# INFO20003 Tutorial – Week 11 Solutions

(Tutorial: Data warehousing)

## Objectives:

This tutorial will cover:

I. Understand the fundamentals of dimensional modelling – 20 mins
II. Design a dimensional model using Kimball's four-step design process – 25 mins
III. Discuss the impact of grain on fact tables – 10 mins

## Key Concepts:

***NOTE for students:*** *This is a brief summary of some of the concepts taught in lecture 20. The lectures contain detailed content related to these and many more concepts. These notes should be considered quick revision instead of a sole resource for the course material.*

- Data warehouse

  A **data warehouse** is a single database of organisational data that allows all of the organisation's data to be stored in a form that supports managers' decision making. The data is integrated from multiple sources internal and external to the organisation by converting it into a common format and validating before storing it to ensure credibility of the warehouse. It keeps historical data and is available to the managerial bodies of the organisation to support high-level decision-making processes and data analysis.

  Unlike an ER model, where data is organised around conceptual entities, the data in a data warehouse is organised around *business processes* such as sales, finance, or marketing.

- Business events

  Data warehouses store information about **business events**. A business event is an event that occurs as part of a business process. For the sales business process, an individual order or sale would be considered a business event. For finance, a payment would be a business event, and for a marketing data warehouse, it might be a view of a webpage or a click on an online ad.

- Dimensions, dimension tables and hierarchies

  A **dimension** is an entity that *describes* and *gives context to* a business event. Some examples of commonly used dimensions are time, customers, products and locations. For example, if 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, the relevant dimensions are Time (quarter of the year), Product (product version) and Customer (customer demographic). Similarly, for an insurance company, an example of a key measurement (fact) is claims, and the dimensions could be agent, policy, customer, and time.

  Dimensions are represented in the data warehouse as **dimension tables**. Within each dimension table, a range of attributes may be stored. A sequence of attributes that describes a dimension across different levels of detail is called a **hierarchy**. For example, for the dimension table 'Location', the data can be stored at various levels such as city, state, or much higher level of country and so on. Similarly, the Time dimension will have a hierarchy of day, week, month,

quarter and year. The hierarchies of dimensions are stored as attributes of the dimensional tables and all the related hierarchies are typically stored in a single dimension table (unless a snowflake schema is used, where parts of the hierarchy can become individual tables). These hierarchies are used for selecting and aggregating data at the desired level of detail. In other words, hierarchies help with *slicing and dicing* the data.

- Facts, fact tables and granularity

  A **fact** is a numeric measurement of a meaningful and significant business event. Consider the scenario where a *customer* buys a *product* at a certain *location* at a certain *time*. The intersection of these four dimensions constitutes a *sale* (the business event). The sale can be measured in terms of the amount of revenue generated, number of items sold, total profit earned, etc – these are all *facts* relating to the sale.

  In a data warehouse, the facts (numerical performance measurements) of a business are stored in a **fact table**. A row in a fact table corresponds to one or more business events.

  A data warehouse fact table is defined as an intersection of the dimensions that describe the business event. In general, the fact table has a PK made up of the foreign keys connecting it to the dimension tables.

  The level of detail present in a fact table is referred to as **grain** or **granularity**. The fact table can store each business event in its own row (for example, preserving each sales event as an individual row) or it can store many business events aggregated together (if sales data is aggregated down to one row per hour or per day). The finer the granularity is, the more precisely a query can extract details from the database.

- Dimensional modelling – the star schema

  The model in which the fact table consisting of numeric measurements is related to all the dimension tables storing descriptive attributes is termed the **dimensional model**. The fact table is at the centre and the dimensional tables are on the sides, making a **star schema**. Figure 1 shows a star schema with Sale as the fact table and Customer, Product, Store and Time as dimensions of the business.

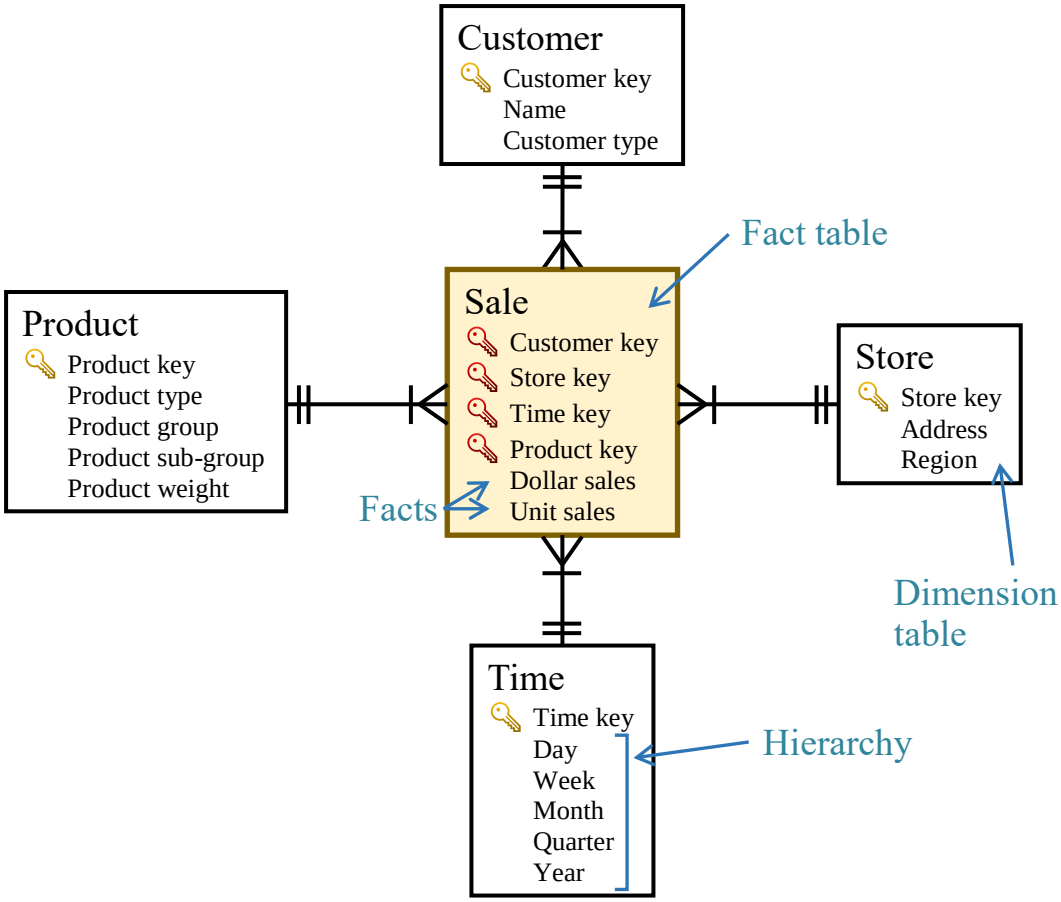


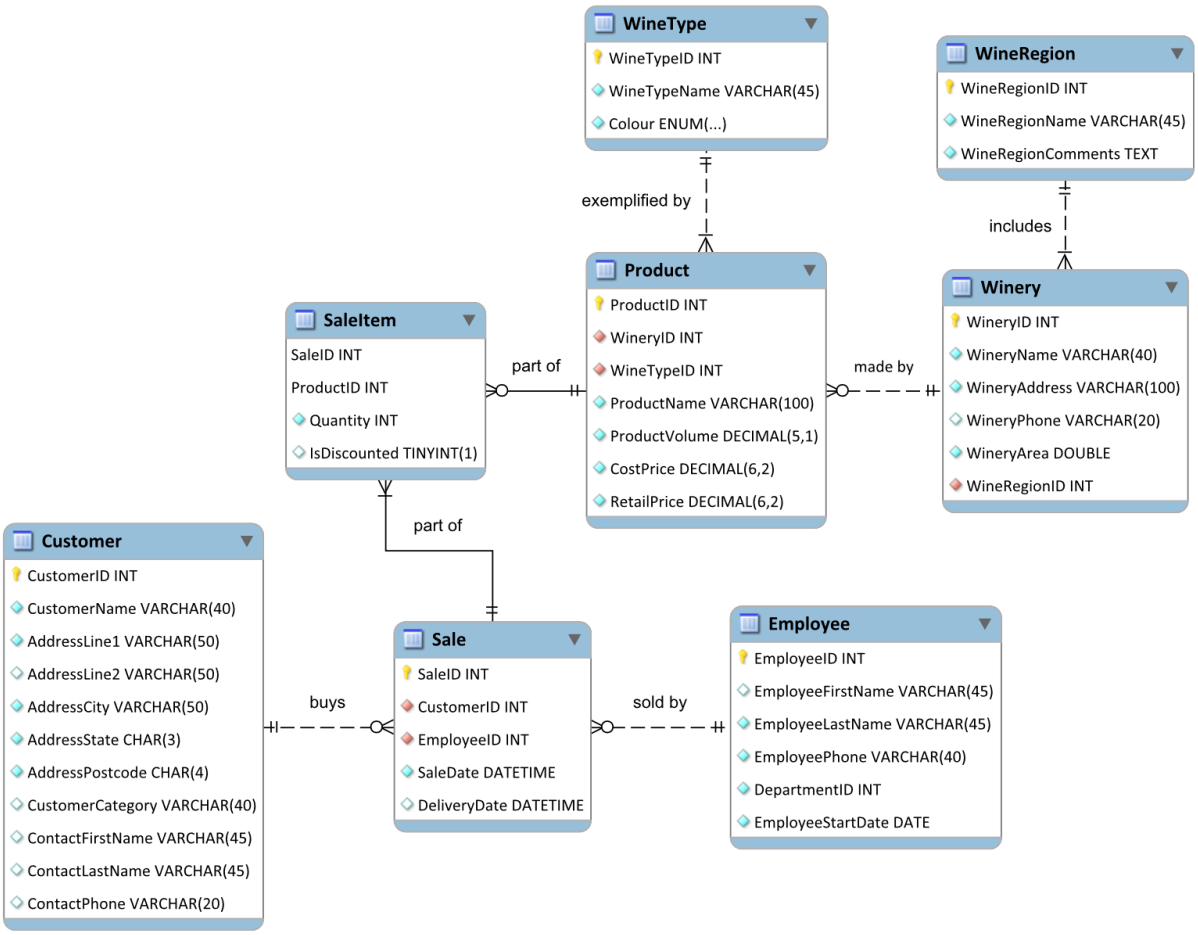

*Figure 1: A simple star schema for a sales data warehouse with four dimension tables.*

## Exercise:

### 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”.

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

- 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)?

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

i. Select and explain the business process.

As stated in the case study, product **sales** is the business process. Analyses related to sales can be of varying natures and may use different measures associated with Sales.

ii. Declare the grain and justify your choice.

Because Wimmera Wines sells to retailers and restaurants, they would not make a large number of sales (but each individual sale can include large quantities of items). It is appropriate to store each sale item as its own row in the fact table with no aggregation. If however we do not require such detailed information, and perhaps weekly sales are sufficient, the data would be aggregated by week.

iii. Identify and explain the dimensions.

Looking at the data available from the given ER model of the existing database, and consider the business process (sales), the following dimensions are relevant for evaluating business measures related to Sales:

- Employee
- Customer
- Time
- Product

iv. 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)

**(Show only if time)** A sample star schema for the Wimmera Wines DW is shown on the final page.

## 2. Fact tables in practice

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?

   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.

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

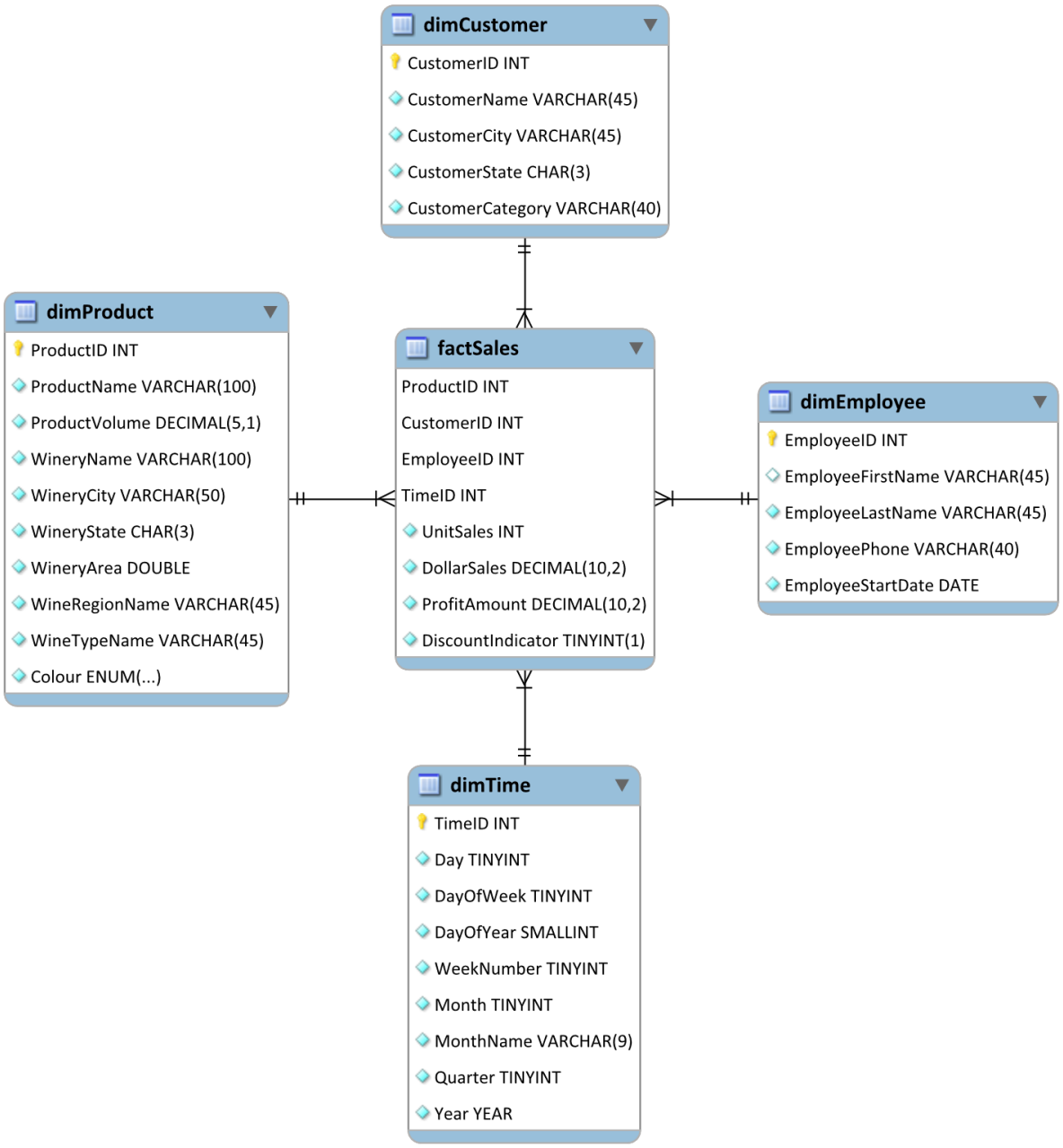
   The first sale-item of sale 54, and the sale-item of sale 67, took place on the same day and relate to the same product. These two rows will be aggregated into a single row in the fact table with Dollar Sales = $30.00 and Quantity = 3. In total, four rows will be inserted into the fact table.

   While there are more products sold on that day, note that we cannot aggregate such records because they are: i) for a different product (two sale-items of sale 54 for products 9644 and 8574), or ii) for a different region (sale 76 is for the Sydney region).

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?

   Information about the hour when a sale was made is not stored if a daily grain is used. Questions about hourly sales can only be answered when the grain is hourly (or finer).

   We can answer questions about daily sales when the grain is daily. We can also answer these questions from an hourly-grain fact table – up to 24 hourly rows can be combined (aggregated) into a single daily row when the fact table is queried, using a GROUP BY clause.

## Appendix: Question 1 star schema solution

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