Chapter 3:
BI using Data Warehousing

- Introduction to DW
- DW architecture [ch7 paulraj] [ch 3.3 Han Kamber]
- ETL Process[chapt 12 paulraj]
- Top-down and bottom-up approaches, characteristics and benefits of data mart[ch 2 paulraj]
- Difference between OLAP[ch 15 paulraj] and OLTP.
- Dimensional analysis[ch 5 paulraj]- Define cubes. Drill-down and roll- up – slice and dice or rotation
- OLAP models- ROLAP and MOLAP[ch 15 paulraj]
- Define Schemas- Star, snowflake and fact constellations [chapt 10&11 paulraj] [ch 3.2 Han Kamber]

Chapter 3: Data Warehousing and OLAP Technology: An Overview

- What is a data warehouse?
- A multi-dimensional data model
- Data warehouse architecture
- Data warehouse implementation
- From data warehousing to data mining

A producer wants to know….

Which are our lowest/highest margin customers ?

What is the most effective distribution channel?

Who are my customers and what products are they buying?

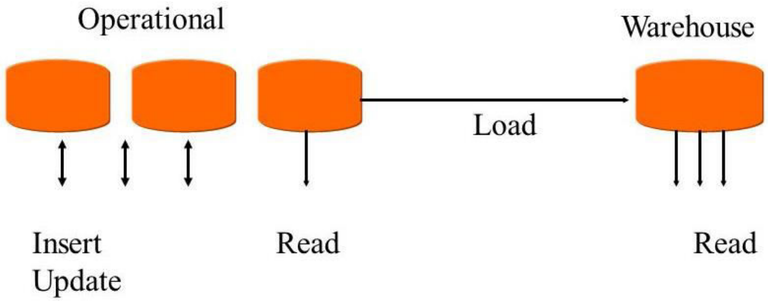
What product prom- -otions have the biggest impact on revenue?

Which customers are most likely to go to the competition ?

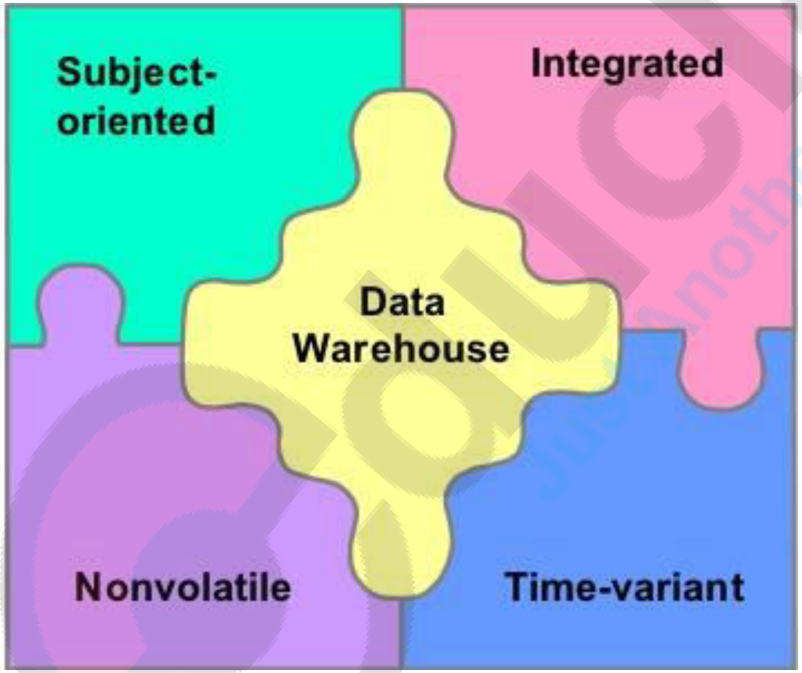
What impact will new products/services have on revenue and margins?

What is Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization’s operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
  - They are static with infrequent updates, mostly read only data.
  - Integrated from several heterogeneous operational databases DW is a standalone repository.
- Data warehousing:
  - the **process** of constructing and using data warehouses

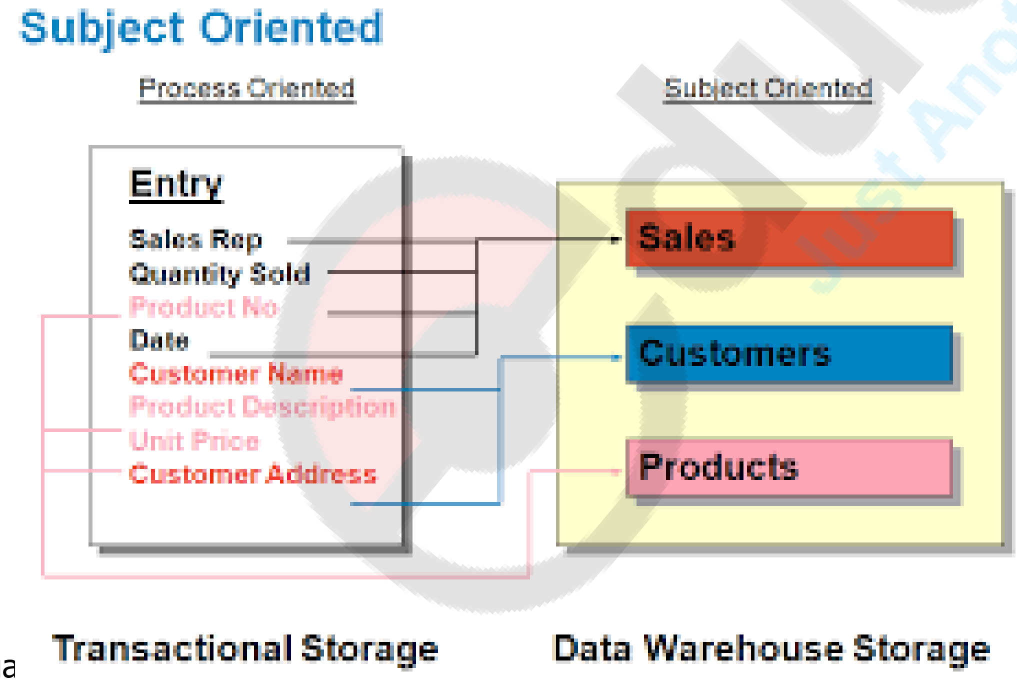


- “A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision-making process.”—W. H. Inmon



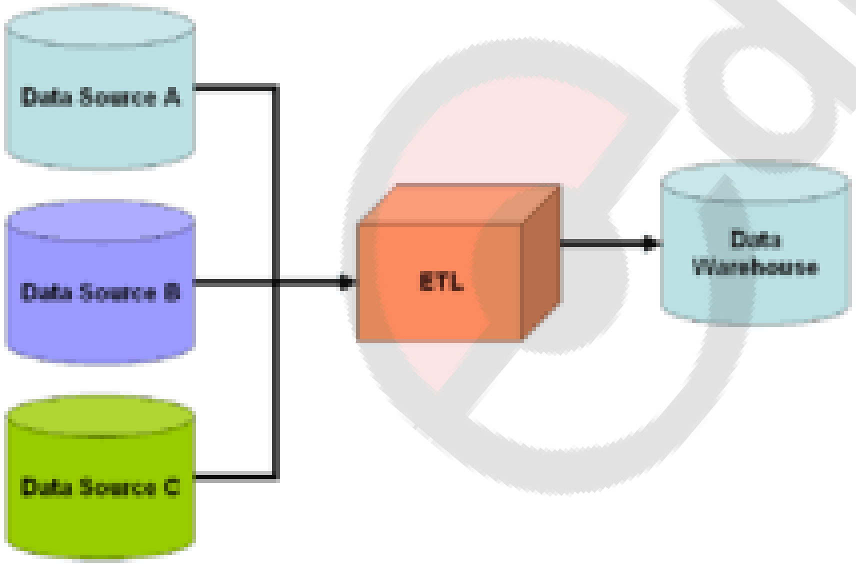
Data Warehouse—Subject-Oriented

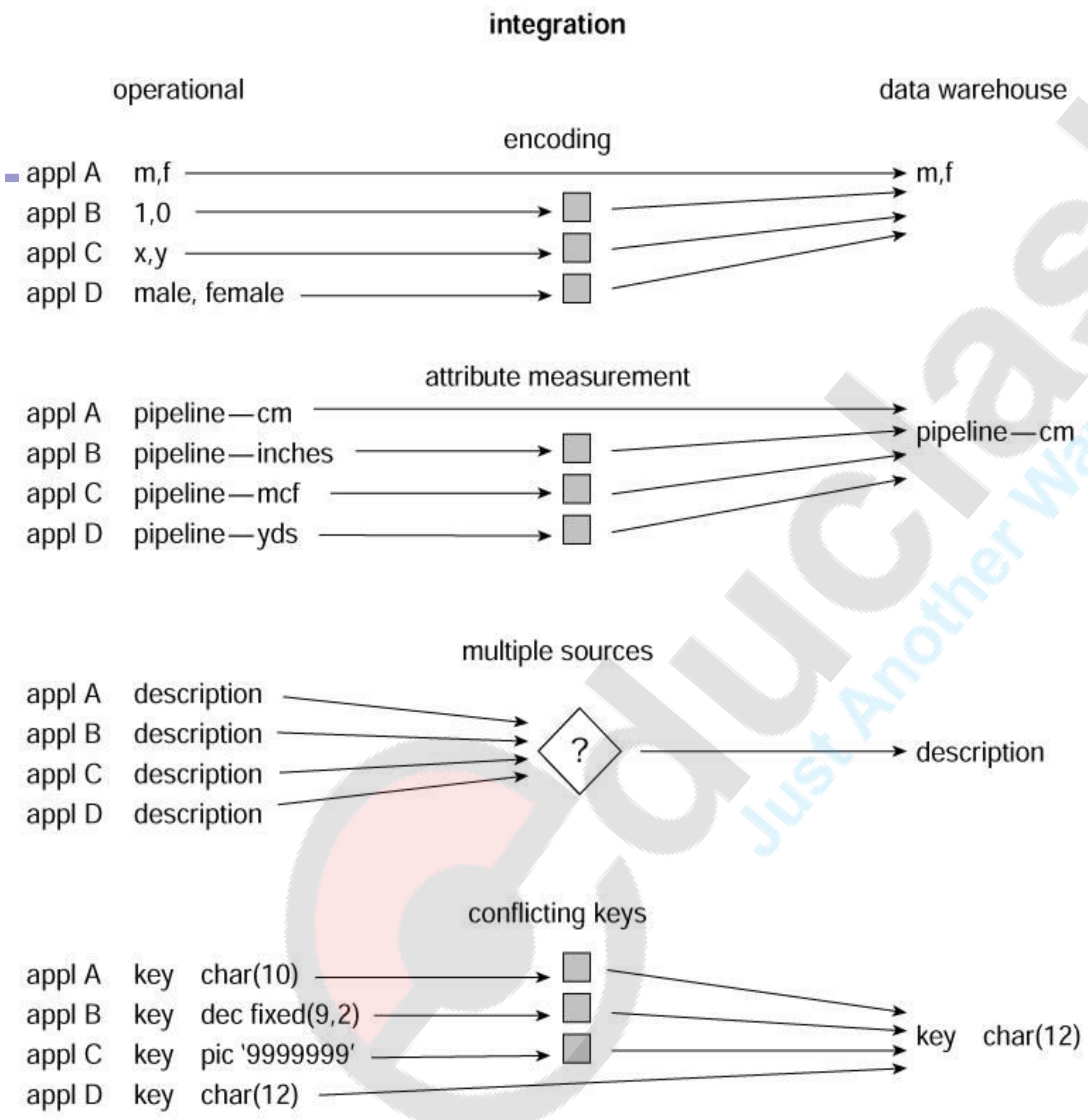
- Organized around major subjects, such as customer, product, sales
- Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process



Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
    - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
    - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
        - E.g., Hotel price: currency, tax, breakfast covered, etc.
    - When data is moved to the warehouse, it is converted.





integration

operational

data warehouse

encoding

appl A m,f

appl B 1,0

appl C x,y

appl D male, female

m,f

attribute measurement

appl A pipeline—cm

appl B pipeline—inches

appl C pipeline—mcf

appl D pipeline—yds

pipeline—cm

multiple sources

appl A description

appl B description

appl C description

appl D description

?

description

conflicting keys

appl A key char(10)

appl B key dec fixed(9,2)

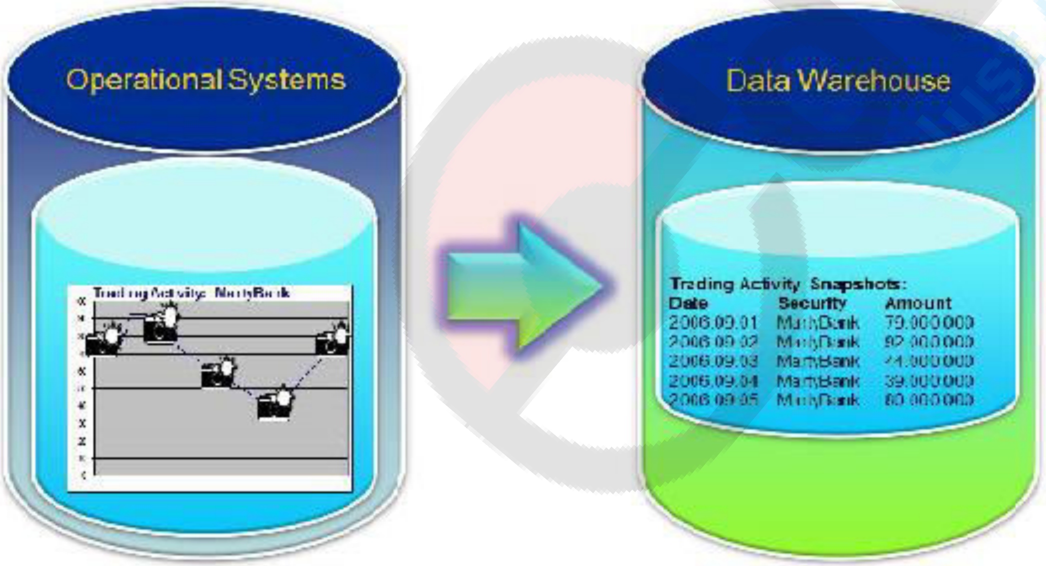
appl C key pic '9999999'

appl D key char(12)

key char(12)

Data Warehouse—Time Variant

- The time horizon for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time, explicitly or implicitly
  - But the key of operational data may or may not contain "time element"



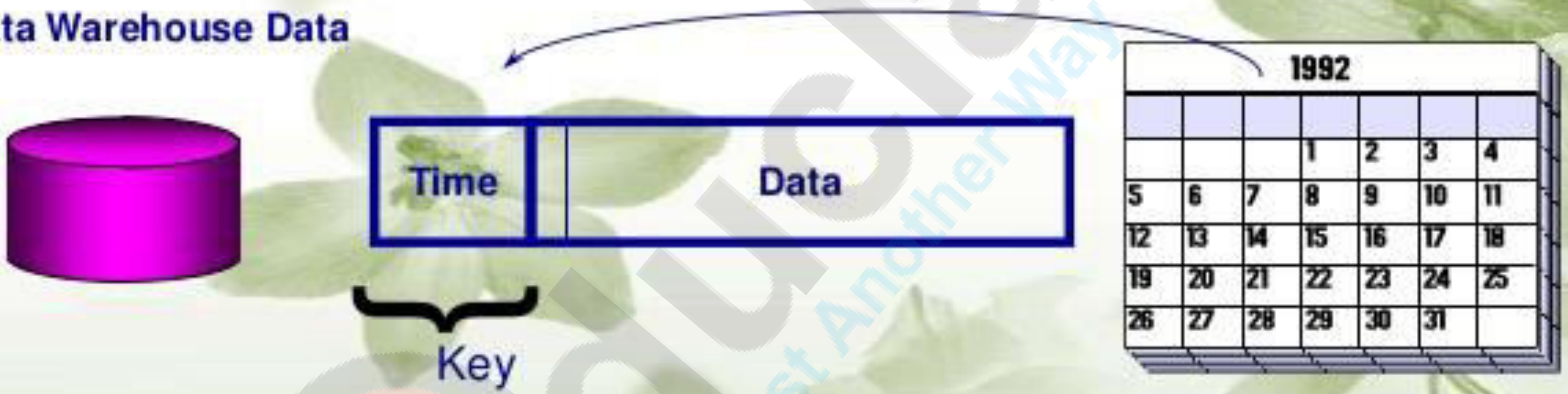
Trading Activity Snapshots:

Date	Security	Amount
2006.09.01	MartyBank	79.000.000
2006.09.02	MartyBank	92.000.000
2006.09.03	MartyBank	71.000.000
2006.09.04	MartyBank	39.000.000
2006.09.05	MartyBank	80.000.000

Time - Variant

- **Data is stored as a series of snapshots or views which record how it is collected across time.**

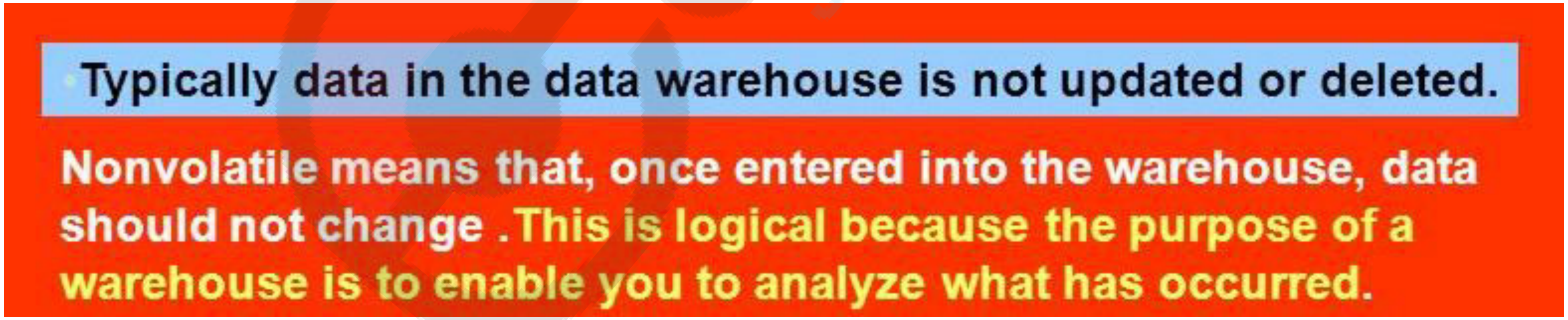
Data Warehouse Data



- Data is tagged with some element of time - creation date, as of date, etc.
- Data is available on-line for long periods of time for trend analysis and forecasting. For example, five or more years

Data Warehouse—Nonvolatile

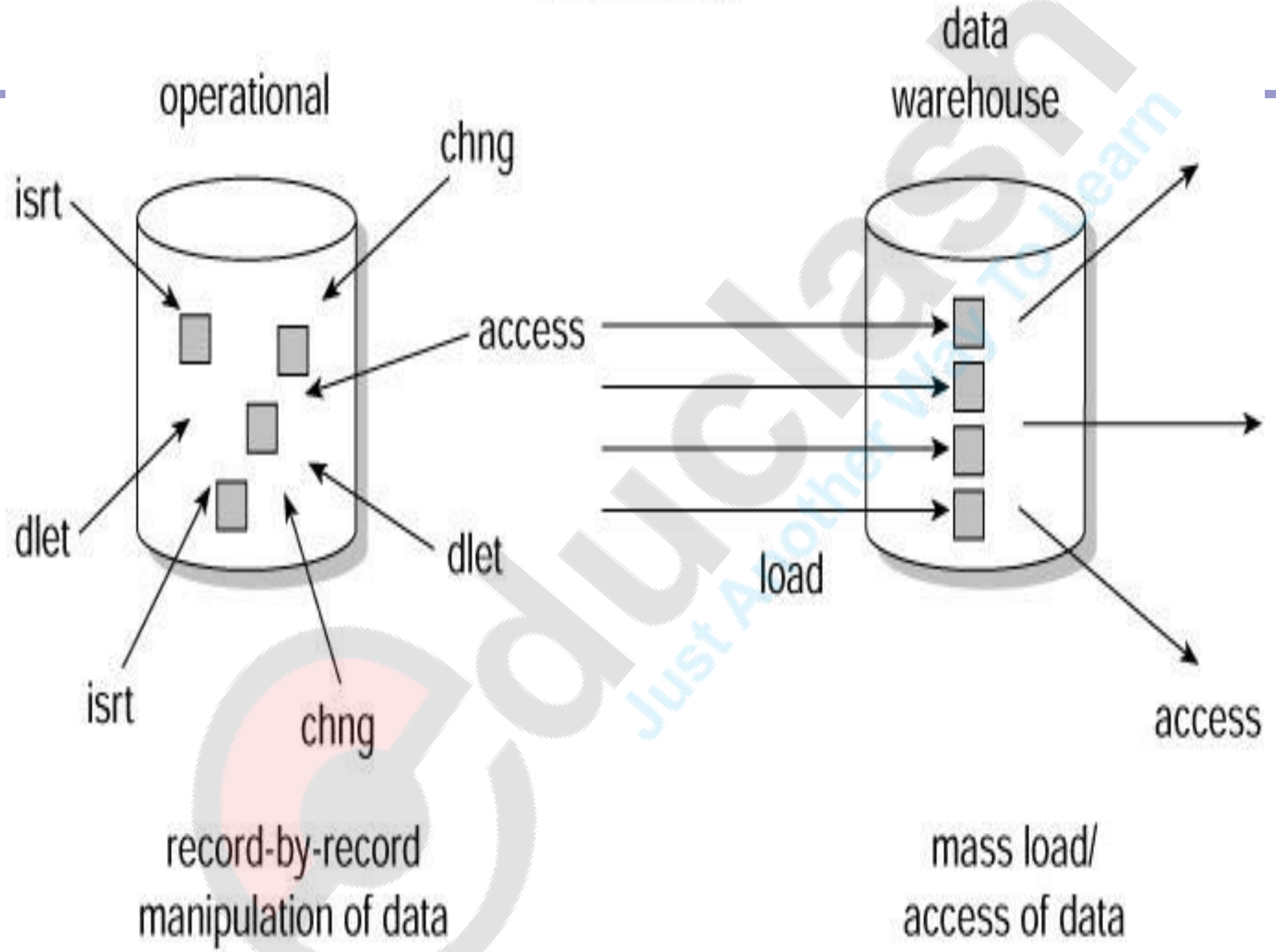
- A physically separate store of data transformed from the operational environment
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*



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

Nonvolatile means that, once entered into the warehouse, data should not change .This is logical because the purpose of a warehouse is to enable you to analyze what has occurred.

nonvolatility



The goals of a Data Warehouse

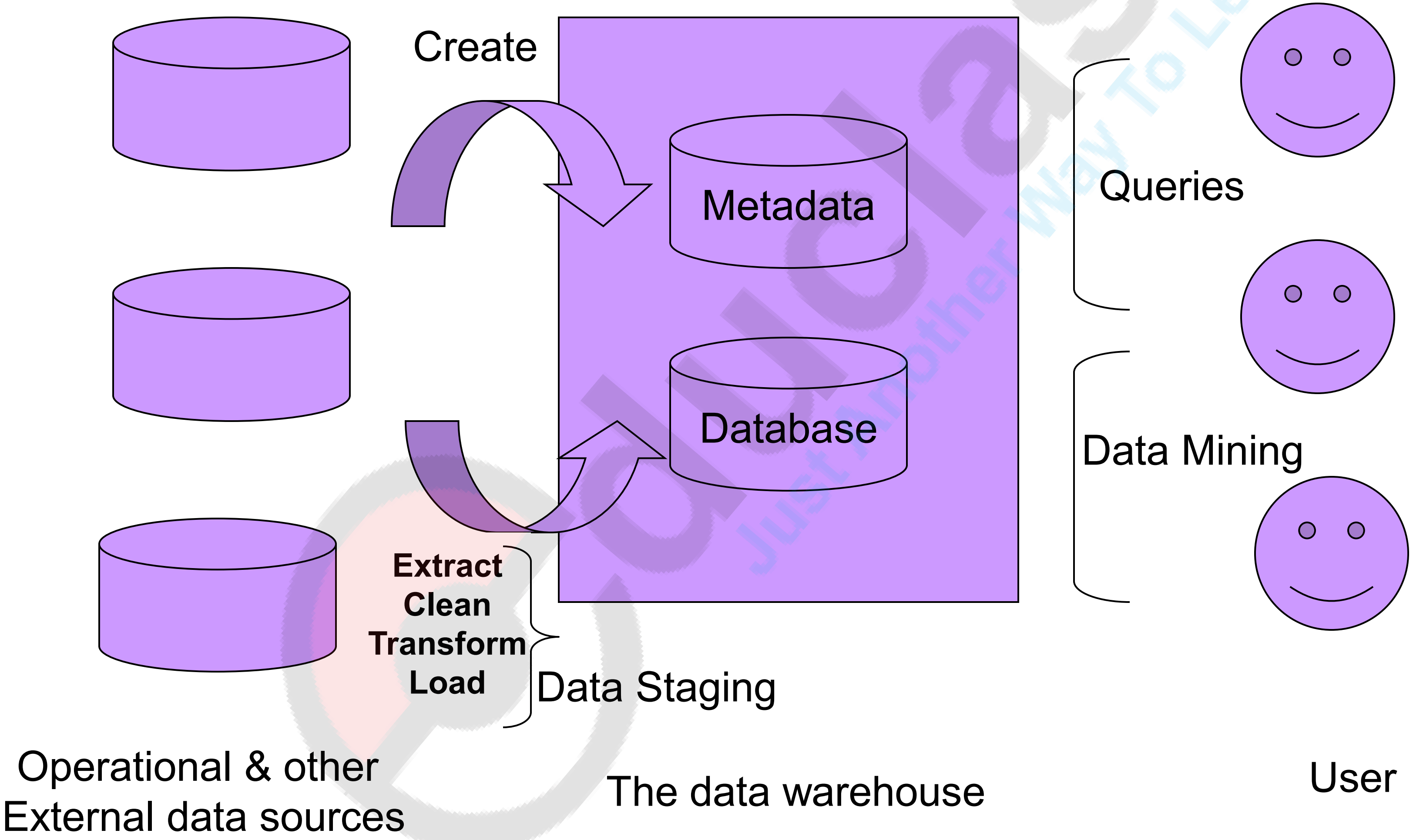
- We have mountains of data in this company, but we can't access it."
- "We need to slice and dice the data every which way."
- "You've got to make it easy for business people to get at the data directly."
- "Just let me know what is important."
- "It drives me crazy to have two people present the same business metrics at a meeting, but with different numbers."
- "We want people to use information to support more fact-based decision making."

The goals of a Data Warehouse

- The data warehouse must make an organization's information easily accessible.
- The data warehouse must present the organization's information consistently.
- The data warehouse must be adaptive and resilient to change.
- The data warehouse must be a secure bastion that protects our information assets.
- The data warehouse must serve as the foundation for improved decision making.
- The business community must accept the data warehouse if it is to be deemed successful.

- Data warehouse is not a single software or hardware product you purchase to provide strategic information.
- it is a computing environment where users can find strategic information to make strategic decisions.

Data Warehouse Architecture



Data Warehouse Architecture

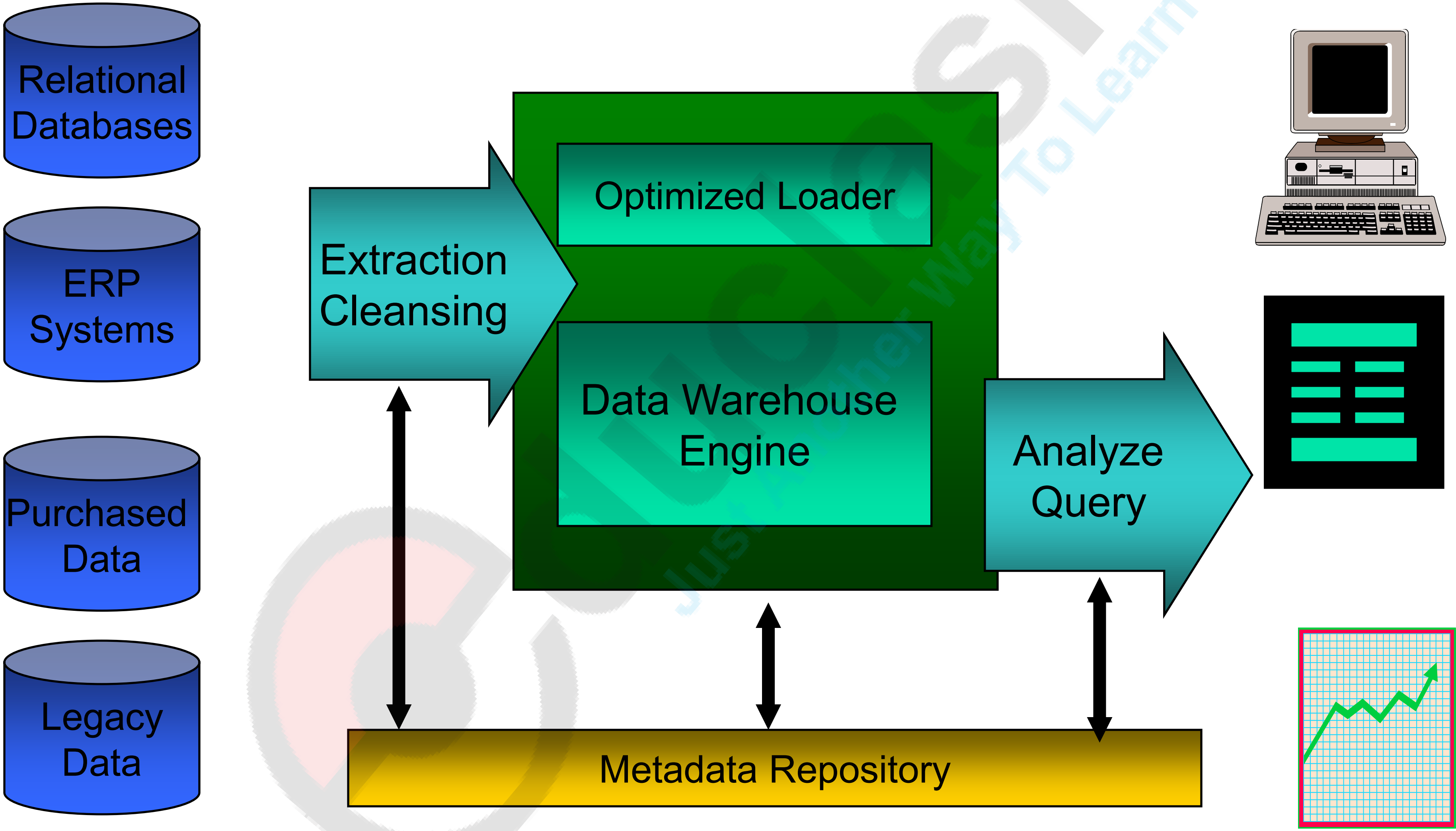
The major elements of a data warehouse and the major external entities with which a data warehouse interacts include:-

- The transaction or other operational databases from which the data warehouse is populated. External data is also fed into some data warehouse.
- A process to extract data from this database and bring it into the data warehouse.
- A process to transform the data into the database structure & internal formats of the warehouse
- A process to cleanse the data, to make sure it is of sufficient quality for the decision making purposes for which it will be used.
- A process to load the cleansed data into the data warehouse database.

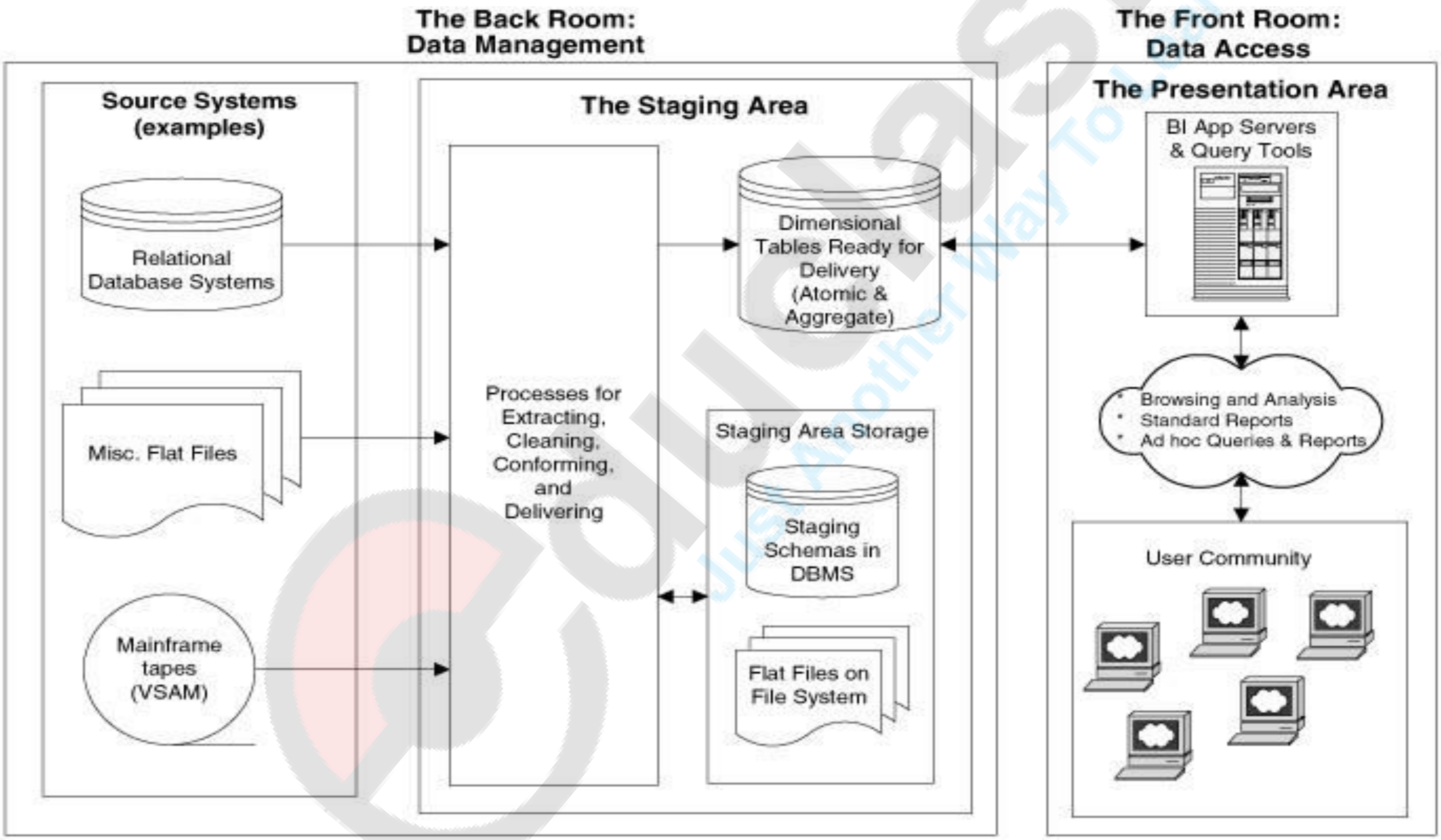
Data Staging Area

- A storage area where extracted data is cleaned, transformed and deduplicated.
- Initial storage for data
- Need not be based on Relational model
- Mainly sorting and Sequential processing
- Does not provide data access to users
- Analogy – kitchen of a restaurant

Data Warehouse Architecture



Data Warehouse Architecture



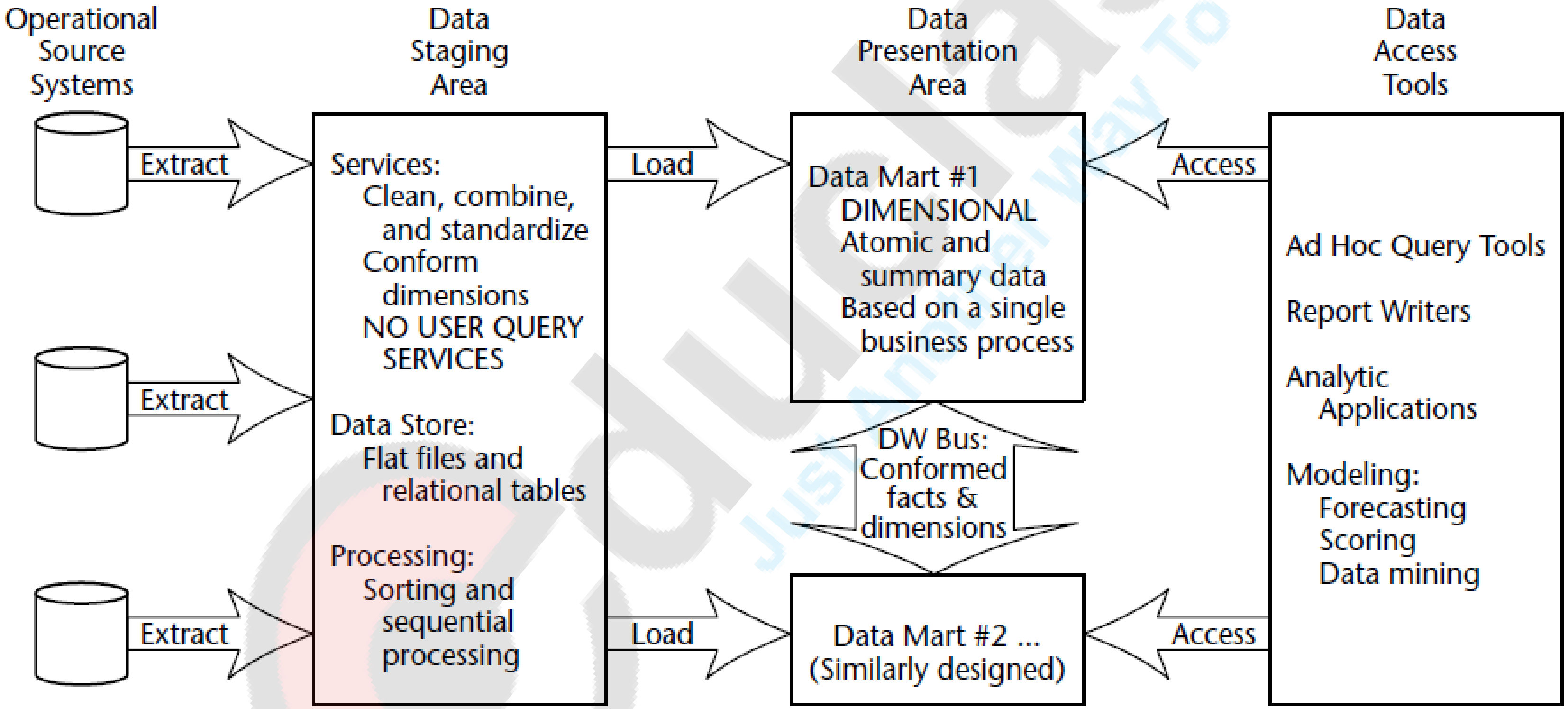


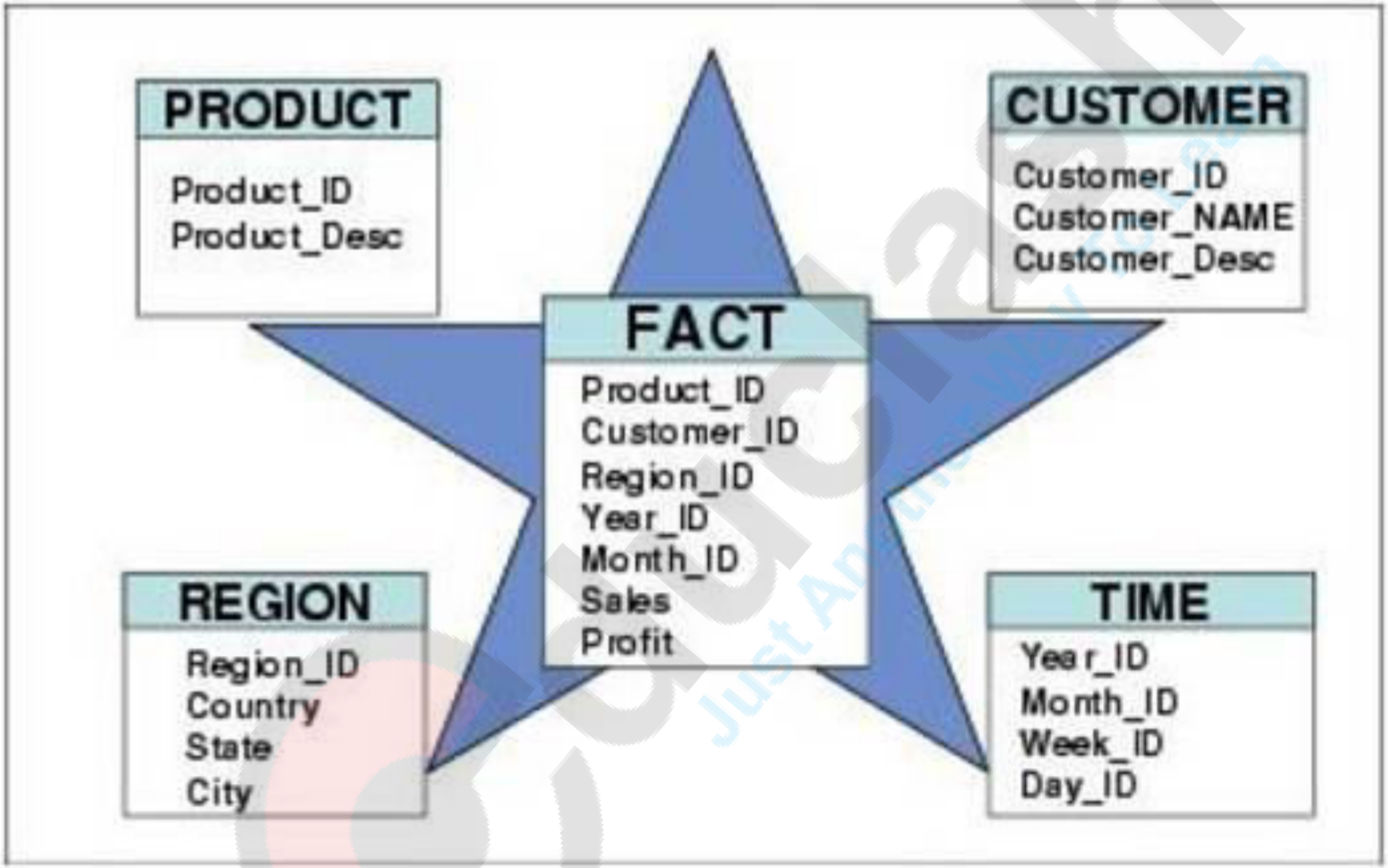
Figure 1.1 Basic elements of the data warehouse.

The benefits of data warehousing

- The potential benefits of data warehousing are high returns on investment.
- substantial competitive advantage.
- increased productivity of corporate decision-makers.

Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation



PRODUCT

Product_ID
Product_Desc

CUSTOMER

Customer_ID
Customer_NAME
Customer_Desc

FACT

Product_ID
Customer_ID
Region_ID
Year_ID
Month_ID
Sales
Profit

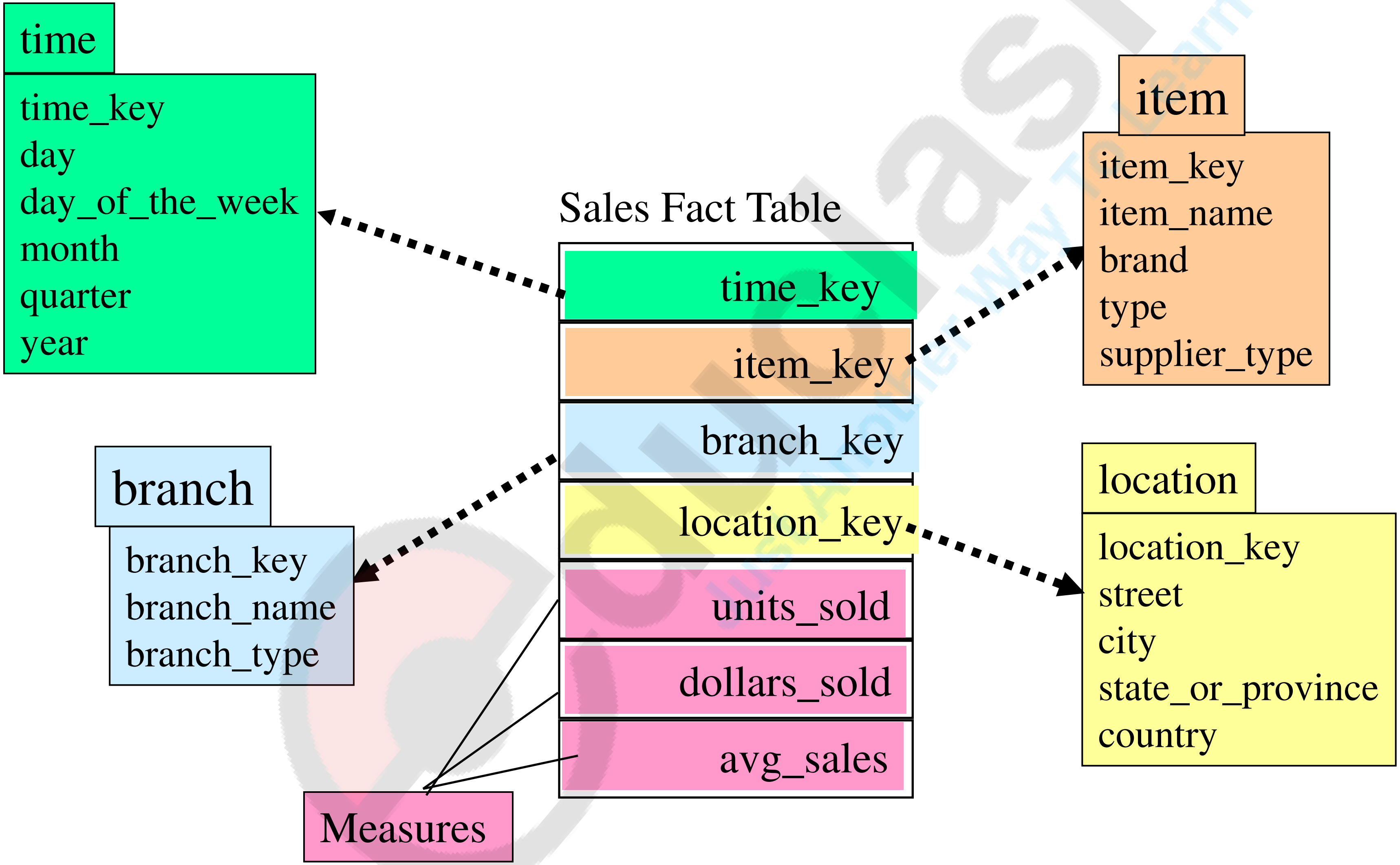
REGION

Region_ID
Country
State
City

TIME

Year_ID
Month_ID
Week_ID
Day_ID

Example of Star Schema



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

item

item_key
item_name
brand
type
supplier_type

branch

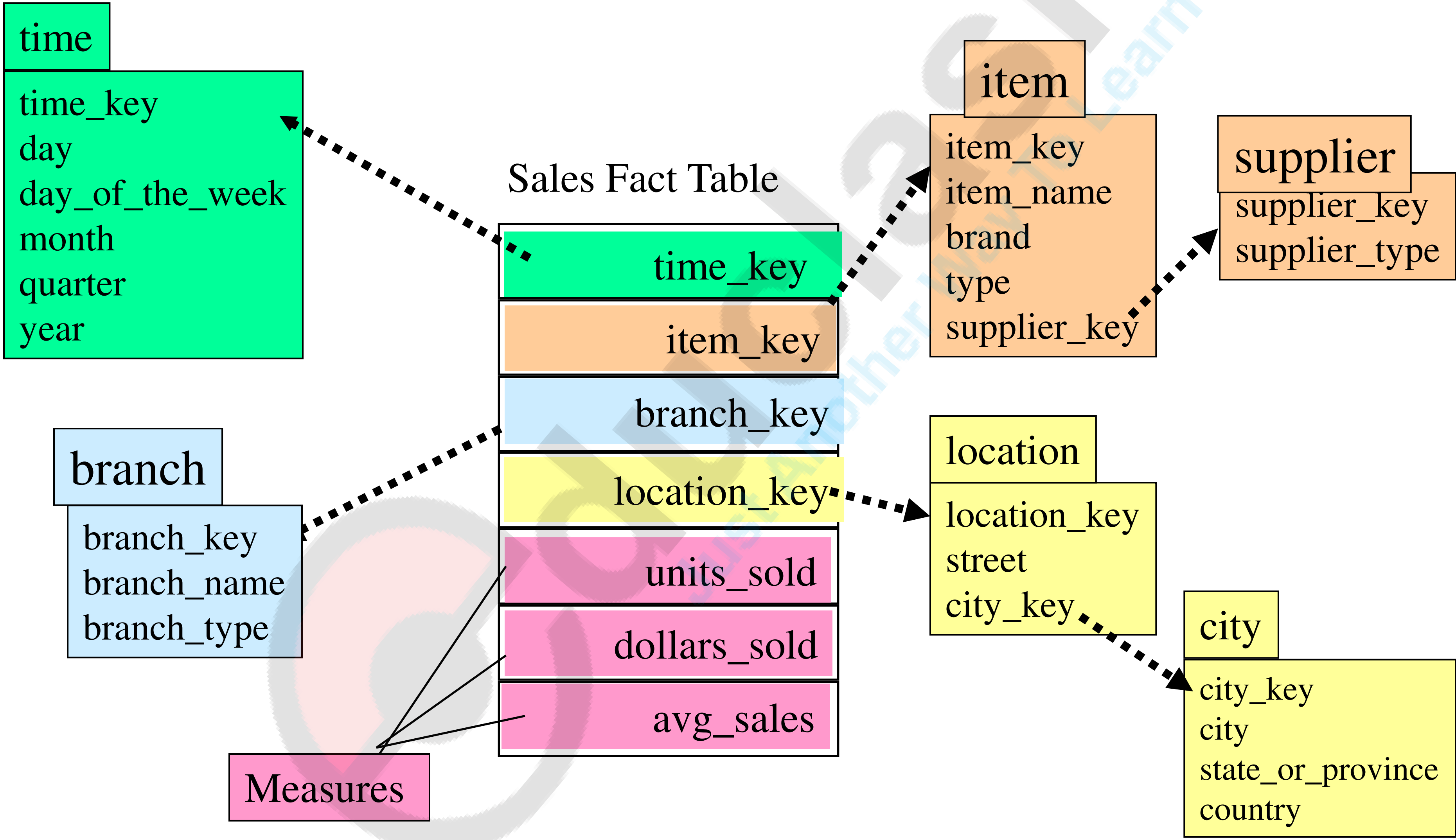
branch_key
branch_name
branch_type

location

location_key
street
city
state_or_province
country

Measures

Example of Snowflake Schema



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

branch

branch_key
branch_name
branch_type

item

item_key
item_name
brand
type
supplier_key

supplier

supplier_key
supplier_type

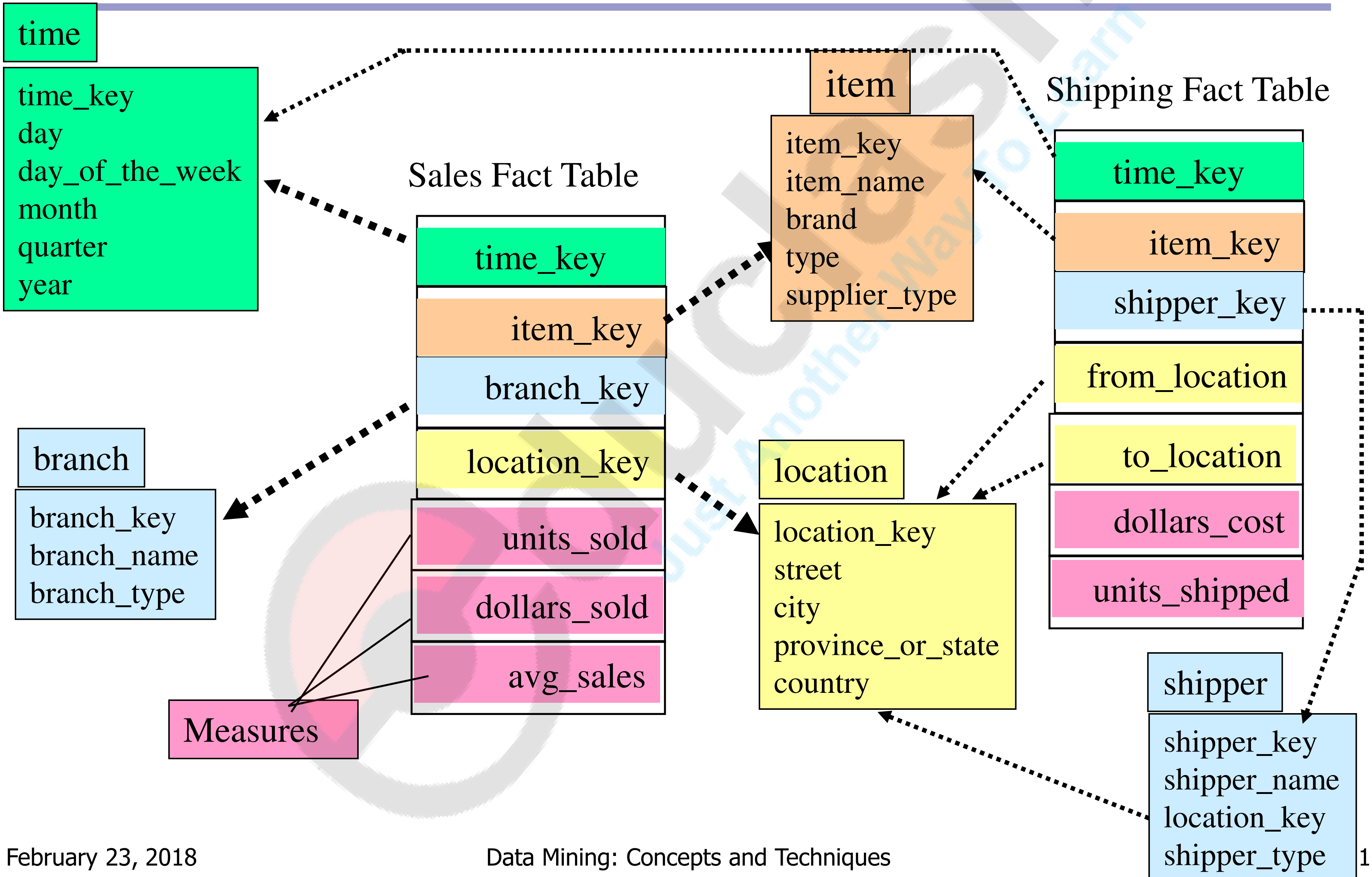
location

location_key
street
city_key

city

city_key
city
state_or_province
country

Example of Fact Constellation



time

time_key
day
day_of_the_week
month
quarter
year

Sales Fact Table

time_key
item_key
branch_key
location_key
units_sold
dollars_sold
avg_sales

Measures

branch

branch_key
branch_name
branch_type

item

item_key
item_name
brand
type
supplier_type

location

location_key
street
city
province_or_state
country

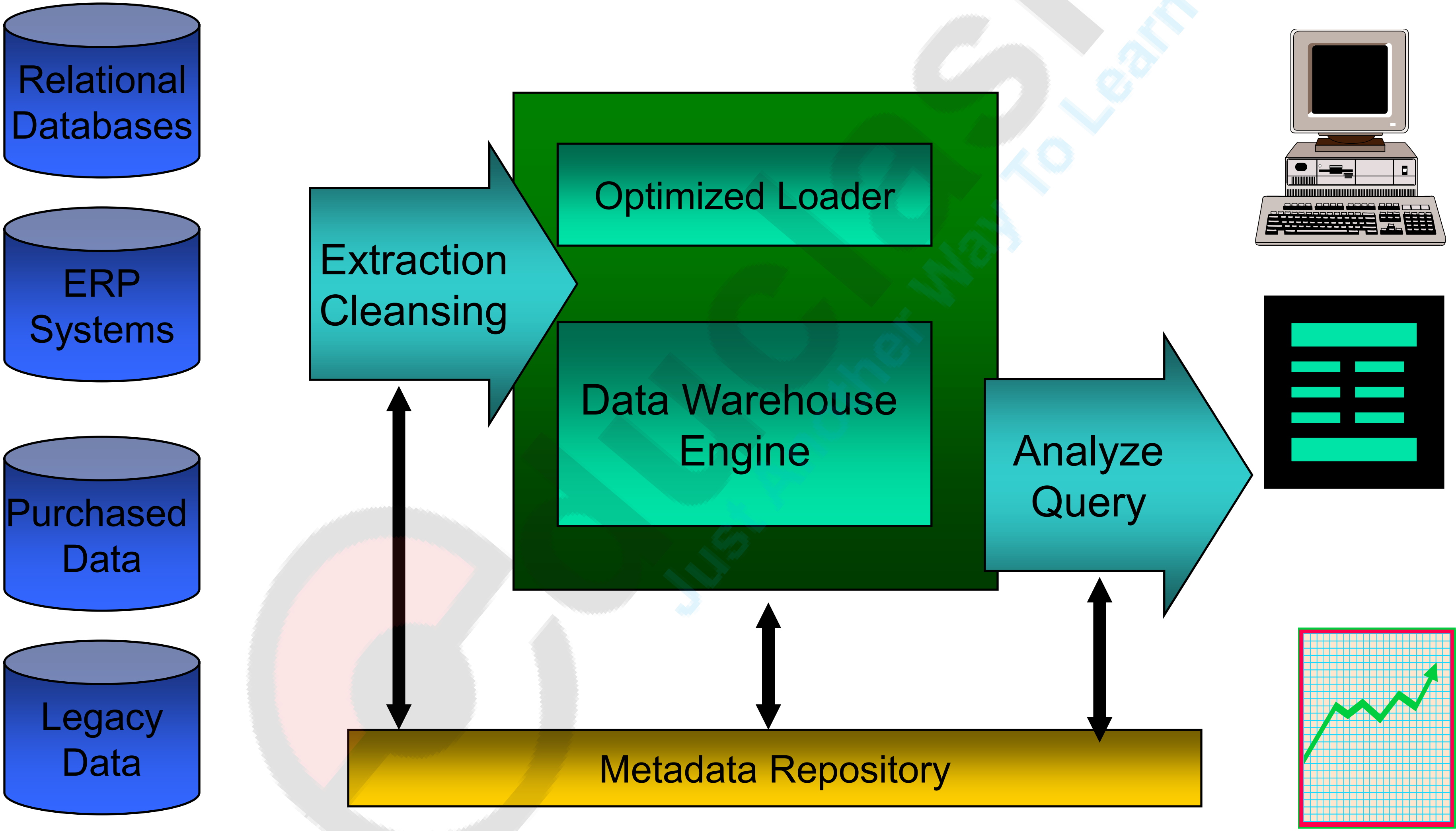
Shipping Fact Table

time_key
item_key
shipper_key
from_location
to_location
dollars_cost
units_shipped

shipper

shipper_key
shipper_name
location_key
shipper_type

Metadata


Relational Databases
ERP Systems
Purchased Data
Legacy Data
Extraction Cleansing
Optimized Loader
Data Warehouse Engine
Analyze Query
Metadata Repository

Metadata Repository

- Meta data is the data defining warehouse objects. It stores:
- Description of the structure of the data warehouse
  - schema, view, dimensions, hierarchies, derived data defn, data mart locations and contents
- Operational meta-data
  - data lineage (history of migrated data and transformation path), currency of data (active, archived, or purged), monitoring information (warehouse usage statistics, error reports, audit trails)
- The algorithms used for summarization
- The mapping from operational environment to the data warehouse
- Data related to system performance
  - warehouse schema, view and derived data definitions
- Business data
  - business terms and definitions, ownership of data, charging policies

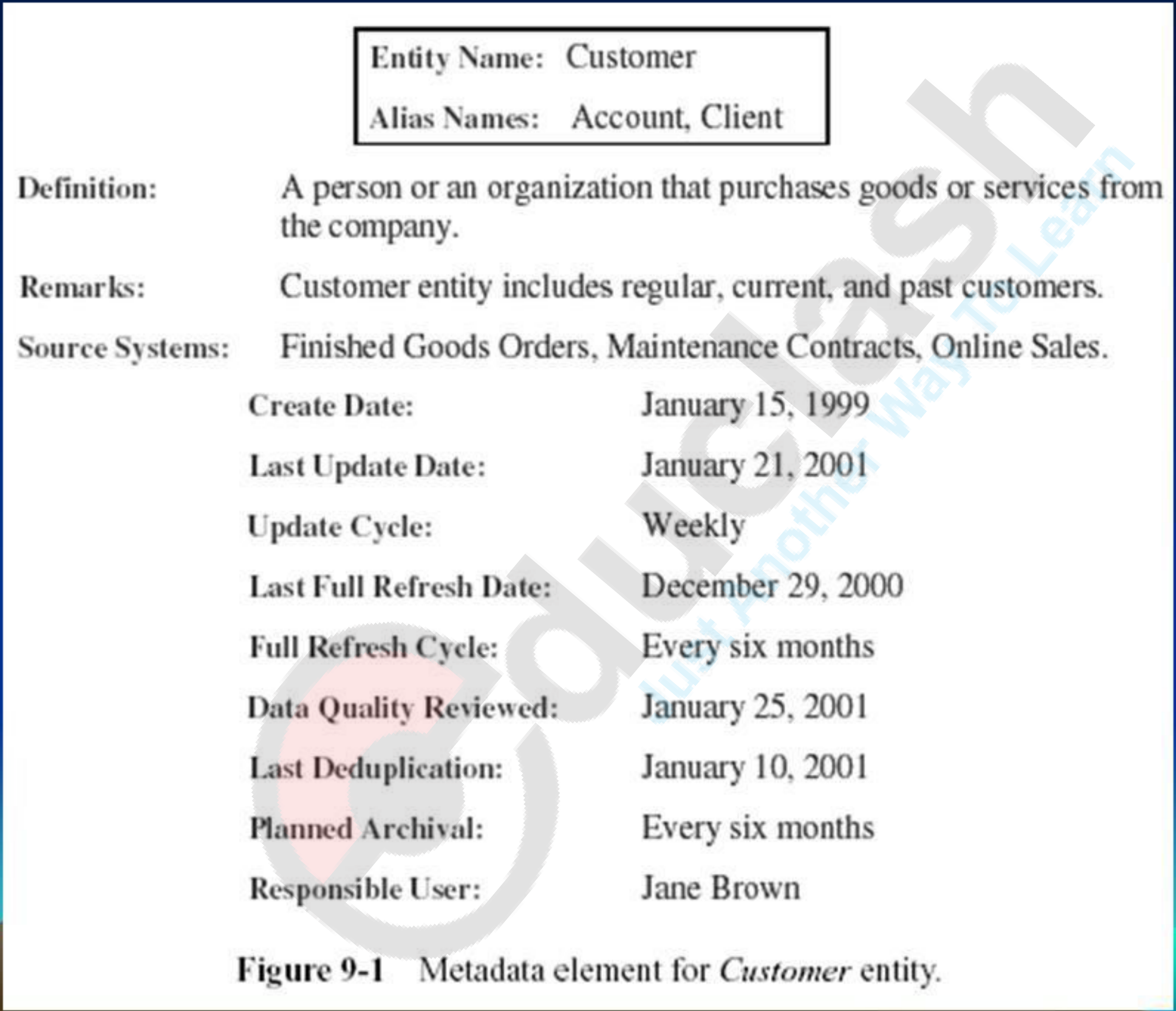
Metadata

- Data about data, data dictionary, data catalog
- Keeps info about the logical data structures, files and addresses , indexes, etc.
- Types are:
  - Operational Metadata:
    - data from various operational sources are combined, records are split, combine parts of records, multiple coding schemes and different fields lengths and data types.
    - To deliver info you need to tie them back together
  - Extraction & transformation metadata:
    - Extraction frequencies, Extraction methods and Extraction business rules need to be recorded. source system info,
    - Contains info about all transformations taking place in staging area.
  - End User Metadata:
    - Navigation map of DW for the end user
    - Allows end user to use its own business terminology and look for info

Metadata

Helps:

- As a glue to connect all parts of DW.
- Provide info to the developer about content and structure *(IT personnel need to know data sources and targets; database, table and column names; refresh schedules; data usage measures; etc.)*
- Content recognizable in end users terms *(Users need to know entity/attribute definitions; reports/query tools available; report distribution information; help desk contact information, etc.)*
- It is useful to have a central information repository to tell users what’s in the data warehouse, where it came from, who is in charge of it etc.
- The metadata can also tell query tools what’s in the data warehouse, where to find it, who is authorized to access it etc.



Entity Name: Customer

Alias Names: Account, Client

Definition: A person or an organization that purchases goods or services from the company.

Remarks: Customer entity includes regular, current, and past customers.

Source Systems: Finished Goods Orders, Maintenance Contracts, Online Sales.

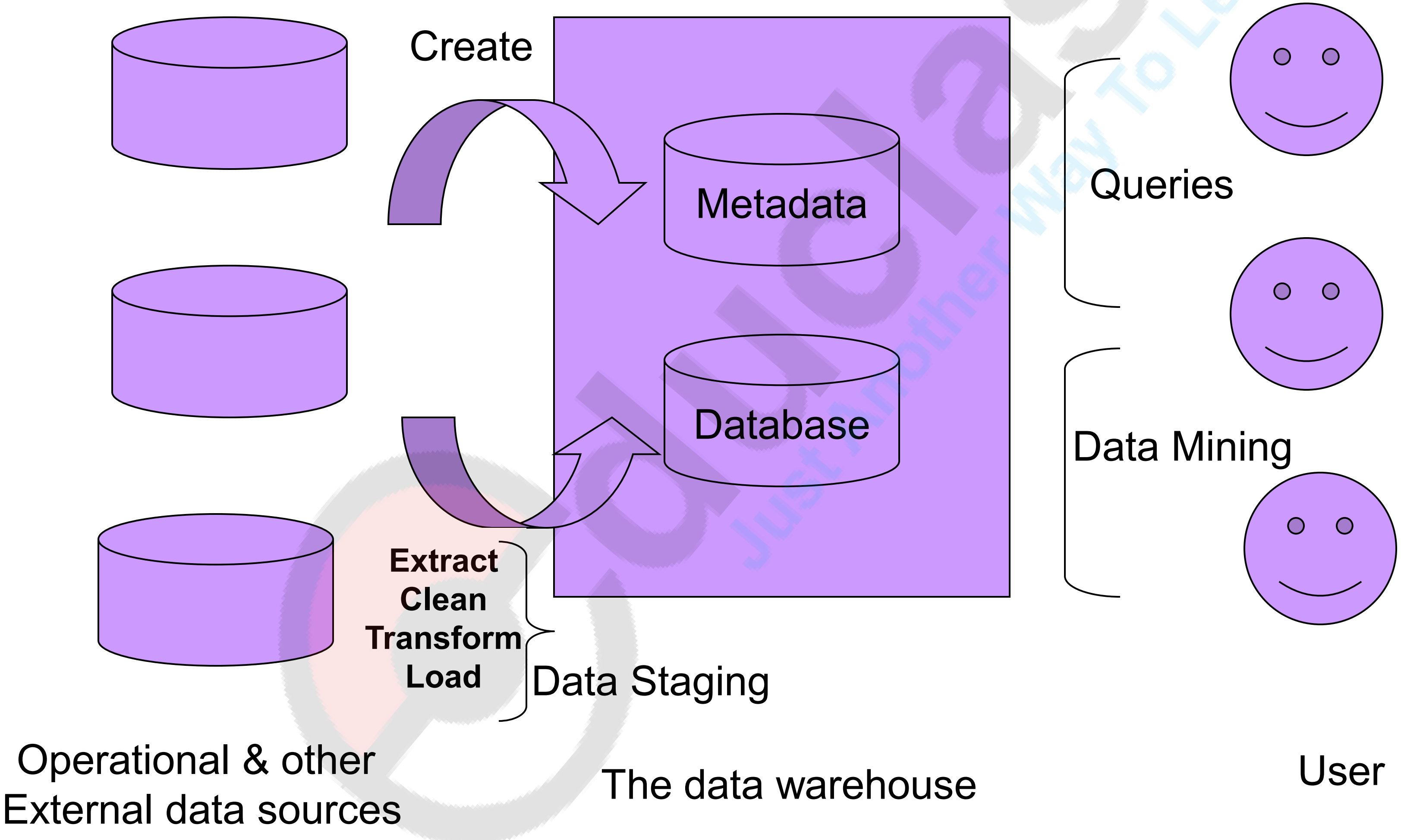
Create Date:	January 15, 1999
Last Update Date:	January 21, 2001
Update Cycle:	Weekly
Last Full Refresh Date:	December 29, 2000
Full Refresh Cycle:	Every six months
Data Quality Reviewed:	January 25, 2001
Last Deduplication:	January 10, 2001
Planned Archival:	Every six months
Responsible User:	Jane Brown

Figure 9-1 Metadata element for *Customer* entity.

Data Staging Area: ETL

- A storage area where extracted data is cleaned, transformed and deduplicated.
- Initial storage for data
- Need not be based on Relational model
- Mainly sorting and Sequential processing
- Does not provide data access to users
- Analogy – kitchen of a restaurant

ETL


Create
Metadata
Database
Queries
Data Mining
Extract
Clean
Transform
Load
Data Staging
Operational & other
External data sources
The data warehouse
User

- Dimensional analysis[ch 5 paulraj]
- Define cubes.
- Drill- down and roll- up – slice and dice or rotation
- OLAP models- ROLAP and MOLAP[ch 15 paulraj]
- Define Schemas- Star, snowflake and fact constellations [chapt 10&11 paulraj] [ch 3.2 Han Kamber]

Data Warehouse vs. Heterogeneous DBMS

- Traditional heterogeneous DB integration: A query driven approach
  - Build wrappers/mediators on top of heterogeneous databases
  - When a query is posed to a client site, a meta-dictionary is used to translate the query into queries appropriate for individual heterogeneous sites involved, and the results are integrated into a global answer set
  - Complex information filtering, compete for resources
- Data warehouse: update-driven, high performance
  - Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why Separate Data Warehouse?

- High performance for both systems
  - DBMS— tuned for OLTP: access methods, indexing, concurrency control, recovery
  - Warehouse—tuned for OLAP: complex OLAP queries, multidimensional view, consolidation
- Different functions and different data:
  - missing data: Decision support requires historical data which operational DBs do not typically maintain
  - data consolidation: DS requires consolidation (aggregation, summarization) of data from heterogeneous sources
  - data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconciled
- Note: There are more and more systems which perform OLAP analysis directly on relational databases

From Tables and Spreadsheets to Data Cubes

- A data warehouse is based on a multidimensional data model which views data in the form of a data cube
- A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions
  - Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

1. Short note on:
    1. Data Mart(DM)
    2. Data Quality -----2015-KT
2. Differentiate between
    1. DW Vs DM -----2015-KT, 2014-KT, 2016-KT
    2. Operational system Vs informational system -----2016-KT
3.
4. Compare and contrast OLTP & DW.
5. What is a data warehouse and a data mart. What are characteristics of a DW? How DW and DM are different from each other.
6. What is DW? Why it is needed? Explain ETL in detail.-----2015, 2014, 2016
7. Explain ETL in DW? ---2015-Rev
8. Explain the architecture of DW with neat diagram. ----2016-KT
9. What is data staging? Explain ETL process in detail. Write detailed architecture of DW. -----2015-KT
10. Define data warehouse. Explain any 3 architectural types of DW. ---2014
11. Explain the top down and bottom up approach in DW and suggest which is better. Explain the practical approach to construct a data warehouse.
12. What is metadata of DW? How it is different from metadata of OLTP systems.
13. Describe steps of DW implementation. (Rob C. 652, Rob C pg-488 2010 print) ---2014 –KT
14. Explain performance improvement techniques of DW.
15. What are the success factors for DW project?
16. Explain functional components of DW project development

- Short note on:
  - Roll up and drill down -----2015
  - Dimensional modeling ---2014
  - MOLAP ----2016-KT , 2016-KT
  - ROLAP
  - Start schema ----2015-Rev
  - Snow flake schema ----2014-Rev-KT
- Compare following:
  - ROLAP and MOLAP -----2015,2015-KT, 2014-Rev-KT
  - OLTP & OLAP -----2015-KT, 2014, 2016-KT
  - Data mining Vs OLAP ----2016
- What is fact and dimension data? Differentiate between fact and dimension table. What are the components of fact and dimension table? (Paulraj- 212, Mallach- 496)
- What is multidimensional data cube of hypercube? How slice and dice technique fits into this model? ---2014, 2015-Rev, 2014-Rev-KT
- What is factless fact table? (Paulraj- 249)
- Write short note on information package diagram.
- What is dimensional analysis and modeling? Explain development phases of dimensional modeling. (Paulraj -204)
- What is dimension modeling? Discuss different dimension modeling techniques in detail. ---2014 –KT
- Explain snowflake schema, star schema and fact constellation schema with suitable example. Mention advantages & disadvantages. (Paulraj -220, 238, 249) -----2015-KT
- What is family of stars/ fact constellation schema? (Paulraj -249) -----2015-KT
- Explain fact constellation schema for inventory management system assuming appropriate information. ----2016-KT
- Explain OLAP architecture with a neat diagram. -----2016-KT
- Explain major functions of OLAP. -----2015-KT
- Define OLAP. Explain MOLAP and ROLAP with suitable diagram. -----2014-KT, 2014
- What is Fundamental difference between MOLAP and ROLAP? -----2016
- Explain OLAP operations on multidimensional cubes with examples .-----2015, 2016
- Explain various OLAP implementation techniques.