
COMP9318 Tutorial 1

Wei WANG

The University of New South Wales

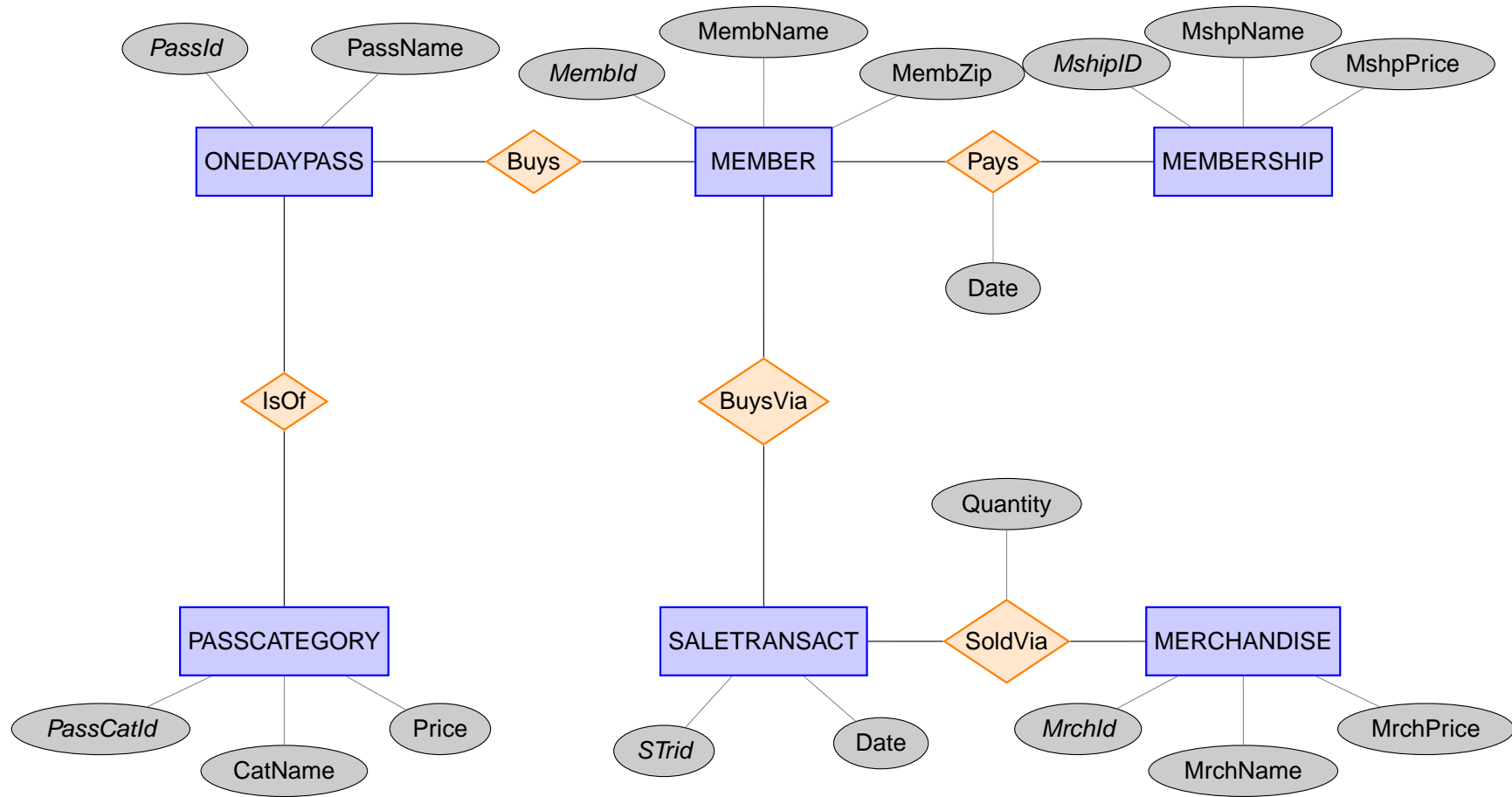
weiw@cse.unsw.edu.au

-
- ① Data Warehouse and OLAP
 - ② Data Preprocessing

Q1

- ① Create a star schema diagram that will enable FIT-WORLD GYM INC. to analyze their revenue.
 - ➔ The fact table will include — for every instance of revenue taken — attribute(s) useful for analyzing revenue.
 - ➔ The star schema will include all dimensions that can be useful for analyzing revenue
 - ➔ The only two data sources are shown below
- ② Appreciate the ETL process involved populating the data warehouse.
- ③ Appreciate the difference of formulating queries: “Find the percentage of revenue generated by members in the last year”.
- ④ How many cuboids are there in the complete data cube?

ER DIAGRAM



DATA INSTANCES

MEMBER

<i>Membid</i>	<i>MembName</i>	<i>MembZip</i>	<i>MshpID</i>	<i>MsDatePayed</i>
111	Joe	60611	M1	1-Jan-04
222	Mary	60640	M3	1-Jan-04
333	Sue	60611	M3	1-Jan-04

ONEDAYPASS

<i>PassID</i>	<i>PassDate</i>	<i>PassCatID</i>	<i>Membid</i>
1-001	1-Jan-04	PSA	111
1-002	1-Jan-04	PSA	333
1-003	2-Jan-04	PSK	333

PASSCATEGORY

<i>PassCatId</i>	<i>CatName</i>	<i>Price</i>
PSA	Adult	\$20
PSS	Senior	\$10
PSK	Kid	\$3

Note: MEMEBERS can bring in non-member guests. For each non-member guest, a member buys a one-day-guest-pass of a certain pass category.

MEMBERSHIP

<i>MshpID</i>	<i>MshpName</i>	<i>MshpPrice</i>
M1	Platinum	\$1,000
M2	Gold	\$800
M3	Value	\$300

MERCHANDISE

<i>MrchID</i>	<i>MrchName</i>	<i>MrchPrice</i>
AP1	T-shirt	\$11
AP2	Hat	\$9
EQ1	Jump Rope	\$12

SOLDVIA

<i>STrid</i>	<i>MrchID</i>	<i>Quatity</i>
11111	AP1	1
11112	AP2	1
11112	AP2	1
11113	EQ1	3

SALESTRANSACT

<i>STrid</i>	<i>Date</i>	<i>Membid</i>
11111	1-Jan-04	333
11112	2-Jan-04	222
11113	3-Jan-04	111

ANOTHER DATA SOURCE

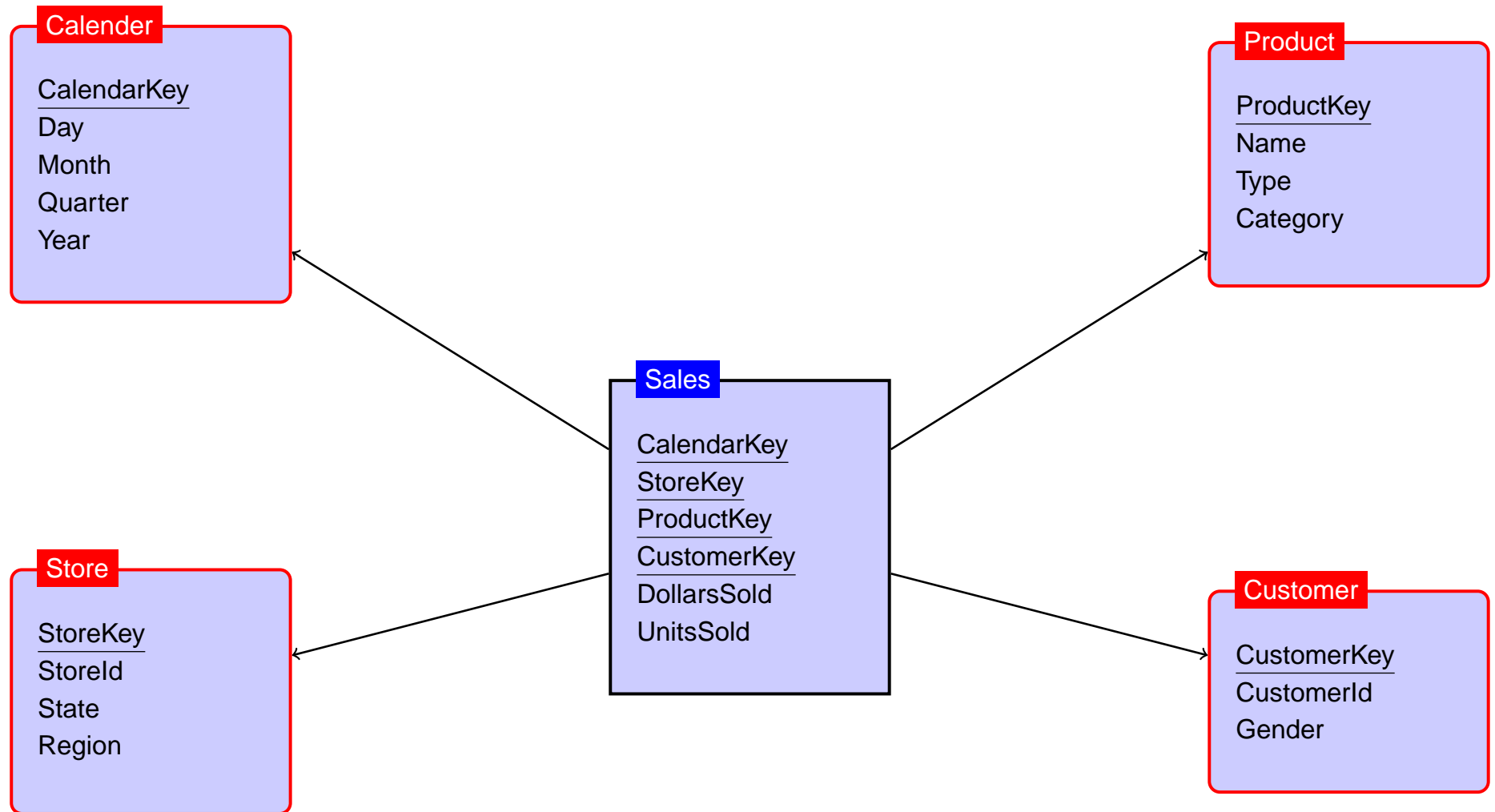
SPECIALEVENT

<i>CorpCustID</i>	<i>CorpCustNameLoc</i>	<i>EventTypeCode</i>	<i>EvetType</i>	<i>EventDate</i>	<i>AmountCharged</i>
CC1	Sears, Chicago 60640	L-A	All Day Rental,	January 4, 2004	\$3500
CC2	Boeing, Chicago 60611	L-H	Half Day Rental,	January 5, 2004	\$2200

Q2

Consider the star schema below

- Write an MDX query that display total DollarsSold for each product category and each store in the State 'CA'.
- create a star schema that has **Month** and **Region** as the finest granularity on the corresponding dimensions. Show all tables in the new data model populated with the data based on the data from the original model.



POPULATED TABLES

CALENDAR

<i>CalendarKey</i>	<i>Day</i>	<i>Month</i>	<i>Quarter</i>	<i>Year</i>
1	1	Jan	1	2003
2	2	Jan	1	2003
3	1	Feb	1	2003

STORE

<i>StoreKey</i>	<i>StoreID</i>	<i>State</i>	<i>Region</i>
1	X1	Maine	East
2	X2	New Jersey	East
3	Y1	Ohio	Midwest

SALES

<i>CalendarKey</i>	<i>ProKey</i>	<i>StoreKey</i>	<i>CustKey</i>	<i>\$Sold</i>	<i>UnitsSold</i>
1	1	1	1	\$15	1
1	2	2	2	\$20	1
1	2	1	1	\$40	2
1	2	2	1	\$20	1
2	2	1	1	\$19	1
2	2	2	1	\$19	1
3	3	3	1	\$9	2
3	3	3	2	\$9	1
3	3	3	3	\$9	1

PRODUCT

<i>ProKey</i>	<i>ProName</i>	<i>ProType</i>	<i>Category</i>
1	Luvs 50	Diapers	Infant Care
2	Huggies 24	Diapers	Infant Care
3	High C	Vitamin	Dietary Supp

CUSTOMER

<i>CustKey</i>	<i>CustID</i>	<i>Gender</i>
1	12	Male
2	23	Male
3	34	Female

Q3

Suppose that the data for analysis include the attribute A having the following values: 1, 7, 7, 19, 13. The order among the data is important and thus you cannot sort them. Number of bins is 3. Ties can be broken arbitrarily.

1. Consider the following table.

# of items in a bin	1	7	7	19	13
1	0	0	0	0	0
2	18	?	?	?	n/a
3	?	?	?	n/a	n/a

Each entry in the table records the SSE (*Sum Square Error*) if a certain number of items are stored in a bin. More formally, for an entry in the i -th column in the j -th row (both i and j start from 1), it records the SSEs if the consecutive j values starting

at column i is put into one bin. For example, every entry in the first row is zero because there is no error (i.e., $SSE = 0$) if only one value is put into a bin. The first column in the second row is 18 because this is the SSE of a bin with values 1 and 7.

Fill in the rest of the entries in the table marked with “?”.

2. Construct the V-optimal histogram (which minimizes the SSE). You need to illustrate your steps.
3. MaxDiff is a heuristic algorithm to obtain a near-optimal histogram. Show an example such that it is possible that MaxDiff algorithm will give a histogram that is **not** optimal.

Q4

Calculate the edit distance between the following pairs of strings:

1. "abcd" and "cdab"
2. "abcdef" and "cdabfe"

Q5

Consider performing the prefixed-based similarity join on the following tables R using an edit distance threshold of 1 and 2-grams (aka. bi-grams) without considering the beginning/end of the string characters (i.e., $\sim/\$$).

R	
1	report
2	repert
3	raport
4	reporter