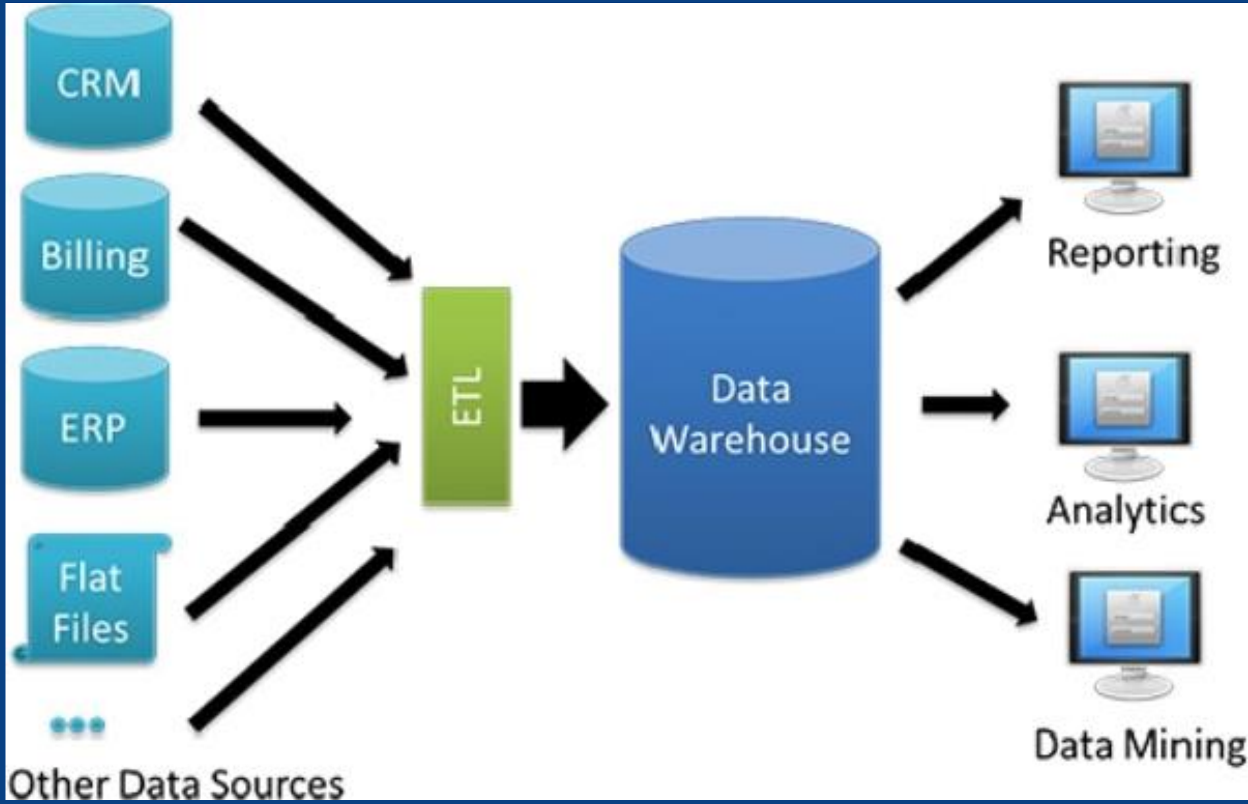
THE UNIVERSITY OF MELBOURNE

Data Warehousing

Database Systems & Information Modelling INFO90002



Week 9 – DW
Dr Tanya Linden
Dr Renata Borovica-Gajic
David Eccles

This Lecture Learning Objectives

By the end of this lecture you should be able to:

Articulate the differences between **transactional** (operational) and **informational** (dimensional) databases

Explain the **characteristics of a Data Warehouse**

Understand and explain the **overall architecture of a Data Warehouse**

Design **Star Schemas**

Motivations: A 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?

Statisticians are fortune tellers but with greater accuracy

What product promotions 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?

BUT HOW?

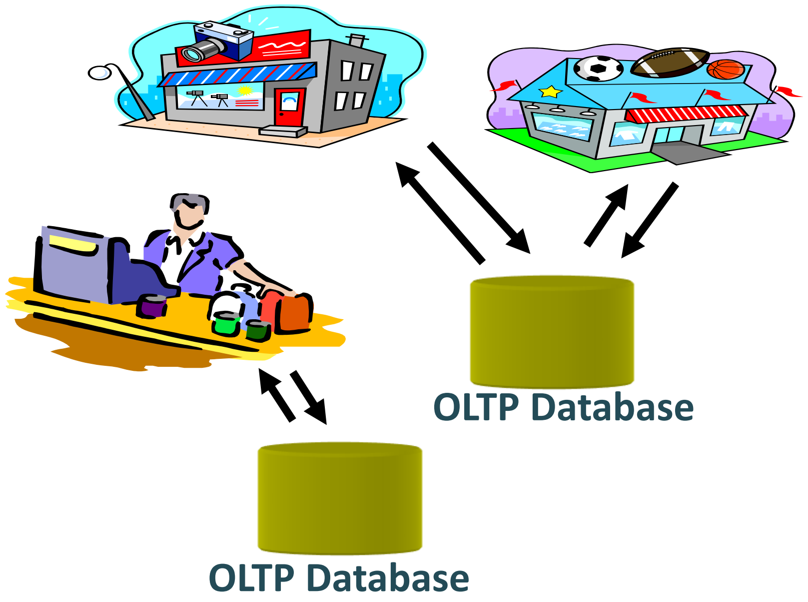
Relational Databases for Operational Processing

Used to run day to day business operations

Automation of routine business processes

- Accounting
- Inventory
- Purchasing
- Sales

Created huge efficiencies



OLTP Databases

OLTP = “OnLine Transaction Processing”

Transaction processing supports daily (routine, repetitive) operations

- Mundane but crucial
- Become even more important with the growth of the internet

Definition:

- Collection of read/write operations
- Processed as one unit
- Reliably and efficiently processed
- No data loss due to interference and failures (operating system, program, disk, …)

OLTP Data Characteristics

Characteristics of data:

- Transaction oriented
  - DML

Inserts Updates Deletes

- May be inconsistent and incomplete
  - Data may not be in its final form
- Volatile – continually changing
  - Data maybe subject to change
- Current
  - Data related to the operation of the business TODAY!

Databases are great, BUT …

Too many of them

- Everybody wanted one, or two, or more
- Production, Marketing, Sales, Accounting …

Everybody got what was best for them

- IBM, Oracle, Access, Microsoft

Eventually this re-created the problem databases were meant to solve

- Duplicated data
- Inaccessible data
- Inconsistent data

But data is useful for analysis and decision making

What can be done about it? SPOT!

Need an integrated way of getting the ENTIRE organisational data

Need an Informational Database, rather than a Transactional Database

- A single database that allows *all* of the organisations data to be stored in a form that can be used to support **organisational decision processes**

A centralised repository for decision making

- Populated from operational databases and external data sources
- Integrated and transformed data
- Optimised for reporting

Single Point of Truth (SPOT) about the data

Data Warehouse: An Informational Database

Data Warehouse:

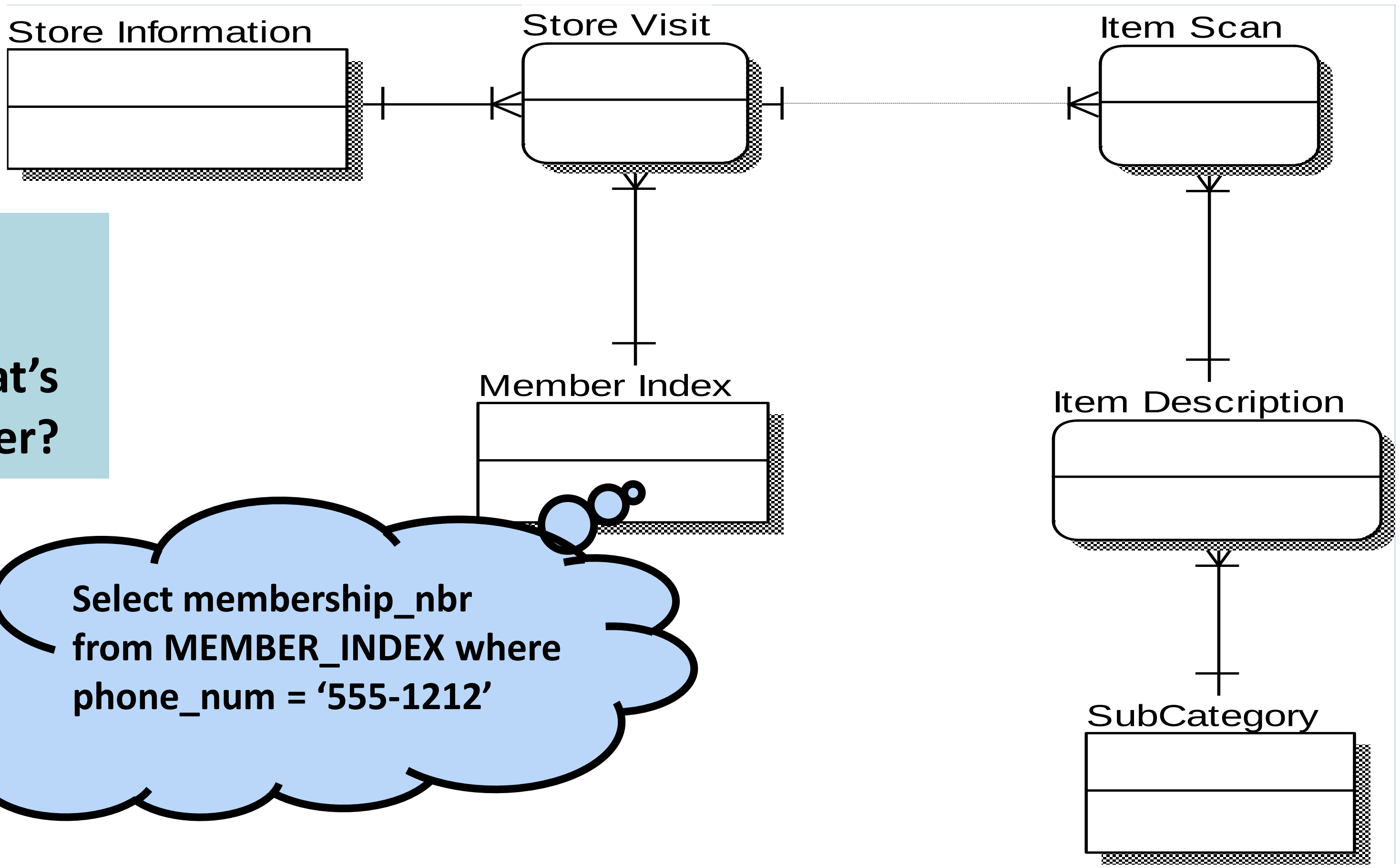
- A single repository of organisational data
- Integrates data from multiple sources
  – Extracts data from source systems, *transforms*, loads into the warehouse
- Makes data available to managers/users
- Supports analysis and decision-making

Involve a large data store (often several Terabytes, Petabytes of data)

Difference between Transactional and Informational Systems

Characteristic	Transactional	Informational
Primary Purpose	Run the day to day business	Support decision making
Type of Data	Current data – representing the state of the business	Historical data – snapshots and predictions
Primary Users	Customers, clerks and other employees	Managers, analysts
Scope of Usage	Narrow, planned, fixed interfaces	Broad, ad hoc, complex interfaces
Design Goal	Performance and availability	Flexible use and data accessibility
Volume	Many constant updates and queries on a few tables or rows	Periodic batch updates, complex querying on multiple or all rows

Transactional (Operational) Questions

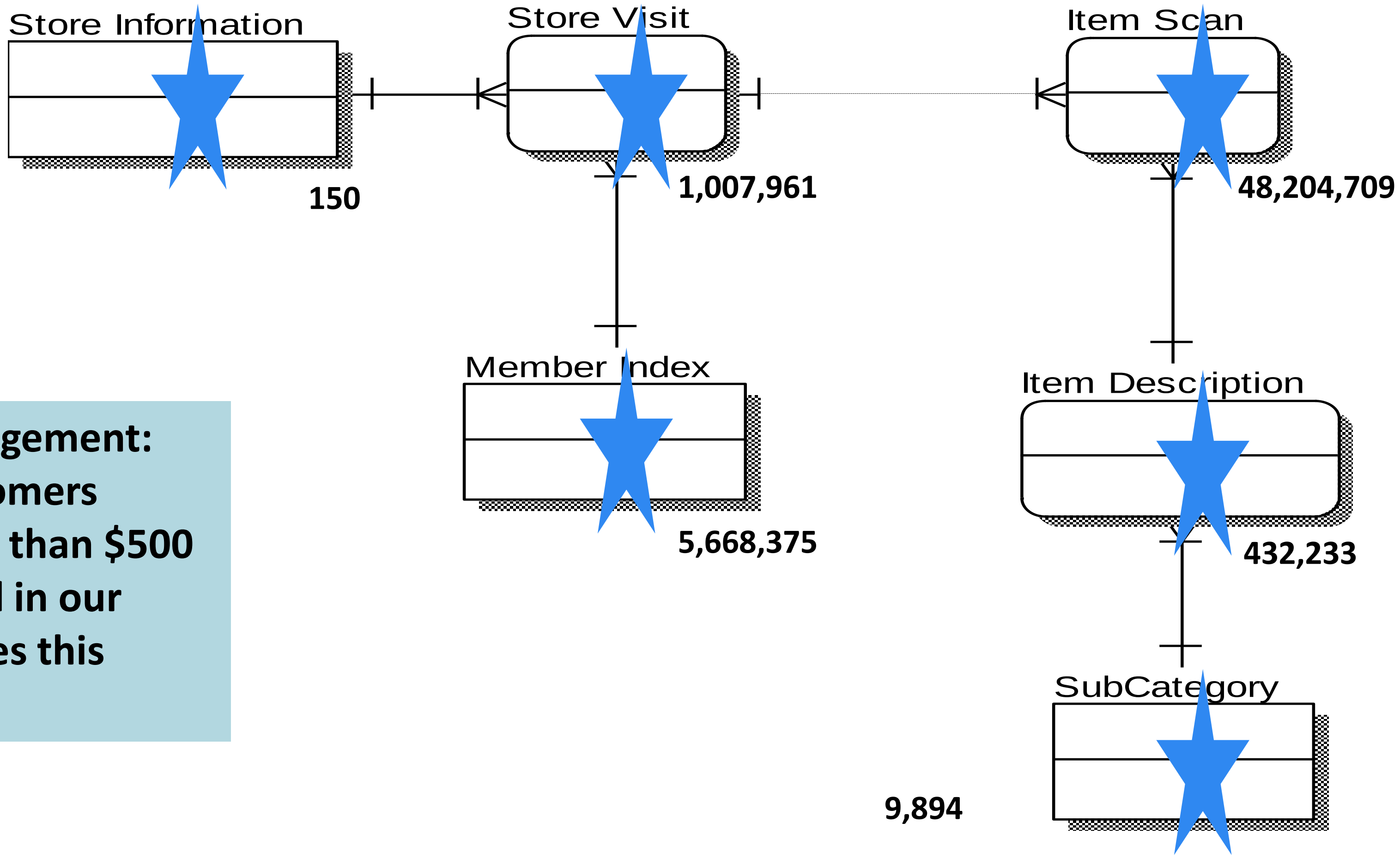


Customer Service:
Help! I forgot my membership card! What's my membership number?

Select membership_nbr from MEMBER_INDEX where phone_num = '555-1212'

How many tables affected? 1
How many rows have to be accessed? 1 (with an index in place)

Analytical Questions



Campaign Management: How many customers purchased more than $500 worth of alcohol in our Melbourne stores this year?

How many tables affected? 6
How many rows? millions

DW Supports Analytical Queries

A manager may be interested in numerical *aggregations*

- How **many**?
- What is the **average**?
- What is the **total cost**?

A manager may be interested in understanding *dimensions*

- Sales **by state by customer type**
- Sales **by product by store by quarter**

DW will help answer these questions

Characteristics of a DW

Subject oriented

- Data warehouses are organised around particular subjects (sales, customers, products)

Validated, Integrated data

- Data from different systems converted to a common format: allows comparison and consolidation of data from different sources
- Data from various sources validated before being sent to a data warehouse

Time variant

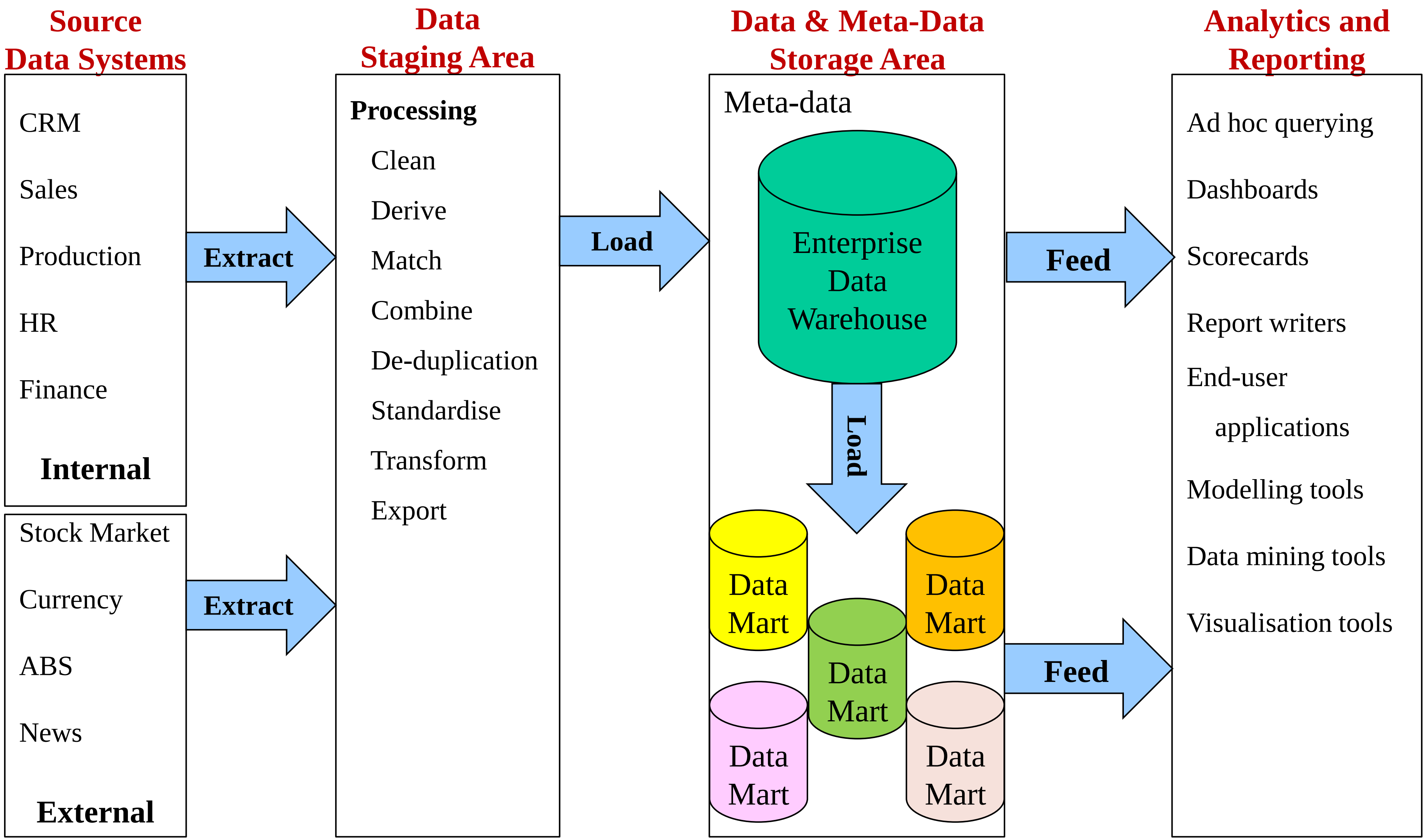
- Historical data
- Trend analysis crucial for decision support: requires historical data
- Data consists of a series of “snapshots” which are time stamped

Non-volatile

- Users have Read access only – all updating done automatically by ETL* process and periodically by a DBA

*ETL stands for “extract, transform, load” - the three processes that, in combination, move data from one or more databases, or other sources to a unified repository, typically a data warehouse

A DW Architecture



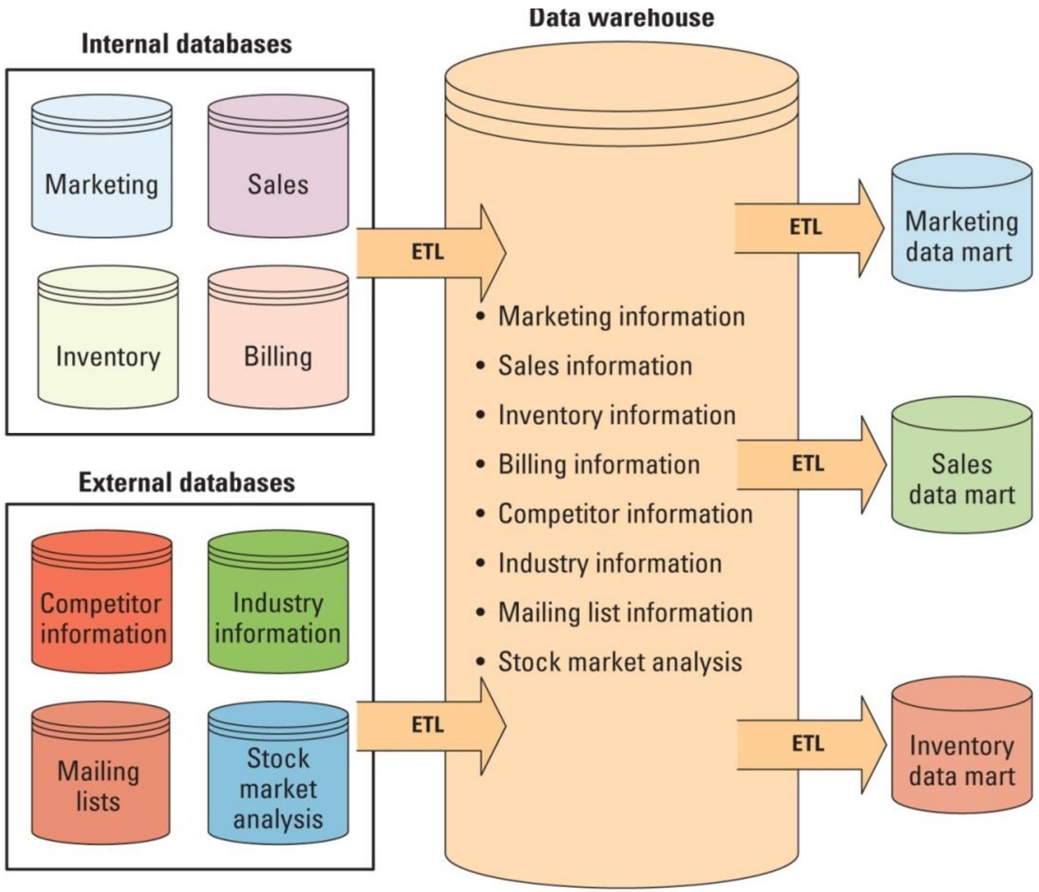
Data marts and data mining

Data mart

- contains a subset of data warehouse information

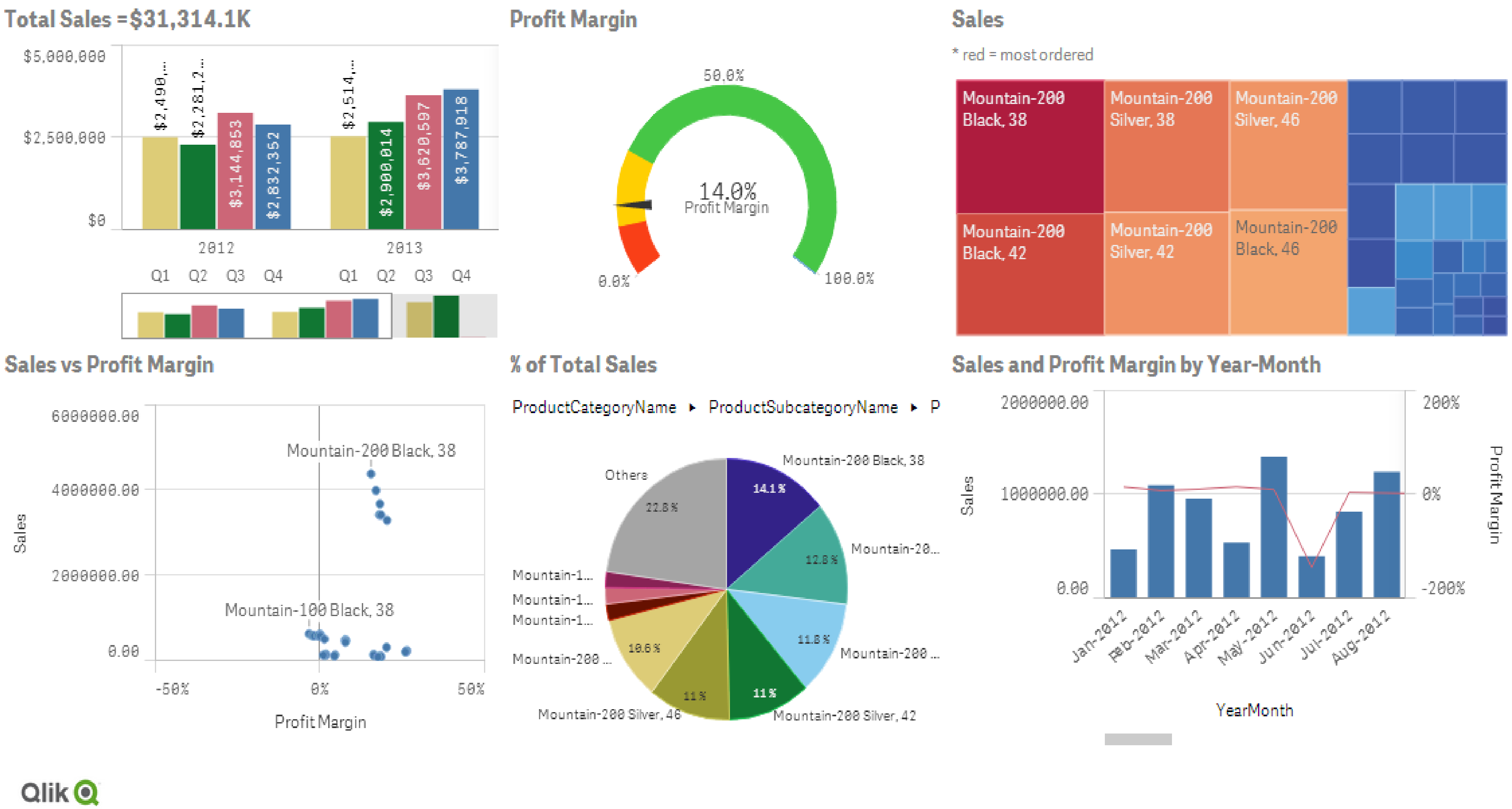
Data-mining

- A process in which algorithms are applied to information to uncover patterns and relationships otherwise difficult to find



Baltzan, Lynch & Fisher 2015, p. 279

Business Intelligence Dashboard



Characteristics of a DW

Subject oriented

- Data warehouses are organised around particular subjects (sales, customers, products)

Validated, Integrated data

- Data from different systems converted to a common format: allows comparison and consolidation of data from different sources
- Data from various sources validated before storing it in a data warehouse

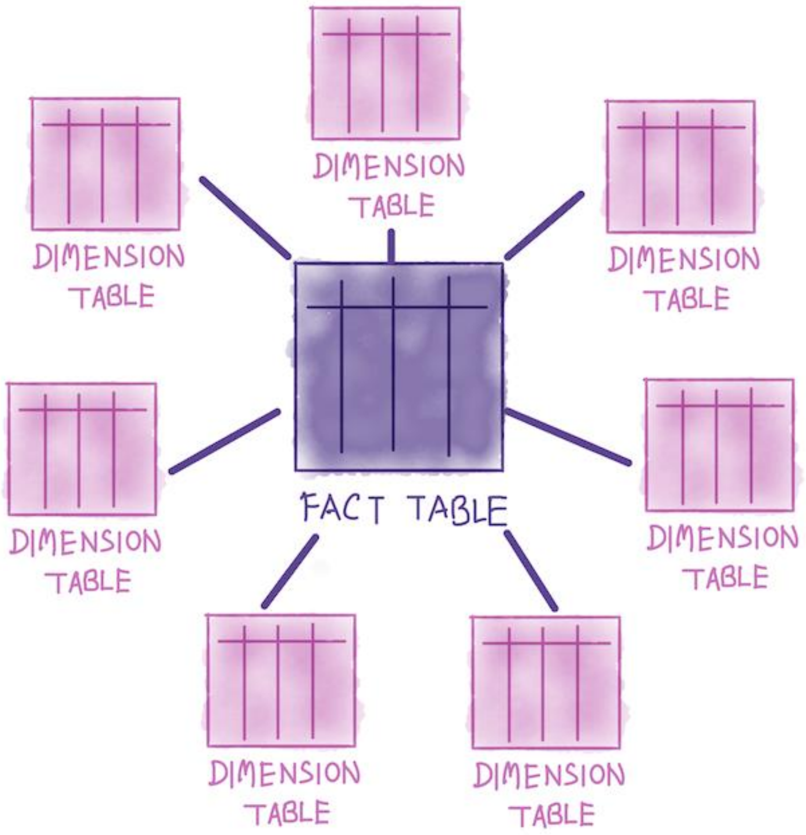
Problems

Incomplete Errors

- Missing Fields
- Records or Fields that, by design, are not recorded, e.g. the type of people that buy Big Issue from Big Issue Vendors when a sale is made

Incorrect Errors

- Wrong data entered into source system
  - E.g. manual entering of data will always have a percentage of incorrect data → human error



Dimensional Modelling

Business Analyst World

Fact Dimension

How much **revenue** did the **product G** generate in the **last three months**, broken down by

Dimension

month for the south eastern sales **region**, by individual **stores**, broken down by

Dimension

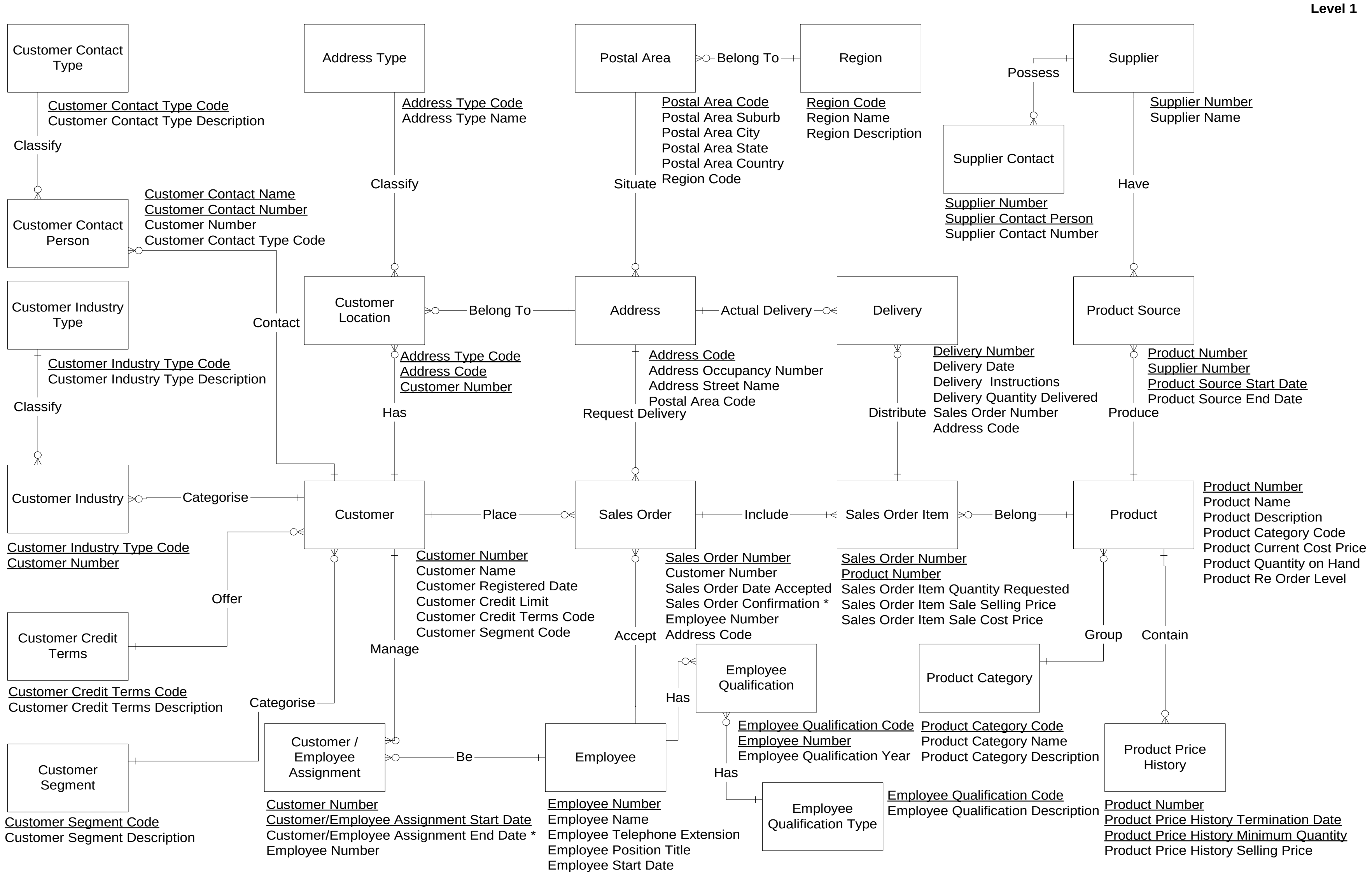
promotions, compared to estimates and to the previous version of the product

- Analysis starts usually with a single indication of something strange, then goes deep into the data, left to a new dimension, right to another, up to the summary, back down and left and right again, until the problem is identified…

- Dimensional Analysis: To support business analysts view
  – Revenue per product per customer per location?

Fact Dimension Dimension Dimension

Example ER model



Introduction to Dimensional Modelling

Popularised by Ralph Kimball in the 1990s

Based on the *multi-dimensional* model of data and designed for retrieval-only databases

Very simple, intuitive, and easily-understood structure

Also known as *star schema* design

A dimensional model consists of:

- Fact table
- Several dimensional tables
- (Sometimes) hierarchies in the dimensions

Essentially a simple and restricted type of ER model

Fact Table

A fact table contains the actual business measures (additive, aggregates), called ***facts***

The fact table also contains *foreign keys* pointing to ***dimensions***

Sale

keys
- Time key
- Store key
- Customer key
- Product key

facts
- Dollar sales
- Unit sales

Fact Table - example

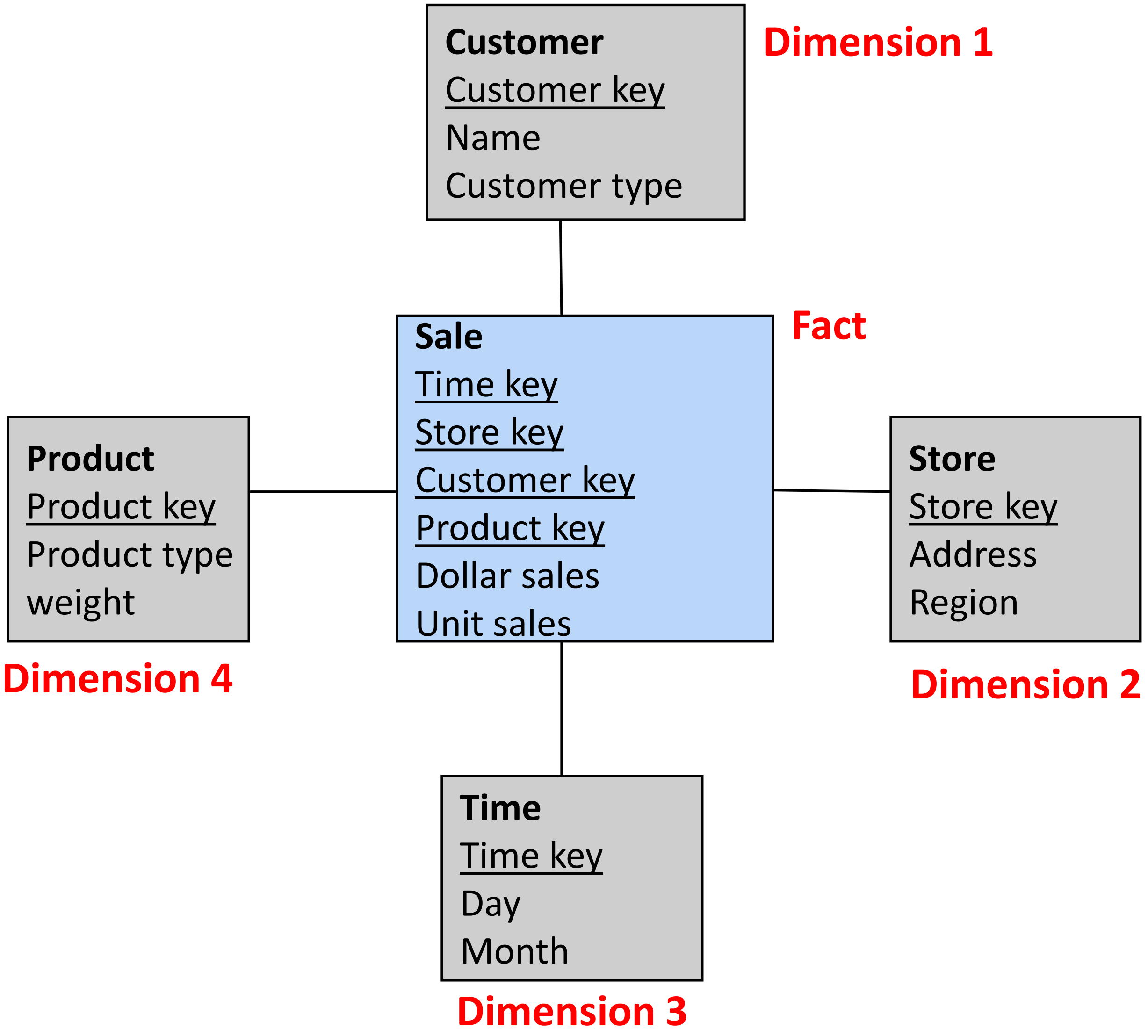
Actual data might look like this

Granularity, or level of detail, is a key issue

- Finest level of detail for a fact table, determined by the finest level of each dimension

Time-id	*Store-id*	*Cust-id*	*Prod-id*	*Dollar sales*	*Unit Sales*
T100	S303	C101	P98	$120,000	5,000
T101	S303	C256	P98	$240000	10,000
T102	S387	C101	P10	$456,000	27,899
T100	S234	C400	P56	$100,200	5,600

Star schema – dimensional model



Dimension Hierarchies

Product
Product key
Product name
Product type
Product group
Product sub-group
weight

Sale
Time key
Store key
Customer key
Product key
Dollar sales
Unit sales

Product name	e.g. Hammer
- Product type	e.g. Tool
- Product group	e.g. Hardware

Dimension Table - example

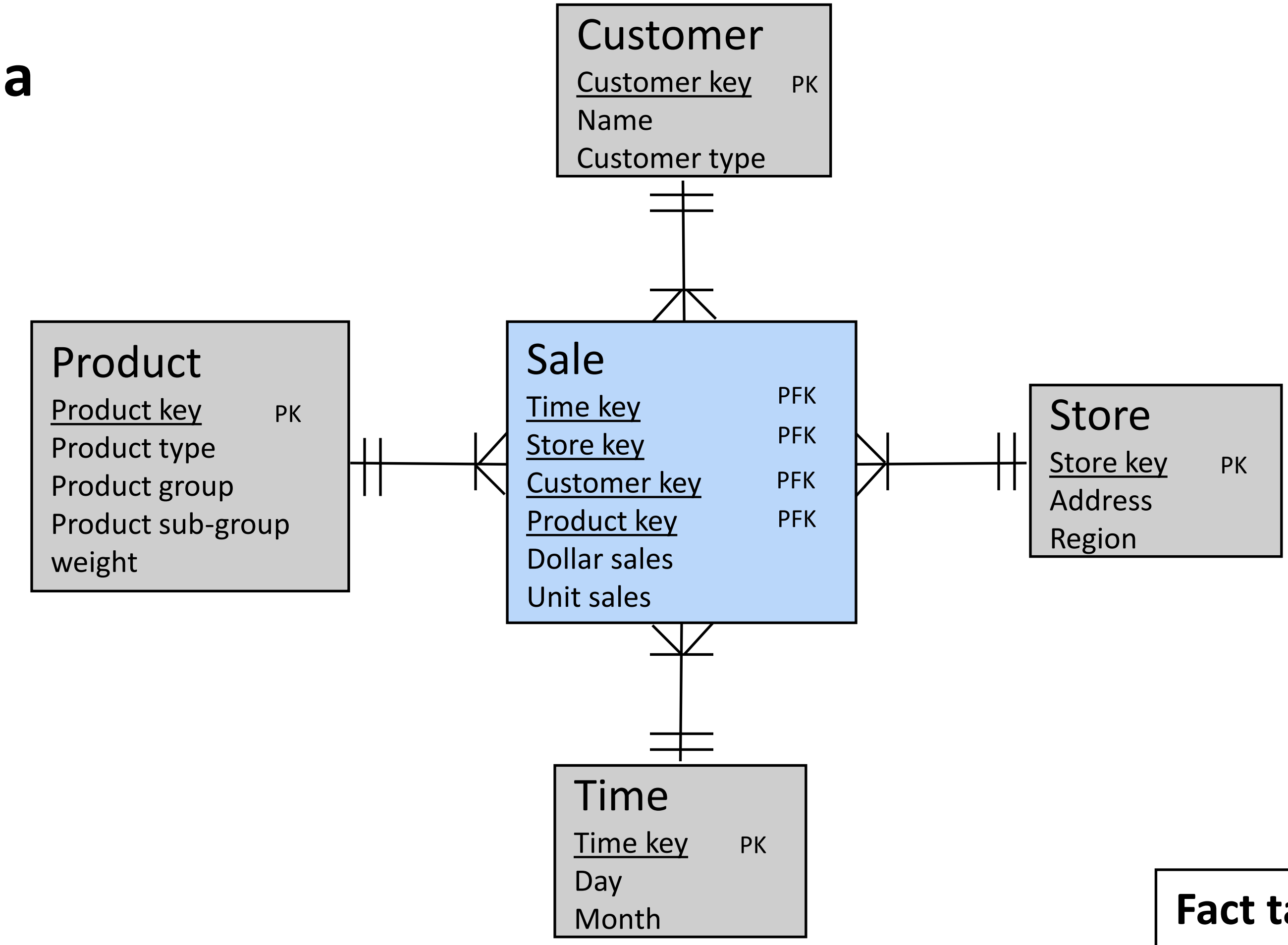
Actual data might look like this

Hierarchy evident in data

Prod-id	*Prod-Name*	*Prod-Group*	*Prod-Subgroup*	*Weight*
P10	Hammer	Hardware	Tool	5kg
P56	10cm Nails	Hardware	Nails	1kg
P98	Plastic Pipe	Plumbing	Pipe	1kg

Dimensional model as an ER model

Star schema



Fact table is an intersection table

Designing a Dimensional Model

Steps:

1. Choose a Business Process
2. Choose the measured facts (usually numeric, additive quantities)
3. Choose the granularity of the fact table
4. Choose the dimensions
5. Complete the dimension tables

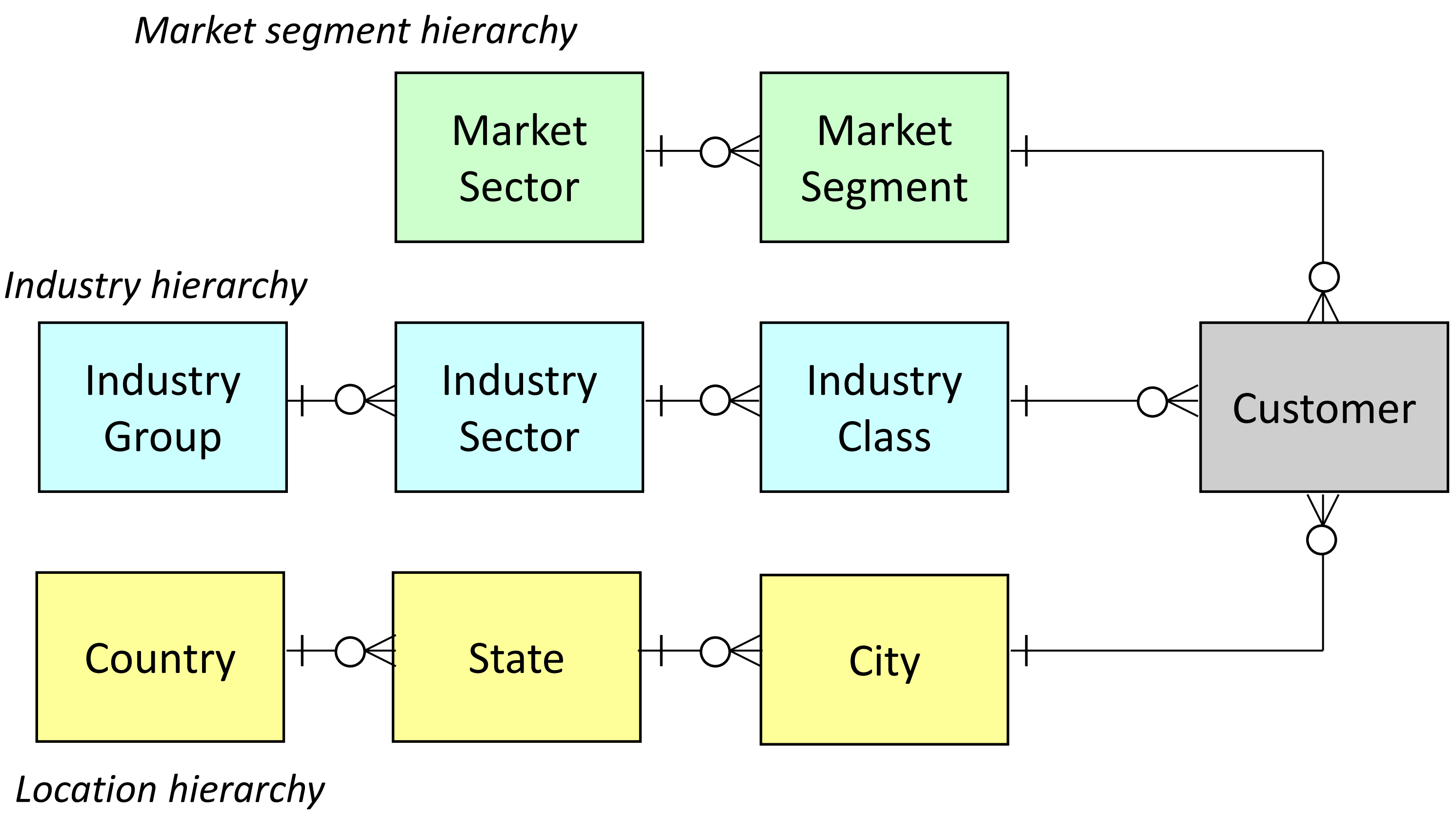
(Kimball, 1996)

Embedded Hierarchies in Dimensional Tables

Customer

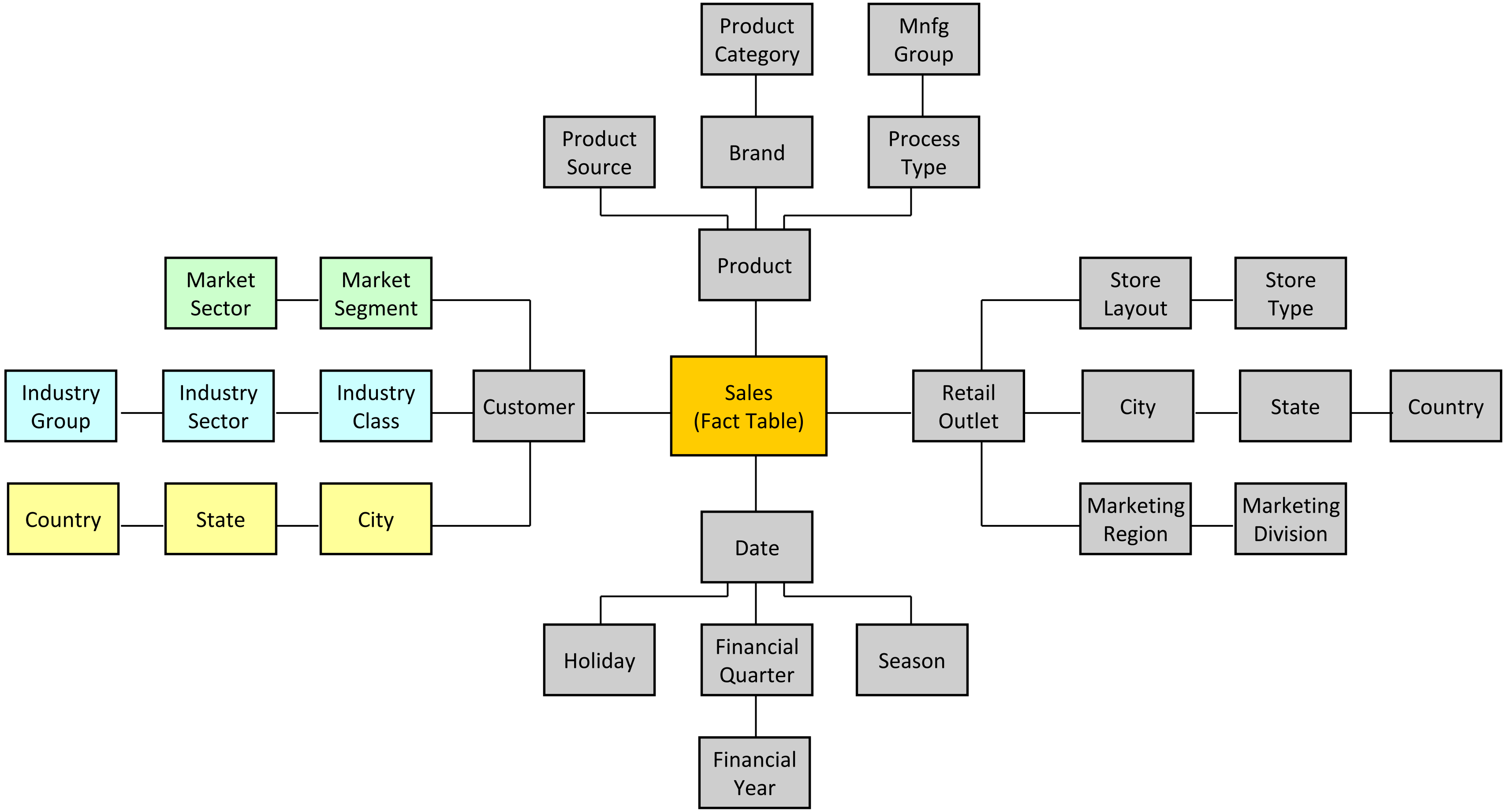
Customer-id
Customer name
Market segment
Market sector
Industry class
Industry sector
Industry group
City
State
Country

Embedded Hierarchies in Dimensional Tables



Customer-id
Customer name
Market segment
Market sector
Industry class
Industry sector
Industry group
City
State
Country

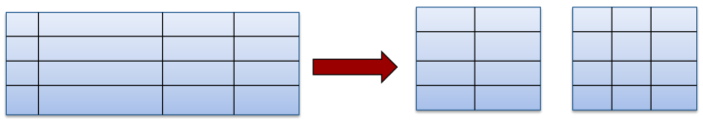
Snowflake Schema: hierarchy in dimensions



Design Outcomes: Normalised or Denormalised?

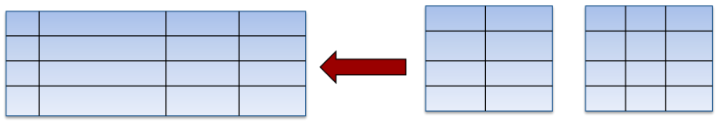
Normalisation

- Eliminates redundancy
- Storage efficiency
- Referential Integrity



Denormalisation

- Fewer tables (fewer joins)
- Fast querying
- Design is tuned for end-user analysis



Exercise

We are making a data warehouse for a real estate agency. The company wants to track information about the **selling** of their properties. This warehouse keeps information about the **agents** (license#, first name, last name, phone #), **buyers** that come in (buyer id, first name, last name, phone #), and **property** (property#, property address, price). The information managers want to be able to find is **the number of times a property is viewed**, **sales price**. The information needs to be broken down **by rental agent, by buyer, by property** and **for different time** (day, week, month, quarter and year).

Draw a star schema to support the design of this data warehouse.

What’s examinable

- Differences between transactional and informational databases
- Modelling a star schema
- Identifying the best grain level
- Defining facts and dimension tables

Thank you