FINAL EXAMINATION
TDT4300

AUTUMN 2022

INFORMATION

- Academic contact during examination: Dhruv Gupta
- E-mail: dhruv.gupta@ntnu.no
- Examination date: 10-08-2022
- Examination time (from-to): 09:00-12:00
- Permitted examination support material: Open book
- Language: English

- Checked By:
- Date:
- Signature:

1 DATAWAREHOUSES AND OLAP OPERATIONS

Exercise 1. The FIDE Chess Olympiad 2022 is currently ongoing with teams from all over the world participating in India. Each game is played between two players from two different teams. FIDE's database for the Olympiad records the following particulars about each player: date, player A name, player A title, player A rating, country A (i.e., team A), player B name, player B title, player B rating, country B (i.e., team B), game location, and outcome. You are hired as a data scientist at FIDE to analyze the performance of players and teams with respect to previous tournaments. FIDE's database contains the same details for all the tournaments and previously held Olympiads.

1. Design concept hierarchies for the different attributes in FIDE's database. State any other assumption you make to model the hierarchies.
2. Based on the concept hierarchies created above, design a star schema for storing the data in a data warehouse.
3. Why would an engineer consider implementing a snowflake schema over a star schema? Would you consider implementing the data warehouse in a snowflake schema here? Why or why not?
4. At the business meeting at the upcoming FIDE World Championship, you are tasked with presenting the results of the tournaments played in 2022 with tournaments held previously. An interesting question to answer is: the number of players winning games from countries not traditionally considered as chess-playing nations (e.g., Myanmar, Japan, and Luxembourg)? In particular, identify the growth of the sport considering following attributes when comparing the results from 2022 to previous historical results:
   - The average rating of Grandmasters in an example country / team in 2022 as compared to previous years.
   - The average rating of Grandmasters in an example country / team in 2022 as compared to average rating of a player in the continent.

   For each of the queries above specify the OLAP operations required to arrive at the results.

Hint: For the OLAP operations we are expecting operations of the type "Roll Up, Drill Down, Slice, Dice, and Pivot" to arrive at the sub-cube or cross-tab to visualize the answer. For example, for the query "what are the total computer sales by Florida for quarter Q1" the answer is:
Roll Up Location: City -> State; Roll Up Time : Weeks -> Quarter;
Dice: State = "Florida" AND Quarter = "Q1";

1 1. Concept Hierarchies:
   a) Date Concept Hierarchy:
   Time $\rightarrow$ Day $\rightarrow$ Month $\rightarrow$ Year $\rightarrow$ Decade $\rightarrow$ ALL.
   b) Game City, Player A Country, and Player B Country Concept Hierarchy:
   city $\rightarrow$ state $\rightarrow$ country $\rightarrow$ continent $\rightarrow$ ALL.
   c) Player A and Player B FIDE Rating Concept Hierarchy:
   candidate master with rating $\geqslant$ 2200 (CM) $\rightarrow$ FIDE Master (FM) with rating of $\geqslant$ 2300 $\rightarrow$ International Master with rating greater than $\geqslant$ 2400 (IM) $\rightarrow$ Grandmaster $\geqslant$ 2500 (GM)

2. Star Schema: See Figure 1.

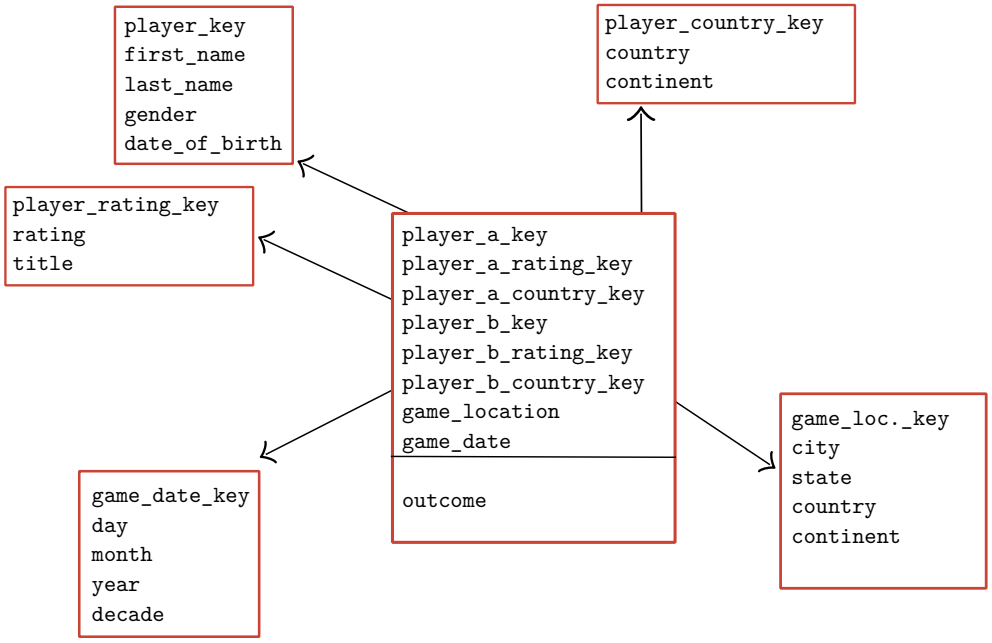


Figure 1: Star Schema.

3. A snowflake schema is beneficial here as it helps normalize a lot of the redundant information stored (e.g., both players information).

4. Following are the queries you will need to compute the trends for an example country, say, Japan:

   a) Example solution for the first query:

ROLL-UP game_date_key	:day → year.		
DICE	:game_date_key="2022"		AND
	player_A_country =	"Japan"	AND
	player_A_rating ⩾ 2500.		

ROLL-UP game_date_key	:day → year.		
DICE	:game_date_key="2022"		AND
	player_B_country =	"Japan"	AND
	player_B_rating ⩾ 2500.		

ROLL-UP game_date_key	:day → ALL.		
DICE	:player_A_country =	"Japan"	AND
	player_A_rating ⩾ 2500.		

ROLL-UP game_date_key	:day → ALL.		
DICE	:player_B_country =	"Japan"	AND
	player_B_rating ⩾ 2500.		

b) Example solution for the second query:

ROLL-UP game_date_key	:day $\rightarrow$ year.		
DICE	:game_date_key="2022"		AND
	player_A_country =	"Japan"	AND
	player_A_rating $\geqslant$ 2500.		

ROLL-UP game_date_key	:day $\rightarrow$ year.		
DICE	:game_date_key="2022"		AND
	player_B_country =	"Japan"	AND
	player_B_rating $\geqslant$ 2500.		

ROLL-UP game_date_key	:day $\rightarrow$ ALL.		
ROLL-UP player_A_country	:country $\rightarrow$ continent.		
DICE	:game_date_key="2022"		AND
	player_A_country =	"Asia"	AND
	player_A_rating $\geqslant$ 2500.		

ROLL-UP game_date_key	:day $\rightarrow$ ALL.		
ROLL-UP player_B_country	:country $\rightarrow$ continent.		
DICE	:game_date_key="2022"		AND
	player_B_country =	"Asia"	AND
	player_B_rating $\geqslant$ 2500.		

2 DATA

Exercise 2. Computation power can be recorded in various ways e.g., in terms of processing of speed of CPU, algorithmic complexity, and turnaround time of an algorithm. What is the attribute type for the following cases:

1. Processing Speed of CPU (e.g., 300 MHz, 1 GHz, 2.5 GHz ...).
2. Algorithmic Time Complexity (e.g., $\mathcal{O}(n), \mathcal{O}(n^2), \log_2(n)$ ...).
3. In an experimental setup for a study, algorithm A is implemented and its runtime is recorded upon completion. Similarly, results for baselines B_1 and B_2 are recorded. To compute the speedup with respect to the new algorithm A the following attribute is computed: $\frac{B_1}{A}$ and $\frac{B_2}{A}$. What is this type of attribute?

2 Types for the above three attributes.

1. Quantitative, numeric, and ratio.
2. Qualitative, categorical, and ordinal.
3. Quantitative, numeric, and ratio.

Exercise 3. You are working as a full stack developer at Google, you are in charge of implementing a new feature for monitoring jobs running on their compute nodes. The feature to implement is to assign a categorical attribute based on the jobs runtime. A job on a compute node finishes roughly around in one hour and fifteen minutes. For a given job you need to assign one of the following states ALMOST_DONE, MIDWAY, and JUST_STARTED. Given the following example log data on the Ganymede compute node at Google how would you implement this feature? Based on the approach you have utilized what are runtime values assigned to the states of ALMOST_DONE, MIDWAY, and JUST_STARTED.

2 An example solution can be arrived at by utilizing one of the discretization techniques such as equal interval width. This would give the following interval values for the three states.

1. ALMOST_DONE = [1 hrs, 12 mins, 43 sec ; 55 min, 16 sec].
2. MIDWAY = [41 mins, 0 sec ; 39 min, 4 sec].
3. JUST_STARTED = [20 mins, 23 sec ; 4 sec].

Process ID	Node:Port	Start Date	Time Elapsed
01	ganymede.google.org:8042	Jan 2 14:09:50 2025	1hrs, 12mins, 43sec
02	ganymede.google.org:8042	Jan 2 14:13:24 2025	1hrs, 9mins, 8sec
03	ganymede.google.org:8042	Jan 2 14:14:28 2025	1hrs, 8mins, 4sec
04	ganymede.google.org:8042	Jan 2 14:15:49 2025	1hrs, 6mins, 43sec
05	ganymede.google.org:8042	Jan 2 14:19:05 2025	1hrs, 3mins, 28sec
06	ganymede.google.org:8042	Jan 2 14:20:32 2025	1hrs, 2mins, 1sec
07	ganymede.google.org:8042	Jan 2 14:22:28 2025	1hrs, 4sec
08	ganymede.google.org:8042	Jan 2 14:23:09 2025	59mins, 23sec
09	ganymede.google.org:8042	Jan 2 14:23:33 2025	58mins, 59sec
10	ganymede.google.org:8042	Jan 2 14:23:58 2025	58mins, 34sec
11	ganymede.google.org:8042	Jan 2 14:26:13 2025	56mins, 19sec
12	ganymede.google.org:8042	Jan 2 14:27:09 2025	55mins, 23sec
13	ganymede.google.org:8042	Jan 2 14:27:16 2025	55mins, 16sec
14	ganymede.google.org:8042	Jan 2 14:41:33 2025	41mins, 0sec
15	ganymede.google.org:8042	Jan 2 14:41:54 2025	40mins, 39sec
16	ganymede.google.org:8042	Jan 2 14:42:24 2025	40mins, 9sec
17	ganymede.google.org:8042	Jan 2 14:43:19 2025	39mins, 14sec
18	ganymede.google.org:8042	Jan 2 14:43:28 2025	39mins, 4sec
19	ganymede.google.org:8042	Jan 2 14:43:28 2025	39mins, 4sec
20	ganymede.google.org:8042	Jan 2 15:02:10 2025	20mins, 23sec
21	ganymede.google.org:8042	Jan 2 15:02:17 2025	20mins, 16sec
22	ganymede.google.org:8042	Jan 2 15:02:30 2025	20mins, 3sec
23	ganymede.google.org:8042	Jan 2 15:03:09 2025	19mins, 24sec
24	ganymede.google.org:8042	Jan 2 15:03:31 2025	19mins, 2sec
25	ganymede.google.org:8042	Jan 2 15:04:27 2025	18mins, 5sec
26	ganymede.google.org:8042	Jan 2 15:05:57 2025	16mins, 35sec
27	ganymede.google.org:8042	Jan 2 15:06:03 2025	16mins, 29sec
28	ganymede.google.org:8042	Jan 2 15:06:13 2025	16mins, 19sec
29	ganymede.google.org:8042	Jan 2 15:12:10 2025	10mins, 23sec
30	ganymede.google.org:8042	Jan 2 15:12:17 2025	10mins, 16sec
31	ganymede.google.org:8042	Jan 2 15:14:45 2025	7mins, 47sec
32	ganymede.google.org:8042	Jan 2 15:15:39 2025	6mins, 53sec
33	ganymede.google.org:8042	Jan 2 15:15:44 2025	6mins, 48sec
34	ganymede.google.org:8042	Jan 2 15:16:36 2025	5mins, 56sec
35	ganymede.google.org:8042	Jan 2 15:18:15 2025	4mins, 17sec
36	ganymede.google.org:8042	Jan 2 15:19:00 2025	3mins, 32sec
37	ganymede.google.org:8042	Jan 2 15:19:43 2025	2mins, 49sec
38	ganymede.google.org:8042	Jan 2 15:20:42 2025	1mins, 50sec
39	ganymede.google.org:8042	Jan 2 15:20:59 2025	1mins, 34sec
40	ganymede.google.org:8042	Jan 2 15:22:29 2025	4sec

Table 1: Example Log Data.

3 ASSOCIATION RULE ANALYSIS

Exercise 4. Compute the frequent itemsets for the transaction database given in Table 2 using the Apriori algorithm with minimum support equal to 50%. Also find all the association rules that can be generated with minimum confidence equal to 75%.

tid	itemset
t_1	ACD
t_2	BCE
t_3	ABCE
t_4	BDE
t_5	ABCE
t_6	ABCD

Table 2: Transaction database.

3 Solution for the Apriori algorithm.

1.

C_1	**Support**
A	4
B	5
C	5
D	3
E	4

L_1	**Support**
A	4
B	5
C	5
D	3
E	4

2.

C_2	**Support**
AB	3
AC	3
AD	2
AE	2
BC	4
BD	2
BE	3
CD	2
CE	3
DE	1

C_2	**Support**
AB	3
AC	3
BC	4
BE	4
CE	3

3.

C_2	**Support**
ABC	3
BCE	3

C_2	**Support**
ABC	3
BCE	3

4.

C_2	**Support**
ABCE	2

C_2	**Support**
$\varnothing$	-

Generation of the association rules.

1. Frequent itemset: $ABC \rightarrow \{AB, AC, BC\}$.

Association Rule	**Support**
$AB \rightarrow C$	1.00
$AC \rightarrow B$	1.00
$BC \rightarrow A$	0.75

Association Rule	**Support**	**Type**
$AB \rightarrow C$	1.00	Strong
$AC \rightarrow B$	1.00	Strong
$BC \rightarrow A$	0.75	Strong

2. Frequent itemset: $BCE \rightarrow \{BC, CE, BE\}$.

Association Rule	**Support**
$BC \rightarrow C$	0.75
$CE \rightarrow B$	1.00
$BE \rightarrow C$	1.00

Association Rule	**Support**	**Type**
$BC \rightarrow C$	0.75	Strong
$CE \rightarrow B$	1.00	Strong
$BE \rightarrow C$	1.00	Strong

3. Frequent itemset: $AB \rightarrow \{A, B\}$.

Association Rule	Support
$A \rightarrow B$	0.75
$B \rightarrow A$	0.60

Association Rule	Support	Type
$A \rightarrow B$	0.75	Strong
$B \rightarrow A$	0.60	Weak

4. Frequent itemset: $AC \rightarrow \{A, C\}$.

Association Rule	Support
$A \rightarrow C$	0.75
$C \rightarrow A$	0.60

Association Rule	Support	Type
$A \rightarrow C$	0.75	Strong
$C \rightarrow A$	0.60	Weak

5. Frequent itemset: $BC \rightarrow \{B, C\}$.

Association Rule	Support
$B \rightarrow C$	0.80
$C \rightarrow B$	0.80

Association Rule	Support	Type
$B \rightarrow C$	0.80	Strong
$C \rightarrow B$	0.80	Strong

6. Frequent itemset: $BE \rightarrow \{B, E\}$.

Association Rule	Support
$B \rightarrow E$	0.60
$E \rightarrow B$	0.75

Association Rule	Support	Type
$B \rightarrow E$	0.60	Weak
$E \rightarrow B$	0.75	Strong

7. Frequent itemset: $CE \rightarrow \{C, E\}$.

Association Rule	Support
$C \rightarrow E$	0.60
$E \rightarrow C$	0.75

Association Rule	Support	Type
$C \rightarrow E$	0.60	Weak
$E \rightarrow C$	0.75	Strong

4 CLUSTERING

Exercise 5. Apply single-link hierarchical agglomerative clustering for the dataset given in Table 4. The distance between the two data points can be computed by counting the number of correspondences between their features. To record these, a contingency matrix can be computed (as shown in Table 3) by computing the matches (i.e., n_{11} and n_{00}) and mismatches (i.e., n_{10} and n_{01}). To compute the clusters use the Simple Matching Coefficient given in Equation 1 as the distance function. Provide the answer in the form of a dendogram as well as show the full distance matrix at each step. You may break ties arbitrarily. Terminate the clustering process when you have 4 clusters.

$$\text{Simple Matching Coefficient} = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}. \tag{1}$$

		x_j	
		1	0
x_i	1	n_{11}	n_{10}
	0	n_{01}	n_{00}

Table 3: Contingency Matrix.

Point / Features	f_1	f_2	f_3	f_4	f_5
a	1	0	1	1	0
b	1	1	0	1	0
c	0	0	1	1	0
d	0	1	0	1	0
e	1	0	1	0	1
f	0	1	1	0	0
g	0	1	0	0	1
h	0	0	1	0	1

Table 4: Data for hierarchical agglomerative clustering.

4 1. Iteration 1

	a	b	c	d	e	f	g	h
a	1.0	0.6	0.8	0.4	0.6	0.4	0.0	0.4
b	0.6	1.0	0.4	0.8	0.2	0.4	0.4	0.0
c	0.8	0.4	1.0	0.6	0.4	0.6	0.2	0.6
d	0.4	0.8	0.6	1.0	0.0	0.6	0.6	0.2
e	0.6	0.2	0.4	0.0	1.0	0.4	0.4	0.8
f	0.4	0.4	0.6	0.6	0.4	1.0	0.6	0.6
g	0.0	0.4	0.2	0.6	0.4	0.6	1.0	0.6
h	0.4	0.0	0.6	0.2	0.8	0.6	0.6	1.0

2. Iteration 2: Merge a and c

	b	{a, c}	d	e	f	g	h
b							
{a, c}	0.6	1.0					
d		0.6					
e		0.6					
f		0.6					
g		0.2					
h		0.6					

3. Iteration 3: Merge d and b

	{d, b}	{a, c}	e	f	g	h
{d, b}	1.0					
{a, c}	0.6					
e	0.2					
f	0.6					
g	0.6					
h	0.2					

4. Iteration 4: Merge e and h

	{d, b}	{a, c}	{e, h}	f	g
{d, b}					
{a, c}					
{e, h}	0.2	0.6	1.0		
f			0.6		
g			0.6		

5. Iteration 5: Merge f and g

	{d, b}	{a, c}	{e, h}	{e, h}
{d, b}				
{a, c}				
{e, h}				
{e, h}	0.6	0.6	0.6	1.0

The dendogram is shown in the figure below:

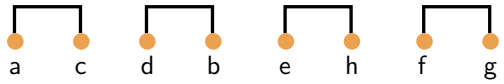


Figure 2: Dendogram.

5 CLASSIFICATION

Exercise 6. Construct a decision tree using the Hunt's Algorithm for the dataset given in Table 5 where the attribute **Class** is the classification label for each record. Use the Gini index for determining the best split points.

Instance	a_1	a_2	a_3	a_4	**Class**
1	F	X	A	P	YES
2	F	X	A	R	YES
3	M	X	C	R	YES
4	F	X	C	R	YES
5	F	X	C	S	NO
6	F	Y	D	S	NO
7	M	Y	A	R	NO
8	F	Y	A	Q	NO

Table 5: Table for decision tree based exercise.

5 1. Determination of root node split.

• Gini at root node:

Class=YES	4
Class=NO	4
Gini Index	0.50

$$\text{Gini} = 1 - \frac{4^2}{8^2} - \frac{4^2}{8^2} = \frac{1}{2} = 0.50.$$

– Split a_1:

$a_1 = M$	
YES	1
NO	1

$a_1 = F$	
YES	3
NO	3

$$\text{Gini} = 1 - \frac{1^2}{2^2} - \frac{1^2}{2^2} = \frac{1}{2} = 0.50.$$

$$\text{Gini} = 1 - \frac{3^2}{6^2} - \frac{3^2}{6^2} = \frac{1}{2} = 0.50.$$

$$\begin{aligned}\text{Gini} &= \frac{2}{8} \cdot \frac{1}{2} + \frac{6}{8} \cdot \frac{1}{2} = \frac{1}{2} \\ &= 0.50.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= 0.50 - 0.50 \\ &= 0.00.\end{aligned}$$

– Split a_2:

$a_2 = X$	
YES	4
NO	1
$a_2 = Y$	
YES	0
NO	3

$$\text{Gini} = 1 - \frac{4^2}{5^2} - \frac{1^2}{5^2} = \frac{8}{25} = 0.32.$$

$$\text{Gini} = 1 - \frac{0^2}{3^2} - \frac{3^2}{3^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{5}{8} \cdot \frac{8}{25} + \frac{3}{8} \cdot 0 = \frac{1}{5} \\ &= 0.20.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= 0.50 - 0.20 \\ &= 0.30.\end{aligned}$$

– Split a_3:

$a_3 = A$	
YES	2
NO	2
$a_3 = C$	
YES	2
NO	1
$a_3 = D$	
YES	0
NO	1

$$\text{Gini} = 1 - \frac{2^2}{4^2} - \frac{2^2}{4^2} = \frac{1}{2} = 0.50.$$

$$\text{Gini} = 1 - \frac{2^2}{3^2} - \frac{1^2}{3^2} = \frac{4}{9}.$$

$$\text{Gini} = 1 - \frac{1^2}{1^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{4}{8} \cdot \frac{1}{2} + \frac{3}{8} \cdot \frac{4}{9} + \frac{1}{8} \cdot 0 = \frac{10}{24} \\ &= 0.41\bar{6}.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= 0.50 - 0.41\bar{6} \\ &= 0.8\bar{3}.\end{aligned}$$

– Split a_4:

$a_4 = P$	
YES	1
NO	0
$a_4 = R$	
YES	3
NO	1
$a_4 = S$	
YES	0
NO	1
$a_4 = Q$	
YES	0
NO	1

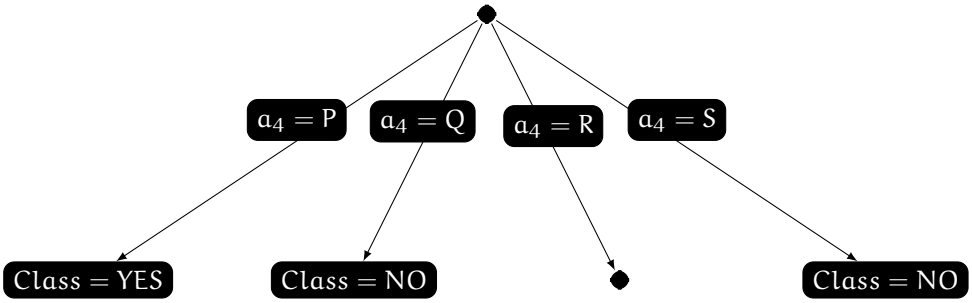
$$\text{Gini} = 1 - \frac{1^2}{1^2} - 0 = 0.00.$$

$$\text{Gini} = 1 - \frac{3^2}{4^2} - \frac{1^2}{4^2} = \frac{3}{8}.$$

$$\text{Gini} = 1 - \frac{1^2}{1^2} = 0.00.$$

$$\text{Gini} = 1 - \frac{1^2}{1^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{1}{8} \cdot 0 + \frac{4}{8} \cdot \frac{3}{8} + \frac{1}{8} \cdot 0 + \frac{1}{8} \cdot 0 = \frac{3}{16} \\ &= 0.1875.\end{aligned}$$



$$\begin{aligned}\text{Gain} &= 0.50 - 0.3125 \\ &= 0.3125.\end{aligned}$$

Split on $a_4 \in \{P, Q, R, S\}$.
• Split on $a_4 = R$:
Gini at root node:

$a_4 = R$	
YES	3
NO	1

$$\text{Gini} = 1 - \frac{3^2}{4^2} - \frac{1^2}{4^2} = \frac{3}{8} = 0.375.$$

– Split a_1:

$a_1 = M$	
YES	1
NO	1

$$\text{Gini} = 1 - \frac{1^2}{2^2} - \frac{1^2}{2^2} = \frac{1}{2} = 0.50.$$

$a_1 = F$	
YES	2
NO	0

$$\text{Gini} = 1 - \frac{2^2}{2^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{2}{4} \cdot \frac{1}{2} + \frac{2}{4} \cdot 0 = \frac{1}{4} \\ &= 0.25.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= \frac{3}{8} - \frac{1}{4} = \frac{1}{2} = 0.125 \\ &= 0.00.\end{aligned}$$

– Split a_2:

$a_2 = X$	
YES	3
NO	0

$$\text{Gini} = 1 - \frac{3^2}{3^2} - 0 = 0.00.$$

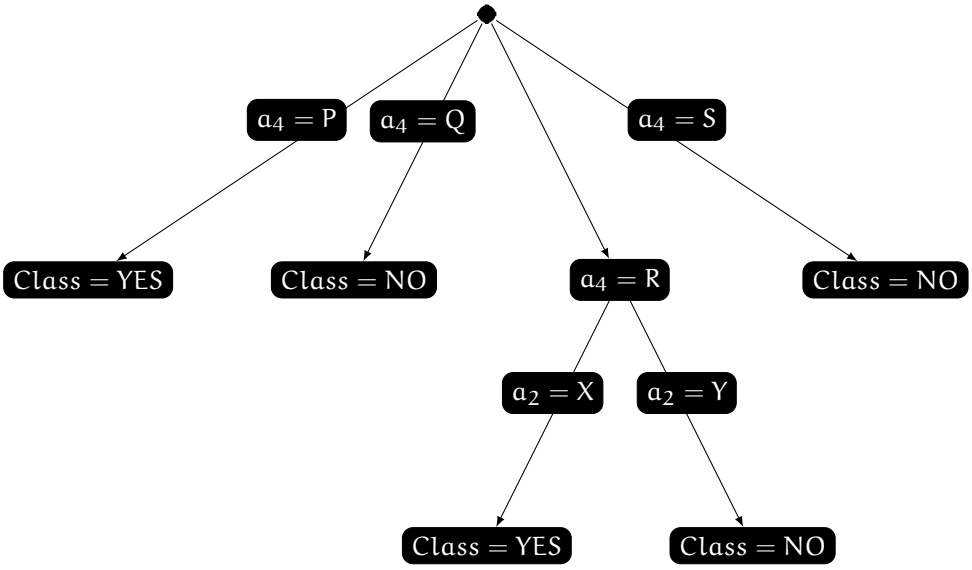
$a_1 = Y$	
YES	0
NO	1

$$\text{Gini} = 1 - \frac{1^2}{1^2} = 0.00.$$

$$\begin{aligned}\text{Gini} &= \frac{3}{4} \cdot 0.00 + \frac{1}{4} \cdot 0 = 0.00 \\ &= 0.50.\end{aligned}$$

$$\begin{aligned}\text{Gain} &= \frac{3}{8} - 0.00 = \frac{3}{8} = 0.375 \\ &= 3.75.\end{aligned}$$

Split on a_2.


$a_4 = P$
$a_4 = Q$
$a_4 = S$
Class = YES
Class = NO
$a_4 = R$
Class = NO
$a_2 = X$
$a_2 = Y$
Class = YES
Class = NO