Master of Technology in Knowledge Engineering

Business Intelligence – Data Warehouse

Brandon NG
brandon.ng@nus.edu.sg

Institute of Systems Science
National University of Singapore

Topics – Data Warehousing Fundamentals

- **Introduction to Data Warehousing**
- Dimensional Modeling
- Data Acquisition

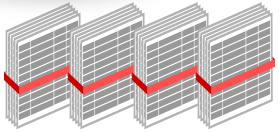
The Goal of a DW / BI solution is to…

Provide Information & Knowledge to support Decision Making at all levels of your organization

Early Decision Support System

Operational Systems

Operational Transactions



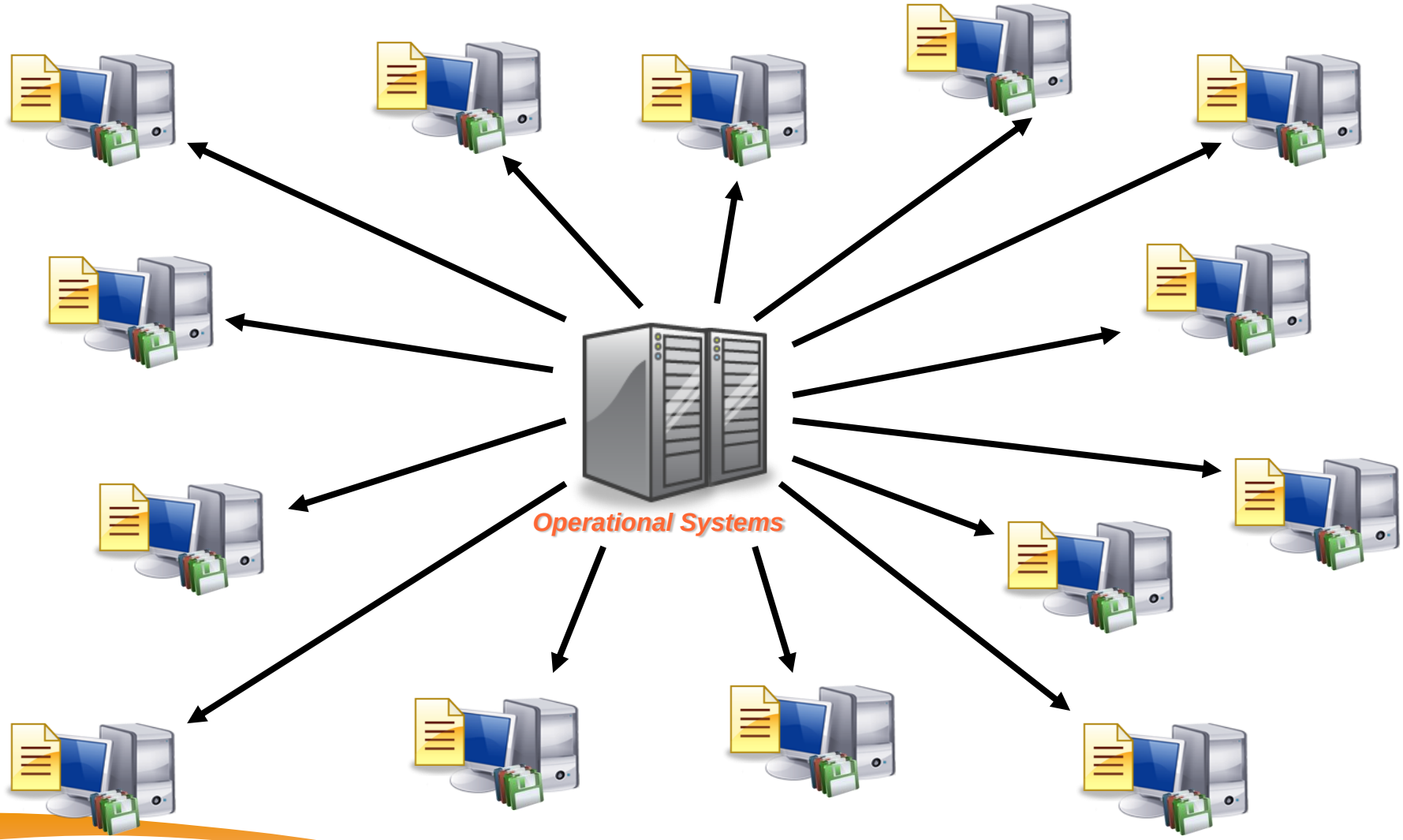
Ad–hoc reports

Ad–hoc Access

Decision Makers

Same database hosting business transactions and analytics

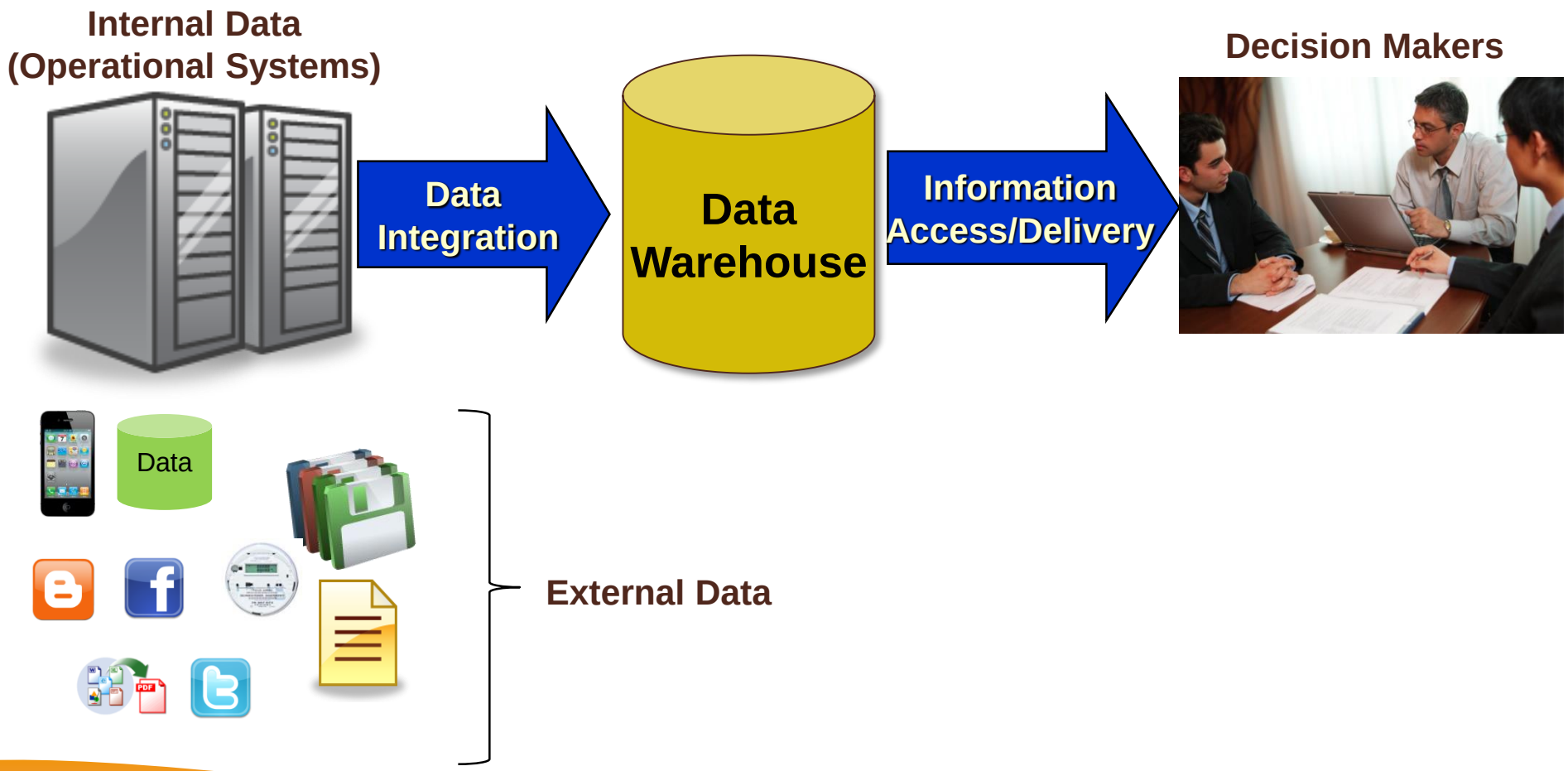
Issues, Issues & More Issues…



What are typical issues encountered?

- Extract explosion
- Duplication of effort
- Inconsistent results
- Reports become obsolete fast
- No common time basis for reports
- No historical trends
- Different levels of granularity
- Very limited sharing & distribution
- Inconsistent data & poor data quality
- No way to tackle missing information
- Different users using different/multiple technologies
- Absence of metadata

The Need for DW / BI Environment…



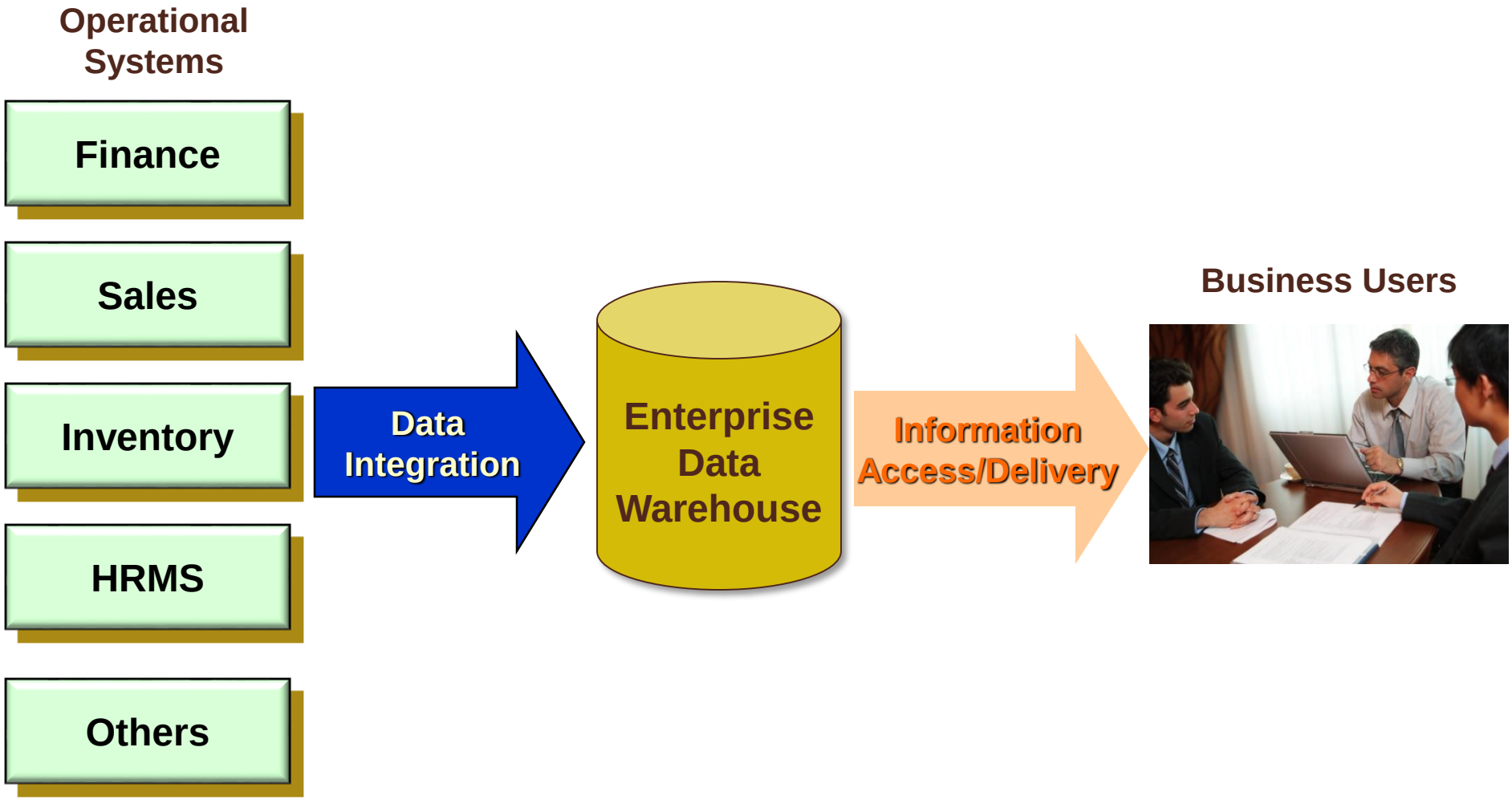
Evolvement of Data Warehousing

Data Warehouse +
Business Intelligence
=
Analytical Systems

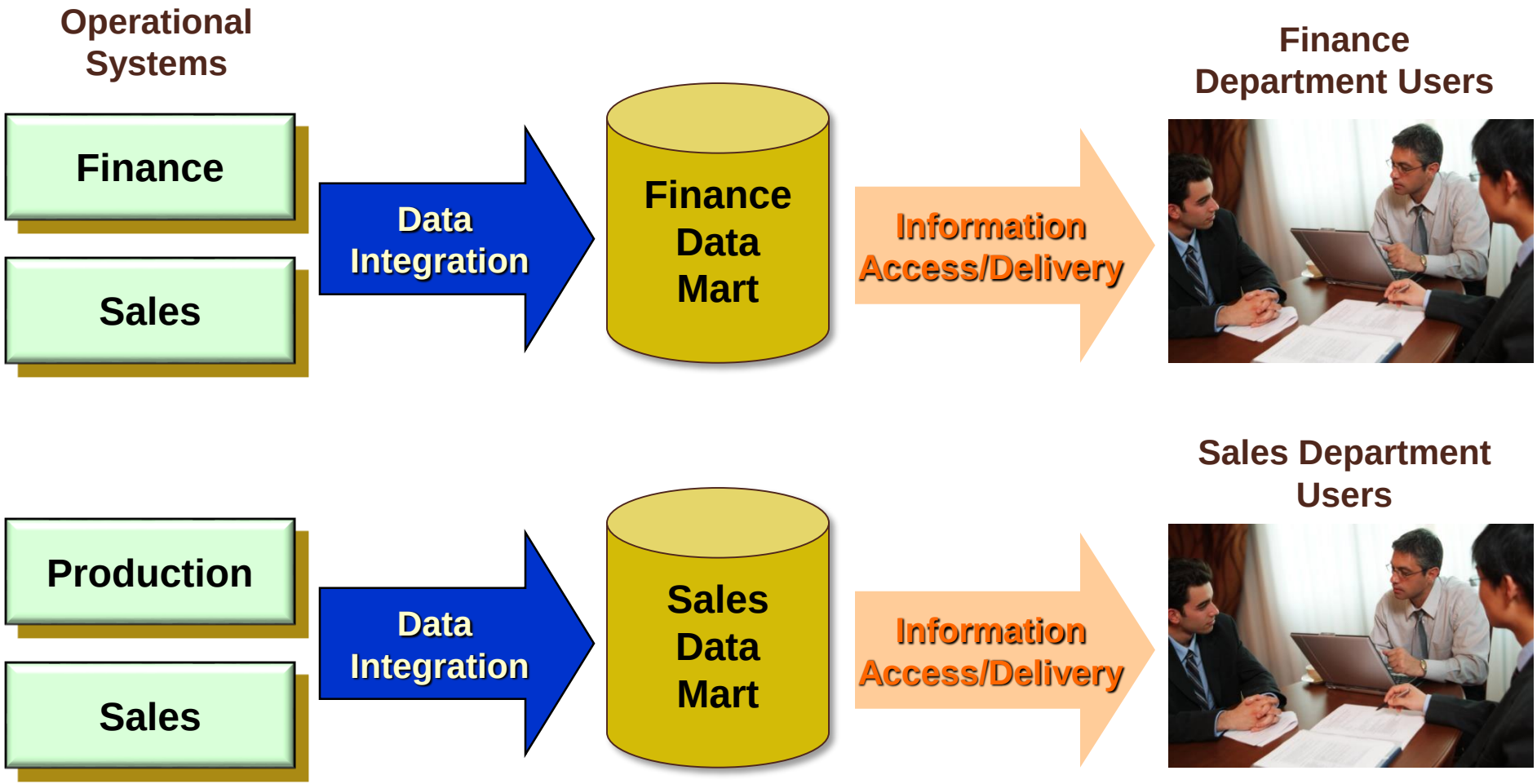
A Few Synonyms..

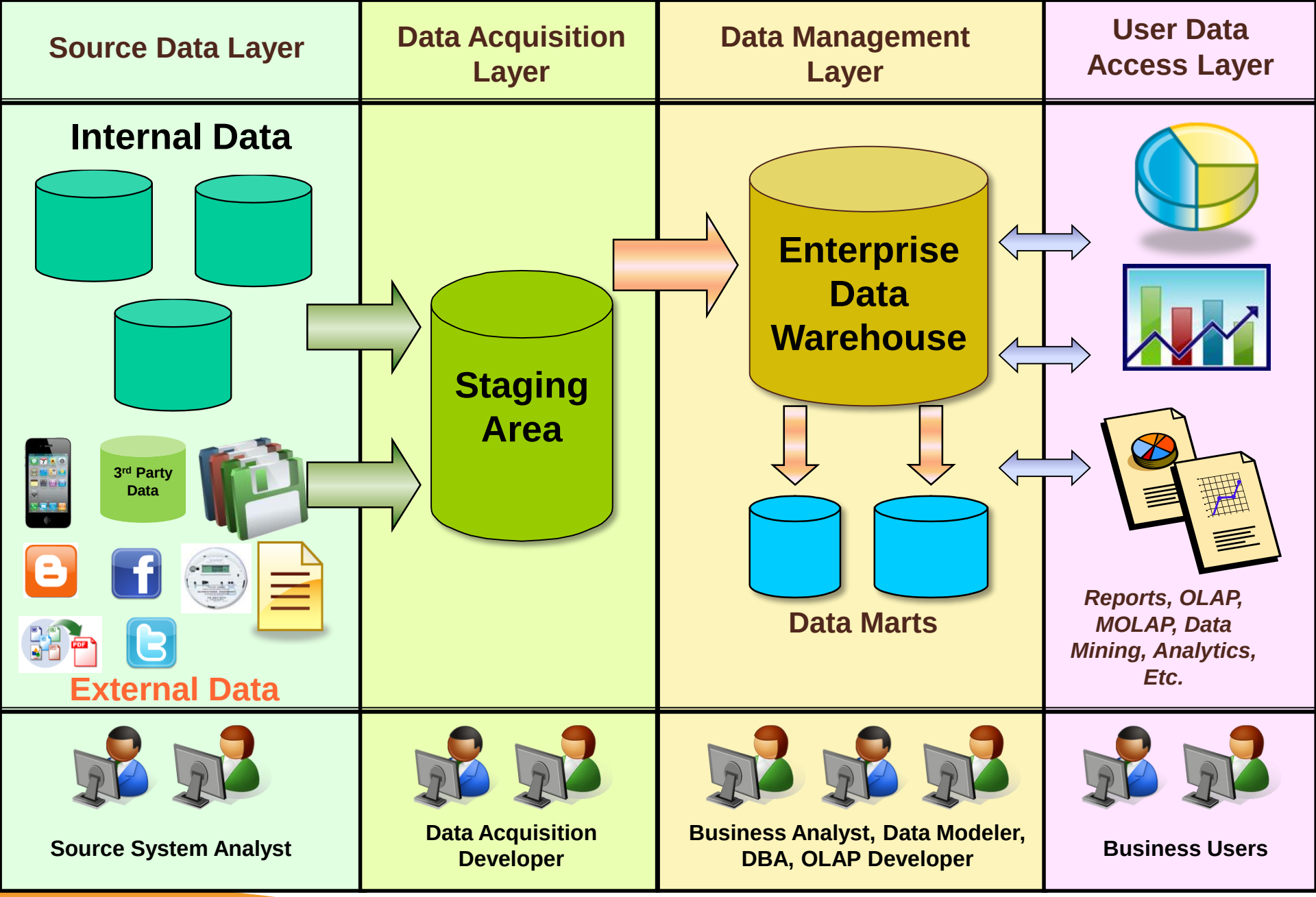
- Decision Support System (DSS)
- Business Intelligence (BI)
- Executive Information System (EIS)
- Management Information System (MIS)
- Data Warehouse (DW)
- Data Mart (DM)
- Analytical Systems
- …

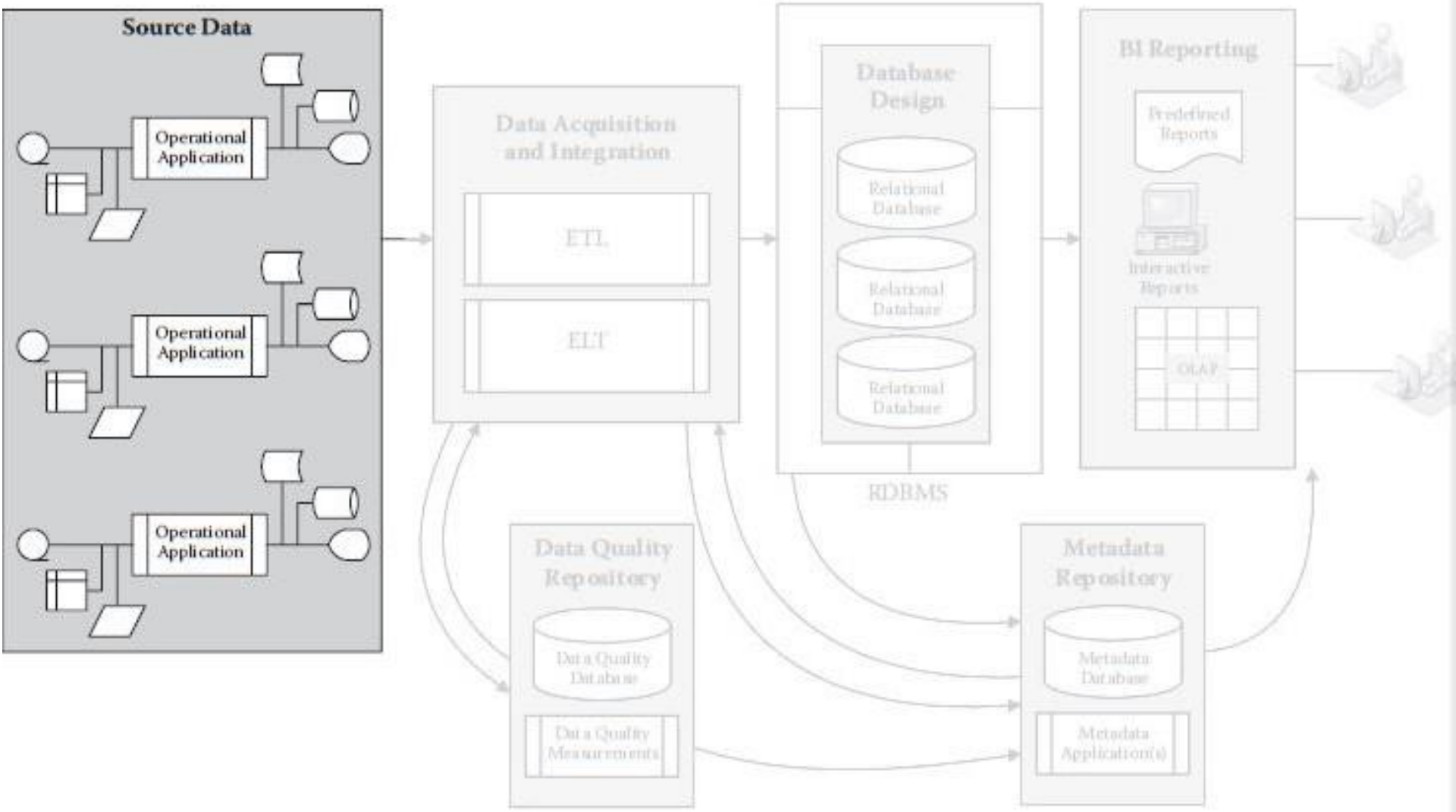
Enterprise Data Warehouse



Independent Data Marts







The Classic Definition…

"A Data Warehouse is a:

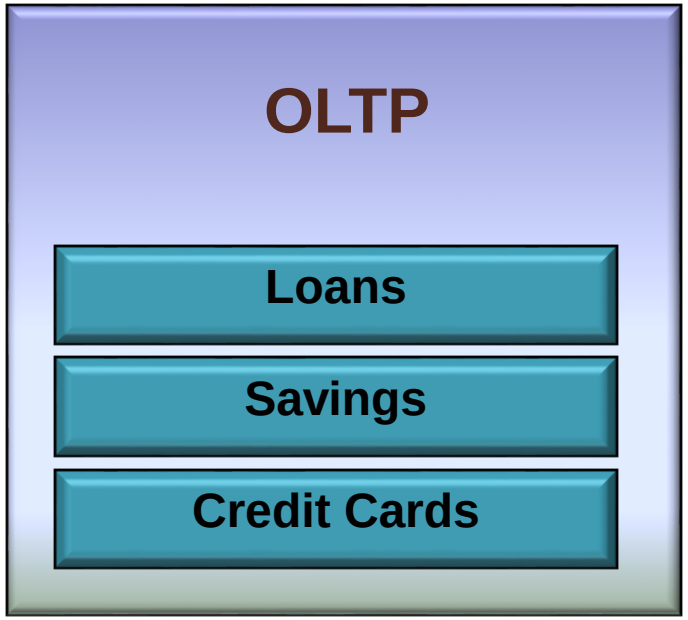
- ***Subject oriented***
- ***Integrated***
- ***Non-volatile***
- ***Time variant***

***collection of data* in support of management's decisions"**

Source: Bill H. Inmon

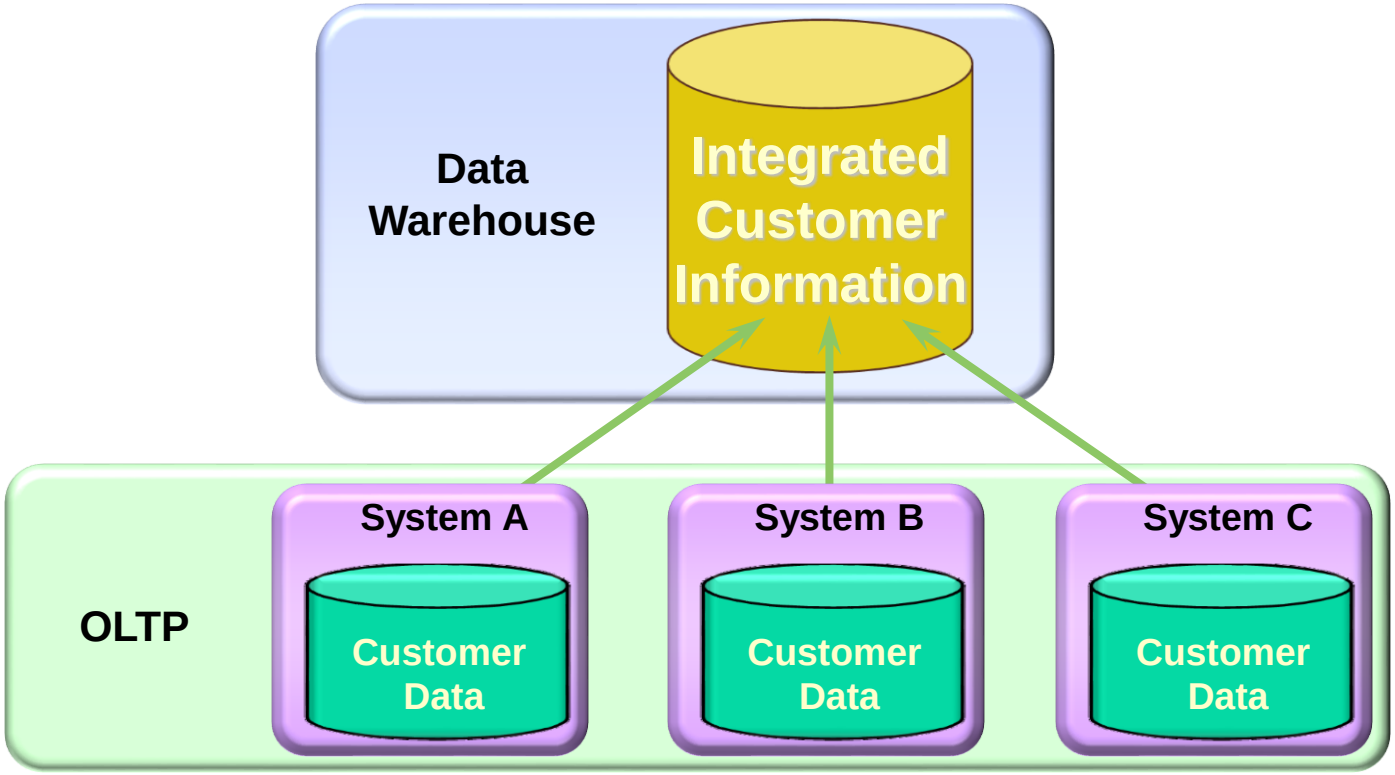
Subject Oriented…

- OLTP data are application oriented
- DW data is organized by business subject areas



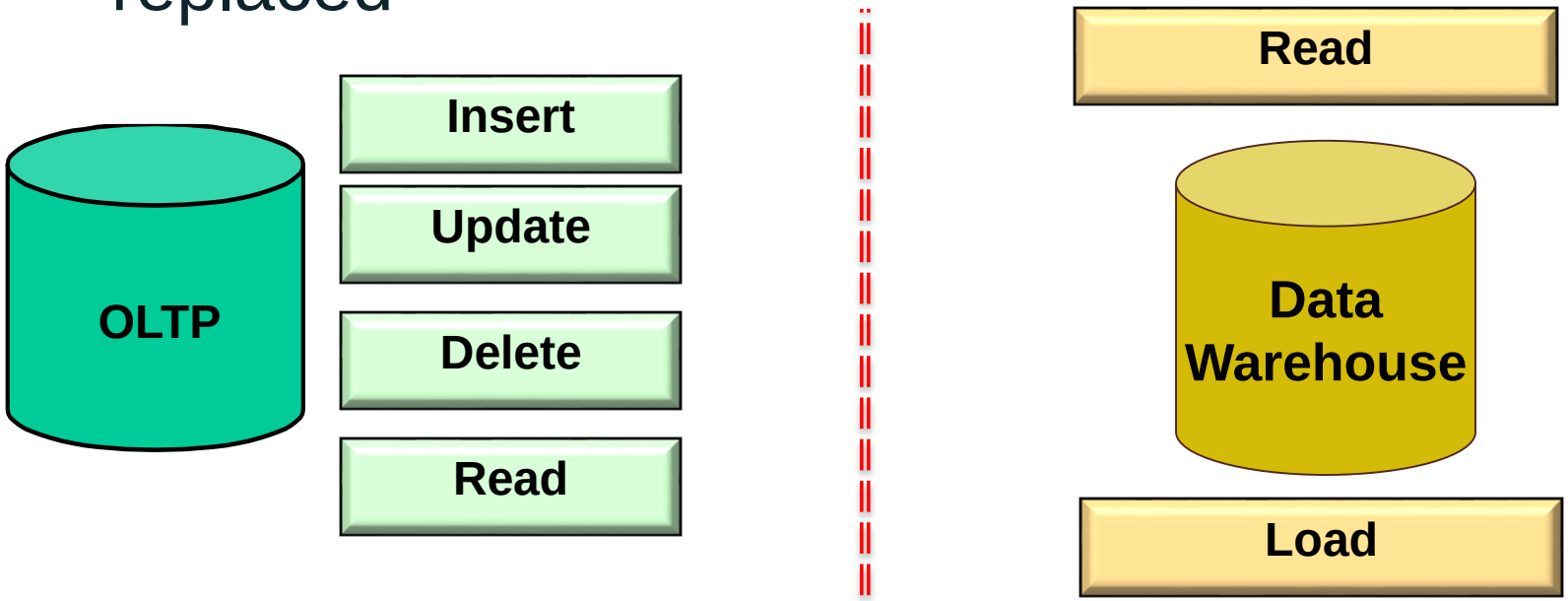
Integrated…

- Provides a single image of business reality by consolidating data from multiple sources



Non-volatile…

- Data in DW remains unchanged between queries (for same time period)
- Data is always added to the DW, seldom replaced



Time Variant…

- Data in DW is stored as series of snapshots each representing a specific time period
  - Snapshots of operational data are added to the DW
  - Data points in the DW are associated with points in time
  - Important for analyzing trends and doing comparisons

OLTP versus DW / BI Comparison…

OLTP System	DW / BI System
Data supports day-to-day operations	Rich historical data for detailed analysis
Data is stored at transaction level	Data is integrated to give a holistic view
Highly Normalized database design	Highly De-normalized database design for query performance

DW / BI and OLTP Parameter Comparison…

Parameter	***Data Warehouse***	***OLTP***
Data Sources	Both Internal & External	Internal (Operational)
DB size	Large to very large	Small to large
Key usage	Analysis & predictive	Business Processes
Organization	Subject oriented	Applications
Age of data	Snapshots over time	Limited (e.g. 6 months)
SQL's	Usually reads	Insert, update, delete & read
Response Time	Medium to very fast	Very fast
Usage pattern	Random	Predictable

Topics – Data Warehousing Fundamentals

- Introduction to Data Warehousing
- **Dimensional Modeling**
- Data Acquisition

Modelling in Different Environment

Internal Data (Operational Systems)

Data Integration

Data Warehouse

Information Access/Delivery

Decision Makers

Database normalisation

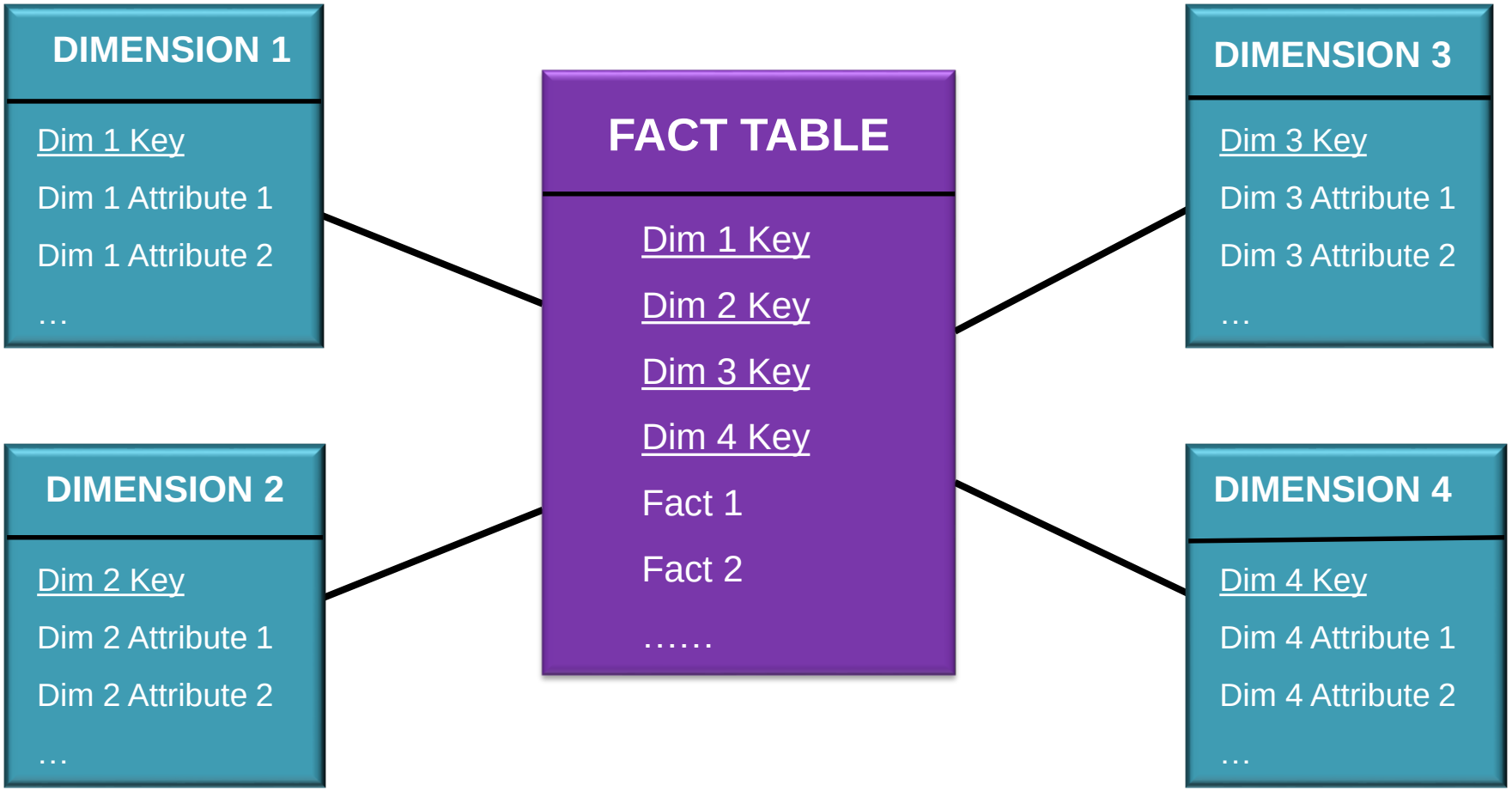
Dimensional Modelling

Common Modelling Techniques

Dimensional Model

- It represents the multi-dimensional business view using:
  - Business perspectives (dimensions)
  - Business measures (facts)
- Primarily used to model:
  - Relational Database (RDMS) for Relational Online Analytical Processing (ROLAP)

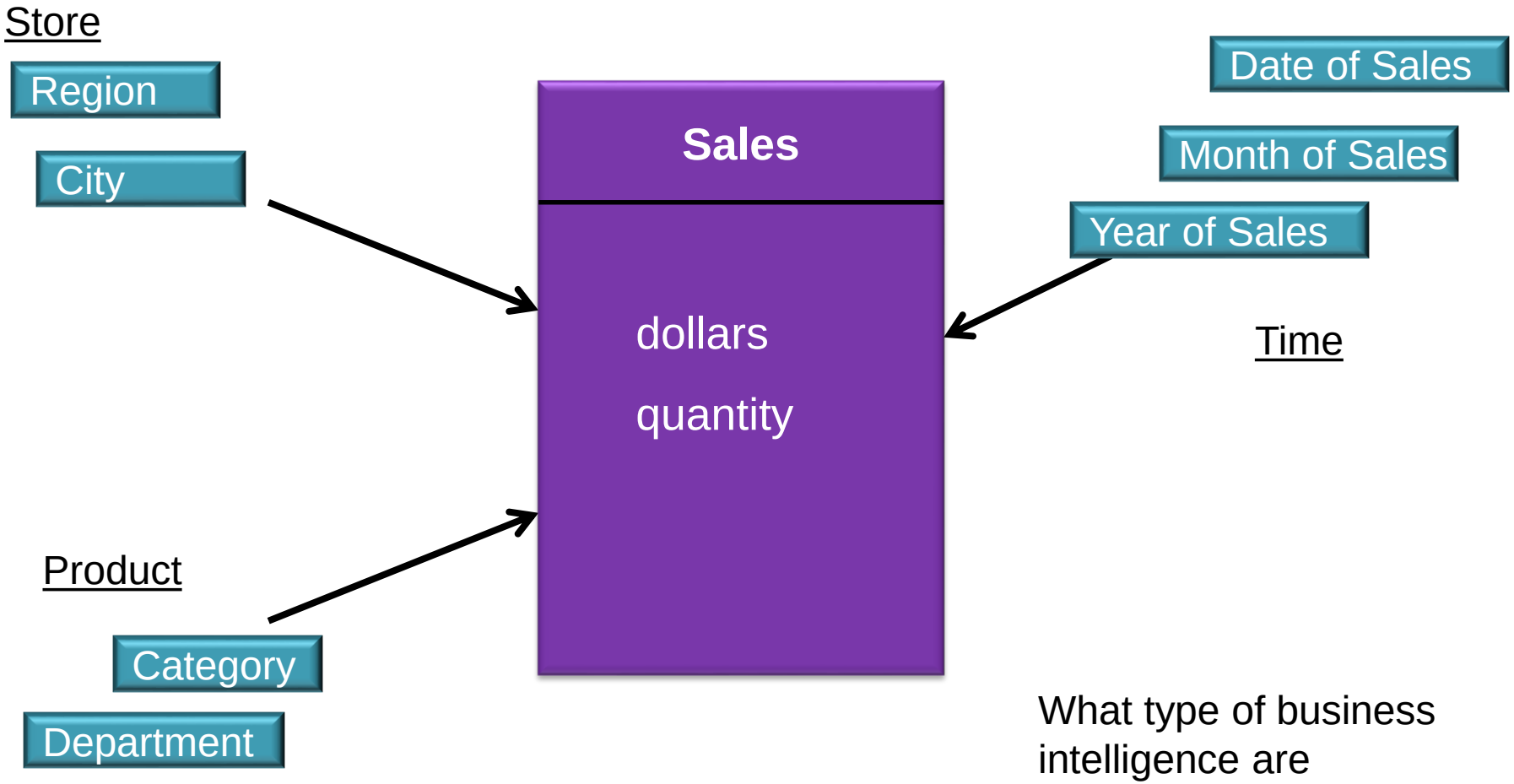
Dimensional Model - Example



Dimensional Model Terminology

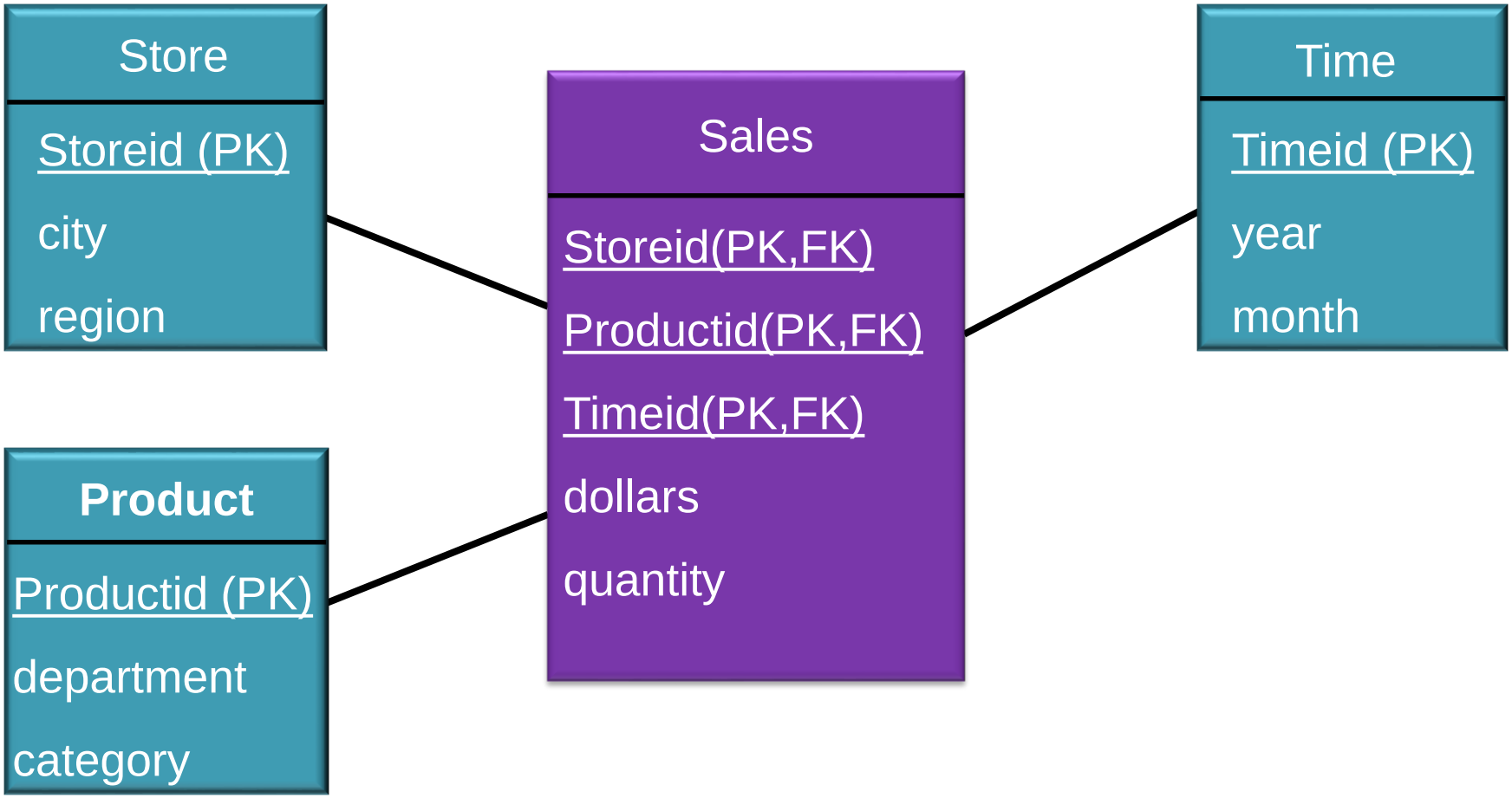
- Dimension table
  - Tables connected to Fact table containing near static data
- Attribute
  - Non key field in Dimensional table
- Fact table
  - Central table in a dimensional model containing facts
- Fact(s)
  - Are business measure, normally numeric, stored in fact table
- Grain
  - The level at which data is kept in the fact table
- Additive
  - Facts that can be added across dimensions

Dimensional Model – Business Driver



What type of business intelligence are needed?

Dimensional Model - Example



Granularity Supported

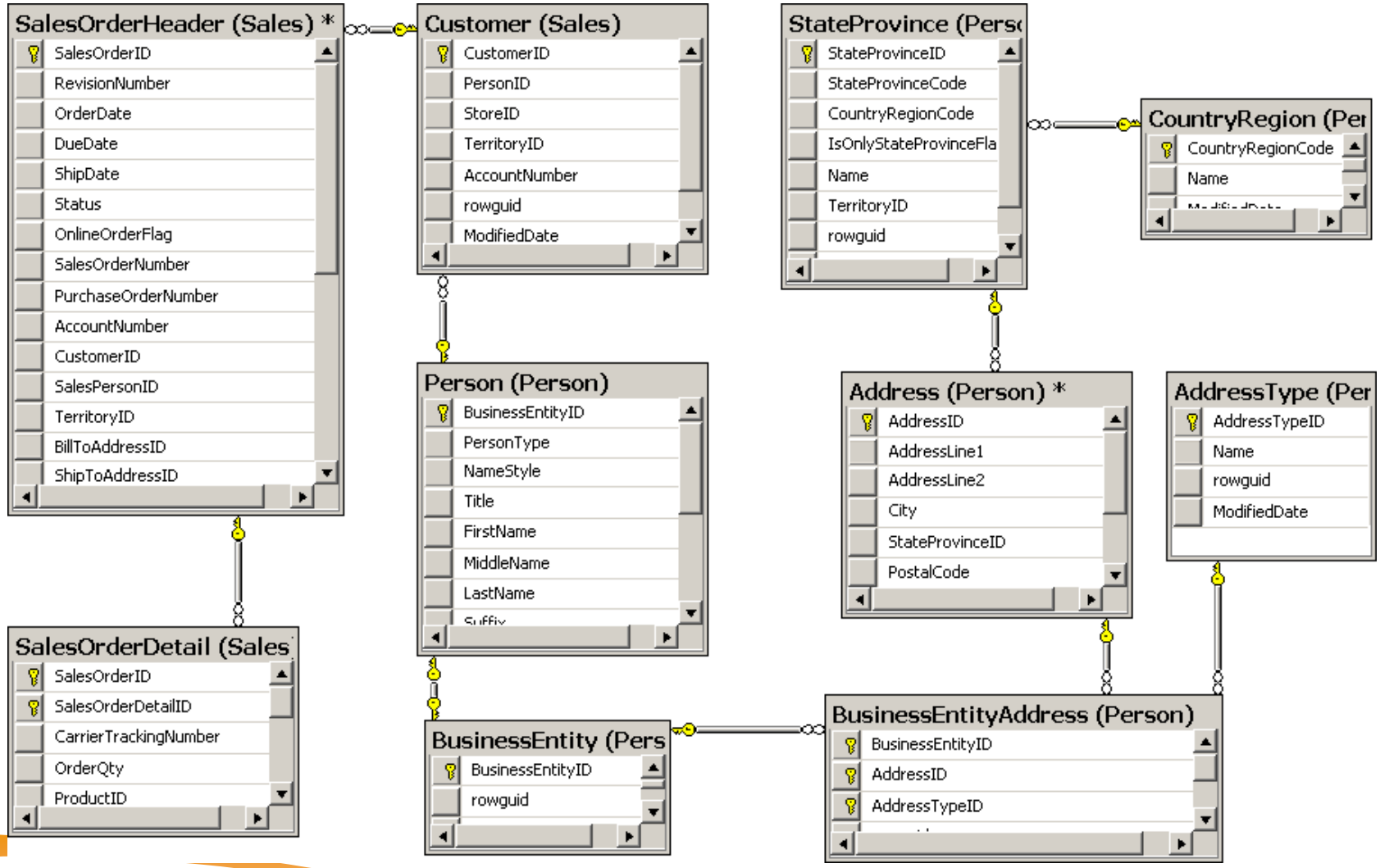
- Sales per Store per Product per Day
- Sales per Product per Day

Other scenarios:

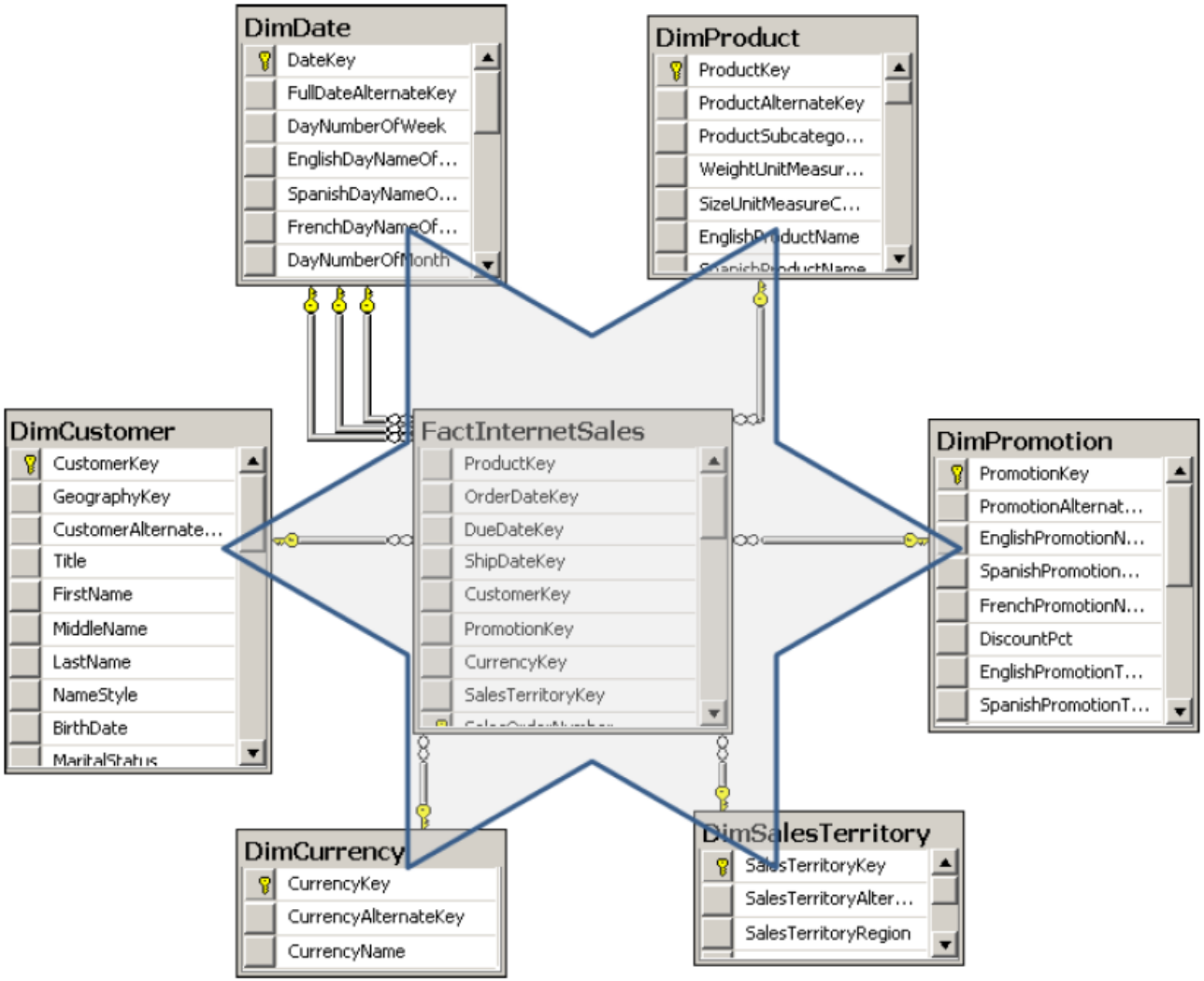
- Deposit summary per Account at end of the Month
- Sales per Product per Customer per Salesperson per Month
- Payments per Patient per Month per Hospital

from
Entity Relational Data Modelling
to
Dimensionality Modelling

ERD



Star Schema



Some Facts About Dimensional Modeling

(1/2)

- Less rigorous when compared to entity/relation modeling
- Allowing the designer more with practical discretion to organizing tables to accommodate database complexity and to improve performance
- It represents the multi-dimensional business views using:
  - Business perspectives (dimensions)
  - Business measures (facts)

•Source: Kimball group

Some Facts About Dimensional Modeling

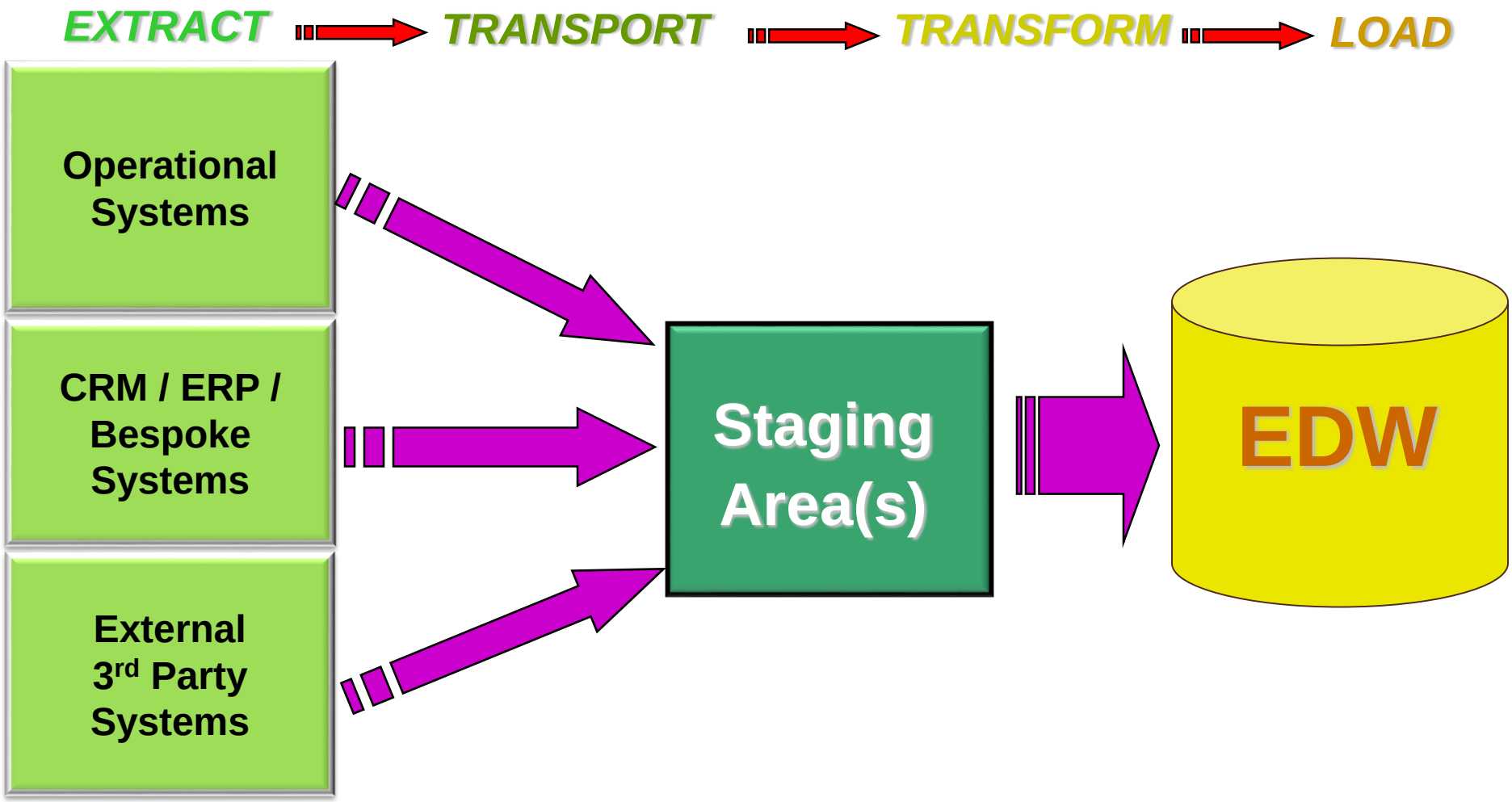
(2/2)

- A fact table is the central table in a star schema of a data warehouse
- A fact table stores quantitative information for analysis
- A fact table is normally denormalized
- A fact table works with dimension tables
- A fact table holds the data to be analyzed
- A dimension table stores data about the ways in which the data in the fact table can be analyzed
- A fact table will consist of two categories of columns
  - The foreign keys column allows joins with dimension tables
  - The measures columns contain the data that is being analyzed

Topics – Data Warehousing Fundamentals

- Introduction to Data Warehousing
- Dimensional Modeling
- **Data Acquisition**

Data Acquisition



What is Data Acquisition ?

- It is the process of capture, integrate, clean, transform, aggregate and load the required data to the DW after assuring data quality
- It goes beyond ETL (Extract, Transform & Load) process
- It includes any means of populating data into the DW
- Non-ETL examples are:
  - Data entry system to capture missing data
  - Data maintenance system to clean dirty data

Source Data Anomalies

- No unique key for the same attribute in different systems
- Data naming and coding anomalies
- Meaning anomalies between various systems
- Spelling and text inconsistencies

CUS_ID	CUS_NAME	CUS_ADDRESS
902234	Hewlett Packard Singapore Pte Ltd	101 Alexandra Road, #01-01, Singapore
102345	HP Pte Ltd	Alexandra Road, S(07955)
A127645	HP Singapore	101 Alexandra Rd, #01-01 HP Building, Singapore

Data Cleansing

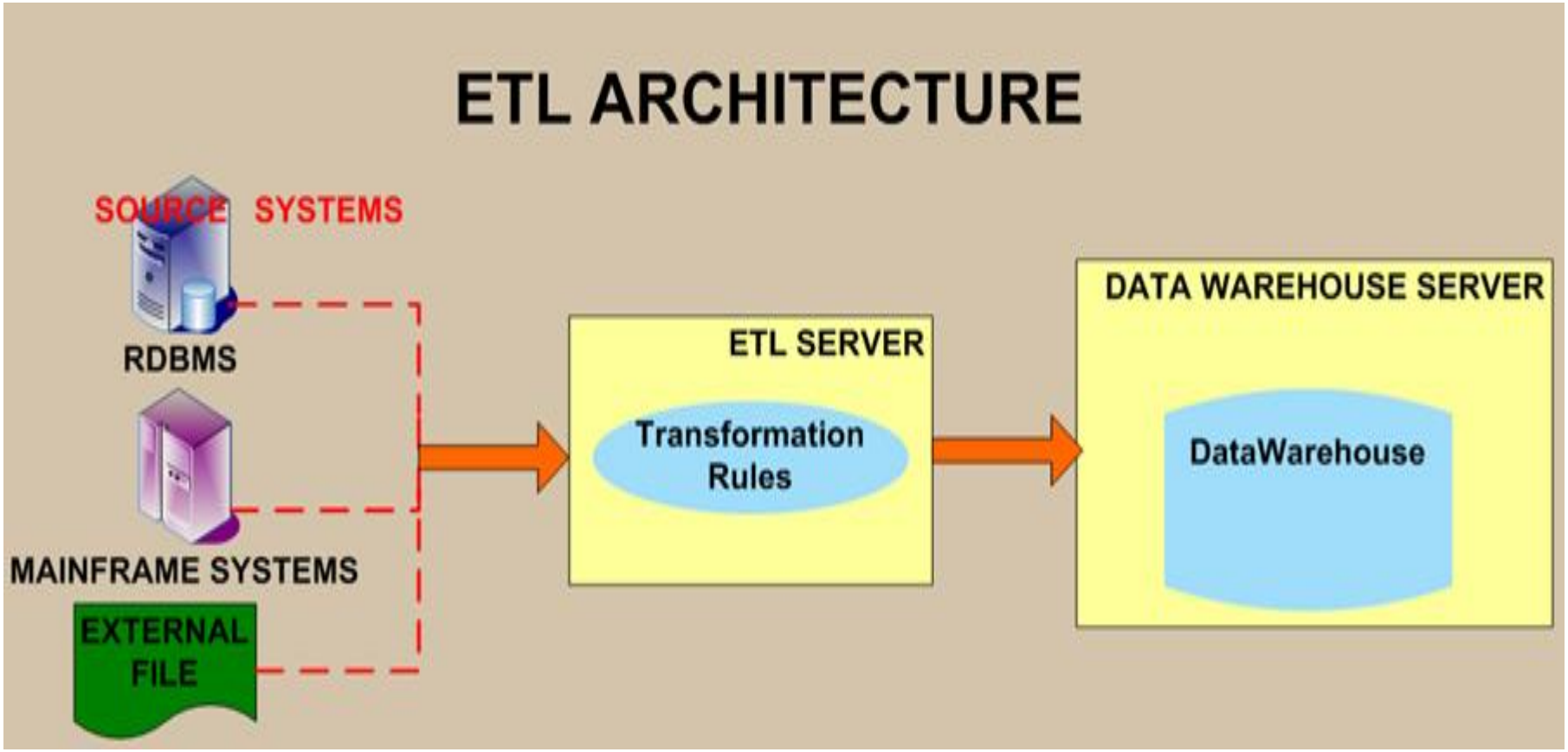
- The DW is only functional if it is accurate
- BI / DW can be inaccurate and unusable if the data in it is not clean or inaccurate
- Operational System data can be dirty
- Do not dismiss the cleanup phase of the DW project

What Makes Data Dirty ?

- Definitions are ambiguous – Age (is it months or years)
- Data accuracy - incorrect values, can be older values that may no longer be accurate
- Data integrity - business and integrity rule violations
- Data value inconsistencies - do multiple sources agree?
- Completeness and missing data - more detail needed
- Versioning conflicts - changes in formats / time periods / categories / codes
- Data heterogeneity - unit incompatibilities, hidden values

ETL Options

- Data can be extracted, transformed and loaded using:
  - Replication facility of standard RDBMS
  - Specialized ETL tools
  - Using custom coded programs.



Summary

- Transactional Database vs Data Warehousing
  - OLTP vs OLAP
- Dimensional Modeling
  - Fact & Dimensional Model
- Data Acquisition
  - ETL

Dimensional Modeling - Exercise

- Break out into your teams
- Business objective:
  - To perform sales analysis to track effectiveness of sales promotions
  - To perform sales analysis to track effectiveness of sales promotions on a customer age group
- Develop a Dimensional Model to enable the stated business objective
  - Include your attributes

~Last Slide ~