INFO20003 Database Systems

Week 11

Data Warehouse

- a single database of organisational data, that allow all the organisation's data to be stored in a form, that support managers' decision making

- Characteristics:
  - Validated & integrated data
  - Time variant
  - Subject oriented
  - Non-volatile

Business Events

- *an event that occurs as part of a business process*

Examples:

- Sales business process: an individual order or sale
- Finance: a payment
- Marketing: a view of a webpage or a click on an online ad

Data warehouse modelling

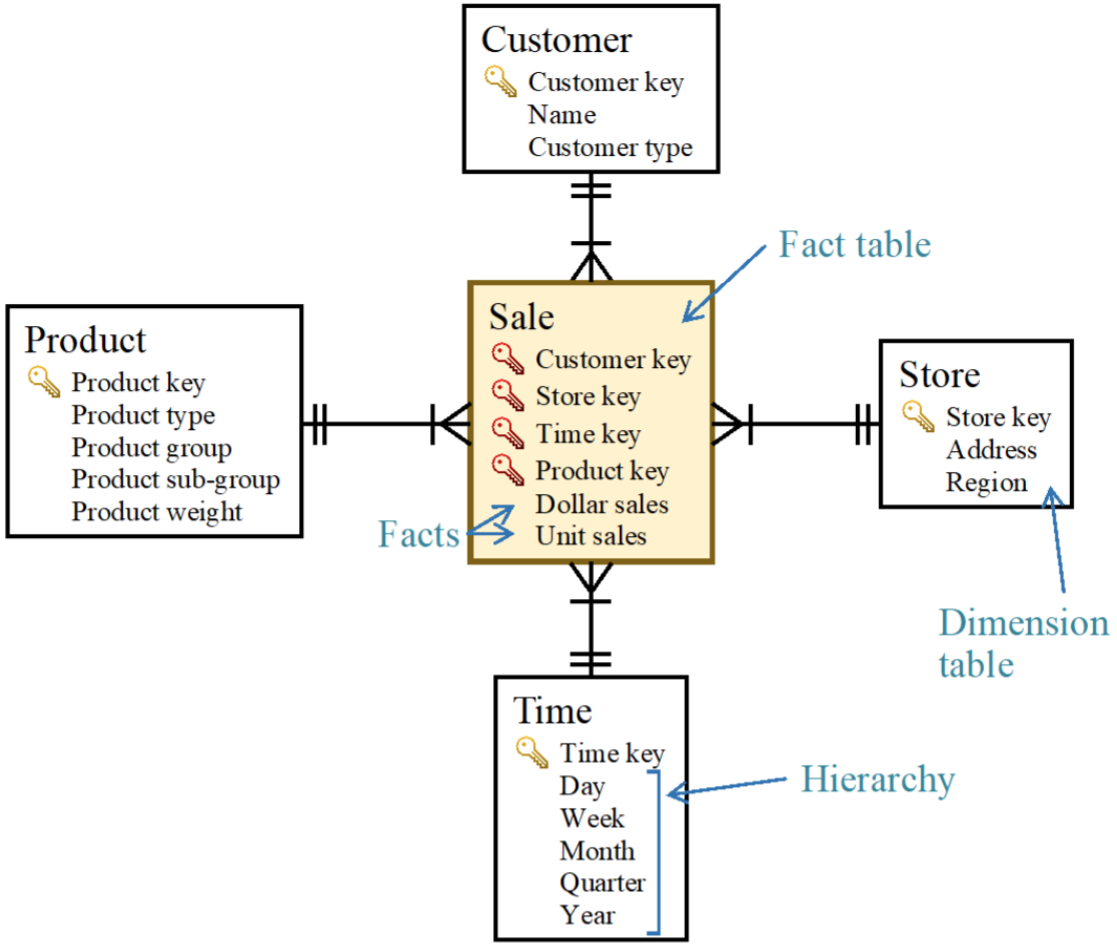


Figure 1: A simple star schema for a sales data warehouse with four dimension tables.

Dimension

- Dimension: *an entity that describes and gives context to a business event*
- Common dimension: time, customers, products, locations

a CEO is interested in a comparison of revenue of a new model of the product with the older model, in every quarter of the year, by customer demographic group

Dimensions: Time, Product, Customer

- represented as dimension tables

Dimension

Hierarchy: A sequence of attributes that describes a dimension across different levels of detail

- E.g. Location: city → state → country
- E.g. Time: day → week → month → quarter → year

- Stored as attributes of the dimensional tables
- Used for selecting and aggregating data at desired level of detail

Dimensions

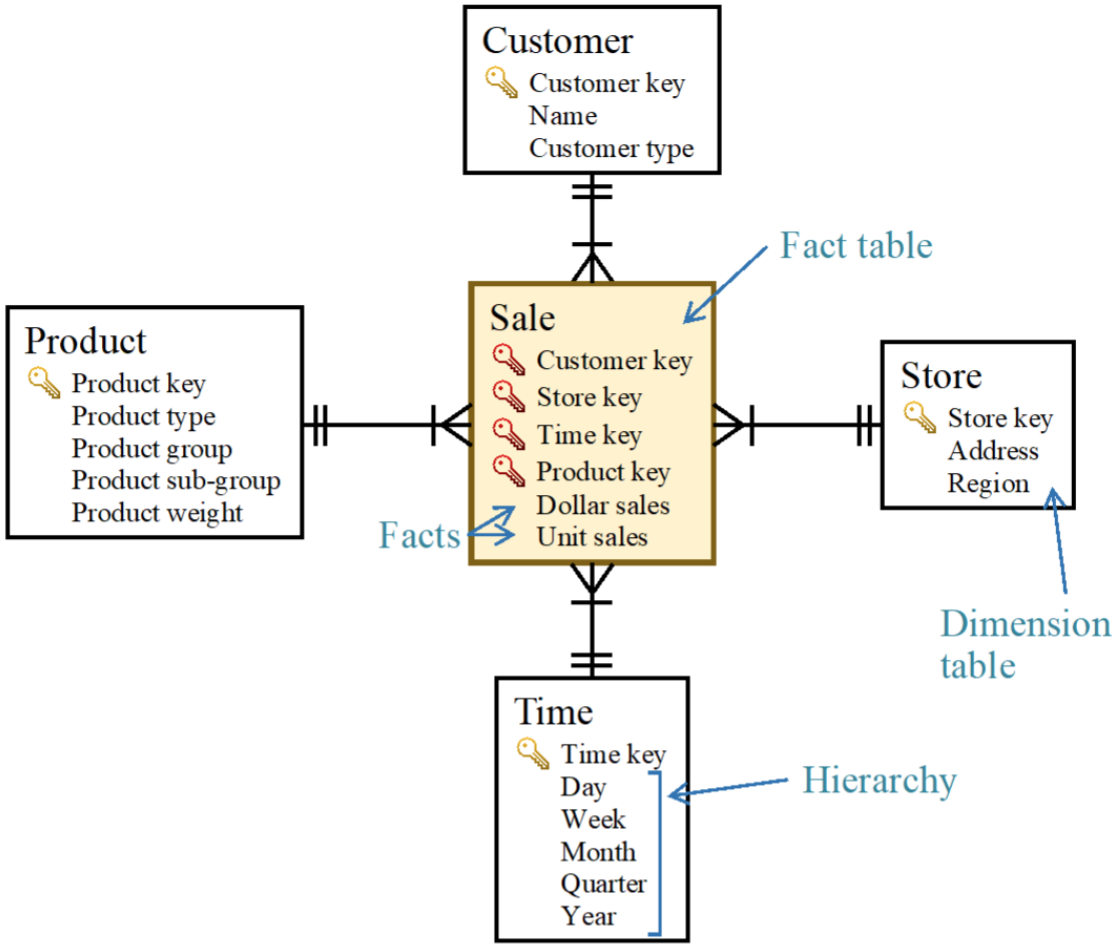


Figure 1: A simple star schema for a sales data warehouse with four dimension tables.

Dimensions: Customer, Store, Time, Product

Examples of aggregated data:

- Sale per Store per Day
- Sale per Product type per Day
- Sale per Product group per Year

Fact

- Fact: is a numeric measurement of a meaningful and significant business event

- E.g. **Customer** buys a **product** at a certain **location** at a certain **time**
  - E.g. facts: revenue, #items sold, total profit earned

- stored in a fact table as attributes

Fact

- Fact table: an intersection of the dimensions
  - PK = foreign keys referencing the dimension tables
  - Other attributes = facts
- grain/granularity: the level of detail present in a fact table
- The fact table can store:
  - each business event in its own row
  - many business events aggregated together
- The finer the granularity is, the more precisely a query can extract details from the database

Fact

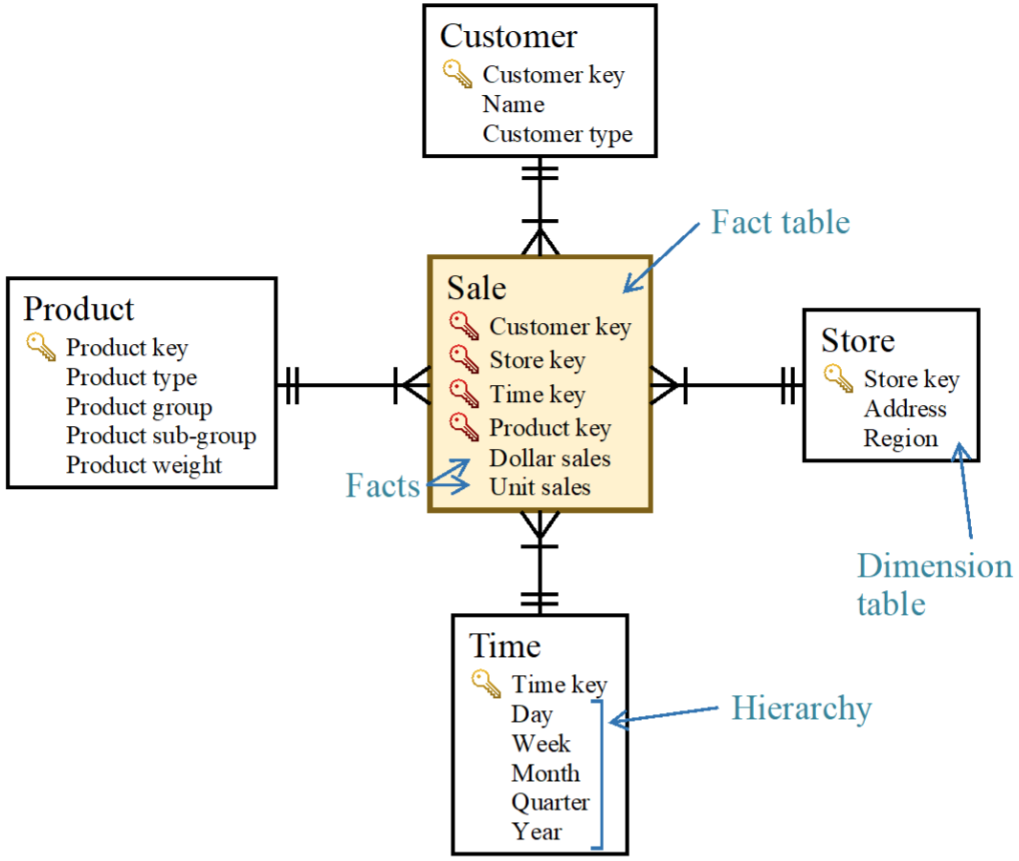


Figure 1: A simple star schema for a sales data warehouse with four dimension tables.

Facts: dollar sales, unit sales

Granularity:

- Product type
- Day
- Store
- Customer

Questions:

- *dollar sale* per customer per product type per day per store?
- *total dollar* sale per customer type, per product group, per year, per store region?
- *average unit sale* per week?

Q1

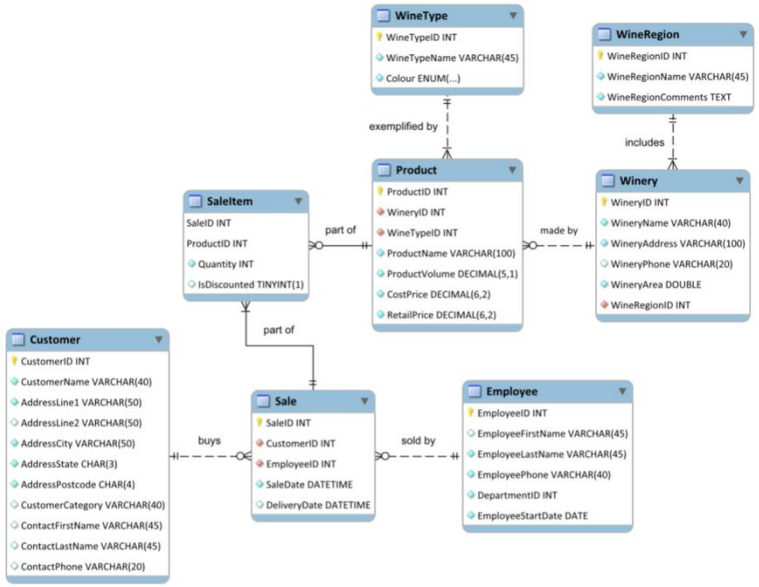
1. Designing a dimensional model

Wimmera Wines is a large company that takes deliveries of grapes from wine growers, produces and bottles wine, and sells those bottles to retailers and restaurants. They produce many different types of wine at a range of price points, from cheap cask wine to top-of-the-range vintage bottles.

Wimmera Wines' day-to-day OLTP database uses the following ER model:

The company is aiming to increase their product sales by 20% in comparison to the last 3 years. To help the business achieve their aim, you have been hired to design a data warehouse that can help business managers analyse data related to the sales theme.

The company is keen to understand all the aspects of their business that contribute to strong sales. For example, two business measures that have been mentioned are "total number of units of each product sold" and "revenue generated by each employee per year".



Q1a)

a. As a class, brainstorm some more business measures that Wimmera Wines managers might need if they are to achieve their aim.

The company is keen to understand all the aspects of their business that contribute to strong sales. For example, two business measures that have been mentioned are "total number of units of each product sold" and "revenue generated by each employee per year".

- Number of products sold per year
- Sales by a particular state
- Sales of a product in a given quarter of a year
- Revenue generated from a particular customer category
- Which product is selling the best (hence generating the most revenue)?

Q1b)

b. Use Kimball's five-step dimensional design process to design a dimensional model for Wimmera Wines' product sales subject area.

i. Select and explain the business process.

product **sales**

Q1b)

b. Use Kimball's five-step dimensional design process to design a dimensional model for Wimmera Wines' product sales subject area.

ii. Identify and explain the facts.

The following sales-related facts can be extracted from the source database:

- Unit sales
- Dollar sales
- Profit amount
- Discount indicator (non-additive fact)

Q1b)

b. Use Kimball's four-step dimensional design process to design a dimensional model for Wimmera Wines' product sales subject area.

iii. Declare the grain and justify your choice.

- Wimmera Wines sells to retailers and restaurants:
  - Would not make a large number of sales
  - each individual sale can include large quantities of items

- Appropriate to store each sale item as its own row
- If do not require such detailed info, and perhaps weekly sales are sufficient → data would be aggregated by week

Q1b)

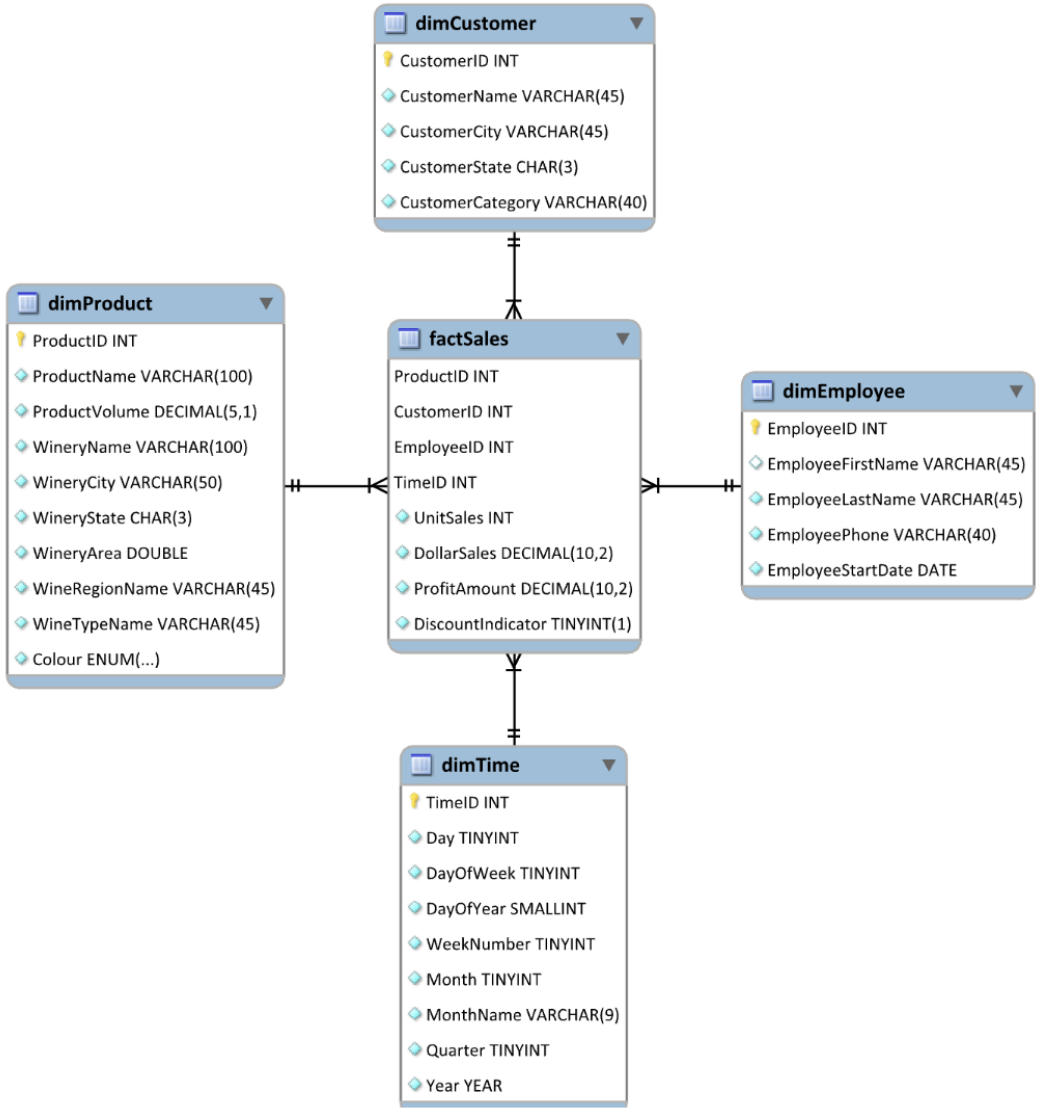
b. Use Kimball's four-step dimensional design process to design a dimensional model for Wimmera Wines' product sales subject area.

iv. Identify and explain the dimensions.

- Employee
- Customer
- Time
- Product

Q1 Data Warehouse design

Notice how the Product dimension is denormalised. It has many transitive functional dependencies, such as WineryName → WineryCity and WineTypeName → Colour.



Q2

Consider the following fact table:

Suppose the following sales data has been extracted from the business's operational database:

SaleID	SaleDate	CustomerID	CustomerCity	ProductID	Price	Quantity
54	2003-12-13 14:13	788	Melbourne	9644	$10.00	2
54	2003-12-13 14:13	788	Melbourne	8574	$15.00	1
67	2003-12-13 15:05	903	Melbourne	9644	$10.00	1
76	2003-12-13 17:26	322	Sydney	9644	$5.00	4
77	2003-12-14 09:58	292	Melbourne	8229	$15.00	2

a. Starting from this source data, how many rows will be inserted into the fact table if an hourly grain is selected?
b. How many rows will be inserted into the fact table if a daily grain is selected?
c. At which level of granularity can we answer questions about hourly sales? At which level of granularity can we answer questions about daily sales?

Q2a)

a. Starting from this source data, how many rows will be inserted into the fact table if an hourly grain is selected?

SaleID	SaleDate	CustomerID	CustomerCity	ProductID	Price	Quantity
54	2003-12-13 14:13	788	Melbourne	9644	$10.00	2
54	2003-12-13 14:13	788	Melbourne	8574	$15.00	1
67	2003-12-13 15:05	903	Melbourne	9644	$10.00	1
76	2003-12-13 17:26	322	Sydney	9644	$5.00	4
77	2003-12-14 09:58	292	Melbourne	8229	$15.00	2

None of these sale-item rows share the same hour, geography and product. No aggregation can be performed. Five rows will be inserted into the fact table.

Q2b)

b. How many rows will be inserted into the fact table if a daily grain is selected?

SaleID	SaleDate	CustomerID	CustomerCity	ProductID	Price	Quantity
54	2003-12-13 14:13	788	Melbourne	9644	$10.00	2
54	2003-12-13 14:13	788	Melbourne	8574	$15.00	1
67	2003-12-13 15:05	903	Melbourne	9644	$10.00	1
76	2003-12-13 17:26	322	Sydney	9644	$5.00	4
77	2003-12-14 09:58	292	Melbourne	8229	$15.00	2

54 & 67 will be aggregated

In total, four rows will be inserted into the fact table.

Q2c)

c. At which level of granularity can we answer questions about hourly sales? At which level of granularity can we answer questions about daily sales?

Hourly sale: a hourly grain or finer

Daily sale: a daily grain or finer