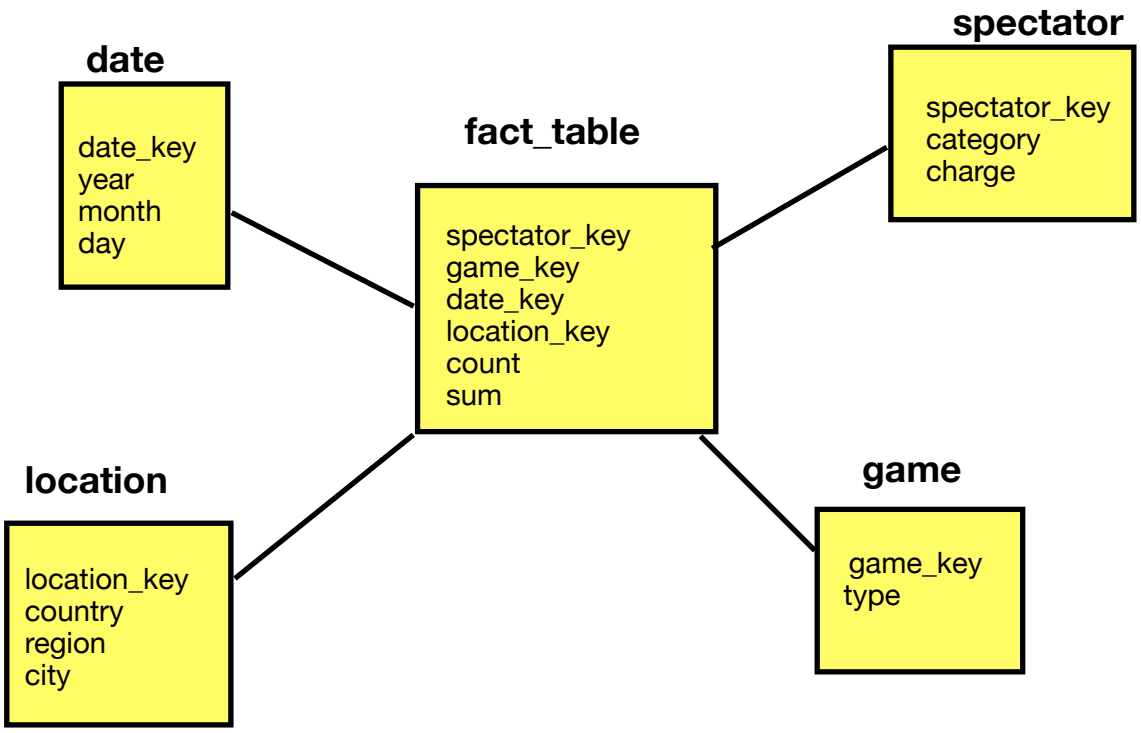
Course: Data Mining
Student: Ekaterina Eremina
Topic: HW1

Part_1: written part,
Exercise_1,

(a) Star schema diagram



(b) OLAP operations

// define diagram
define cube fact_table star [date, spectator, game, location]:
sum = sum(charge*count), count = count(*)
define dimension date as (date_key, day, month, year)
define dimension game as (game_key, type)
define dimension spectator as (spectator_key, category, charge)
define dimension location as (location_key, country, region, city)

//extract selected data
select sum(sum) as result
where (location,city = «Los Angeles», spectator,category = «student»)

(c) Bitmap indexing

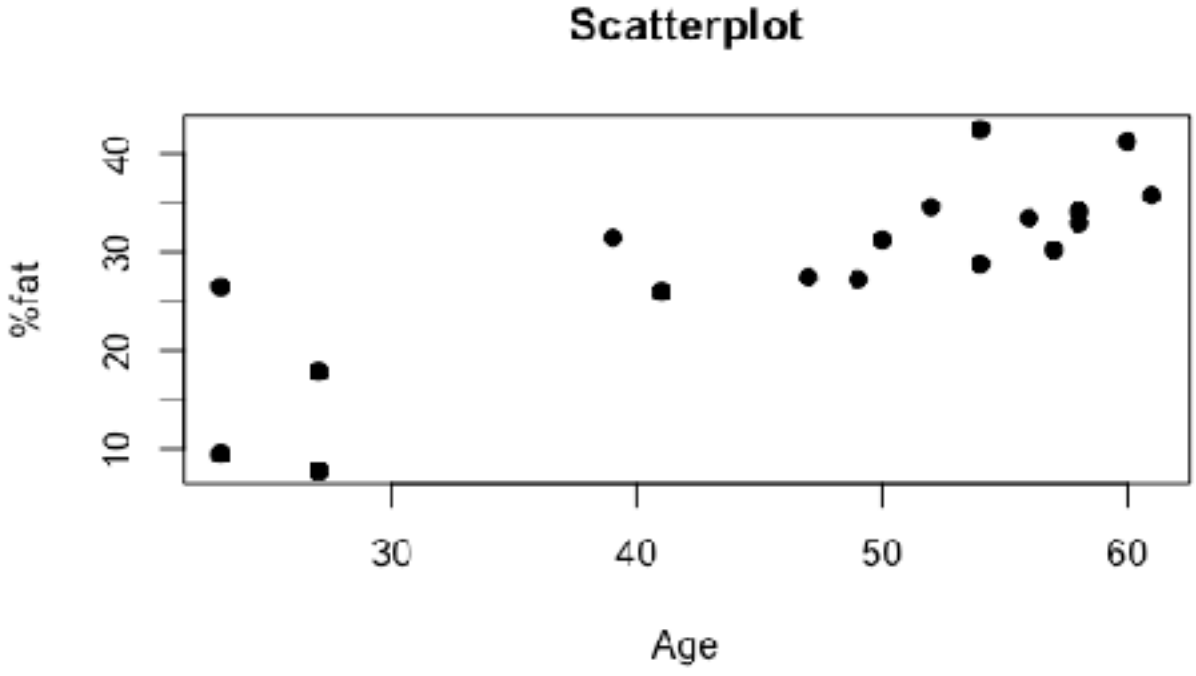
I think, that in these particular DWH there is no need to use bitmap indexing, None of the fields information can be coded in 1/0 way, For example, we can look at charge rates, If they were different for different types of games and spectators, than bitmap indexing might optimize counting,

Exercise_2,

(a) Summary statistics

	age	%fat
mean	46,44	28,78
median	51,00	30,70
standart deviation	13,21862	9,254395

(b, c) Plots



(d) Normalization

age	%fat	norm_age
23	9,5	0,0000000
23	26,5	0,0000000
27	7,8	0,1052632
27	17,8	0,1052632
39	31,4	0,4210526
41	25,9	0,4736842
47	27,4	0,6315789
49	27,2	0,6842105
50	31,2	0,7105263
52	34,6	0,7631579
54	42,5	0,8157895
54	28,8	0,8157895
56	33,4	0,8684211
57	30,2	0,8947368
58	34,1	0,9210526
58	32,9	0,9210526
60	41,2	0,9736842
61	35,7	1,000000

(e) Correlations

	X,fat	norm_age
X,fat	1,0000000	0,8176188
norm_age	0,8176188	1,0000000

Age and %fat are highly positively correlated,

(f,g) Smoothing

- by bin means, bin depth = 6

bin_1, mean = 19,12	bin_2, mean = 30,32	bin_3, mean = 36,92
7,8	27,4	33,4
9,5	28,8	34,1

17,8	30,2	34,6
25,9	31,2	35,7
26,5	31,4	41,2
27,2	32,9	42,5

- by bin boundaries, bin depth = 6

bin_1	**bin_2**	**bin_3**
7,8	31	31
7,8	31	31
19,4	31	31
19,4	31	31
19,4	31	42,6
19,4	31	42,6

Interval width = (max-min)/k = (42,5-7,8)/3 = 11,6
Bin intervals = [7,8, 19,4), [19,4, 31), [31, 42,6],

Exercise_3,

(a) Apriori

A	4	100 %
B	4	100 %
C	2	50 %
D	3	75 %
E	2	50 %

For the next step we take {A, B, D},

A, B	4	100 %
A, D	3	75 %
B, D	3	75 %
A, B, D	3	75 %

(b) Confidence

For all frequent item sets we count confidence,
It is easy to see, that this isn't symmetric measure,

A, B	If A then B	100/100 = 100%
	If B then A	100/100 = 100%

A, D	If A then D	75/100 = 75%
	if D then A	75/75 = 100%
B, D	If B then D	75/100 = 75%
	If D then B	75/75 = 100%
A, B, D	If A, B then D	75/100 = 75%
	If A, D then B	75/75 = 100%
	if B, D then A	75/75 = 100%

(c) Association Rules

buys{X, A}&buys{X, B} -> buys{X,D}
buys{X, A}&buys{X, D} -> buys{X,B}
buys{X, B}&buys{X, D} -> buys{X,A}

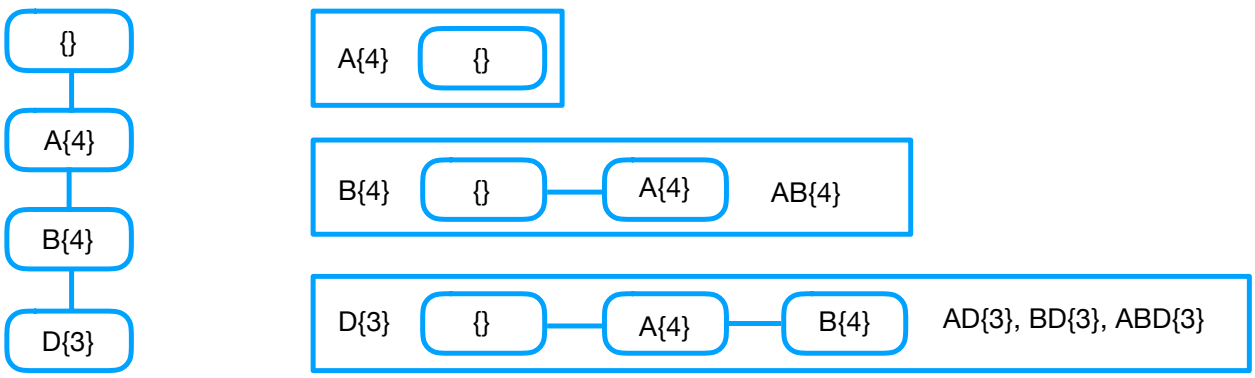
Exercise_4,

(a) FP_Growth

A	4	100 %
B	4	100 %
C	2	50 %
D	3	75 %
E	2	50 %

Frequent item sets {A, B, D}

T1	ABDC
T2	ABDCE
T3	ABE
T4	ABD



(b) Comparison

Results of Apriori and FP-Growth algorithms are the same, so let's look how many times both algorithms scan dataset,

Apriori	5*4 + 3*4 + 4 = 36
FP-Growth	5*4 + 4 = 24

So FP-Growth worked 1,5 times faster than Apriori-algoithm,

Part_2: lab part,

To find association rules for this dataset, let's construct different models, And then compare their results.

MODEL_1	**Max = 4**	**Min = 7%**	**Min = 45%**
consequent	antecedent		
milk	water and pasta	10,168	62,887
milk	coffee	14,675	58,571
milk	beer	7,338	57,143
milk	tomato souce	10,692	54,902
milk	water	24,633	54,043
MODEL_2	**Max = 4**	**Min = 10%**	**Min = 50%**
consequent	antecedent		
milk	water and pasta	10,168	62,887
milk	coffee	14,675	58,571
milk	tomato souce	10,692	54,902
milk	water	24,633	54,043
milk	pasta	37,107	51,412
MODEL_3	**Max = 4**	**Min = 5%**	**Min = 50%**
consequent	antecedent		
milk	coffee and pasta	6,289	71,667
milk	water and pasta	10,168	62,887
milk	tunny	6,813	61,538
milk	tomato souce and pasta	5,136	61,224
milk	juices	6,289	60,000
MODEL_4	**Max = 4**	**Min = 6%**	**Min = 60%**
consequent	antecedent		
milk	water and pasta	10,168	62,887
milk	tunny	6,813	61,538
milk	juices	6,289	60,000
milk	coffee and pasta	6,289	71,667

MODEL_5	Max = 4	Min = 6%	Min = 45%
consequent	antecedent		
milk	pasta	37,107	51,412
milk	water	24,633	54,043
milk	biscuits	19,182	49,727
milk	coffee	14,675	58,571
pasta	water and milk	13,312	48,031

From their comparison we can conclude, that most likely this transactions were obtained from Italy :) Because products, that consumers buy are needed for Italian food - it was a small joke. The most frequent items are {milk, water, coffee, pasta}, top-five rules will be as follows:
{milk} -> {water, pasta}
{milk} -> {coffee}
{milk} -> {tomato sause}
{milk} -> {tunny}
{milk} -> {juices}.