INFO20003 Database Systems

Xiuge Chen

Tutorial 11
2020.06.01

1. **Notice - 2min**

2. **Understand the fundamentals of dimensional modeling - 20min**

3. **Design a dimensional model using Kimball's four-step design process - 25min**

4. **Discuss the impact of grain on fact tables - 10min**

5. **Lab - 1 hour**

1. Assignment 3 has released - LMS Assessments

2. due date: **6:00pm Friday 05 Jun**

3. Tips:

   - Follow the submission instruction and format

   - Exercises on tut 6, 7, 8 might be helpful

   - Also Resources - L14 Homework Solutions

   - Lots of calculation - double check

1. Data warehouse

- A **single database** of **organizational** data

- Allows all of the organization's data to be stored in a form that supports managers' **decision making**.

- Integrated from multiple sources internal and external to the organization by converting it into a **common format** and **validating** before storing it to ensure credibility of the warehouse.

- Keeps **historical data** and is available to the managerial bodies of the organization to support high-level decision-making processes and data analysis.

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

2. Business events

- 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

- For a marketing data warehouse, it might be a view of a webpage or a click on an online ad.

3. Dimensions, dimension tables and hierarchies

- A **dimension** is an entity that describes and gives context to a business event.

- Like time, customers, products and locations

- Example 1: 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).

- Example 2: for an insurance company, an example of a key measurement (fact) is claims, and the dimensions could be agent, policy, customer, and time

3. Dimensions, dimension tables and hierarchies

- Dimensions are represented in the data warehouse as **dimension tables**

- A sequence of **attributes** that describes a dimension across different levels of detail is called a **hierarchy**

- Example 1: 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

- Example 2: Time dimension will have a hierarchy of day, week, month, quarter and year.

- 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.

4. Facts, fact tables and granularity

- A **fact** is a **numeric measurement** of a meaningful and significant **business event**

- Example: 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

- 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.

4. Facts, fact tables and granularity

- 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.

4. Dimensional modeling – 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**
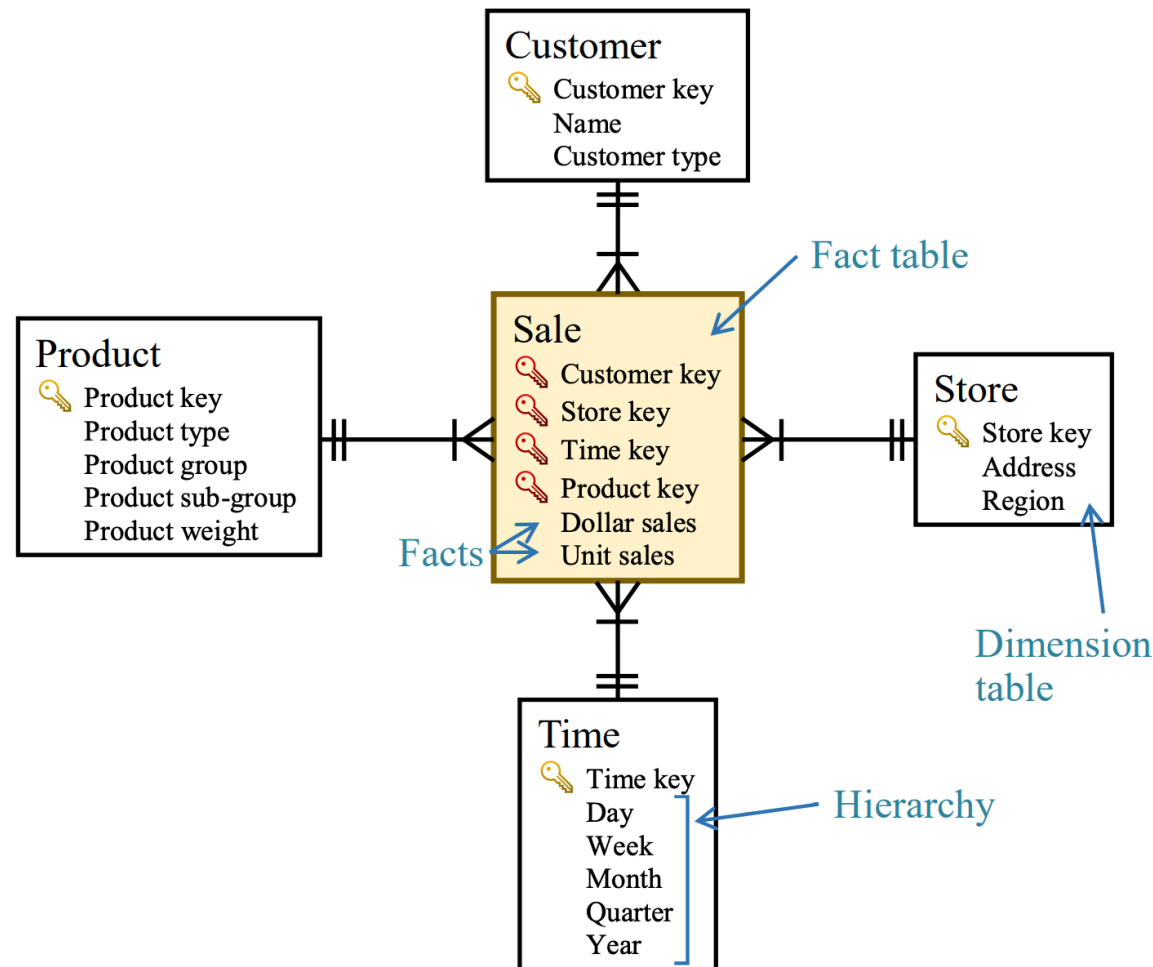
4. Dimensional modeling – the star schema

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

Any questions?

© University of Melbourne

1. Designing a dimensional model Exercise

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

Design a dimensional model

1. Designing a dimensional model Exercise
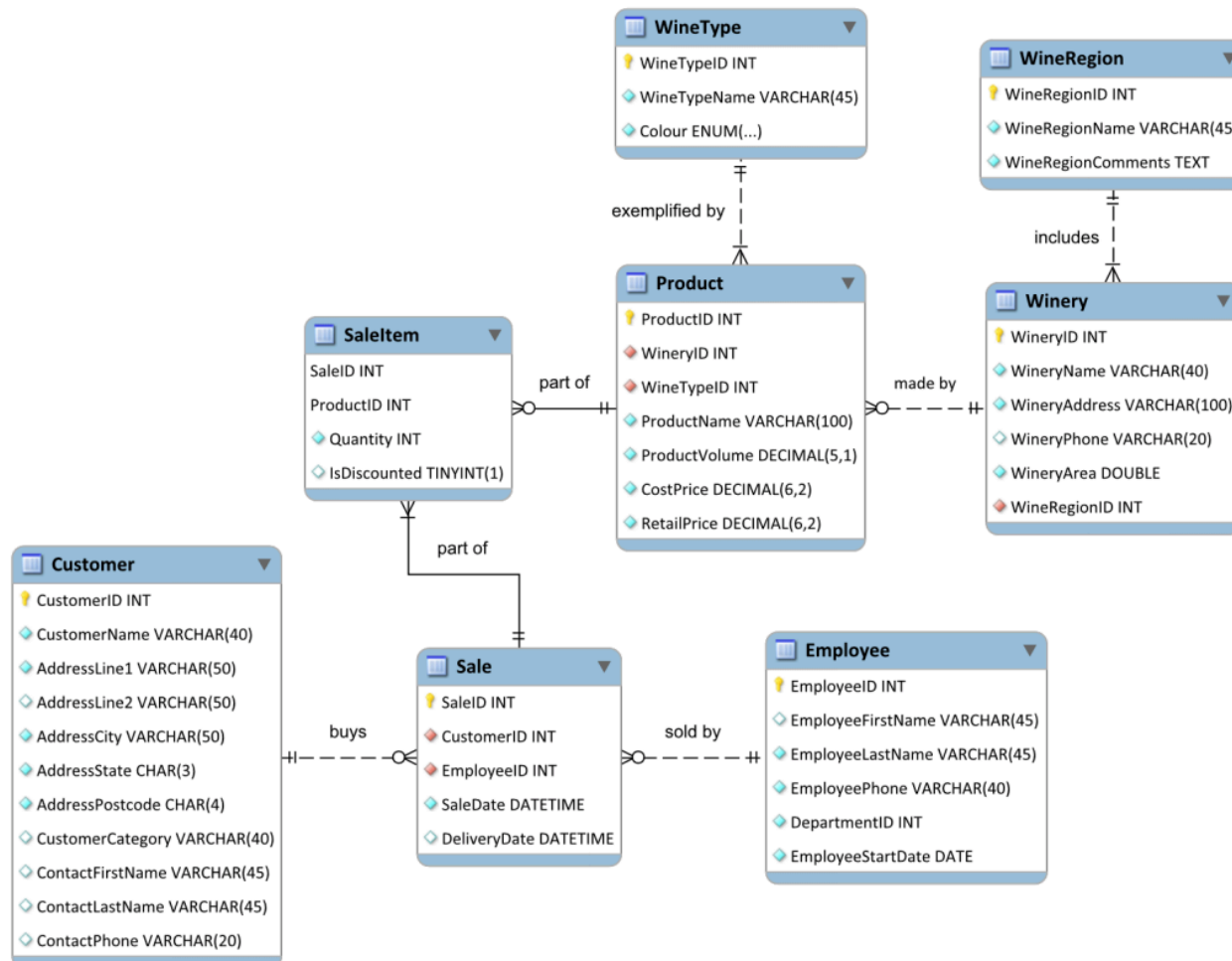
1. Designing a dimensional model Exercise

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".

1. Designing a dimensional model Exercise

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)?

1. Designing a dimensional model Exercise

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.

ii. Declare the grain and justify your choice.

iii. Identify and explain the dimensions.

iv. Identify and explain the facts.

1. Designing a dimensional model Exercise

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.

1. Designing a dimensional model Exercise

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

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.

1. Designing a dimensional model Exercise

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

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
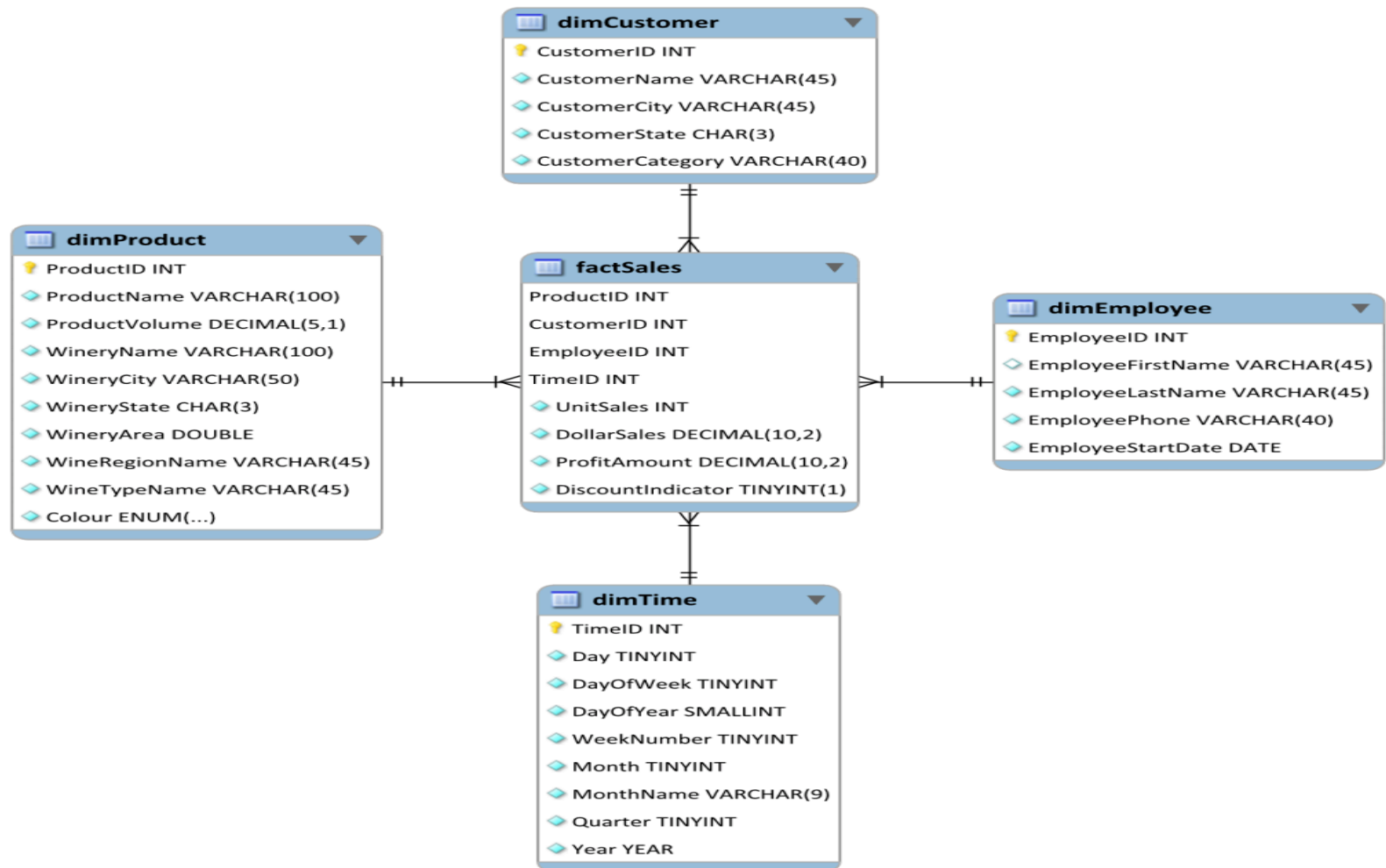- Product

1. Designing a dimensional model Exercise

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

iii. 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)

1. Designing a dimensional model Exercise

Any questions?

1. Fact tables in practice

Consider the following fact table:

Sale
- Time key
- Geography key
- Product key
- Dollar sales
- Unit sales

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

1. Fact tables in practice

Sale
- 🔑 Time key
- 🔑 Geography key
- 🔑 Product key
- Dollar sales
- Unit sales

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.

1. Fact tables in practice

Sale
- 🔑 Time key
- 🔑 Geography key
- 🔑 Product key
- Dollar sales
- Unit sales

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

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).

1. Fact tables in practice

Sale
- 🗝 Time key
- 🗝 Geography key
- 🗝 Product key
- Dollar sales
- Unit sales

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

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.

Any questions?