



THE UNIVERSITY OF  
MELBOURNE

# INFO90002 Database Systems & Information Modelling

Lecture 19

## Data Warehousing

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

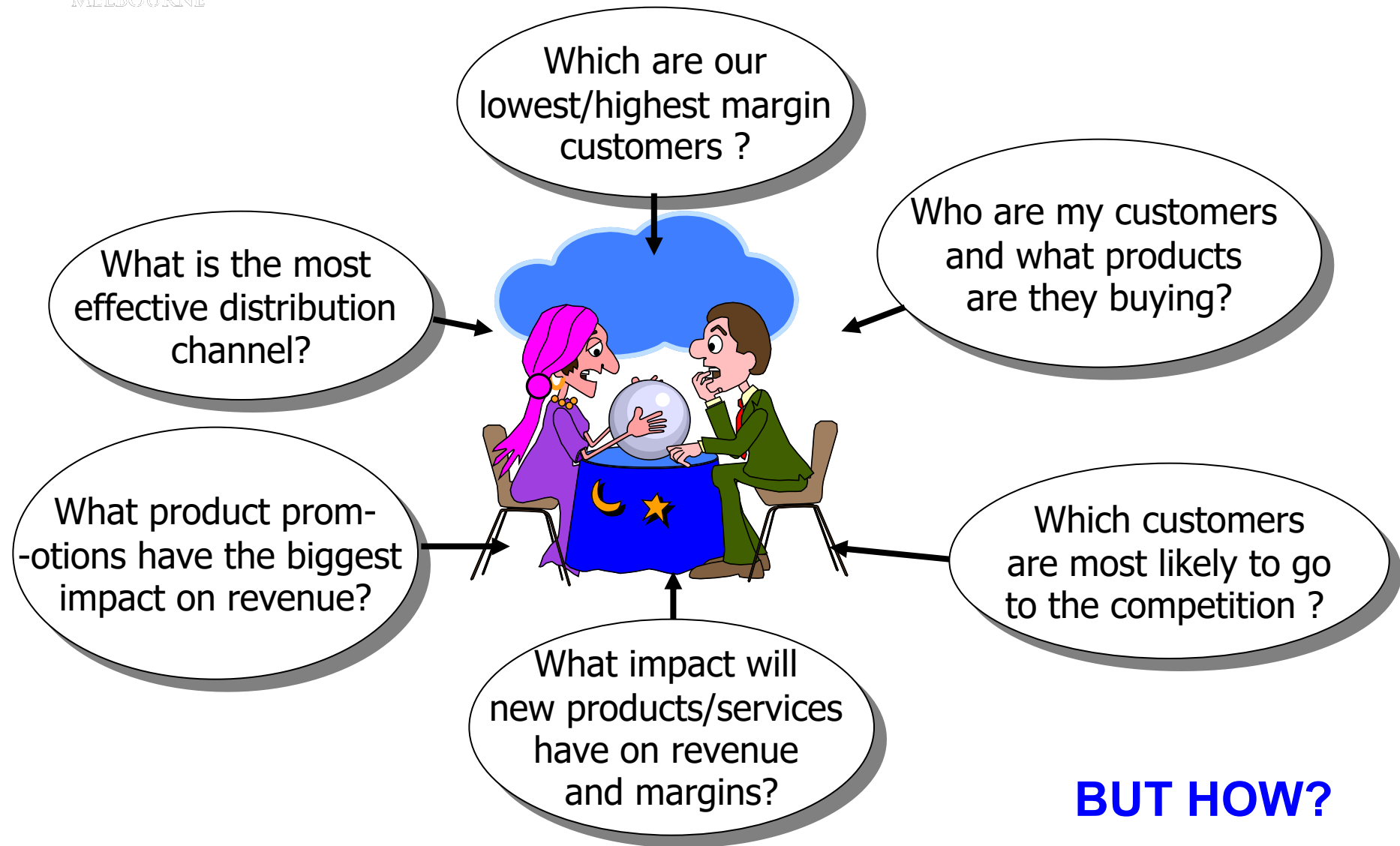
- Articulate the differences between **transactional** (operational) and **informational** (dimensional) databases
- Explain the **characteristics of a DW**
- Understand and explain the **overall architecture of a DW**
- Design **Star Schemas**



THE UNIVERSITY OF  
MELBOURNE

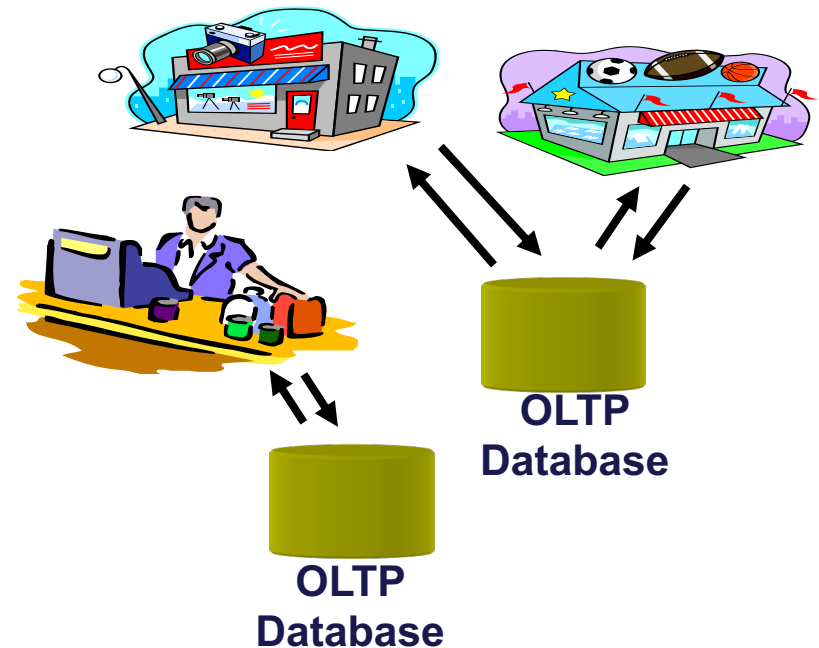
# Part 1: Introduction to Data Warehousing

# Motivations: A manager wants to know....



# 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



MELBOURNE

- 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, ...)

MELBOURNE

- 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!

MELBOURNE

- 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 we do about it?

MELBOURNE

- Need an integrated way of getting the ENTIRE organisational data
- Its really 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 about the data (SPOT)

MELBOURNE

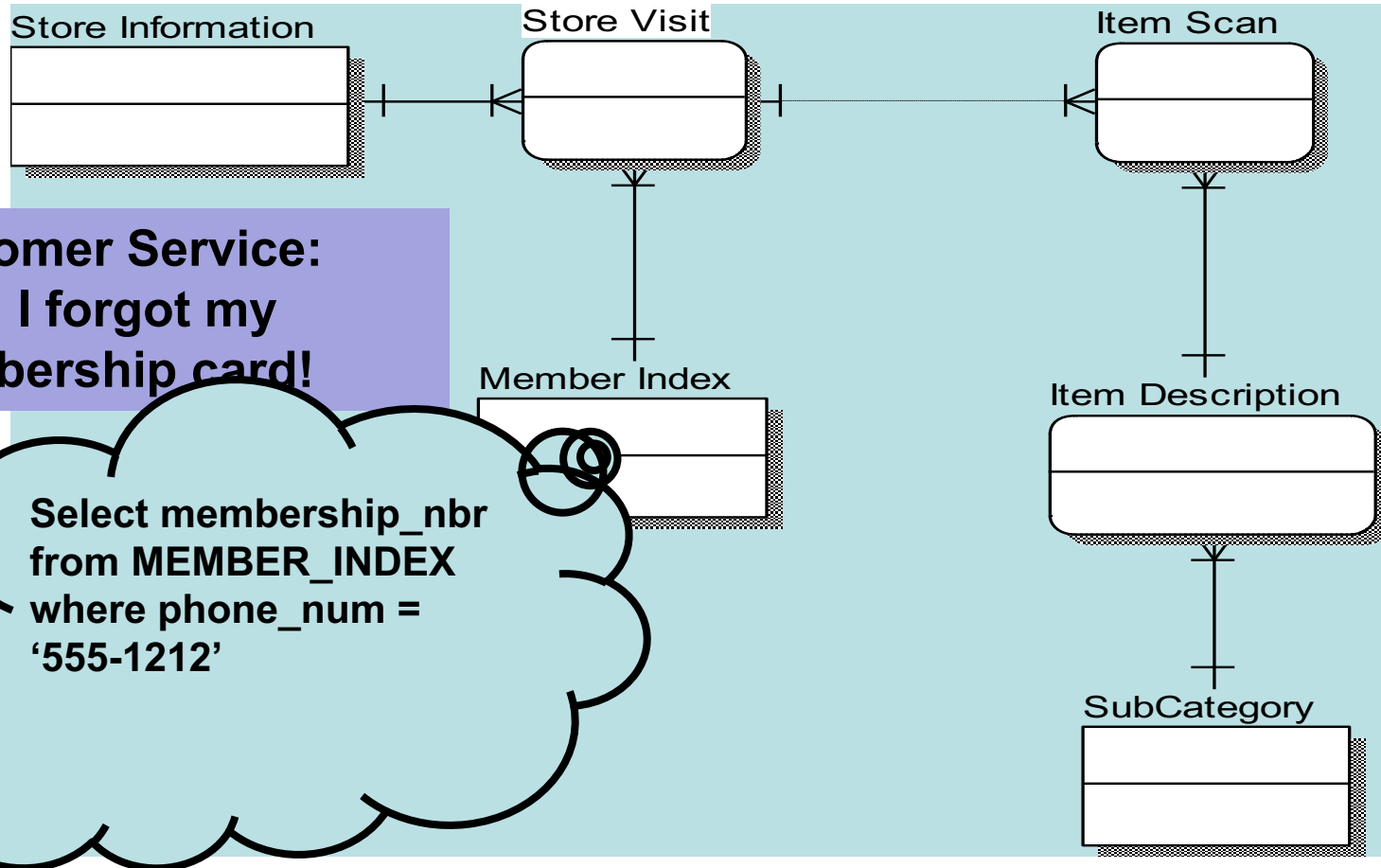
- **Data Warehouse:**
  - A single repository of organisational data
  - Integrates data from multiple sources
    - Extracts data from source systems, transforms, loads into the warehouse
  - Makes data available to managers/users
  - Supports analysis and decision-making
- Involve a large data store (often several Terabytes, Petabytes of data)

# Difference between Transactional & Informational Systems

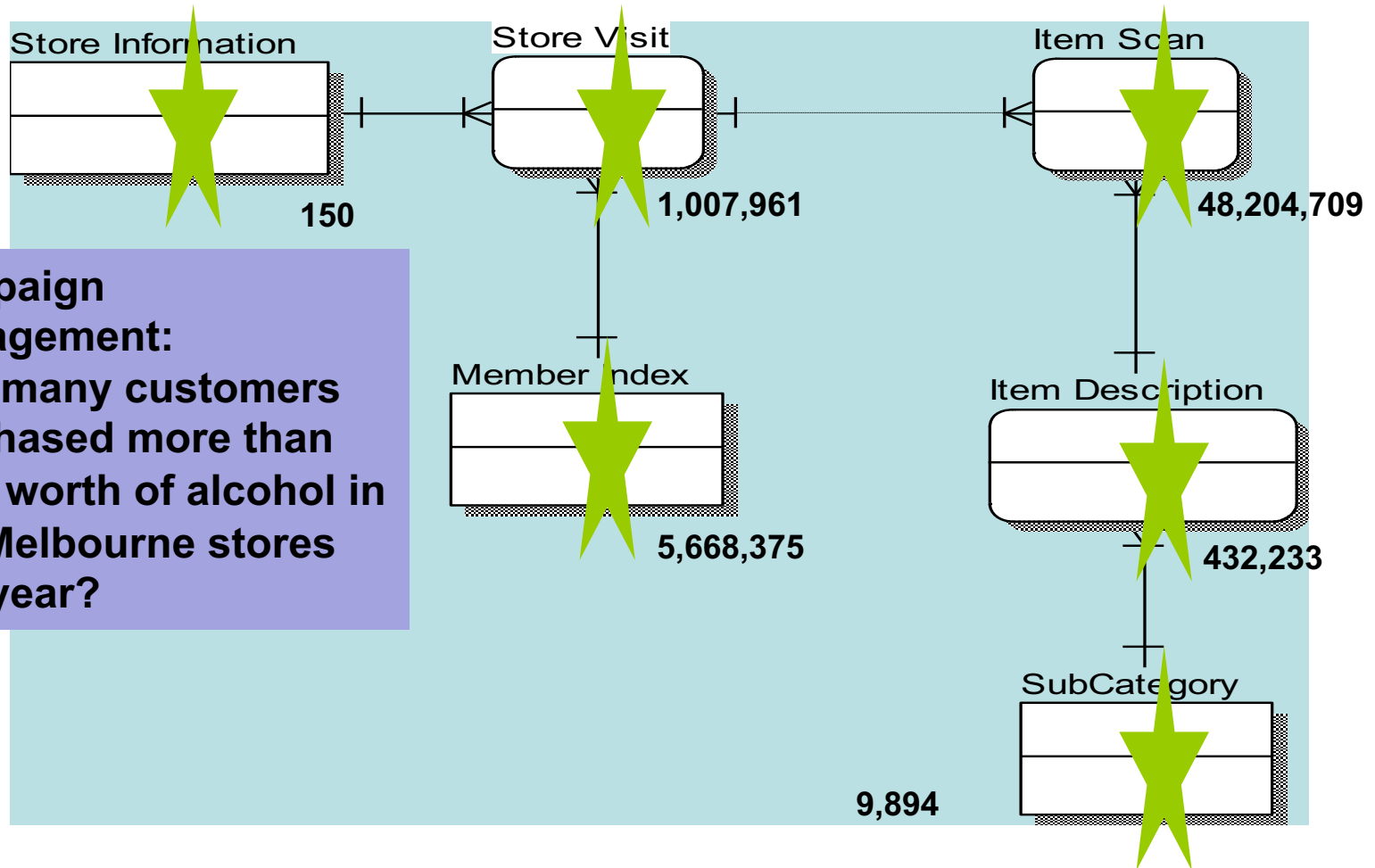
MELBOURNE

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

MELBOURNE



MELBOURNE



MELBOURNE

- One is interested in numerical *aggregations*
  - How **many**?
  - What is the **average**?
  - What is the **total cost**?
- One is interested in understanding *dimensions*
  - Sales **by state by customer type**
  - Sales **by product by store by quarter**

DW will help answer these questions

MELBOURNE

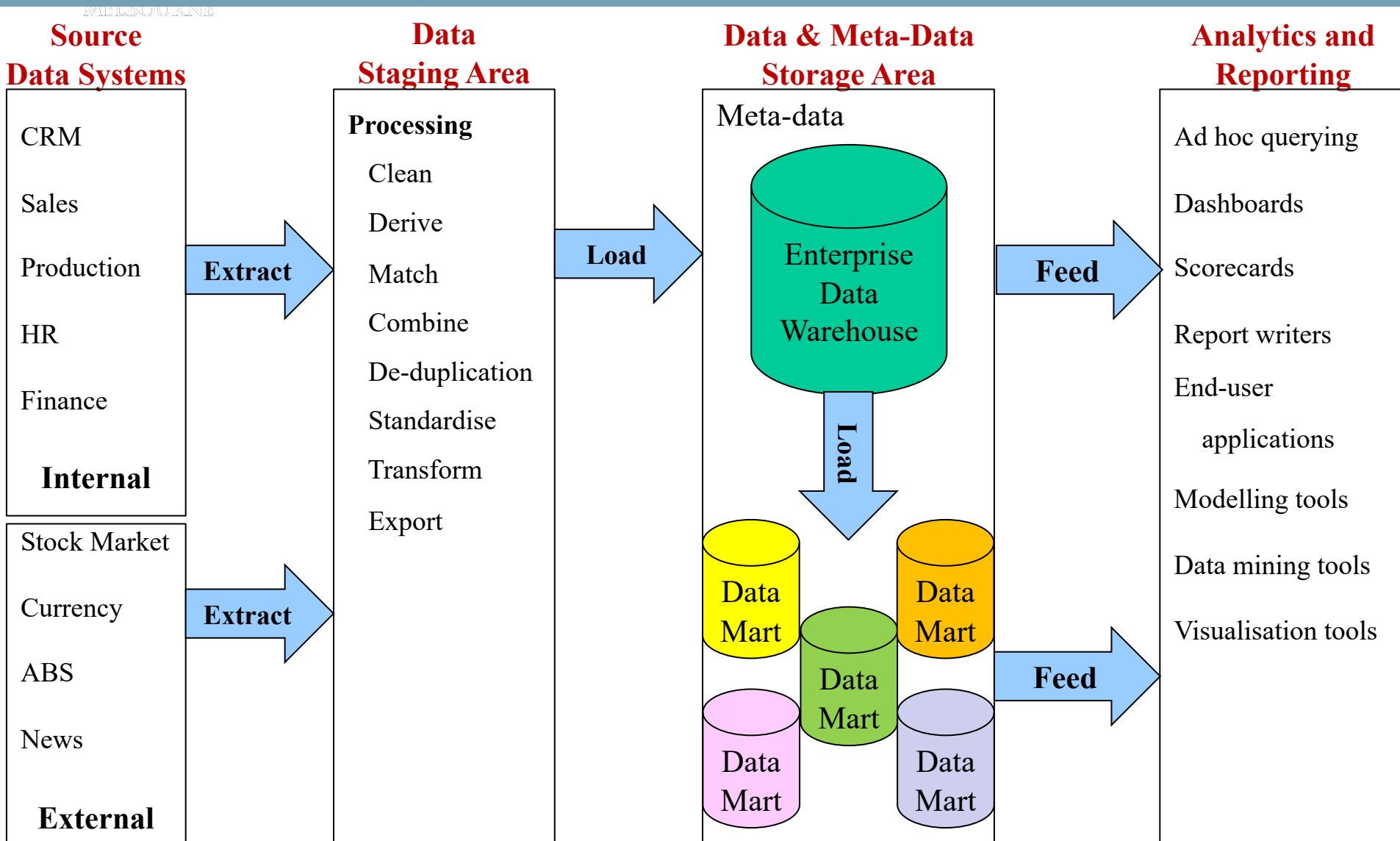
- Subject oriented
  - Data warehouses are organised around particular subjects (sales, customers, products)
- Validated, Integrated data
  - Data from different systems converted to a common format: allows comparison and consolidation of data from different sources
  - Data from various sources validated before storing it in a data warehouse



MELBOURNE

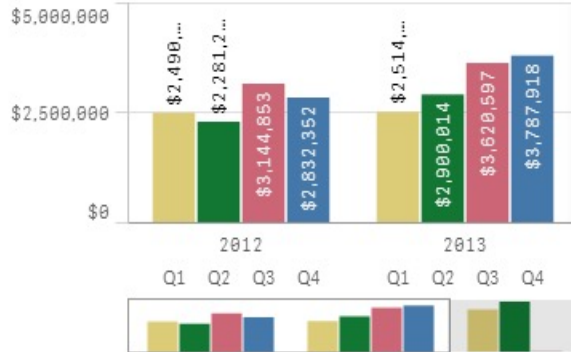
- Time variant
  - Historical data
  - Trend analysis crucial for decision support: requires historical data
  - Data consists of a series of “snapshots” which are time stamped
- Non-volatile
  - Users have read access only – all updating done automatically by ETL process and periodically by a DBA

# A DW Architecture

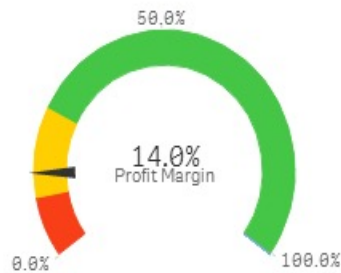


MELBOURNE

Total Sales = \$31,314.1K

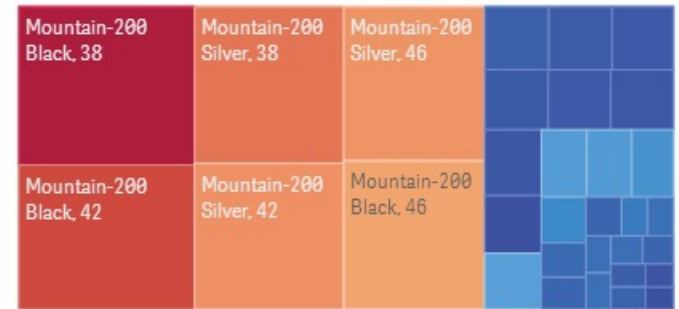


Profit Margin

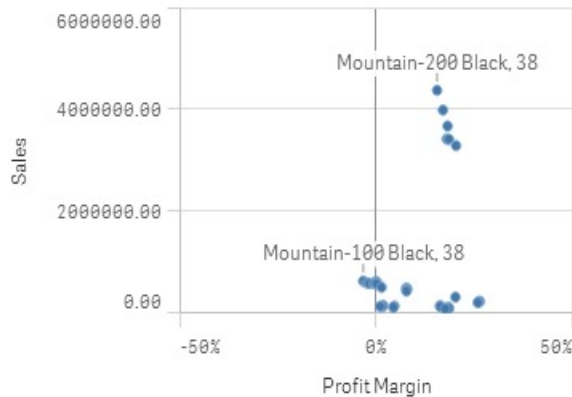


Sales

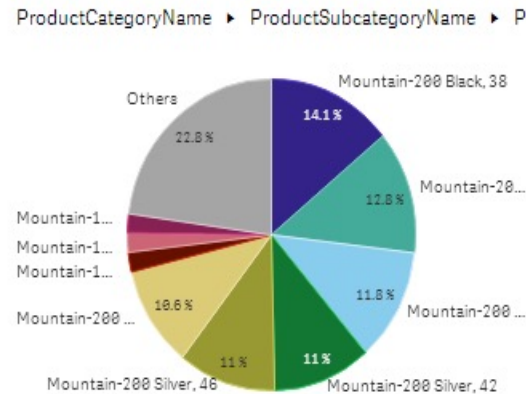
\* red = most ordered



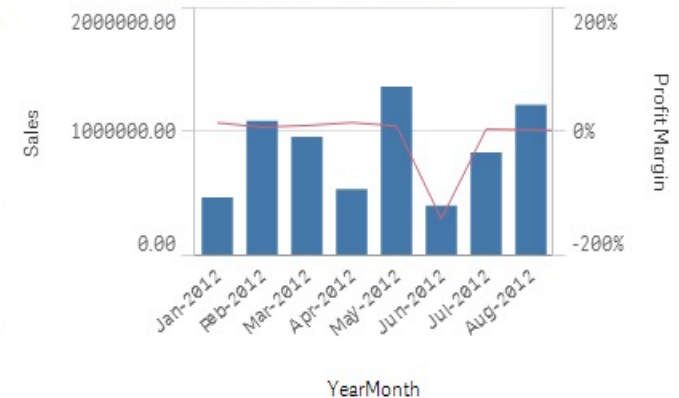
Sales vs Profit Margin



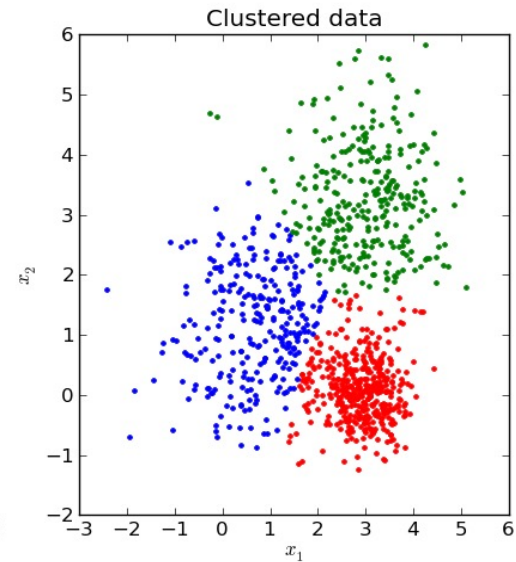
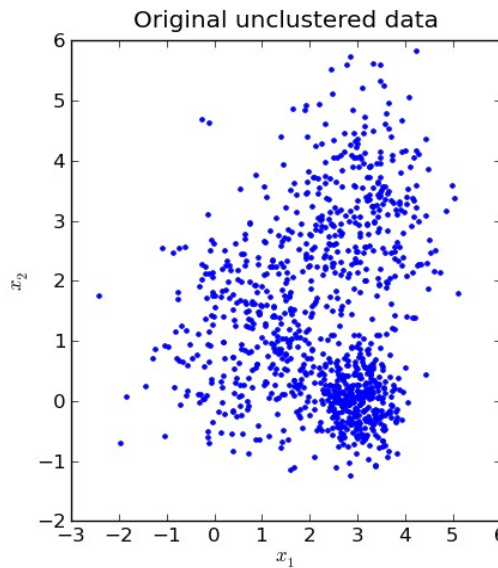
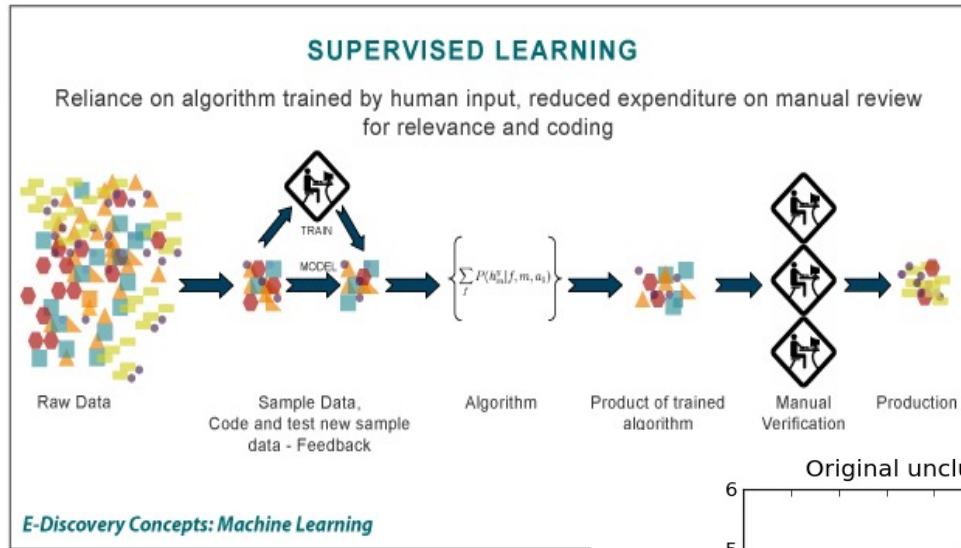
% of Total Sales



Sales and Profit Margin by Year-Month



MELBOURNE



<http://us.hudson.com/legal/blog/postid/513/predictive-analytics-artificial-intelligence-science-fiction-e-discovery-truth>  
<http://pypr.sourceforge.net/kmeans.html>

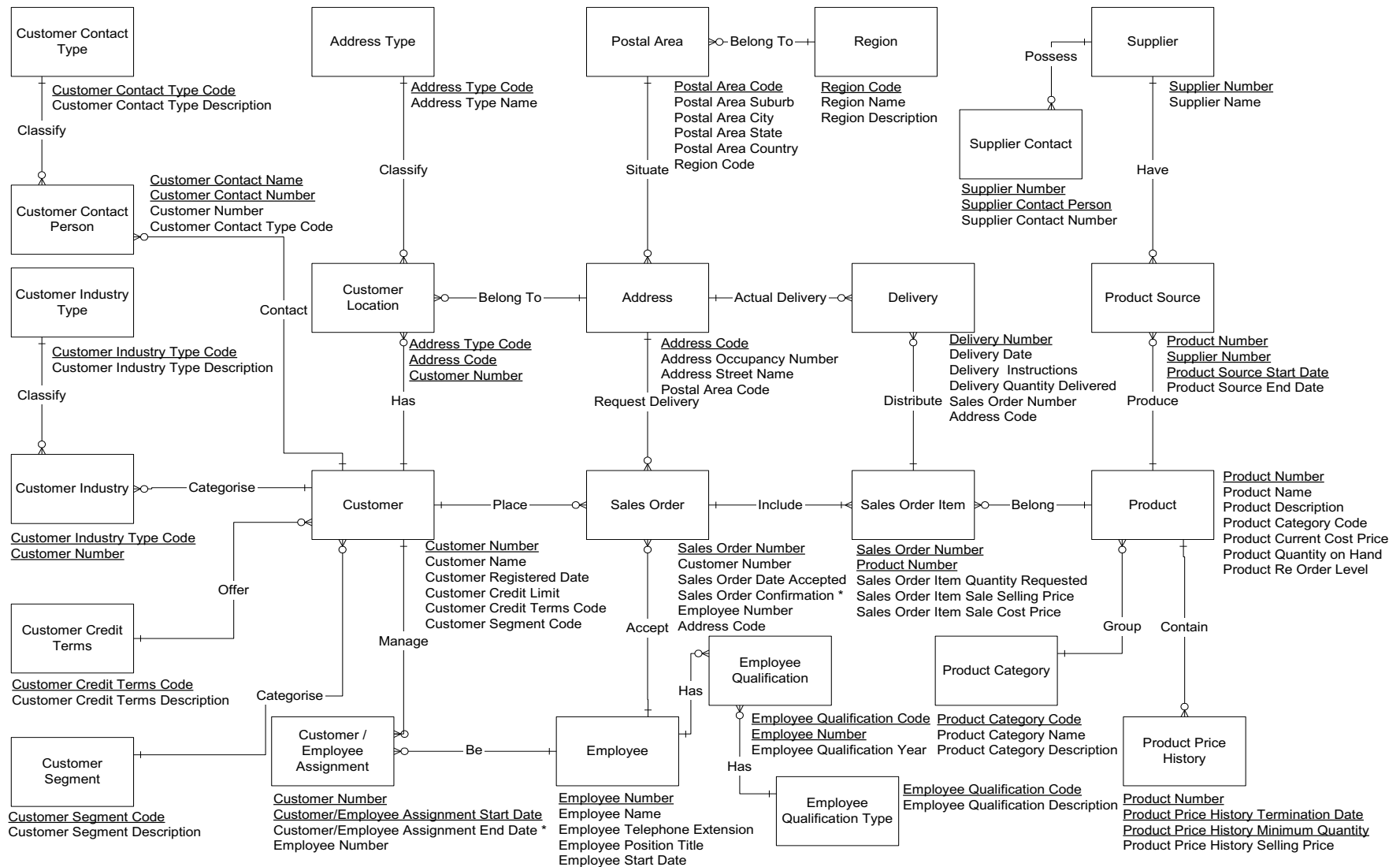


THE UNIVERSITY OF  
MELBOURNE

# Part 2- Dimensional Modelling

MELBOURNE

- How much **revenue** <sup>Fact</sup> did the **product G** generate in the **last three months**, broken down by month for the south eastern sales **region**, <sup>Dimension</sup> by individual **stores**, broken down by **promotions**, <sup>Dimension</sup> 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?  
<sup>Fact</sup> <sup>Dimension</sup> <sup>Dimension</sup> <sup>Dimension</sup>



MELBOURNE

- 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



MELBOURNE

- 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*

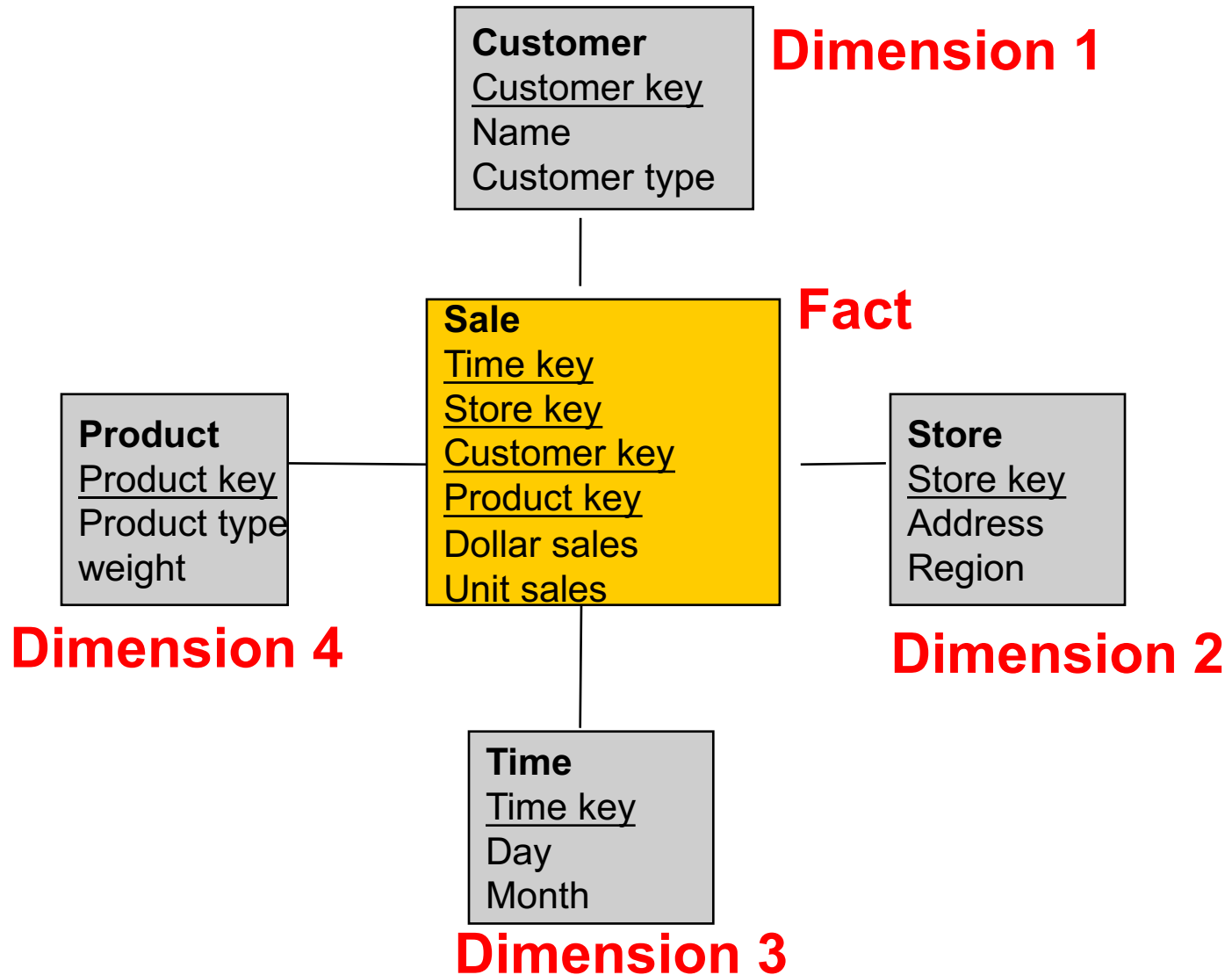


# Fact Table - example

- 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

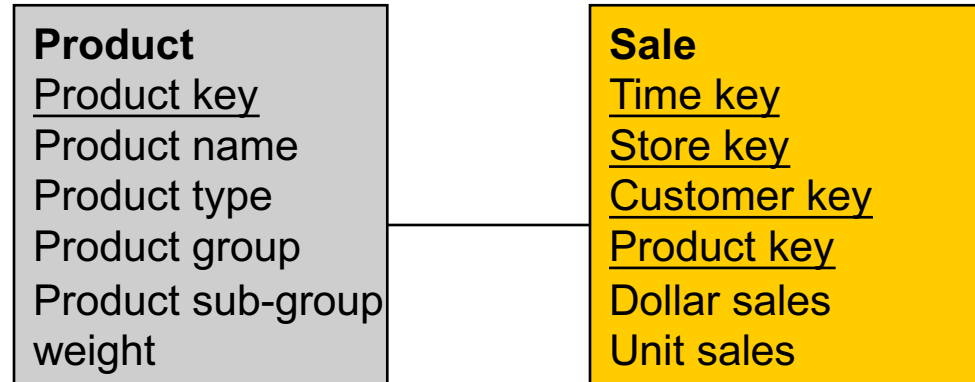
| <i>Time-id</i> | <i>Store-id</i> | <i>Cust-id</i> | <i>Prod-id</i> | <i>Dollar sales</i> | <i>Unit Sales</i> |
|----------------|-----------------|----------------|----------------|---------------------|-------------------|
| T100           | S303            | C101           | P98            | \$120,000           | 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

MELBOURNE



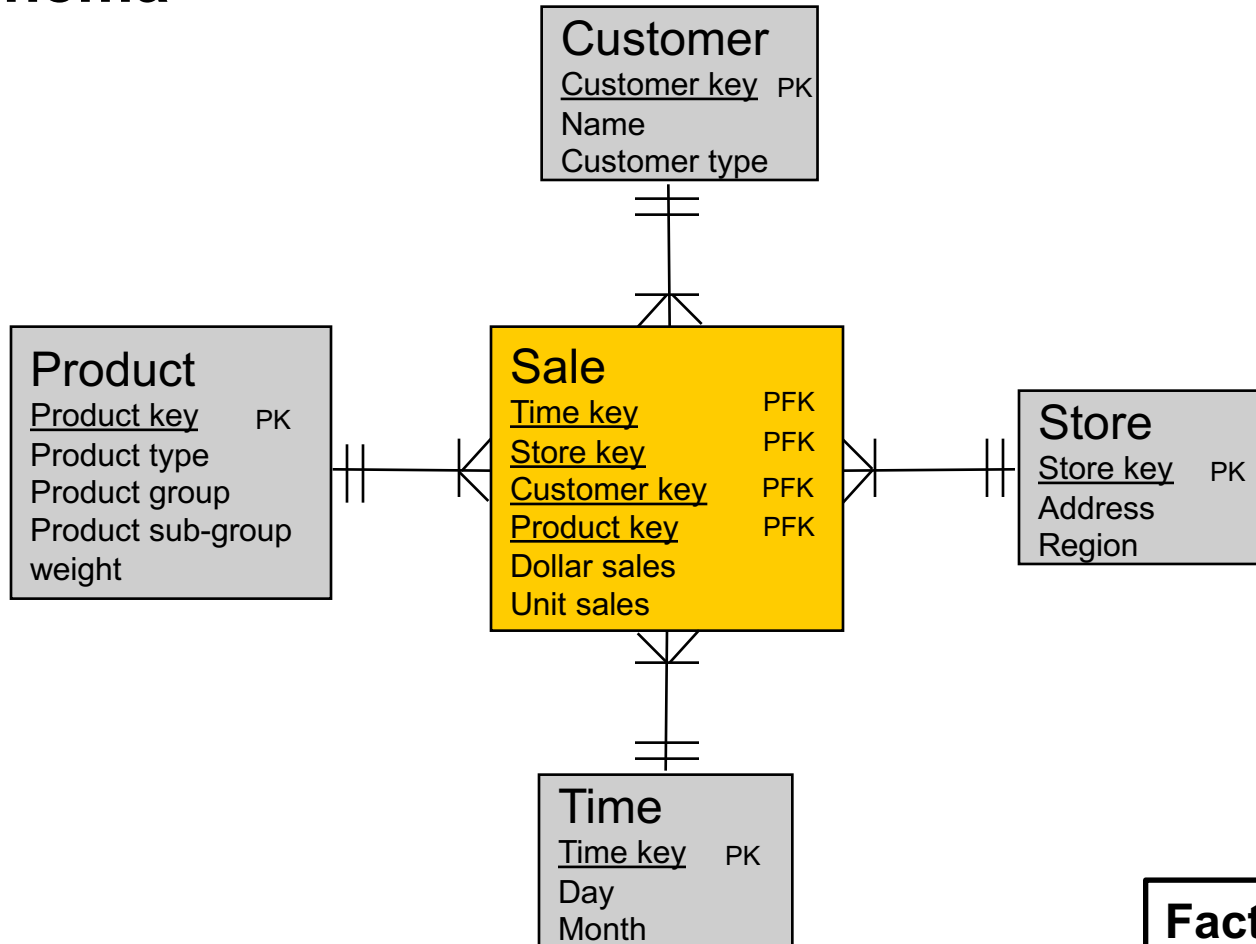
|                 |               |
|-----------------|---------------|
| Product name    | e.g. Hammer   |
| - Product type  | e.g. Tool     |
| - Product group | e.g. Hardware |

# Dimension Table - example

- Actual data might look like this
- Hierarchy evident in data

| <i>Prod-id</i> | <i>Prod-Name</i> | <i>Prod-Group</i> | <i>Prod-Subgroup</i> | <i>Weight</i> |
|----------------|------------------|-------------------|----------------------|---------------|
| P10            | Hammer           | Hardware          | Tool                 | 5kg           |
| P56            | 10cm Nails       | Hardware          | Nails                | 1kg           |
| P98            | Plastic Pipe     | Plumbing          | Pipe                 | 1kg           |

## Star schema



**Fact table is an  
intersection table**

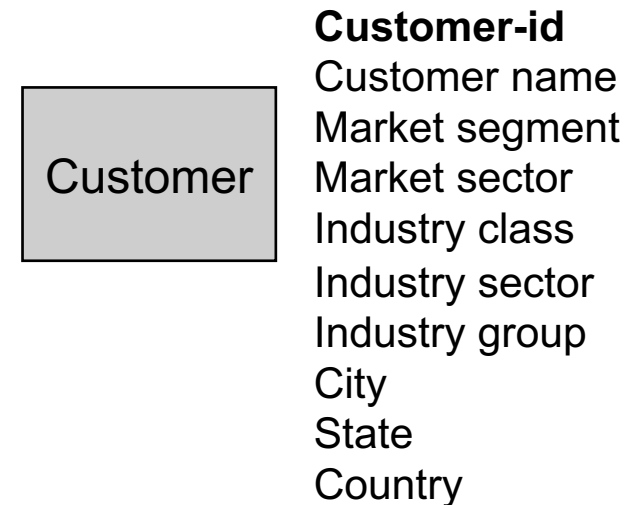
MELBOURNE

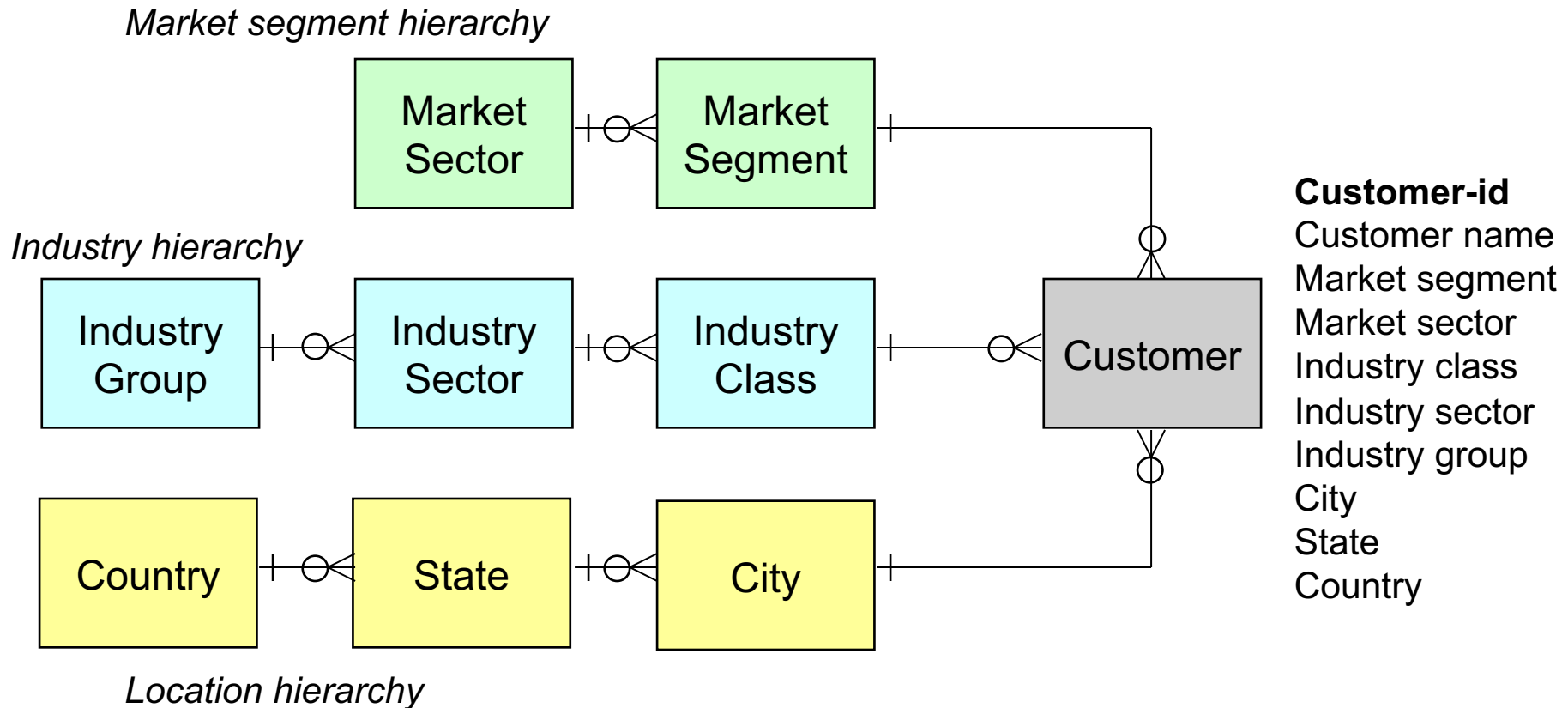
## Steps:

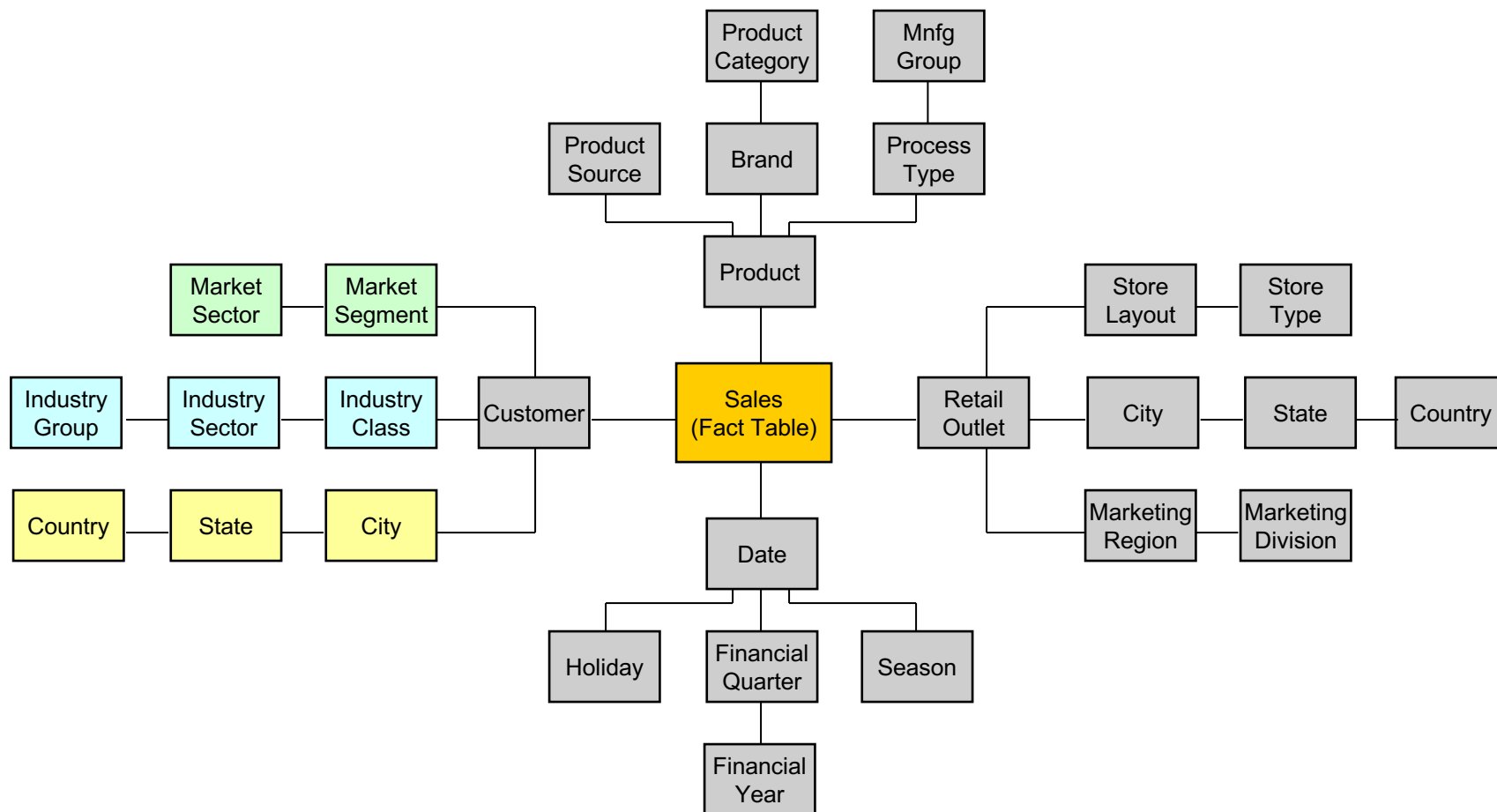
1. Choose a Business Process
2. Choose the measured facts (usually numeric, additive quantities)
3. Choose the granularity of the fact table
4. Choose the dimensions
5. Complete the dimension tables

(Kimball, 1996)





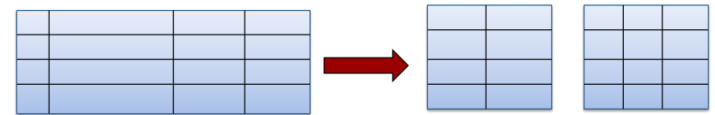




# Design Outcomes: Normalised or Denormalised?

- Normalisation

- Eliminates redundancy
- Storage efficiency
- Referential Integrity



- Denormalisation

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



MELBOURNE

- 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 accessible **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 is Examinable?

MELBOURNE

- Differences between transactional and informational databases
- Designing a star schema
- Defining facts and dimension tables

More technical details (Not assessed)

<https://www.youtube.com/watch?v=w-S0fj0fmqg&list=PLdQddgMBv5zHcEN9RrhADq3CBColhY2hl&index=17>

MELBOURNE

- Security & Ethics