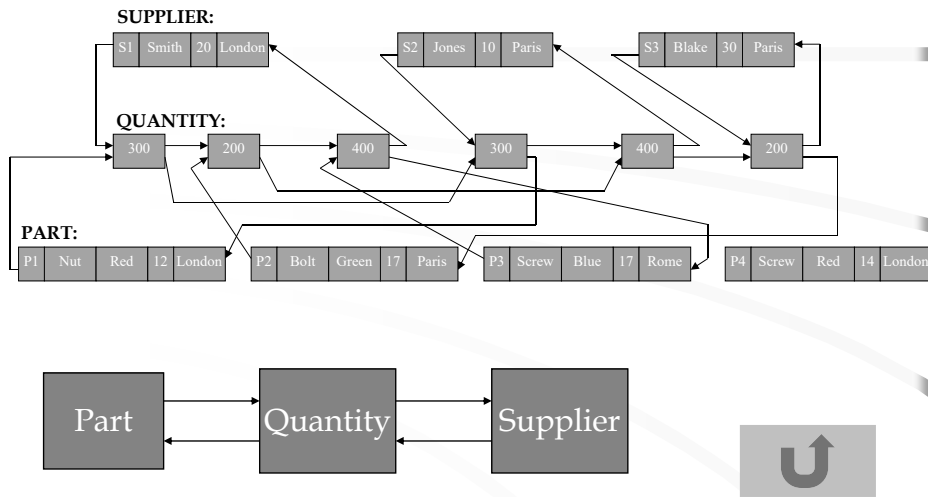
Relational Database Models



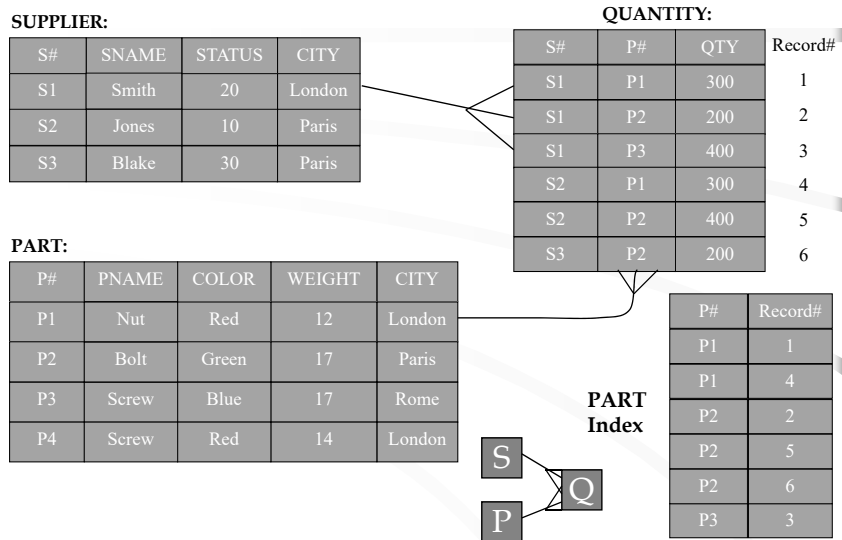
Hierarchical Database Model



Network Database Model



Relational Database Model



Benchmarking Database Models

Queries Models	Q1: Find supplier numbers for suppliers who supply part P2.	Q2: Find part numbers for parts supplied by supplier S2.
Relational	Next: Get (next) shipment where P# = P2. Shipment found? If not, exit. Print S#. Go to Next.	Next: Get (next) shipment where S# = S2. Shipment found? If not, exit. Print P#. Go to Next.
Hierarchical	Next: Get (next) part where P# = P2. Get (next) supplier for this part. Supplier found? If not, exit. Print S#. Go to Next.	Next: Get (next) part. Part found? If not, exit. Get (next) supplier for this part where S# = S2. Supplier found? If not, go to next. Print P#. Go to Next.
Network	Next: Get (next) part where P# = P2. Get (next) connector for this part. Connector found? If not, exit. Get superior supplier for this connector. Print S#. Go to Next.	Next: Get (next) supplier where S# = S2. Get (next) connector for this supplier. Connector found? If not, exit. Get superior part for this Connector. Print P#. Go to Next.

R p.27

Comparison of DB Organizations

Data Base System	Technical Advantages	Managerial Usage Suited for
Hierarchical	Natural relationships are predefined and provide easy access paths.	Routine reporting; can be designed to serve key tasks; development of ad hoc reports.
Network	Natural relationships or any other combination of records can be linked with pre-established pointers.	Key tasks; ad hoc reporting; specialized circumstances in which unusual data associations are needed.
Relational	New tables and files can be quickly constructed.	Routine reporting; creation of new data bases for DSS systems and "what if..." models.

Recommendations

- Hierarchical model is most efficient for a VLDB under structured queries and top-down access pattern.
- Relational model is most efficient and flexible for ad hoc queries when it is properly indexed.
- Network model is flexible and has stable performance, but not as efficient as relational model.
- Overall, relational model is the best choice.

Conceptual Data Model: Entity-Relationship Diagram (ERD)



Figure 4.1 Entity-Relationship Diagram

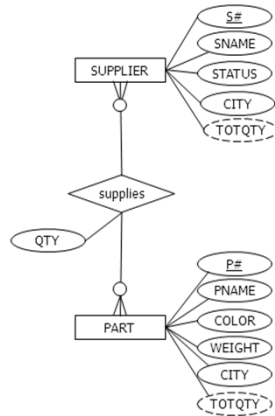
Entities = things about which data are collected
(e.g., Customer, Order, Product)

Attributes = actual elements of data to be collected

Relationships = associations between entities
(e.g., Submits, Includes)

From Conceptual Model to Logical Design

ERD Example:



Convert ERD to relations:

SUPPLIER

S#	SNAME	STATUS	CITY
S1	Smith	20	London
S2	Jones	10	Paris
S3	Blake	30	Paris
S4	Clark	20	London
S5	Adams	30	Athens

supplies

S#	P#	QTY
S1	P1	300
S1	P2	200
S1	P3	400
S1	P4	200
S1	P5	100
S1	P6	100
S2	P1	300
S2	P2	400
S3	P2	200
S4	P2	200
S4	P4	300
S4	P5	400

PART

P#	PNAME	COLOR	WEIGHT	CITY
P1	Nut	Red	12	London
P2	Bolt	Green	17	Paris
P3	Screw	Blue	17	Rome
P4	Screw	Red	14	London
P5	Cam	Blue	12	Paris
P6	Cog	Red	19	London

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.31

Enterprise Data Modeling

- Divide work into major functions
- Divide each function into processes
- Divide processes into activities (e.g., forecast sales for next quarter)
- List data entities assigned to each activity
- Check for consistency in names

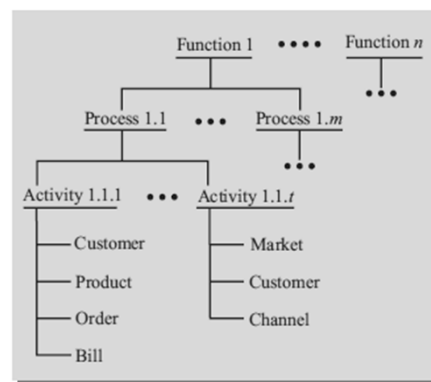
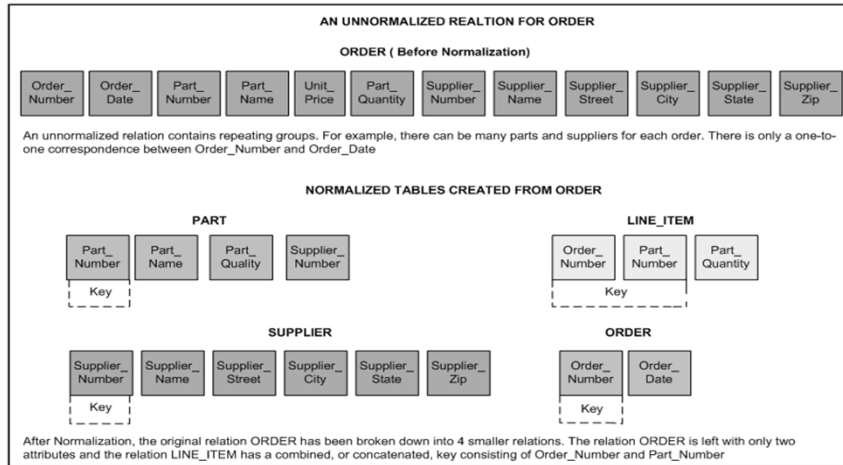


Figure 4.2 Enterprise Decomposition for Data Modeling

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.32

- The process of creating simple data structures from more complex ones using a set of rules that yields a stable structure.



Big Data Management



Learning Objectives

- What is a data warehouse (DW)?
- Why do we need a DW?
- Why is DW getting popular?
- How does DB evolve into DW?
- What does a DW architecture look like?
- What are the characteristics of data in DW?
- How do we create a DW?
- Which data model does a DW have?
- How does a user interface with a DW?
- What is OLAP? What is data mining?

Definition

- **Data Warehouse:** An integrated and consistent store of subject-oriented data that is obtained from a variety of sources and contains current and historical values to support decision making in an organization.
- “A data warehouse is a subject-oriented, integrated, nonvolatile, time-variant collection of data used in support of management’s decisions.” (W.H. Inmon, *Building the Data Warehouse*, 1996)

Need for Data Warehousing

- Lack of credibility in data quality: Need to reconcile data before it is used.
- Data volatility: Need to keep historical time-variant data.
- Fragmented heterogeneous data (see Figure D-1 ): Need integrated, company-wide view of high-quality information.
- Mixing operational and decision support systems and data (see Table D-1  for differences): Need to separate these two data systems.

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.37

Figure D-1:
Examples of
heterogeneous
data (regarding
student's
personal
information)



STUDENT_DATA

Student_No	Last_Name	MI	First_Name	Telephone	Status	•••
123-45-6789	Enright	T	Mark	483-1967	Soph	
389-21-4062	Smith	R	Elaine	283-4195	Jr	

STUDENT_EMPLOYEE

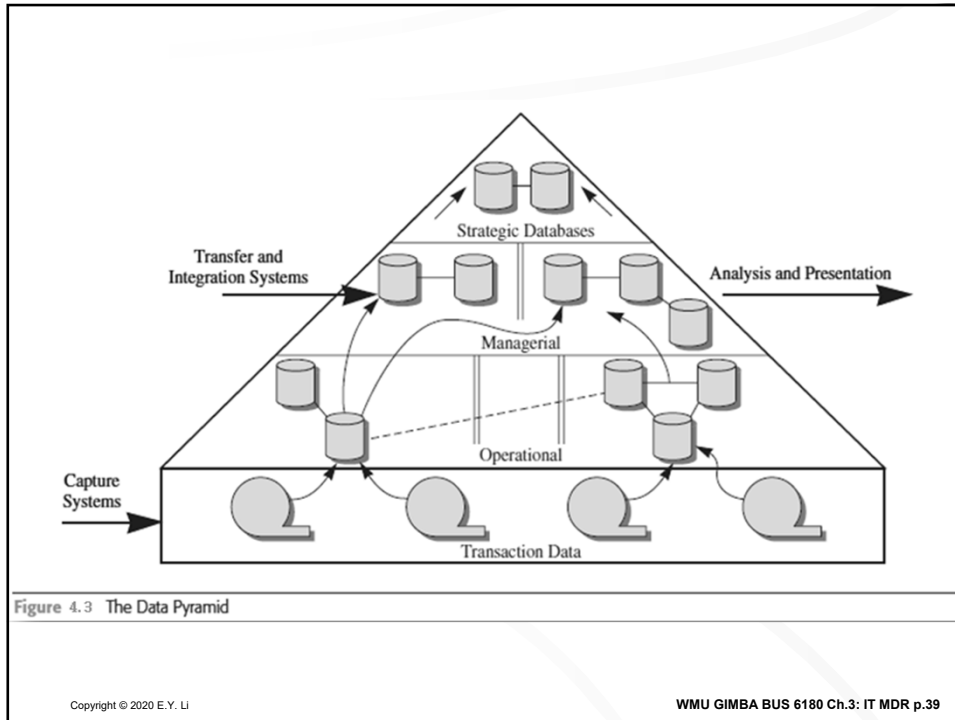
Student_ID	Address	Dept	Hours	•••
123-45-6789	1218 Elk Drive, Phoenix, AZ 91304	Soc	8	
389-21-4062	134 Mesa Road, Tempe, AZ 90142	Math	10	

STUDENT_HEALTH

Name	Telephone	Insurance	ID	•••
Mark T. Enright	483-1967	Blue Cross	123-45-6789	
Elaine R. Smith	555-7828	?	389-21-4062	

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.38



	Operational Data	DSS Data
Table D-1. Operational vs. DSS data	Application oriented	Subject oriented
	Detailed	Summarized, otherwise refined
	Accurate, as of the moment of access	Represents values over time, snapshots
	Run repetitively	Run periodically
	Current data, can be updated	Historic data, is not updated
	Serves the clerical community	Serves the managerial community
	Compatible with SDLC	Completely different life cycle
	High availability	Relaxed availability
	Accessed a unit/record at a time	Accessed a set/group at a time
	Control of update a major concern in terms of ownership	Control of update no issue
	Performance sensitive	Performance relaxed
	Requirements for processing understood a priori	Requirements for processing not understood a priori
	Managed in its entirety	Managed by subsets (or subjects)
	Small amount of data used in a process	Large amount of data used in a process
	High probability of access	Low to modest probability of access
	Non-redundancy	Redundancy is a fact of life
Static structure; variable contents	Flexible structure	
Supports day-to-day operations	Supports managerial needs	
Transaction driven	Analysis driven	

Source: Adapted from Inmon, W.H. *Building the Data Warehouse*, 2nd Ed. New York: John Wiley & Sons, 1996.

Copyright © 2020 E.Y. Li WMU GIMBA BUS 6180 Ch.3: IT MDR p.40

Factors Making Data Warehousing Possible

- Relational DBMS.
- Advances in hardware: speed and storage capacity.
- Available end-user computing interfaces and tools.
- Further reading: IS7-5

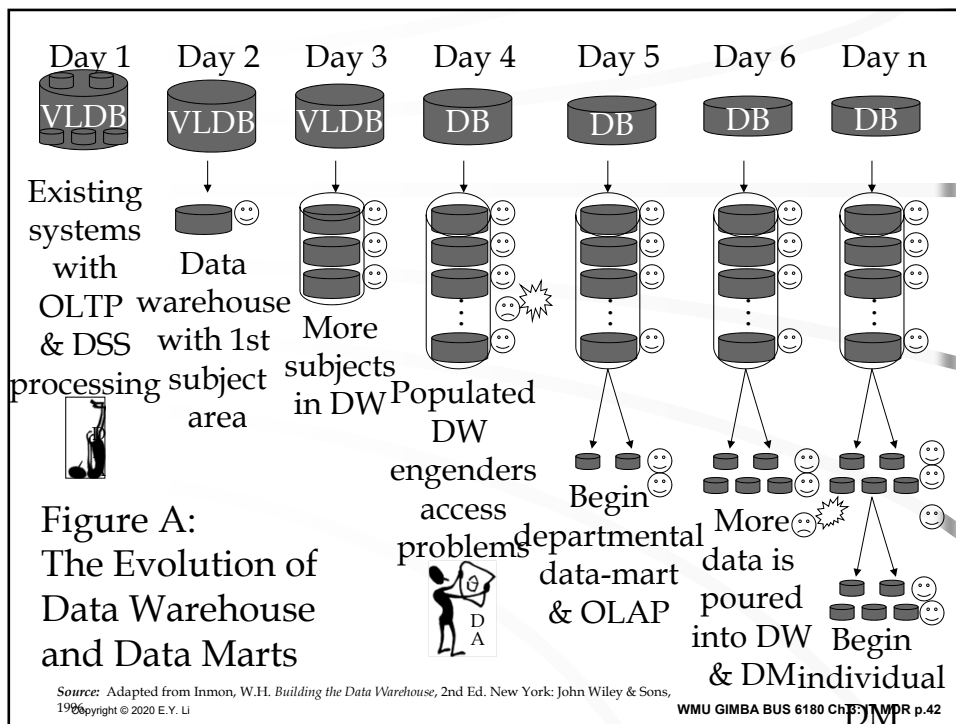
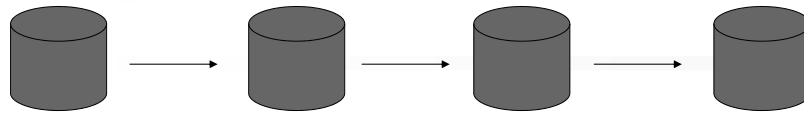


Figure B: The Four Levels of the Architecture



Operational	Atomic/EDW	Departmental	Individual
<ul style="list-style-type: none"> • detailed • day to day • current valued • high probability of access • application oriented 	<ul style="list-style-type: none"> • most granular • time variant • integrated • subject oriented • some summary 	<ul style="list-style-type: none"> • parochial • some derived; some primitive • typical depts • accounting • marketing 	<ul style="list-style-type: none"> • temporary • ad hoc • heuristic • non-repetitive • PC, workstation-based

Source: Adapted from Inmon, W.H. *Building the Data Warehouse*, 2nd Ed. New York: John Wiley & Sons, 1996.

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.43 (Data Mart)

Data Warehouse Architectures

- Two-layer architecture (See Fig. D-2.) ▶
- Three-layer architecture.
 - Operational databases/files
 - Enterprise data warehouse (EDW)- single source of data for decision making.
 - Data marts (DM) - limited scope; data selected from EDW.
 - See Fig. D-3. ▶

Figure D-2:
Generic data warehouse architecture

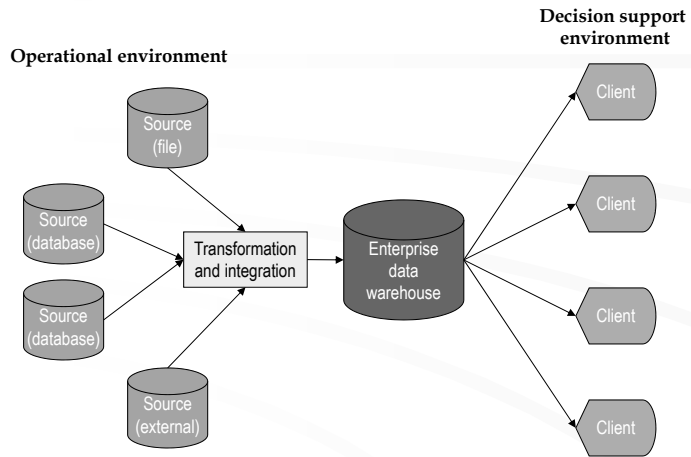
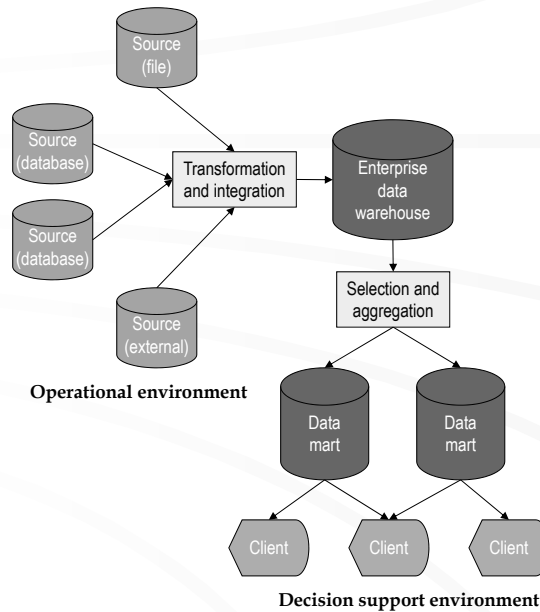


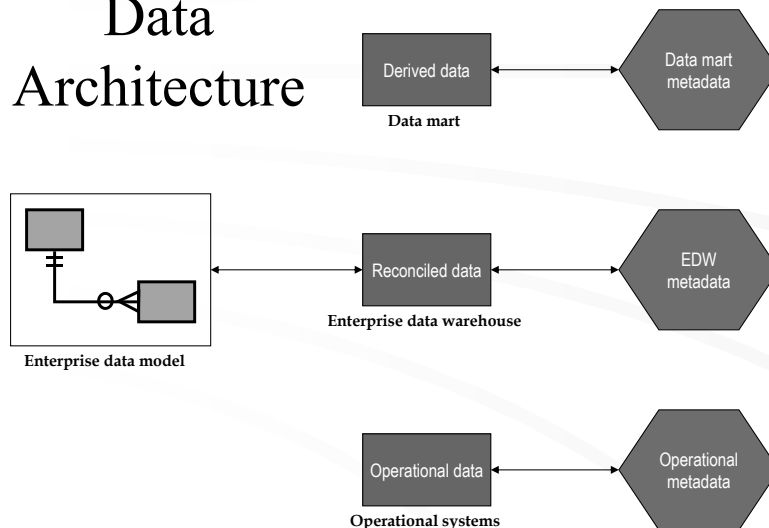
Figure D-3:
Three-layer architecture



Reasons for the Three-Level DW Architecture

- EDW and data marts have different purposes and data architectures.
- Data transformation process is very complex and is best performed in two steps.
- Data marts customized decision support for different groups.
- Data marts facilitate distributed processing.



Three-Level Data Architecture



The Role of DM Metadata

- What subjects are described in the data mart? (Customer? patient? student? product? course?)
- What dimensions and facts are included in the data mart? What is the grain of the fact table?
- How are the data in the data mart derived from the EDW data? What rules are used?
- How are the data in the EDW derived from operational data? What rules are used?
- What reports and predefined queries are available to view the data?
- What drill-down and other data analysis techniques are available?
- Who is responsible for the quality of data in the data marts, and to whom are requests for changes made?

Data Characteristics

- Status vs. Event data.
 - Fig. D-5. 
- Transient vs. Periodic data.
 - Fig. D-6 & Fig. D-7 

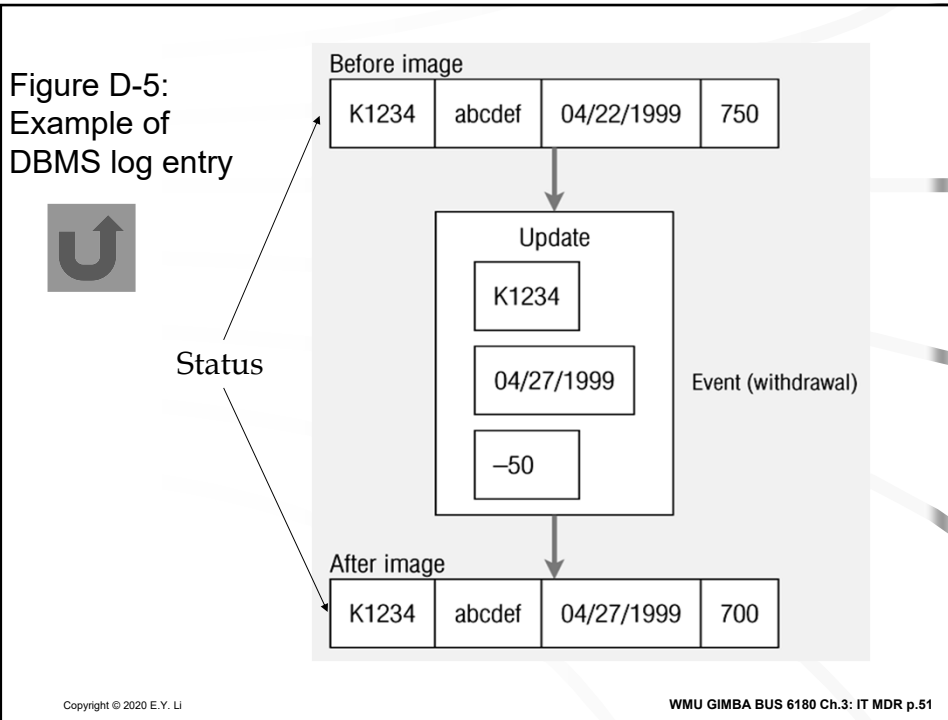


Figure D-6: Transient warehouse data
(Current data is overwritten by new data)

Table X (10/03)

Key	Date	A	B	Action
001	10/01	a	b	C
002	10/02	x	d	U
003	10/03	e	z	U
005	10/02	m	n	C

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.52

Figure D-7: Periodic warehouse data

Table X (10/01)

Key	Date	A	B	Action
001	10/01	a	b	C
002	10/01	c	d	C
003	10/01	e	f	C
004	10/01	g	h	C

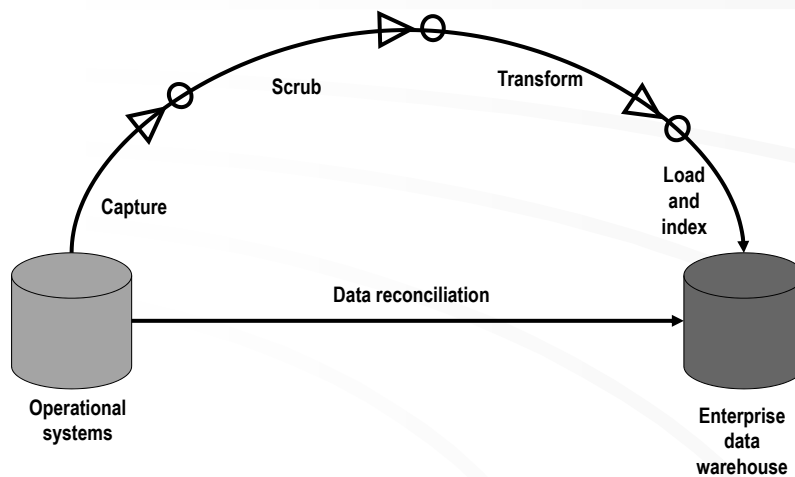
Table X (10/02)

Key	Date	A	B	Action
001	10/01	a	b	C
002	10/01	c	d	C
002	10/02	x	d	U
003	10/01	e	f	C
004	10/01	g	h	C
004	10/02	y	h	U
005	10/02	m	n	C

Table X (10/03)

Key	Date	A	B	Action
001	10/01	a	b	C
002	10/01	c	d	C
002	10/02	x	d	U
003	10/01	e	f	C
003	10/03	e	z	U
004	10/01	g	h	C
004	10/02	y	h	U
004	10/03	y	h	D
005	10/02	m	n	C

The Data Reconciliation Process



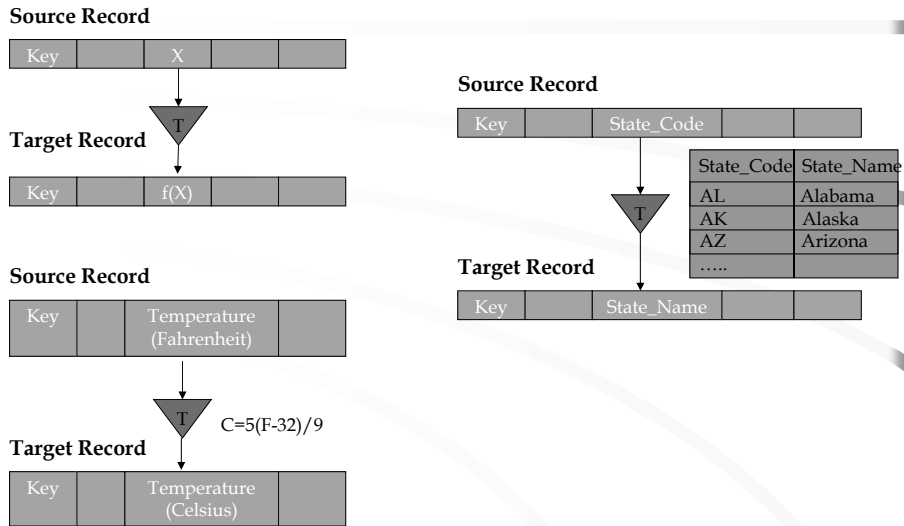
The Data Reconciliation Process

- Capture
 - Static - initial load of a snapshot.
 - Incremental - ongoing update with after-image logs.
- Scrub (or data cleansing)
 - Uses pattern recognition and other artificial intelligence techniques to upgrade the data quality.

The Data Reconciliation Process

- Transform
 - Convert the data format from the source to the target system.
 - Record-Level Functions
 - Selection.
 - Joining.
 - Normalization.
 - Aggregation (for data marts).
 - Field-Level Functions
 - Single-field transformation (see next Fig. D-9)
 - Multi-field transformation (see next Fig. D-10)

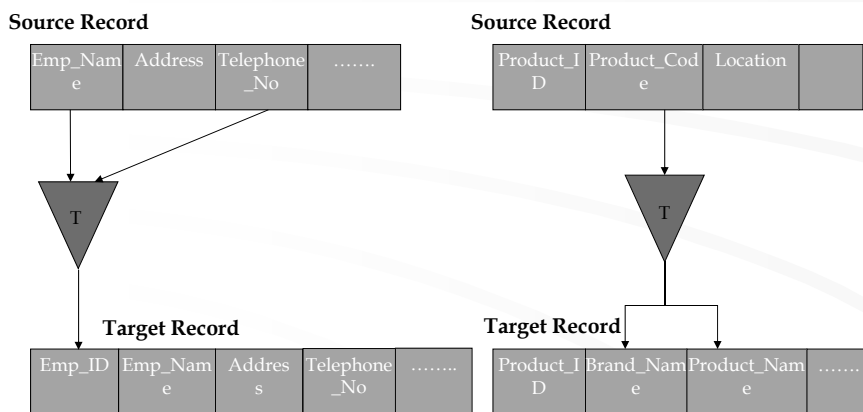
Figure D-9: Single-field Transformation



Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.57

Figure D-10: Multifield Transformation



Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.58

The Data Reconciliation Process

- Load and Index
 - Refresh Mode
 - When the warehouse is first created.
 - Static data capture.
 - The whole warehouse is rewritten periodically
 - Update Mode
 - Ongoing update of the warehouse.
 - Incremental data capture.
 - The new records are written to the warehouse without overwriting or deleting the old records.

Reconciliation Tools

Product Name	Company	Description
Analyze	QDB Solutions, Inc. http://www.qdb.com (7/97 acquired by Prism Solutions, Inc.)	Data quality assessment
WizRules	WizSoft, Inc. http://www.wizsoft.com	Rules discovery
Extract	Evolutionary Technologies International http://www.evtech.com	Extract, transform, load and index
InfoSuite	Platinum Technology, Inc. http://www.platinum.com (3/99 acquired by Computer Associate)	Extract, transform, load and index
Passport	Carleton Corp. http://www.carleton.com (11/99 acquired by Oracle Corp.)	Extract, scrub, transform, load and index
Prism	Prism Solutions, Inc. http://www.prismsolutions.com (4/99 acquired by Ardent Software, Inc. http://www.ardentsoftware.com) (12/99 acquired by Informix, Inc. http://www.informix.com)	Extract, transform, load and index
Integrity	Vality Technology, Inc. http://www.vality.com	Quality analysis, data scrubbing

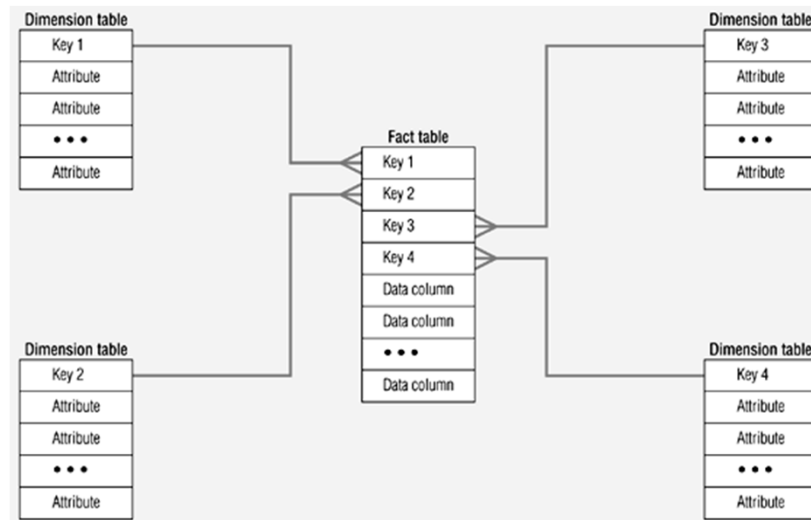
Reconciled Data Characteristics in Enterprise Data Warehouse

- Detailed
- Historical/Periodic
- Normalized
- Comprehensive/Enterprise-wide
- Quality controlled

Derived Data Characteristics in Data Mart

- Type of data
 - Detailed, possibly periodic.
 - Aggregated.
- Distributed to departmental servers.
- Implemented in star schema
(see next Fig.D-11).

Figure D-11:
Components of a star schema



Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.63

Star Schema

- Also called the dimensional model.
- Contains fact tables and dimension tables
- Fig. D-12, 13. ▶
- Grain of a fact table - time period for each record in the table.
- Multiple Fact Tables - Fig. D-14. ▶
- Snowflake Schema - Fig. D-15. ▶

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.64

Figure D-12:
Star schema example

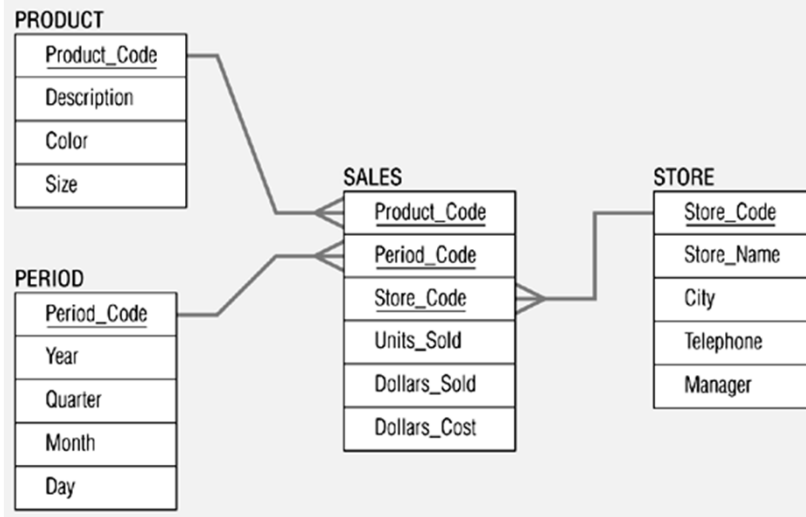


Figure D-13:
Star schema
with sample
data

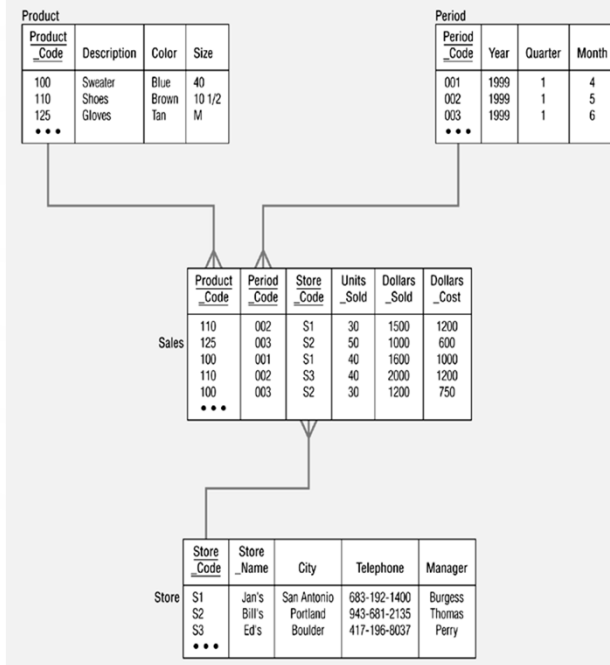


Fig.D-14:
Star
schema
with two
fact tables

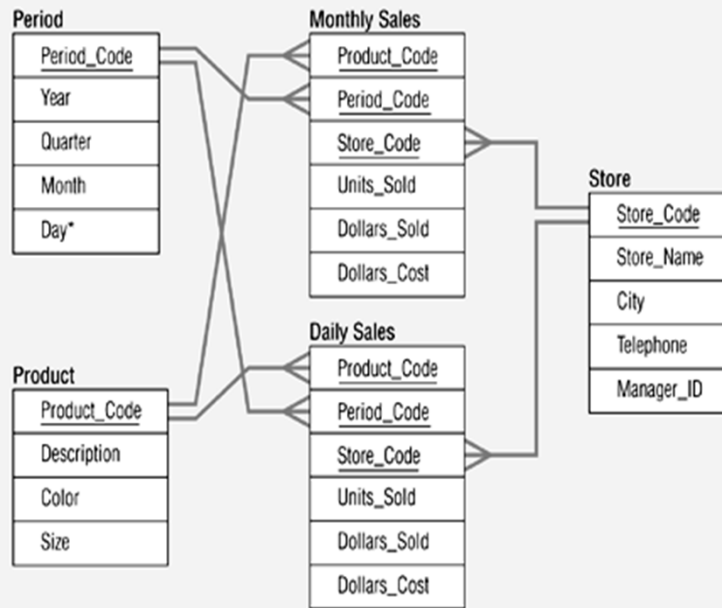
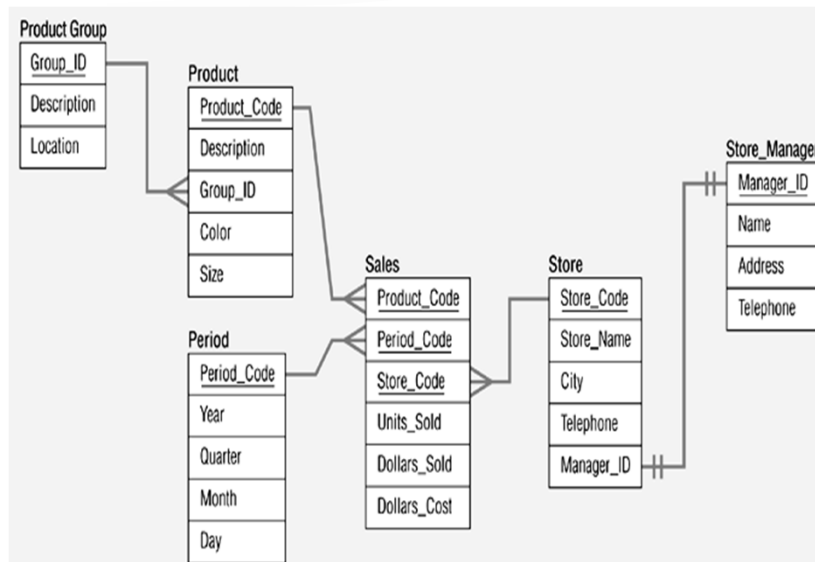


Figure D-15:
Example of snowflake sample



Types of Data Marts

- Dependent - Populated from the EDW.
- Independent - Data taken directly from the operational databases.

The User Interface

- Traditional query and reporting tools.
- On-line analytical processing (OLAP).
 - The use of a set of graphical tools that provides users with multidimensional views of their data and allows them to analyze the data using simple windowing techniques.

The User Interface

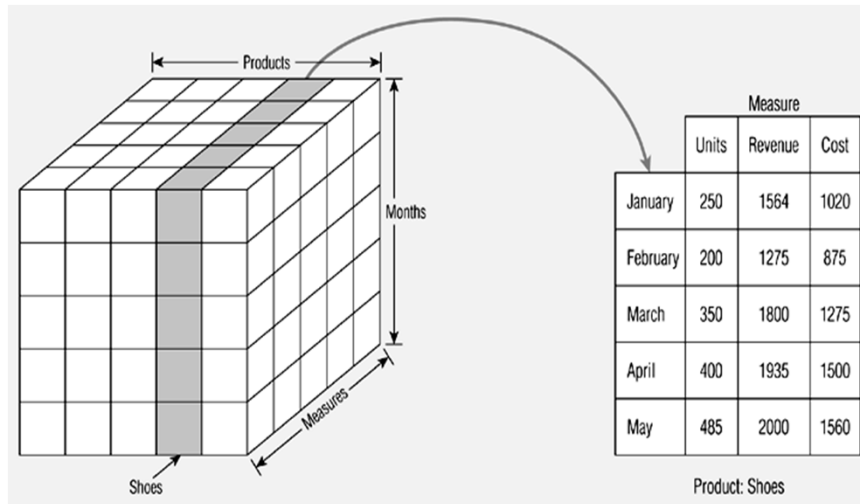
- Slicing a cube - Fig. D-16. ▶
- Pivot
 - Rotate the view for a particular data point to obtain another perspective.
 - For example, take a value from the units column and obtain by-store values.
- Drill-down vs. Aggregate

The User Interface

- Data Mining
 - Knowledge discovery.
 - Search for clusters and patterns in the data.
 - Data visualization: DEMO
 - Data mining techniques - Table D-3. ▶
 - Data mining applications - Table D-4. ▶

Next slide

Figure D-16:
Slicing a data cube



Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.73

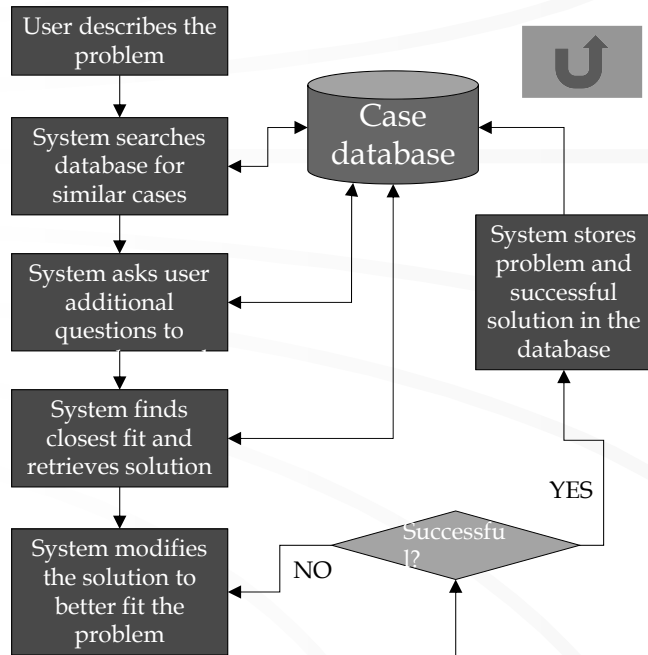
Table D-3:
Data Mining Techniques

Technique	Function
<u>Case-based reasoning</u>	Derives rules from real-world case examples
Rule discovery	Searches for patterns and correlations in large data sets
Signal processing	Identifies clusters of observations with similar characteristics and classifies observations into different pre-defined groups.
<u>Neural nets</u>	Develops predictive models based on principles modeled after the human brain
Fractals	Compresses large databases without losing information

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.74

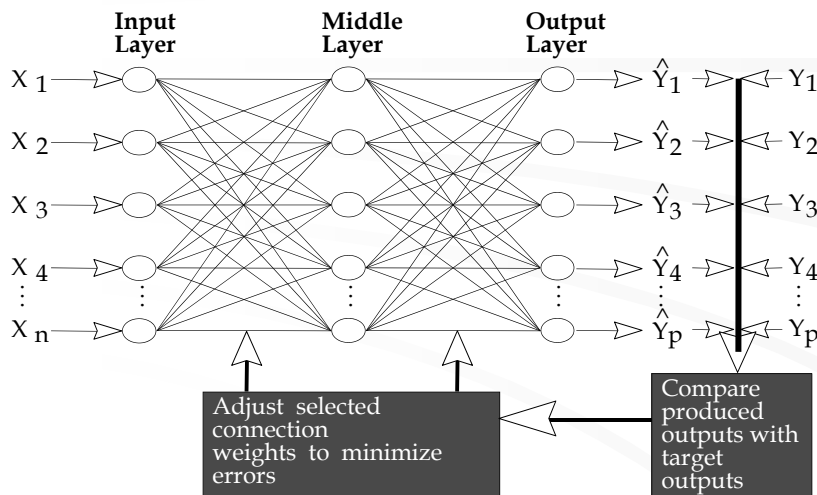
Figure D-17:
Case-Based Reasoning



Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.75

Figure D-18. Neural Network



Note: This feed-back neural network uses supervised learning with training pairs of $(X_1, X_2, \dots, X_n; Y_1, Y_2, \dots, Y_p)$

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.76

Artificial Intelligence

Software that automates and mimics or improves upon tasks that would otherwise require human intelligence.

Machine Learning

Subset of AI. Results improved without explicit programming

Deep Learning

Type of ML. Includes several layers of analysis between input data and output result.

AI can be found in:



Pattern Recognition



Medical diagnosis



Speech Recognition



Computer Vision



Self-driving Automobiles



Natural Language Processing

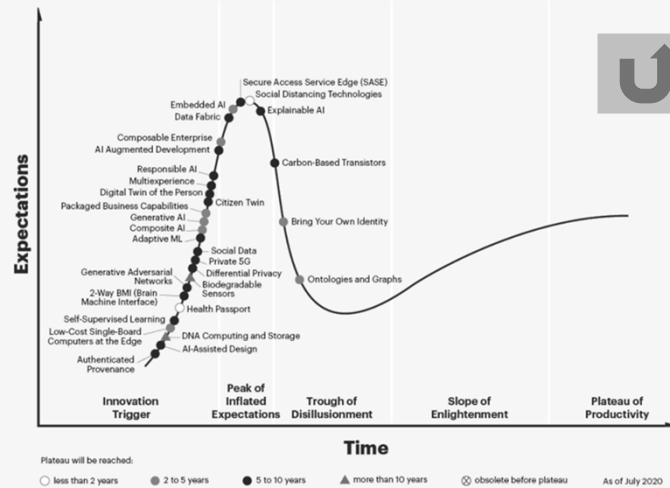
Source: John Gallaugher.



Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.77

Hype Cycle for Emerging Technologies, 2020



[gartner.com/SmarterWithGartner](https://www.gartner.com/SmarterWithGartner)

Source: Gartner
 © 2020 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and Hype Cycle are registered trademarks of Gartner, Inc. and its affiliates in the U.S.

Gartner.

<https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020/>

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.78

Table D-4:
Typical Data Mining Applications



Type of Application	Example
Profiling populations	Developing profiles of high-value customers, credit risks, and credit-card fraud.
Analysis of business trends	Identifying markets with above average (or below average) growth.
Target marketing	Identifying customers (or customer segments) for promotional activity.
Usage analysis	Identifying usage patterns for products and services.
Campaign effectiveness	Comparing campaign strategies for effectiveness.
Product affinity (or market basket)	Identifying products that are purchased concurrently, or the characteristics of shoppers for certain product groups.

For further readings

– Data Mining: TUTORIAL

BOOK

– Data Visualization: TUTORIAL

Data Management Roles

- Data Administration Unit (DAU) is the unit accountable for data management in an organization.
- Database Administrator (DBA) is the position with the responsibility for managing an organization's electronic databases.

Data Management Functions

Data Administration Unit	Database Administrator
<ol style="list-style-type: none">1. Promote and control data sharing,2. Analyze the impact of changes to application systems when data definitions change,3. Maintain metadata,4. Reduce redundant data and processing,5. Reduce system maintenance costs and improve systems development productivity,6. Improve quality and security of data,7. Insure data integrity.	<ol style="list-style-type: none">1. Tuning database management systems,2. Selection and evaluation of and training on database technology,3. Physical database design,4. Design of methods to recover from damage to databases.5. Physical placement of databases on specific computers and storage devices,6. The interface of databases with telecommunications and other technologies.

Data is a corporate resource. Much of our corporate data is stored electronically. Excellence in data management is key to achieving many of our business goals. The following statements constitute our electronic data access policy:

- Corporate data will be shared internally. Data are not owned by a particular individual or organization, but by the whole organization.
- Data will be managed as a corporate resource. Data organization and structure will be planned at the appropriate levels and in an integrated fashion.
- Data quality will be actively managed. Explicit criteria for data accuracy, availability, accessibility, and ease of use will be written by the IS department.
- Data will be safeguarded. As a corporate asset, data will be protected from deliberate or unintentional alteration, destruction, or inappropriate disclosure.
- Data will be defined explicitly. Standards will be developed for data representation.
- Databases will be logically designed to satisfy broad business functions.

Figure 4.8 Example Data Access Policy

Recommendations

- For most organizations today, it is essential to separate informational processing from operational processing by creating a data warehouse (DW).
- Large organizations with many heterogeneous data sources should adopt a three-level DW architecture.
- A successful DW effort requires that a formal program in TQM be implemented as part of the data management effort.

Topic Review

- What is a data warehouse (DW)?
- Why do we need a DW?
- Why is DW getting popular?
- How does DB evolve into DW?
- What does a DW architecture look like?
- What are the characteristics of data in DW?
- How do we create a DW?
- Which data model does a DW have?
- How does a user interface with a DW?
- What is OLAP? What is data mining?

Your Turn

- What are the steps in data reconciliation?
- Do you know of a company that has a data warehouse for big data management?
- Do you use data mining or OLAP tools?

Q & A

Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.87

Case Study II-4 California Franchise Tax Board INC System Project



Copyright © 2020 E.Y. Li

WMU GIMBA BUS 6180 Ch.3: IT MDR p.88

Fiscal Year	NPA's Issued ¹	Returns Filed ²	Total Assessments (millions) ³
2000/2001	87,647	99,376	\$ 261
2001/2002	294,216	151,102	\$1,669
2002/2003	594,212	258,629	\$4,122
2003/2004	499,602	252,103	\$2,986
2004/2005	528,856	248,766	\$2,115

Notes: 1. Notices of Proposed Assessment.
2. The system tracks non-filer accounts from issuance of the demand for a return until account resolution.
3. Total assessments include tax, penalties, fees, and interest.

EXHIBIT 3 Non-filers Detected Through the INC System
Source: http://www.ftb.ca.gov/aboutFTB/taxpayer_advocate/2006_BillRghtsAnnIRpt.pdf. Accessed July 9, 2010.

Proposed Source	New Taxpayers ¹	Expected Value	Explanation
City Business Tax	14,287	\$1,271,543	Self-employed in cities with license
Community Care Licensing	4,312	\$ 866,712	Self-employed care facility providers
Alcoholic Beverage Control	3,569	\$ 717,369	Self-employed seller of liquor/wine
Motor Fuel Data	1,664	\$ 334,866	Self-employed truckers

Notes: 1. "New taxpayers" are non-filers identified via this source.
2. Example calculation: Community Care Licensing: The California Department of Social Services licenses more than 88,000 care facilities for children, adults, and the elderly. Applying the typical self-employed non-filer rate of 4.9% × 88,000 = 4,312 contracts × \$201 taxes owed = \$866,712.

EXHIBIT 6 Expected Value of Proposed New Indirect Data Sources

Copyright © 2020 E.Y. Li WMU GIMBA BUS 6180 Ch.3: IT MDR p.89

	Revenue Per Case	Data Provider
Federal Data Sources		
1099-INT (Interest income)	\$1,784	Internal Revenue Service (IRS)
K-1 Sub S (Partnership income)	\$1,436	IRS
1099-G (Tuition program payments)	\$1,322	IRS
1099-PATR (Income from cooperatives)	\$1,265	IRS
K-1 P/S (Partnership income)	\$1,253	IRS
1099-OID (Original issue discount)	\$1,119	IRS
1099R (Pensions or profit sharing)	\$ 837	IRS
1099-MISC (Miscellaneous income)	\$ 749	IRS
IRS listing of Californians filing Federal returns	\$ 453	IRS
California Data Sources		
CA Sales Tax Return	\$ 993	Board of Equalization
CA EDD Wage data	\$ 626	Employment Development Department
CA EDD Employer data	\$ 555	Employment Development Department

EXHIBIT 5 Direct Income Sources Utilized in the INC System

Copyright © 2020 E.Y. Li WMU GIMBA BUS 6180 Ch.3: IT MDR p.90

Case Analysis Tasks

1. Identify the most important or the most critical issue that leads to the problem in the minicase;
2. Analyze this most important or most critical issue; then, provide your suggestions about what should be done;
3. Identify any additional issues;
4. Analyze these additional issues; then, provide your suggestions about what should be done and set the priority for each of them.

