Data Science and Database Technology

Exam 2023-06-16

Iniziato	venerdì, 16 giugno 2023, 08:10
Terminato	venerdì, 16 giugno 2023, 09:45
Tempo impiegato	1 ora 34 min.
Valutazione	**18,97** su un massimo di 32,00 (**59**%)

Domanda 1

Completo

Punteggio ottenuto 2,50 su 3,00

Extended SQL query (3 points)

The following data warehouse describes the performance of certification courses delivered by an international certification company. Certification courses are delivered in certification offices, which are distributed around the world. The offices are also characterized by the number of classrooms and teachers available. The metrics to be analyzed are the total revenue, the number of delivered hours, the number of enrolled students, and the number of students who passed the certification exam at the end of the course (NumberPassed).

The data warehouse stores the main characteristics of courses and related certifications in terms of topic, delivery mode (attendance, online, blended), type and level (basic, intermediate, or advanced) of certification, and presence or absence of prerequisites (PresenceAdmissionRequirements, Boolean attribute). Enrolled students are characterized by age range (<=20, >21 and <=30, >31 and <=45, >45), profession, and gender (M, F).

The data warehouse is characterized by the following conceptual schema and corresponding logical schema.

FEATURES-COURSE-CERTIFICATION (IDFeatCourseCert, Topic, CertificationType, CertificationLevel, PresenceAdmissionRequirements,DeliveryMode)
ENROLLED-CHARACTERISTICS (IDEnrolledChar, AgeRange, Profession, Gender)
CERTIFICATION-OFFICE (IDCertOffice, CityOffice, RegionOffice, CountryOffice, NumberClassrooms, NumberTeachers)
TEMPO (IDTime, Month, 2-Months, 3-Months, 4-Months, 6-Months, Year)
COURSES-DELIVERY (IDFeatCourseCert, IDEnrolledChar, IDCertOffice, IDTime, TotalRevenue, NumberHoursDelivered, NumberEnrolled, NumberPassed)

Considering certification offices located in Italy, separately by course delivery mode and 6-months period, display

- The total number of hours delivered
- The average number of passed students per month
- The ratio of the total number of enrolled students over the total number of enrolled students separately by year and delivery mode
- The position in a ranking (rank) in descending order with respect to the total number of hours delivered.

Conduct the analysis separately by topic.

---

select delivery mode ,6m,

sum(numberhoursdeliver),

sum(numberpassed)/count(distinct month),

sum(numberenrolled)/sum(sum(numberenrolled)) over (partition by year ,delivery mood),

rank() over (order by sum (numberhoursdelivery)desc)

from coursesdelivery c ,tempo t ,features course certificatoopn f ,

where c.idtime=t.idtime and c.idfeatcourseset=f.idfeatcourseset

and country office =italy

group by deleviry mode ,6m,year

SELECT DeliveryMode, 6-Months, Topic, Year

SUM(NumberHoursDelivered), SUM(NumberPassed)/ COUNT (DISTINCT Month),

SUM(NumberEnrolled)/SUM(SUM(NumberEnrolled)) OVER (PARTITION BY year, DeliveryMode, Topic),

RANK() OVER (PARTITION BY Topic ORDER BY SUM(NumberHoursDelivered) DESC)

FROM COURSES-DELIVERY CD, TIME T, ENROLLED-CHARACTERISTICS EC, FEATURES-COURSE-CERTIFICATION FC, CERTIFICATION-OFFICE CO

WHERE CD.IDTime=T.IDTime AND CD.IDEnrolledChar=EC.IDEnrolledChar AND CD.IDFeatCourseCert =FC.IDFeatCourseCert

AND CD.IDCertOffice =CO.IDCertOffice AND CountryOffice = “Italy”

GROUP BY DeliveryMode, 6-Months, Topic, Year

Commento:

select delivery mode ,6m,

sum(numberhoursdeliver),

sum(numberpassed)/count(distinct month),

sum(numberenrolled)/sum(sum(numberenrolled)) over (partition by year ,delivery mood **, Topic**),

rank() over (**PARTITION By Topic** order by sum (numberhoursdelivery)desc)

from coursesdelivery c ,tempo t ,features course certificatoopn f ,

where c.idtime=t.idtime and c.idfeatcourseset=f.idfeatcourseset

and country office =italy

group by deleviry mode ,6m,year**, Topic**

Domanda 2

Risposta errata

Punteggio ottenuto -0,15 su 1,00

Conceptual schema (1 point, -15% penalty for each wrong answer)

A market analysis institute wants to analyze information collected on sales of confectionery products produced by a multinational company in recent years.

The confectionery products are made in factories located in different geographical areas. The multinational company produces confectionery products of different types (e.g., snacks, cakes, cookies), and in ways that meet the needs of different dietary regimes (e.g., vegetarian or vegan).

The multinational company sells the products through several stores distributed in various geographical areas.

To analyze the sales of products with respect to different market segments, the multinational company keeps track, through loyalty cards, of some information about the customers who purchased the products, such as customer gender, age range, and type of employment.

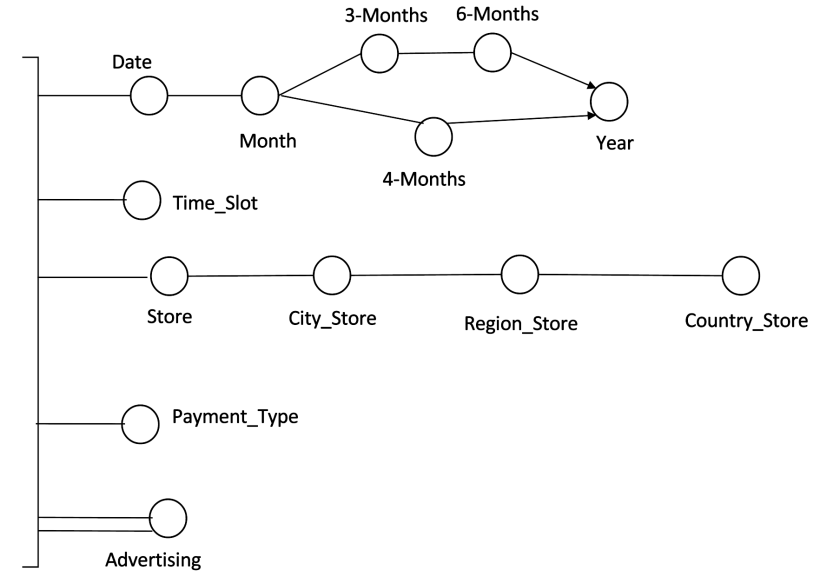
The multinational company uses various communication channels to advertise its products through radio and TV commercials or web banners.

The market analysis institute wants to analyze the average cost of advertising actions and overall sales profit based on:

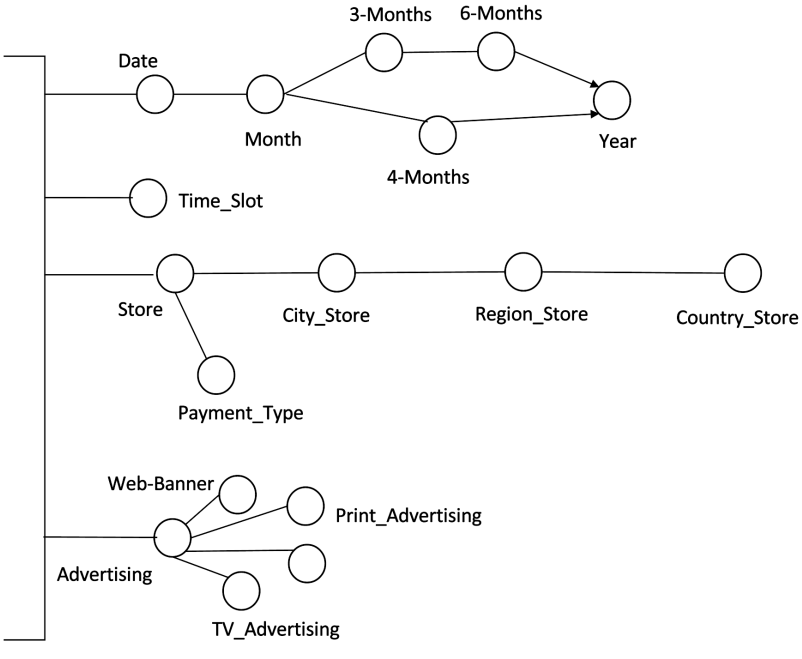
- Production factory, characterized by the information on the city, region, country where it is located, and the person managing the factory
- Product, characterized by the name, type of product (e.g., cake, snack, bar, etc.), indication of compatible diets (one or more values between vegetarian, vegan, celiac disease), ingredient list, and allergen list.
- Gender, age range, and type of employment of the customer who made the purchase. Age range is a value among < 20 years, between 20 and 29 years, between 30 and 49 years, and greater than 50 years. Employment type is a value between freelance, employee, other.
- Store where the product was purchased, characterized by the city region and country where the store is located.
- The type of payment made (a value between cash, credit card, and ATM)
- Date, month, 3.months, 4-Months, 6-months,, year in which the purchase occurred and time slot (a value between morning, afternoon, evening)
- Communication channels used to advertise the product (one or more values from the following: radio advertising, print advertising, television advertising, web banner)

Select, from the dimensions proposed below, those that meet the requirements described in the problem specifications.

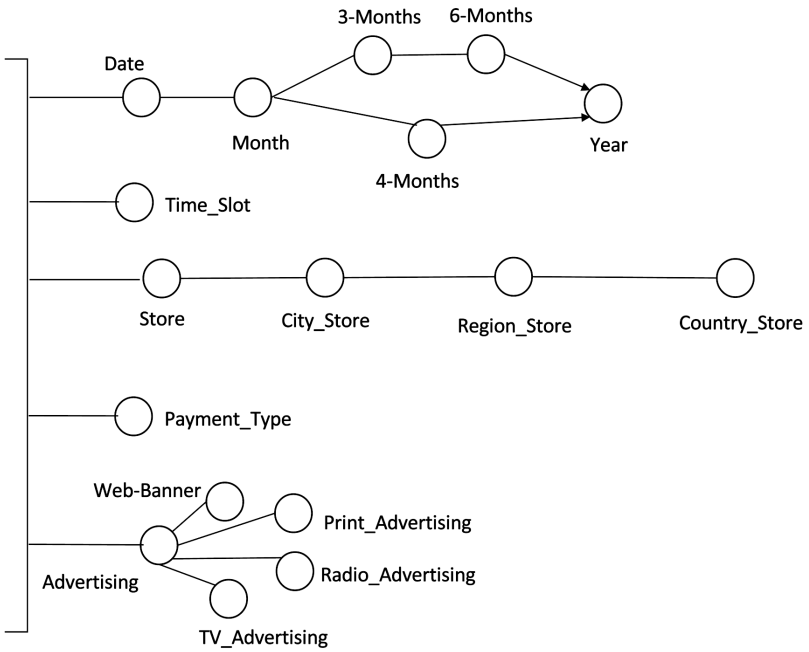
(a)


3-Months
6-Months
Date
Month
Year
4-Months
Time_Slot
Store
City_Store
Region_Store
Country_Store
Payment_Type
Advertising

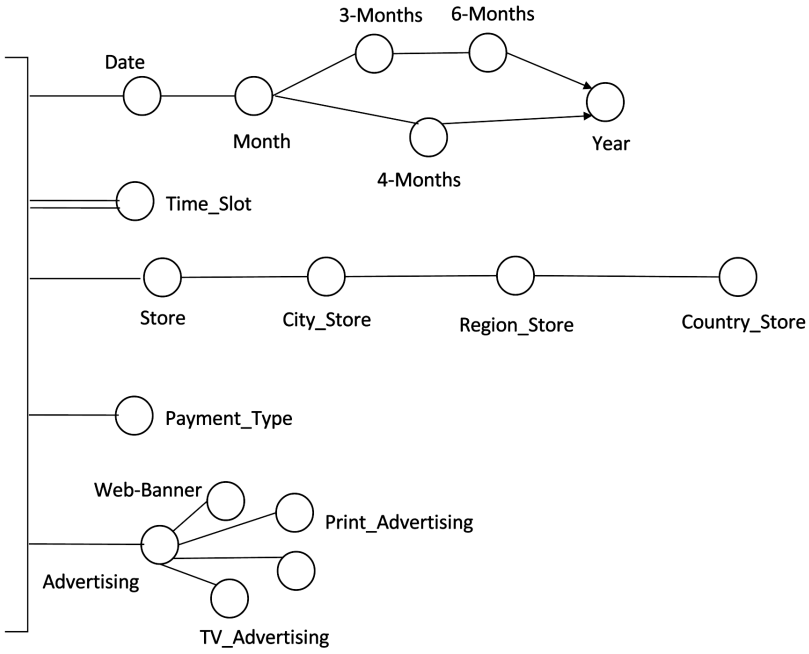
(b)


3-Months
6-Months
Date
Month
Year
4-Months
Time_Slot
Store
City_Store
Region_Store
Country_Store
Payment_Type
Web-Banner
Print_Advertising
Advertising
TV_Advertising

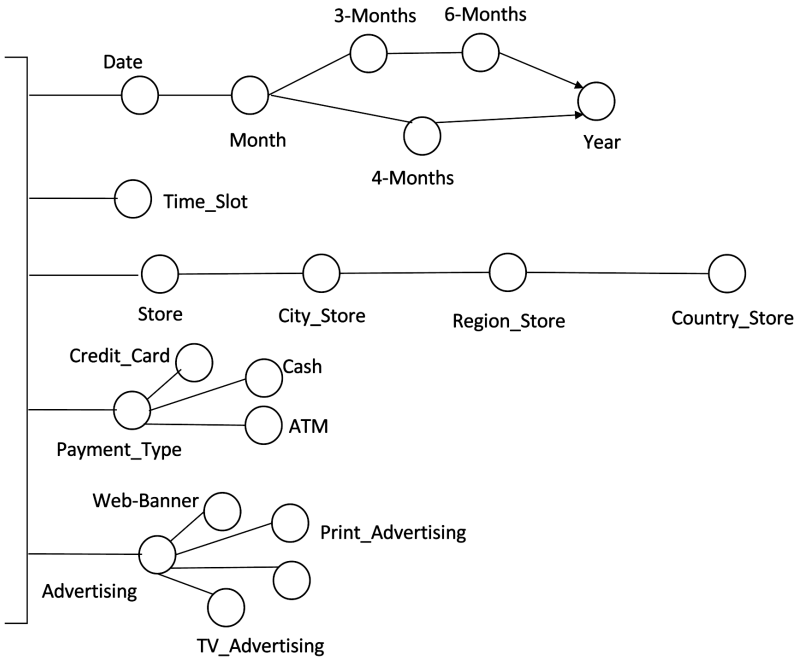
(c)


3-Months
6-Months
Date
Month
Year
4-Months
Time_Slot
Store
City_Store
Region_Store
Country_Store
Payment_Type
Web-Banner
Print_Advertising
Advertising
Radio_Advertising
TV_Advertising

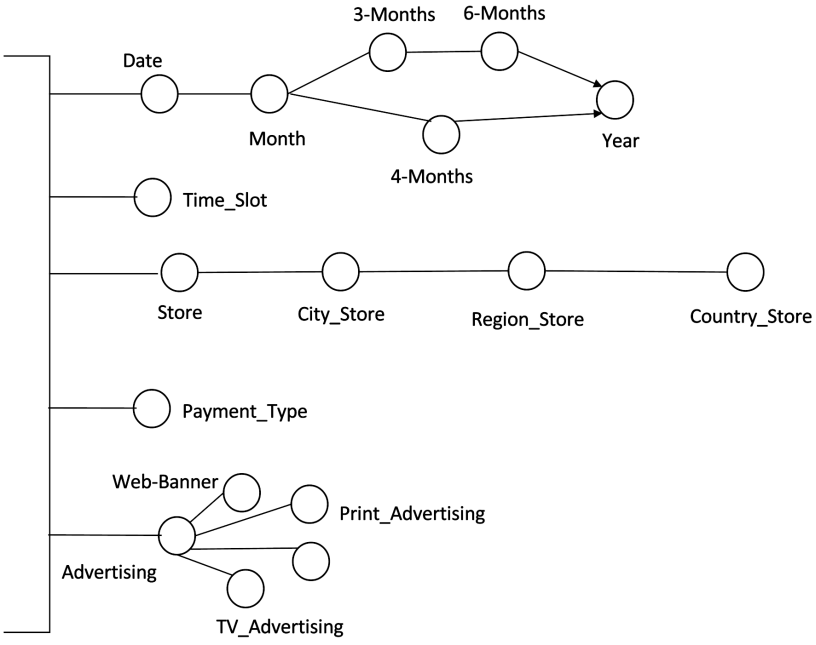
(d)

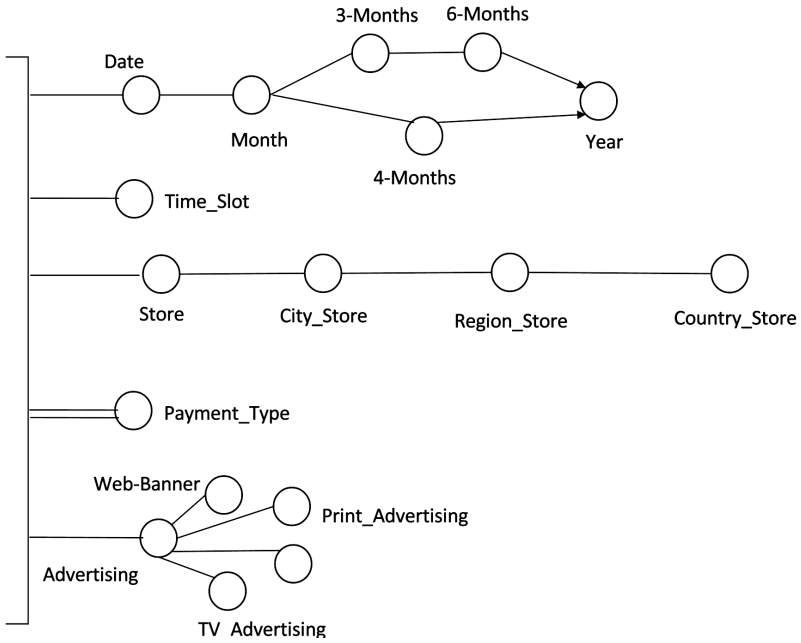

3-Months
6-Months
Date
Month
Year
4-Months
Time_Slot
Store
City_Store
Region_Store
Country_Store
Payment_Type
Web-Banner
Print_Advertising
Advertising
TV_Advertising

(e)


3-Months
6-Months
Date
Month
Year
4-Months
Time_Slot
Store
City_Store
Region_Store
Country_Store
Credit_Card
Cash
ATM
Payment_Type
Web-Banner
Print_Advertising
Advertising
TV_Advertising

(f)

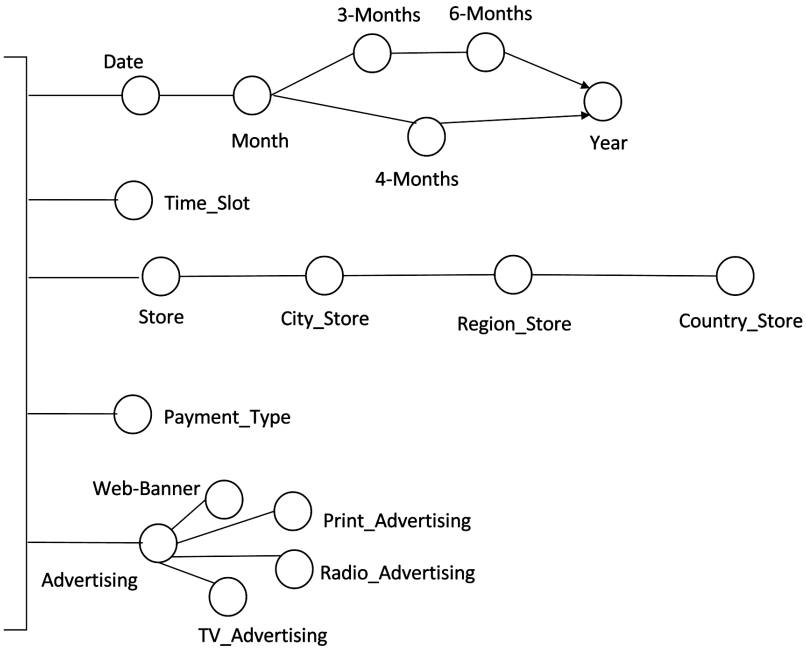

3-Months
6-Months
Date
Month
Year
4-Months
Time_Slot
Store
City_Store
Region_Store
Country_Store
Payment_Type
Web-Banner
Print_Advertising
Advertising
TV_Advertising



(g)

Risposta errata.

La risposta corretta è:



Domanda 3

Risposta corretta

Punteggio ottenuto 1,50 su 1,50

Indexes (1.5 points, -15% penalty for each wrong answer)

The following tables are provided:

STAFF(StaffId, Name, Surname, Email, Telephone, BirthDate, YearsOfExperience, SpId)
SPECIALIZATION(SpId, Name, SpecializationCategory)
FLATCOMPLEX(FCId, Name, Address, Region, Category)
INTERVENTION(Date, FCId, StaffId, Amount)

Assume the following cardinalities:

- card(STAFF) = 10^5 tuples
  - MIN(YearsOfExperience) = 1, MAX(YearsOfExperience) = 30
  - MIN(BirthDate) = 1/1/1930, MAX(BirthDate) = 31/12/1999
- card(SPECIALIZATION) = 10^2 tuples
  - distinct values of SpecializationCategory = 10
- card(FLATCOMPLEX) = 10^6 tuples
  - distinct values of Category = 10
- card(Intervention)= $2 \cdot 10^9$ tuples
  - MIN(Amount) = 100, MAX(Amount) = 100000,
  - MIN(Date) = 1/1/2003, MAX(Date) = 31/12/2022

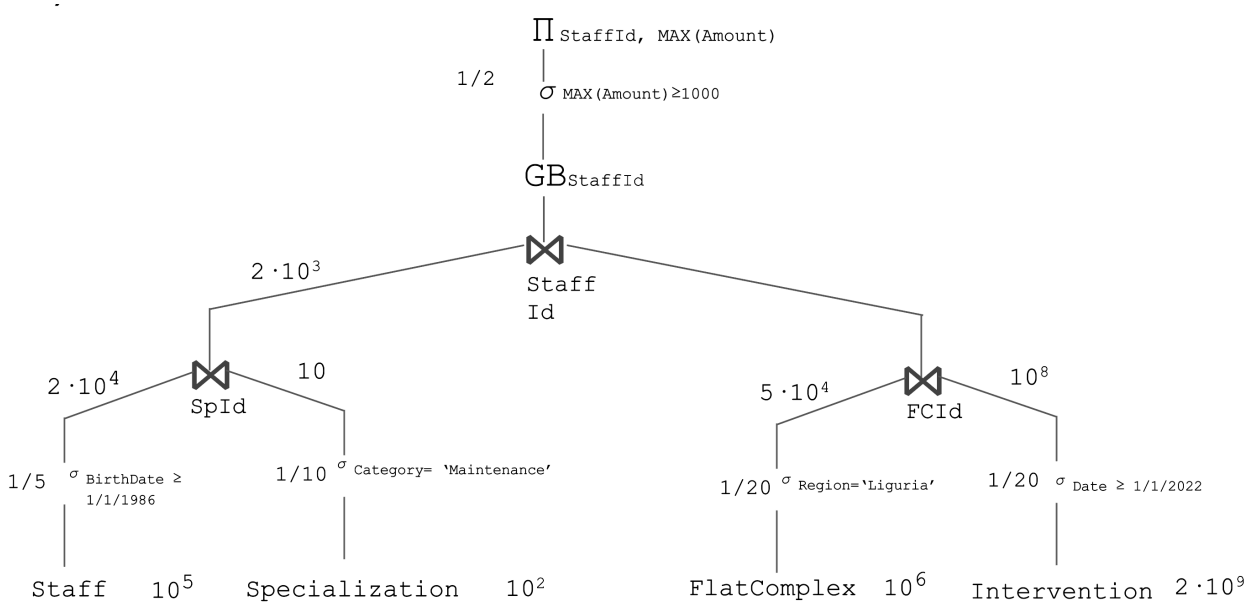
Furthermore, assume the following reduction factor for the having clauses:

Having MAX(Amount)≥1000= 1/2

Consider the following query:

```
SELECT I.StaffId, MAX(Amount)
FROM Intervention I, STAFF ST, SPECIALIZATION S, FLATCOMPLEX FC
WHERE I.StaffId=ST.StaffId and ST.SpId=S.SpId and I.FCId=FC.FCId
and Category='Maintenance' and FC.Region='Liguria' and Date ≥ 1/1/2022
and BirthDate ≥ 1/1/1986)
GROUP BY I.StaffId
HAVING MAX(Amount)≥1000
```

The figure below represents the query tree for the query above.



Select one or more secondary physical structures to increase query performance (if possible) among the options below. You can select multiple correct answers.

Scegli una o più alternative:

- [ ] (a) CREATE INDEX IndexB ON STAFF(BirthDate) - B+-Tree
- [x] (b) CREATE INDEX IndexE ON FLATCOMPLEX(Region) - HASH ✓
- [ ] (c) None - secondary physical structures would not increase query performance.
- [ ] (d) CREATE INDEX IndexC ON SPECIALIZATION(Category) - HASH
- [ ] (e) CREATE INDEX IndexD ON SPECIALIZATION(Category) - B+-Tree
- [ ] (f) CREATE INDEX IndexF ON FLATCOMPLEX(Region) - B+-Tree
- [ ] (g) CREATE INDEX IndexA ON STAFF(BirthDate) - HASH
- [x] (h) CREATE INDEX IndexH ON Intervention(Date) - B+-Tree ✓
- [ ] (i) CREATE INDEX IndexG ON Intervention(Date) - HASH

Risposta corretta.

La risposta corretta è: CREATE INDEX IndexE ON FLATCOMPLEX(Region) - HASH, CREATE INDEX IndexH ON Intervention(Date) - B+-Tree

Domanda 4

Completo

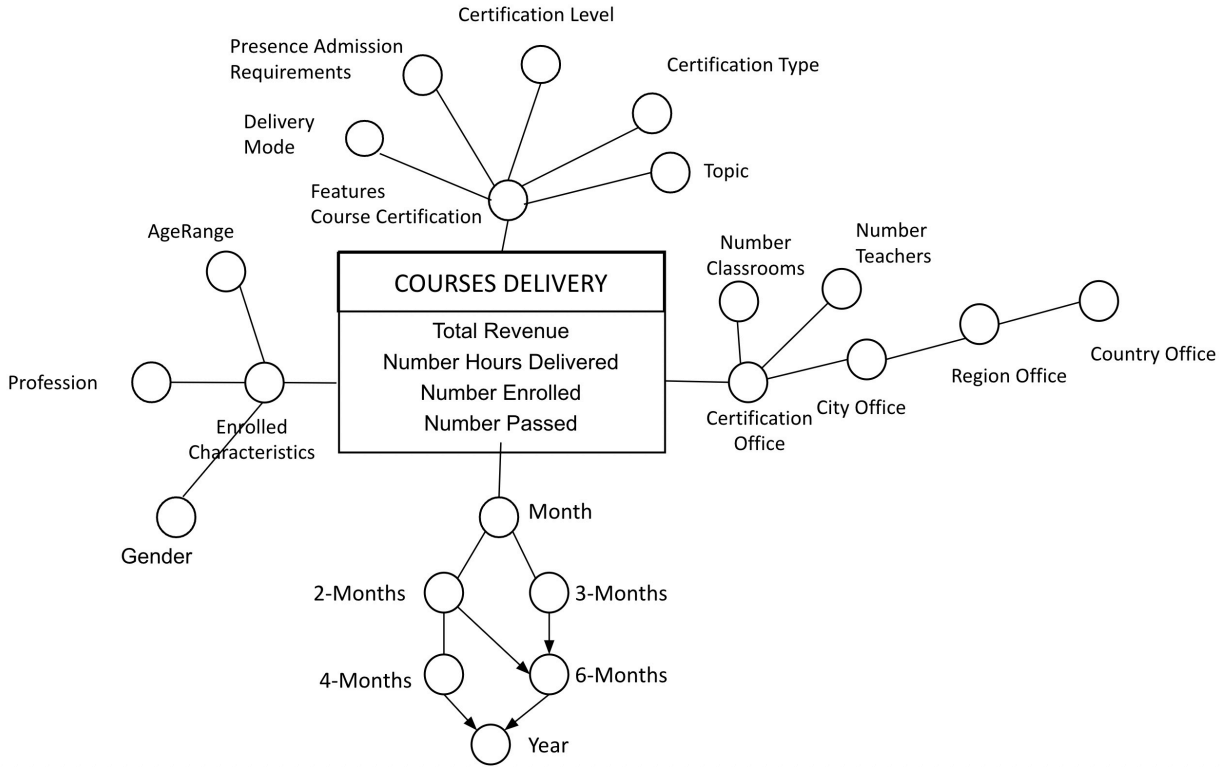
Punteggio ottenuto 3,25 su 5,00

Materialized view (5 points)

The following data warehouse describes the performance of certification courses delivered by an international certification company. Certification courses are delivered in certification offices, which are distributed around the world. The offices are also characterized by the number of classrooms and teachers available. The metrics to be analyzed are the total revenue, the number of delivered hours, the number of enrolled students, and the number of students who passed the certification exam at the end of the course (NumberPassed).

The data warehouse stores the main characteristics of courses and related certifications in terms of topic, delivery mode (attendance, online, blended), type and level (basic, intermediate, or advanced) of certification, and presence or absence of prerequisites (PresenceAdmissionRequirements, Boolean attribute). Enrolled students are characterized by age range (<=20, >21 and <=30, >31 and <=45, >45), profession, and gender (M, F).

The data warehouse is characterized by the following conceptual schema and corresponding logical schema.



FEATURES-COURSE-CERTIFICATION (IDFeatCourseCert, Topic, CertificationType, CertificationLevel, PresenceAdmissionRequirements,DeliveryMode)
ENROLLED-CHARACTERISTICS (IDEnrolledChar, AgeRange, Profession, Gender)
CERTIFICATION-OFFICE (IDCertOffice, CityOffice, RegionOffice, CountryOffice, NumberClassrooms, NumberTeachers)
TIME (IDTime, Month, 2-Months, 3-Months, 4-Months, 6-Months, Year)
COURSES-DELIVERY (IDFeatCourseCert, IDEnrolledChar, IDCertOffice, IDTime, TotalRevenue, NumberHoursDelivered, NumberEnrolled, NumberPassed)

Given the above logic schema, consider the following queries of interest:

a) Considering only male enrolled students in the age group > 45, separately by certification office region and profession, display the total number of enrolled students, the total number of passed students, and total revenue.
b) Separately by country of certification office and month, display the total monthly revenue and the cumulative annual revenue as months pass.
c) Separately by profession, considering only certification office region Piedmont, display the average monthly number of hours delivered and the average monthly revenue.

Given the above logical scheme, answer the following requests:

1. Define a materialized view with CREATE MATERIALIZED VIEW, so as to reduce the response time of the queries of interest (a) to (c) above. Specifically, specify the SQL query associated with Block A in the following statement:

```
CREATE MATERIALIZED VIEW ViewCourses
BUILD IMMEDIATE
REFRESH FAST ON COMMIT
AS
  Block A
```

2. Assume that the management of the materialized view (derived table) is carried out by means of triggers. Write the trigger to propagate to the ViewCourses materialized view the changes due to the insertion of a new record into the COURSES-DELIVERY table.

---

blook A

select gender , age group ,region office ,profession ,country office,month

sum(numberenrolled),

sum(numberpassed),

sum(total revenue),

sum(numberhours deliverd)

from coursesdelivery c ,time t ,certifaction office co ,enrolled charatristic e ,

where c.idtime=t.idtime and c.idcertoffice=co.idcertoffice and c.idenrolledchar=e.idenrolled char

group by gender ,age group ,region office ,profession ,country office ,month

TRIGGER

creat or replase trigger view courses

after inser on courses delivery

for each row

declare

var month date:

var gender, var profession , var region office ,var country office var char (10)

var age group number ,

n integer

begin

select month into var month

from time

where idtime=:new.idtime

select gender ,age group,profeesion into var gender ,var age group ,var proffession

from enrolled charactristic

where idenrolledchar=:new.idenrolledchar

select region office ,countryoffice into var region ,var country

from certifaction office

where idcertoffice=:new.idcertoffice

select count(*) into n

from veiw courses

where month =var month ,and gender=var gender and bage group =var age group and proffesion =var profession

region office=var region and country office =var country

if n >0 then

update veiw courses

set total revenue =tot revenue+:new tot revenue,

numberenrolled= numenrolled+:new numenroll,

numberpassed=num passed+:new num pass,

number hours deliver=num h deliver +:new num h deliver

where month =var month and region =vgar region and country =var country and gernder=var gender and agegroup=var age group and profession =var profession

else inser into veiw () values (var month ,var region,var country ,var gender ,var agegrop,var profession, :newtot revenue,:new numenroll,:newnumpass,:new num h deliver),

end if,

end

1. Block A

SELECT Gender, Profession, AgeRange, RegionOffice, CountryOffice, Month, Year,
SUM(TotalRevenue) AS TotRevenue, SUM(NumberHoursDelivered) AS TotHours, SUM(NumberEnrolled) AS TotEnrolled, SUM(NumberPassed) AS TotPassed
FROM CERTIFICATION-OFFICE CO, ENROLLED-CHARACTERISTICS CI, TIME T, COURSES-DELIVERY EA
WHERE CO.IDCertOffice = EA.IDCertOffice
AND CI.IDEnrolledChar = EA.IDEnrolledChar
AND T.IDTime = EA.IDTime
GROUP BY Gender, Profession, AgeRange, RegionOffice, CountryOffice, Month, Year;

2. Identifier

Gender, Profession, AgeRange, RegionOffice, Month

```
CREATE OR REPLACE TRIGGER TriggerViewCourses
AFTER INSERT ON COURSES-DELIVERY
FOR EACH ROW
DECLARE

VarY DATE,  VarM DATE;
VarRegion,  VarCountry varchar(10);
INTO VarGender, VarProfession, VarAgeRange varchar(10);
N INTEGER;
BEGIN
SELECT Month, Year INTO VarM, varY
FROM TIME
WHERE IDTime = :NEW. IDTime;

SELECT Gender, Profession, AgeRange INTO VarGender, VarProfession, VarAgeRange
FROM ENROLLED-CHARACTERISTICS
WHERE IDEnrolledChar = :NEW.IDEnrolledChar;

SELECT RegionOffice, CountryOffice INTO VarRegion,  VarCountry
FROM CERTIFICATION-OFFICE
WHERE IDCertOffice = :NEW.IDCertOffice;

SELECT COUNT(*) INTO N
FROM ViewCourses
WHERE Month = VarM AND Gender = VarGender
AND Profession= VarProfession AND AgeRange = VarAgeRange
AND RegionOffice = varRegion;

IF N>0 THEN
        UPDATE ViewCourses
        SET TotRevenue = TotRevenue + :NEW.TotalRevenue,
TotHours = TotHours + :NEW.NumberHoursDelivered,
TotEnrolled = TotEnrolled + :NEW.NumberEnrolled,
TotPassed = TotPassed + :NEW.NumberPassed
        WHERE Month = VarM AND Gender= VarGender
AND Profession= VarProfession AND AgeRange= VarAgeRange
AND RegionOffice = varRegion;
ELSE
        INSERT INTO ViewCourses (…) VALUES (VarM, VarY, VarGender, VarProfession, VarAgeRa
nge, VarRegion,  VarCountry,:NEW.TotalRevenue, :NEW.NumberHoursDelivered,
:NEW.NumberEnrolled, :NEW.NumberPassed);
END IF;
END
```

Commento:

blook A

select gender , age group ,region office ,profession ,country office,month **MISSING Year**

sum(numberenrolled) **AS TotEnrolled**,

sum(numberpassed) **AS TotPassed**,

sum(total revenue) **AS GlobalRevenue**,

sum(numberhours deliverd) **AS TotHours**

from coursesdelivery c ,time t ,certifaction office co ,enrolled charatristic e ,

where c.idtime=t.idtime and c.idcertoffice=co.idcertoffice and c.idenrolledchar=e.idenrolled char

group by gender ,age group ,region office ,profession ,country office ,month **MISSING Year**

TRIGGER

creat or replase trigger view courses

after inser on courses delivery

for each row

declare

var month date:

var gender, var profession , var region office ,var country office var char (10)

var age group number ,

n integer

begin

select month, **Year** into var month, **VarYear**

from time

where idtime=:new.idtime

select gender ,age group,profeesion into var gender ,var age group ,var proffession

from enrolled charactristic

where idenrolledchar=:new.idenrolledchar

select region office ,countryoffice into var region ,var country

from certifaction office

where idcertoffice=:new.idcertoffice

select count(*) into n

from veiw courses

where month =var month ,and gender=var gender and bage group =var age group and proffesion =var profession

region office=var region **~~and country office =var country~~**

if n >0 then

update veiw courses

set **GlobalRevenue = Global Revenue** **~~total revenue =tot revenue~~**+:new tot revenue,

TotEnrolled = TotEnrolled **~~numberenrolled= numenrolled~~**+:new numenroll,

TotPassed = TotPassed **~~numberpassed=num passed~~**+:new num pass,

TotHours =TotHours **~~number hours deliver=num h deliver~~** +:new num h deliver

where month =var month and region =vgar region **~~and country =var country~~** and gernder=var gender and agegroup=var age group and profession =var profession

else inser into veiw () values (var month ,var region,var country ,var gender ,var agegrop,var profession, **varyear,** :newtot revenue,:new numenroll,:newnumpass,:new num h deliver),

end if,

end

Domanda 5

Risposta corretta

Punteggio ottenuto 1,00 su 1,00

Recovery (1 point, -15% penalty for a wrong answer)

It is given the following sequence of operations in a log file:

B(T1) B(T2) I2(o1) Commit (T2) CK(T1) B(T3) I3(o2) I1(o1) D3(o2) Commit(T3) U1(o1) FAILURE

Notation:

- Tn = Id of transaction n
- B(Tn) = begin of Tn
- CK = checkpoint
- Un(ox) = update executed by Tn on the object ox; same notation for I (insert) and D (delete)

Which operations are performed for a warm restart?

- (a) None of the other answers are correct
- (b) UNDO = {T2}, REDO = {T3}
- (c) UNDO = {T1}, REDO = {T3} ✓
- (d) UNDO = {T1, T2}, REDO = {}
- (e) UNDO = {}, REDO = {T3}
- (f) UNDO = {T1, T2}, REDO = {T3}
- (g) UNDO = {T3}, REDO = {T1}

Risposta corretta.

La risposta corretta è: UNDO = {T1}, REDO = {T3}

Domanda 6

Risposta errata

Punteggio ottenuto -0,30 su 2,00

2 points (15% penalty for incorrect answer)

The transactional database shown below is given. Let the Apriori algorithm be applied for the extraction of frequent itemsets.

	Transactions
0	C E
1	A B C E
2	A C D
3	B C E
4	A B C
5	A E
6	C D E
7	C D E
8	A B C
9	A C

The value of minsup is 2 (an itemset is frequent if it appears in at least 2 transactions).

Which are the candidate itemsets of length 3 that are generated in the join step and then pruned in the prune step?

Alphabetical sorting is used to establish an order among the itemsets.

- (a) {}
- (b) { ABD, ACD, ADE, BCD, BDE }
- (c) { ABE, ACE }
- (d) { ABC, BCE, CDE }
- (e) { ABC, ABD, ABE, ACD, ACE, ADE, BCD, BCE, BDE, CDE }
- (f) { ABC, ABE, ACE, BCE, CDE } ✗

Risposta errata.

La risposta corretta è: {}

Domanda 7

Risposta corretta

Punteggio ottenuto 2,00 su 2,00

Group by anticipation (2 points, -15% penalty for a wrong answer)
The following tables are provided:

STAFF(StaffId, Name, Surname, Email, Telephone, BirthDate, YearsOfExperience, SpId)
SPECIALIZATION(SpId, Name, SpecializationCategory)
FLATCOMPLEX(FCId, Name, Address, Region, Category)
INTERVENTION(Date, FCId, StaffId, Amount)

Assume the following cardinalities:

- card(STAFF) = 10^5 tuples
  - MIN(YearsOfExperience) = 1, MAX(YearsOfExperience) = 30
  - MIN(BirthDate) = 1/1/1930, MAX(BirthDate) = 31/12/1999
- card(SPECIALIZATION) = 10^2 tuples
  - distinct values of SpecializationCategory = 10
- card(FLATCOMPLEX) = 10^6 tuples
  - distinct values of Category = 10
- card(Intervention)= $2 \cdot 10^9$ tuples
  - MIN(Amount) = 100, MAX(Amount) = 100000,
  - MIN(Date) = 1/1/2003, MAX(Date) = 31/12/2022

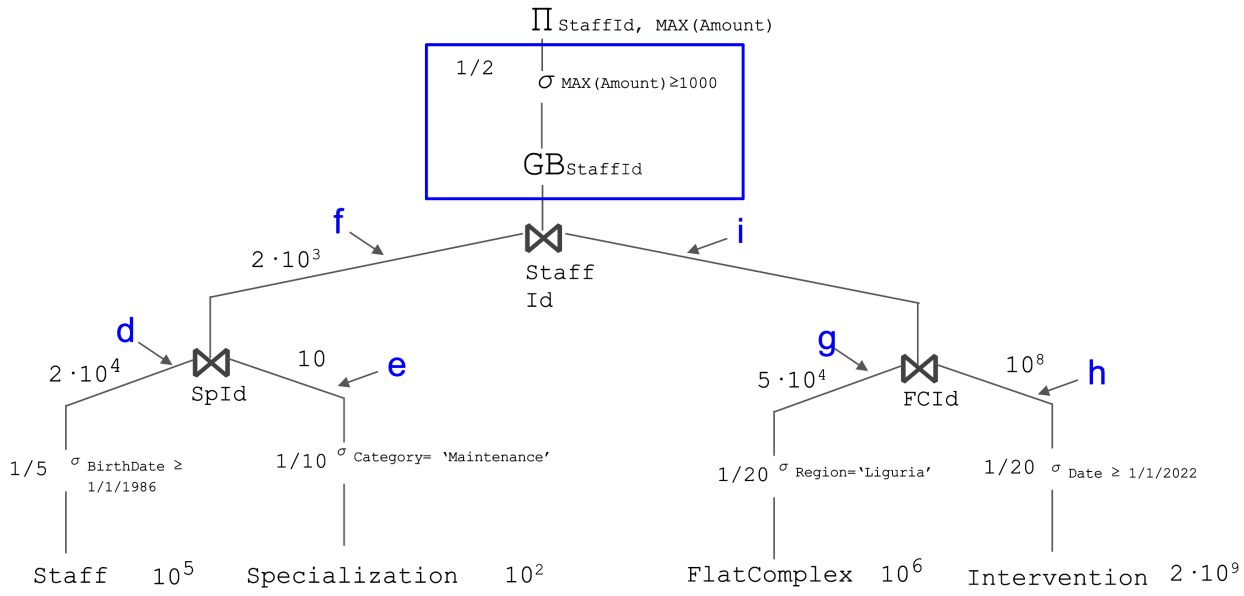
Furthermore, assume the following reduction factor for the having clauses:

Having MAX(Amount)≥1000= 1/2

Consider the following query:

```
SELECT I.StaffId, MAX(Amount)
FROM Intervention I, STAFF ST, SPECIALIZATION S, FLATCOMPLEX FC
WHERE I.StaffId=ST.StaffId and ST.SpId=S.SpId and I.FCId=FC.FCId
and Category='Maintenance' and FC.Region='Liguria' and Date ≥ 1/1/2022
and BirthDate ≥ 1/1/1986)
GROUP BY I.StaffId
HAVING MAX(Amount)≥1000
```

The figure below represents the query tree for the query above.



Analyze the group by anticipation of GROUP BY GROUP BY I.Pid HAVING MAX(Amount)≥1000 represented in the box. Select the solution that **allows maximum efficiency** in executing the query (if any).

- (a) It is possible to anticipate it in branch d
- (b) It is possible to anticipate it in branch h
- (c) It is possible to anticipate it in branch j
- (d) It is possible to anticipate it in branch f
- (e) It is not possible to anticipate the Group BY GROUP BY I.Pid HAVING MAX(Amount)≥1000
- (f) It is possible to anticipate it in branch e
- (g) It is possible to anticipate it in branch i ✓
- (h) It is possible to anticipate it in branch g

Risposta corretta.

La risposta corretta è: It is possible to anticipate it in branch i

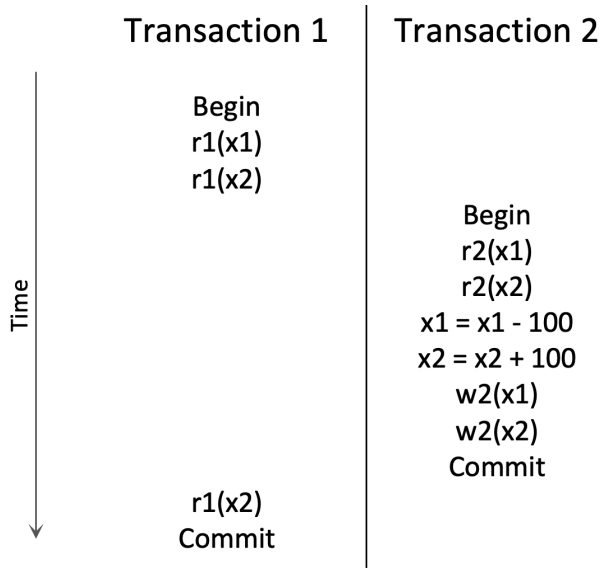
Domanda 8

Risposta non data

Punteggio max.: 1,00

Concurrent access (1 point, -15% penalty for a wrong answer)

The following diagram shows the execution of two transactions in parallel in the absence of a scheduler. What type of anomaly occurs?



Notation:

- rX(o): Read operation by transaction X of object o
- wX(o): Write operation by transaction X of the object o

- (a) Lost update
- (b) No anomaly occurs
- (c) Inconsistent read
- (d) Ghost update (b)
- (e) Ghost update (a)
- (f) Dirty read

Risposta errata.

La risposta corretta è: Inconsistent read

Domanda 9

Parzialmente corretta

Punteggio ottenuto 0,67 su 1,00

Measures (1 point, -15% penalty for each wrong answer)

A market analysis institute wants to analyze information collected on sales of confectionery products produced by a multinational company in recent years.

The confectionery products are made in factories located in different geographical areas. The multinational company produces confectionery products of different types (e.g., snacks, cakes, cookies), and in ways that meet the needs of different dietary regimes (e.g., vegetarian or vegan).

The multinational company sells the products through several stores distributed in various geographical areas.

To analyze the sales of products with respect to different market segments, the multinational company keeps track, through loyalty cards, of some information about the customers who purchased the products, such as customer gender, age range, and type of employment.

The multinational company uses various communication channels to advertise its products through radio and TV commercials or web banners.

The market analysis institute wants to analyze the average cost of advertising actions and overall sales profit based on:

- Production factory, characterized by the information on the city, region, country where it is located, and the person managing the factory
- Product, characterized by the name, type of product (e.g., cake, snack, bar, etc.), indication of compatible diets (one or more values between vegetarian, vegan, celiac disease), ingredient list, and allergen list.
- Gender, age range, and type of employment of the customer who made the purchase. Age range is a value among < 20 years, between 20 and 29 years, between 30 and 49 years, and greater than 50 years. Employment type is a value between freelance, employee, other.
- Store where the product was purchased, characterized by the city region and country where the store is located.
- The type of payment made (a value between cash, credit card, and ATM)
- Date, month, 3.months, 4-Months, 6-months,, year in which the purchase occurred and time slot (a value between morning, afternoon, evening)
- Communication channels used to advertise the product (one or more values from the following: radio advertising, print advertising, television advertising, web banner)

Select from the list below all and only the attributes required to correctly model the requests in the specifications for the fact table (multiple answers can be correct, since multiple measurements can be required)

Scegli una o più alternative:

- [ ] (a) Number of advertising actions
- [ ] (b) Average cost of advertising
- [x] (c) Total cost of advertising ✓

- [ ] (d) Maximum cost of advertising
- [ ] (e) Number of sold products
- [x] (f) Total profit
- [ ] (g) Average profit

Risposta parzialmente esatta.

Hai selezionato correttamente 2.
La risposta corretta è: Number of advertising actions, Total cost of advertising, Total profit

Domanda 10

Completo

Punteggio ottenuto 4,00 su 4,00

Extended SQL query (4 points)

The following data warehouse describes the performance of certification courses delivered by an international certification company. Certification courses are delivered in certification offices, which are distributed around the world. The offices are also characterized by the number of classrooms and teachers available. The metrics to be analyzed are the total revenue, the number of delivered hours, the number of enrolled students, and the number of students who passed the certification exam at the end of the course (NumberPassed).

The data warehouse stores the main characteristics of courses and related certifications in terms of topic, delivery mode (attendance, online, blended), type and level (basic, intermediate, or advanced) of certification, and presence or absence of prerequisites (PresenceAdmissionRequirements, Boolean attribute). Enrolled students are characterized by age range (<=20, >21 and <=30, >31 and <=45, >45), profession, and gender (M, F).

The data warehouse is characterized by the following conceptual schema and corresponding logical schema.

FEATURES-COURSE-CERTIFICATION (IDFeatCourseCert, Topic, CertificationType, CertificationLevel, PresenceAdmissionRequirements,DeliveryMode)
ENROLLED-CHARACTERISTICS (IDEnrolledChar, AgeRange, Profession, Gender)
CERTIFICATION-OFFICE (IDCertOffice, CityOffice, RegionOffice, CountryOffice, NumberClassrooms, NumberTeachers)
TIME (IDTime, Month, 2-Months, 3-Months, 4-Months, 6-Months, Year)
COURSES-DELIVERY (IDFeatCourseCert, IDEnrolledChar, IDCertOffice, IDTime, TotalRevenue, NumberHoursDelivered, NumberEnrolled, NumberPassed)

Considering courses with online delivery mode, separately by profession, 3-month periods, and certification office country, display

- the total number of enrolled students and the total number of passed students
- the average revenue by number of enrolled students
- the average monthly revenue
- the total revenue separately by profession of enrolled students, certification office country, and year,
- the position in a ranking (rank) in descending order of the total number of passed students, separately by year.

---

select profession,3m,office country,

sum(numberenrooled)/sum(numberpassed),

sum(total revenue)/sum(numberenrooled),

sum(total revenue)/count(distinct month),

sum(sum(total revenue)) over (partition by profession,office contry,year),

rank() over (partition by year order by sum(numberpassed)desc),

from coursesdelivery c, time t ,certifactionoffice co ,enrolled characterectic e ,

where c.idtime=t.idtime and c.idcertoffice=co.idsertoffice and c.idenrolledchar=e.idenrolledchar

delivery moode= online

group by profession, 3m,office country,year

SELECT Profession, 3-Months, CountryOffice

SUM(NumberEnrolled), SUM(NumberPassed), SUM(TotalRevenue)/SUM(NumberEnrolled)

SUM(TotalRevenue)/ COUNT (DISTINCT Month)

SUM(SUM(TotalRevenue)) OVER (PARTITION BY A Profession, Year, CountryOffice),

RANK() OVER (PARTITION BY Year ORDER BY SUM(NumberPassed) DESC)

FROM COURSES-DELIVERY CD, TIME T, CERTIFICATION-OFFICE CO, FEATURES-COURSE-CERTIFICATION FC, ENROLLED-CHARACTERISTICS EC

WHERE CD.IDTime=T.IDTime AND CD.IDCertOffice=CO.IDCertOffice AND CD.IDFeatCourseCert =FC.IDFeatCourseCert

AND CD.IDEnrolledChar=ED.IDEnrolledChar AND DeliveryMode="online"

GROUP BY Profession, 3-Months, OfficeCountry, Year

Commento:

Domanda 11

Risposta non data

Non valutata

This question is not a part of the exam

You can use the text area below to write any note or draft (e.g. intermediate steps of an exercise).

Any text written below will not be considered toward the correction of the exam.

Domanda 12

Completo

Punteggio ottenuto 3,00 su 4,00

Extended SQL query (4 points)

The following data warehouse describes the performance of certification courses delivered by an international certification company. Certification courses are delivered in certification offices, which are distributed around the world. The offices are also characterized by the number of classrooms and teachers available. The metrics to be analyzed are the total revenue, the number of delivered hours, the number of enrolled students, and the number of students who passed the certification exam at the end of the course (NumberPassed).

The data warehouse stores the main characteristics of courses and related certifications in terms of topic, delivery mode (attendance, online, blended), type and level (basic, intermediate, or advanced) of certification, and presence or absence of prerequisites (PresenceAdmissionRequirements, Boolean attribute). Enrolled students are characterized by age range (<=20, >21 and <=30, >31 and <=45, >45), profession, and gender (M, F).

The data warehouse is characterized by the following conceptual schema and corresponding logical schema.

FEATURES-COURSE-CERTIFICATION (IDFeatCourseCert, Topic, CertificationType, CertificationLevel, PresenceAdmissionRequirements,DeliveryMode)
ENROLLED-CHARACTERISTICS (IDEnrolledChar, AgeRange, Profession, Gender)
CERTIFICATION-OFFICE (IDCertOffice, CityOffice, RegionOffice, CountryOffice, NumberClassrooms, NumberTeachers)
TIME (IDTime, Month, 2-Months, 3-Months, 4-Months, 6-Months, Year)
COURSES-DELIVERY (IDFeatCourseCert, IDEnrolledChar, IDCertOffice, IDTime, TotalRevenue, NumberHoursDelivered, NumberEnrolled, NumberPassed)

Separately by certification type and 2-months period, display:

- the total revenue,
- the ratio of the total number of passed students over the total number of enrolled students,
- the percentage of the number of passed students with respect to the total number of passed students separately by 6-months period and certification type
- the cumulative total of revenue as 2-months pass, separately by 6-months period and certification type.

Conduct the analysis separately by student gender.

---

select certifaction type,2m,

sum(total revenue),

sum(numberpassed)/sum(numberenrolled),

100*sum (numberpassed)/sum(sum(numberpassed)) over (partition by 6m, certifaction type),

sum(sum(total revenue)) over (partition by 6m ,certifacation type order by 2m rows unbounded preceding)

from courses delivery c ,time t , feauturs course certifaction f ,

where c.idtime=t .idtime and c.id feat coursecert =f.id feat coursecert

group by certifaction type,2m ,6m

SELECT CertificationType, 2-Months, Gender

SUM(TotalRevenue), SUM(NumberPassed)/SUM(NumberEnrolled),

100* Sum(NumberPassed)/SUM(Sum(NumberPassed)) OVER (PARTITION BY 6-Months, CertificationType, Gender),

SUM(SUM(TotalRevenue)) OVER (PARTITION BY 6-Months, CertificationType, Gender ORDER BY 2-Months ROWS UNBOUNDED PRECEDING)

FROM COURSES-DELIVERY CD, TIME T, FEATURES-COURSE-CERTIFICATION FC, ENROLLED-CHARACTERISTIC EC

WHERE CD.IDTime=T.IDTime AND CD.IDFeatCourseCert =FC.IDFeatCourseCert AND EC.IDEnrolledChar =EC. DEnrolledChar

GROUP BY CertificationType, 2-Months, Gender, 6-Months

Commento:

select certifaction type,2m,

sum(total revenue),

sum(numberpassed)/sum(numberenrolled),

100*sum (numberpassed)/sum(sum(numberpassed)) over (partition by 6m, certifaction type**, gender**),

sum(sum(total revenue)) over (partition by 6m ,certifacation type**, Gender** order by 2m rows unbounded preceding)

from courses delivery c ,time t , feauturs course certifaction f ,

where c.idtime=t .idtime and c.id feat coursecert =f.id feat coursecert

group by certifaction type,2m ,6m**, Gender**

Domanda 13

Risposta non data

Punteggio max.: 2,00

2 points (15% penalty for an incorrect answer)

A clustering can be represented as a set of n clusters $\{C_1, C_2, \ldots, C_n\}$. Each cluster C_i is a set of points that have been assigned to the specific cluster.
Given a point $x \ in C_i$, its silhouette can be calculated as:
$silh(x) = \frac{b(x)-a(x)}{max(b(x),a(x))}$
Where:
- a(x) is the average distance between x and all other points in C_i, i.e.
$a(x) = \frac{1}{|C_i|-1} \sum_{j \in C_i, j \in x} dist(j, x)$
- b(x) is the minimum distance between x and clusters to which x does not belong (the distance betwe en a point and a cluster is calculated as the average distance between the point and all points in the cl uster)
$b(x) = min_{C_k, k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} dist(j, x)$
(dist(w,z) is some distance between the points w and z.)

The following is a matrix of Manhattan distances between 6 2-dimensional points.

	a	b	c	d	e	f
a	0	9	9	10	10	7
b	9	0	2	5	9	12
c	9	2	0	7	9	12
d	10	5	7	0	4	7
e	10	9	9	4	0	3
f	7	12	12	7	3	0

A clustering algorithm was used on this dataset, obtaining 3 clusters (0, 1, 2). The cluster labels assigned to the 6 points are as follows:

a: 0
b: 0
c: 1
d: 1
e: 0
f: 2

Which are the silhouettes of points a and c?

The options are reported as (silhouette a, silhouette c) using 4 significant figures.

- (a) (+0.0476, +0.2632)
- (b) None of the options is correct.
- (c) (+0.2632, +0.0476)
- (d) (-0.6111, -0.0952)
- (e) (-0.0476, -0.2632)
- (f) (+0.6111, +0.0952)
- (g) (-0.2632, -0.0476)

- (h) (-0.0952, -0.6111)
- (i) (+0.0952, +0.6111)

Risposta errata.

La risposta corretta è: (-0.2632, -0.0476)

Domanda 14

Risposta non data

Punteggio max.: 1,00

Conceptual schema (1 point, -15% penalty for each wrong answer)

A market analysis institute wants to analyze information collected on sales of confectionery products produced by a multinational company in recent years.

The confectionery products are made in factories located in different geographical areas. The multinational company produces confectionery products of different types (e.g., snacks, cakes, cookies), and in ways that meet the needs of different dietary regimes (e.g., vegetarian or vegan).

The multinational company sells the products through several stores distributed in various geographical areas.

To analyze the sales of products with respect to different market segments, the multinational company keeps track, through loyalty cards, of some information about the customers who purchased the products, such as customer gender, age range, and type of employment.

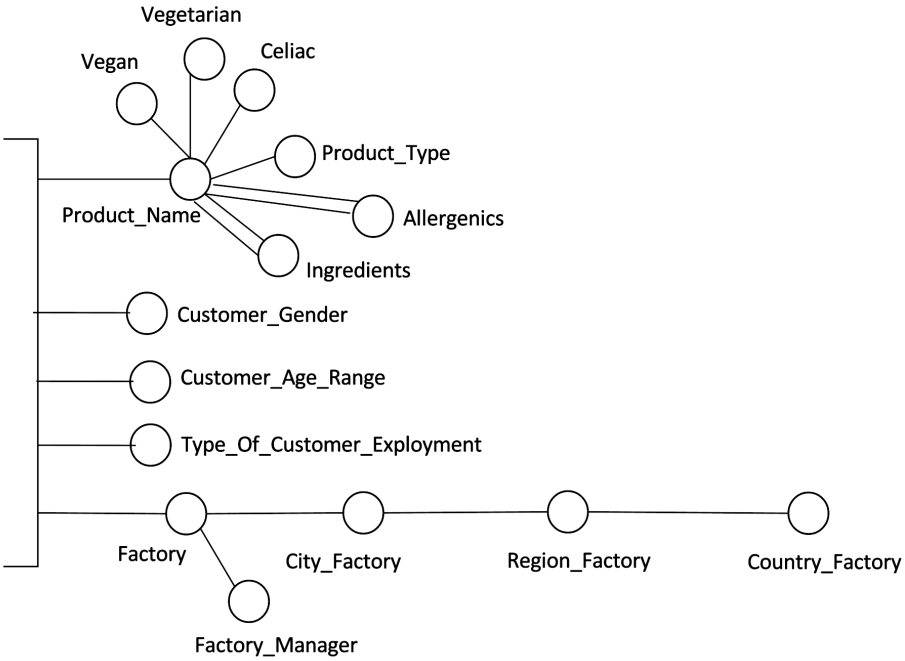
The multinational company uses various communication channels to advertise its products through radio and TV commercials or web banners.

The market analysis institute wants to analyze the average cost of advertising actions and overall sales profit based on:

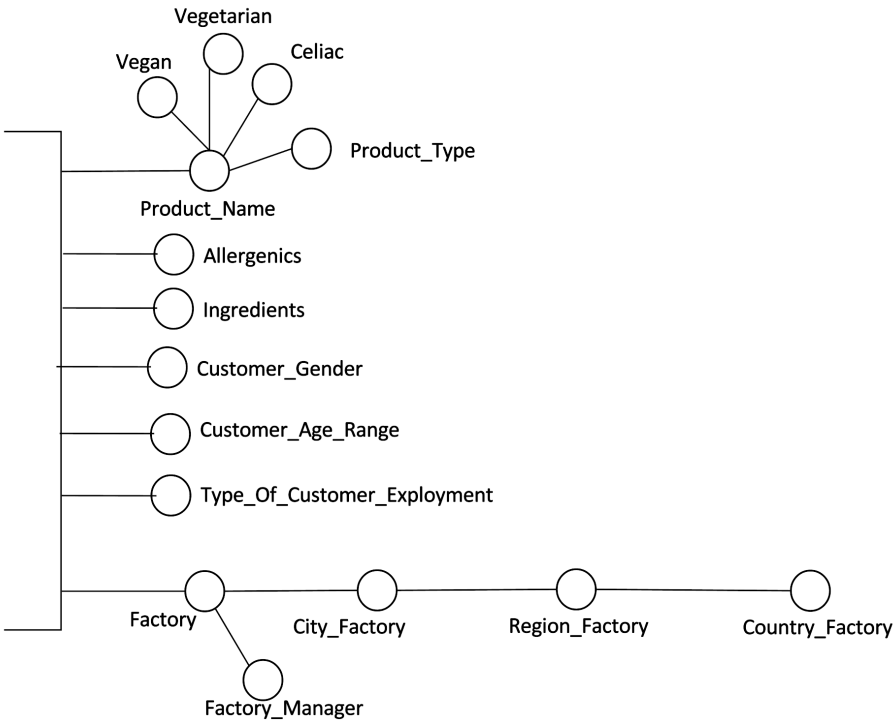
- Production factory, characterized by the information on the city, region, country where it is located, and the person managing the factory
- Product, characterized by the name, type of product (e.g., cake, snack, bar, etc.), indication of compatible diets (one or more values between vegetarian, vegan, celiac disease), ingredient list, and allergen list.
- Gender, age range, and type of employment of the customer who made the purchase. Age range is a value among < 20 years, between 20 and 29 years, between 30 and 49 years, and greater than 50 years. Employment type is a value between freelance, employee, other.
- Store where the product was purchased, characterized by the city region and country where the store is located.
- The type of payment made (a value between cash, credit card, and ATM)
- Date, month, 3.months, 4-Months, 6-months,, year in which the purchase occurred and time

slot (a value between morning, afternoon, evening)
- Communication channels used to advertise the product (one or more values from the following: radio advertising, print advertising, television advertising, web banner)

Select, from the dimensions proposed below, those that meet the requirements described in the problem specifications.

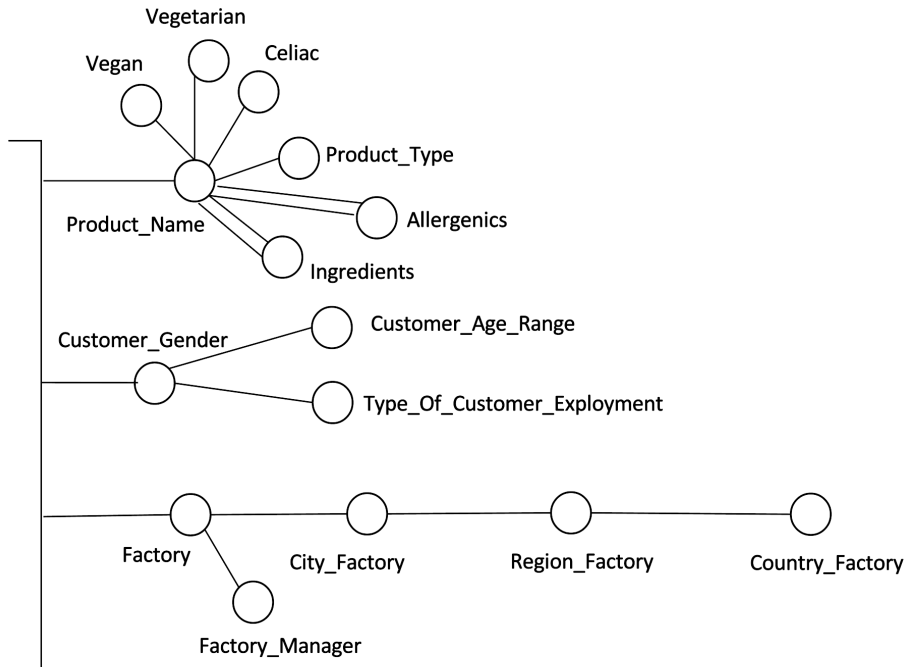


(a)

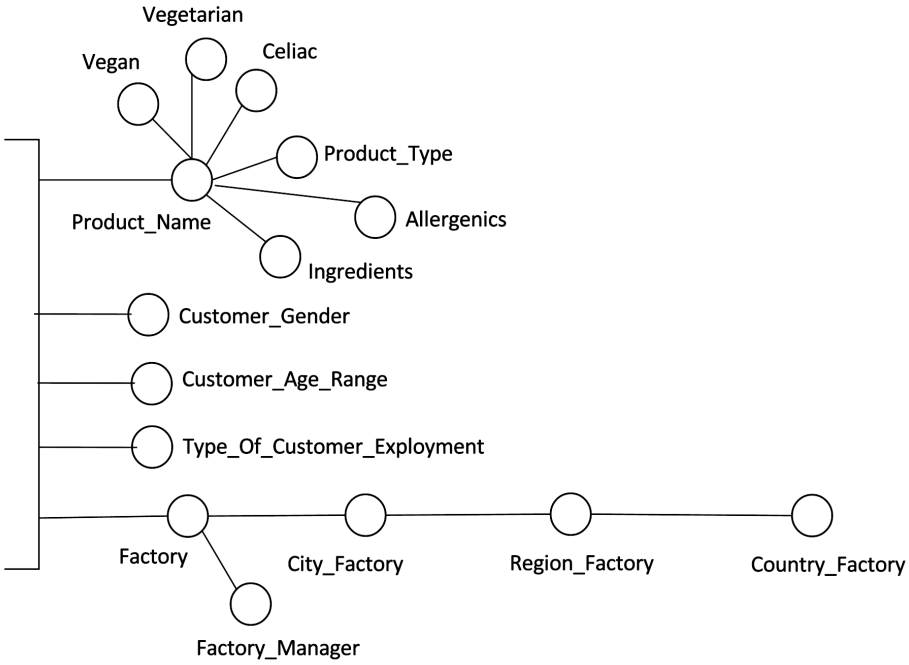


(b)

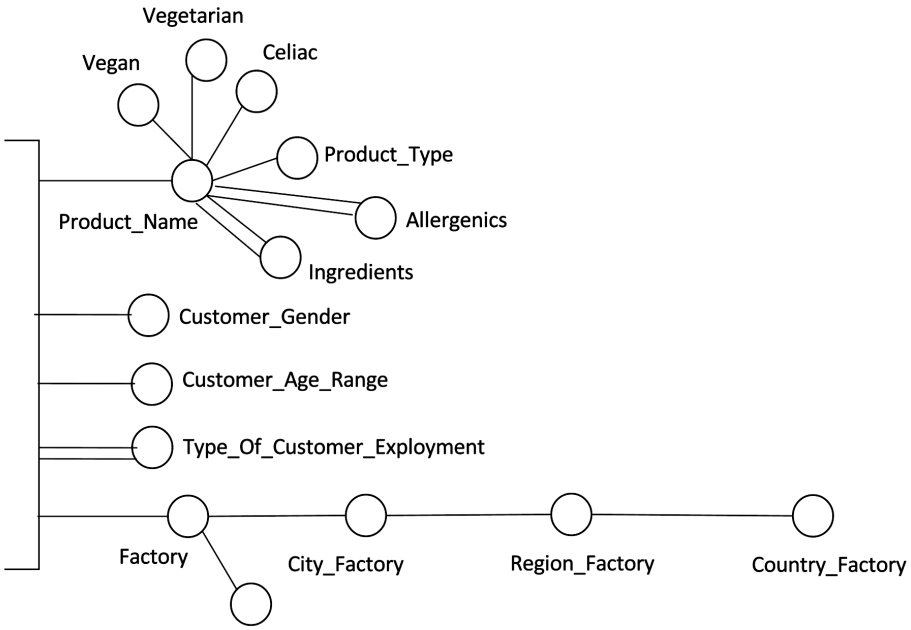
(c)


Vegetarian
Vegan
Celiac
Product_Type
Product_Name
Allergenics
Ingredients
Customer_Gender
Customer_Age_Range
Type_Of_Customer_Exployment
Factory
City_Factory
Region_Factory
Country_Factory
Factory_Manager

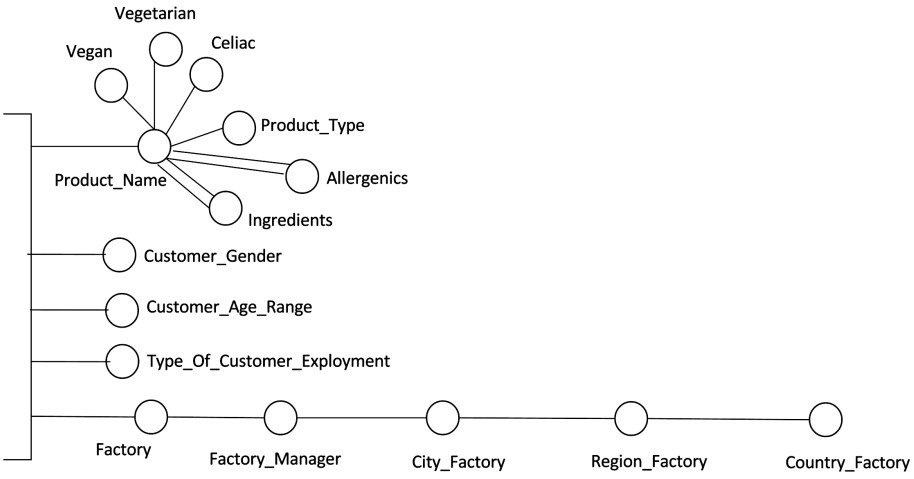
(d)


Vegetarian
Vegan
Celiac
Product_Type
Product_Name
Allergenics
Ingredients
Customer_Gender
Customer_Age_Range
Type_Of_Customer_Exployment
Factory
City_Factory
Region_Factory
Country_Factory
Factory_Manager

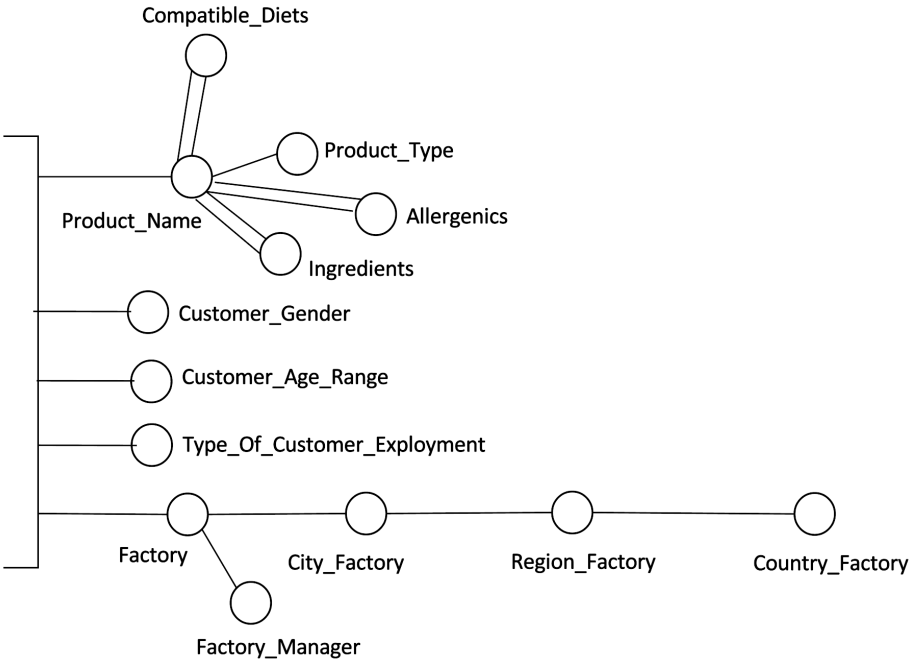
(e)


Vegetarian
Vegan
Celiac
Product_Type
Product_Name
Allergenics
Ingredients
Customer_Gender
Customer_Age_Range
Type_Of_Customer_Exployment
Factory
City_Factory
Region_Factory
Country_Factory
Factory_Manager

(f)

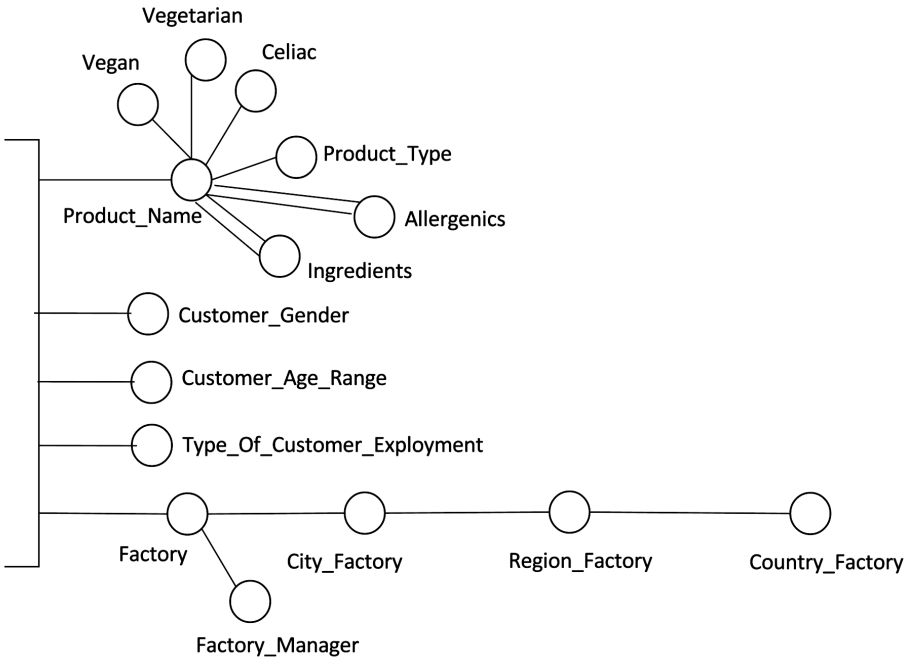

Vegetarian
Vegan
Celiac
Product_Type
Product_Name
Allergenics
Ingredients
Customer_Gender
Customer_Age_Range
Type_Of_Customer_Exployment
Factory
Factory_Manager
City_Factory
Region_Factory
Country_Factory

(g)


Compatible_Diets
Product_Type
Product_Name
Allergenics
Ingredients
Customer_Gender
Customer_Age_Range
Type_Of_Customer_Exployment
Factory
City_Factory
Region_Factory
Country_Factory
Factory_Manager

Risposta errata.

La risposta corretta è:



Domanda 15

Risposta corretta

Punteggio ottenuto 1,50 su 1,50

Cardinalities (1.5 points, -15% penalty for each wrong answer)
The following tables are provided:

STAFF(StaffId, Name, Surname, Email, Telephone, BirthDate, YearsOfExperience, SpId)
SPECIALIZATION(SpId, Name, SpecializationCategory)
FLATCOMPLEX(FCId, Name, Address, Region, Category)
INTERVENTION(Date, FCId, StaffId, Amount)

Assume the following cardinalities:

- card(STAFF) = 10^5 tuples
  - MIN(YearsOfExperience) = 1, MAX(YearsOfExperience) = 30
  - MIN(BirthDate) = 1/1/1930, MAX(BirthDate) = 31/12/1999
- card(SPECIALIZATION) = 10^2 tuples
  - distinct values of SpecializationCategory = 10
- card(FLATCOMPLEX) = 10^6 tuples
  - distinct values of Category = 10

- card(Intervention)= $2 \cdot 10^{9}$ tuples
  - MIN(Amount) = 100, MAX(Amount) = 100000,
  - MIN(Date) = 1/1/2003, MAX(Date) = 31/12/2022

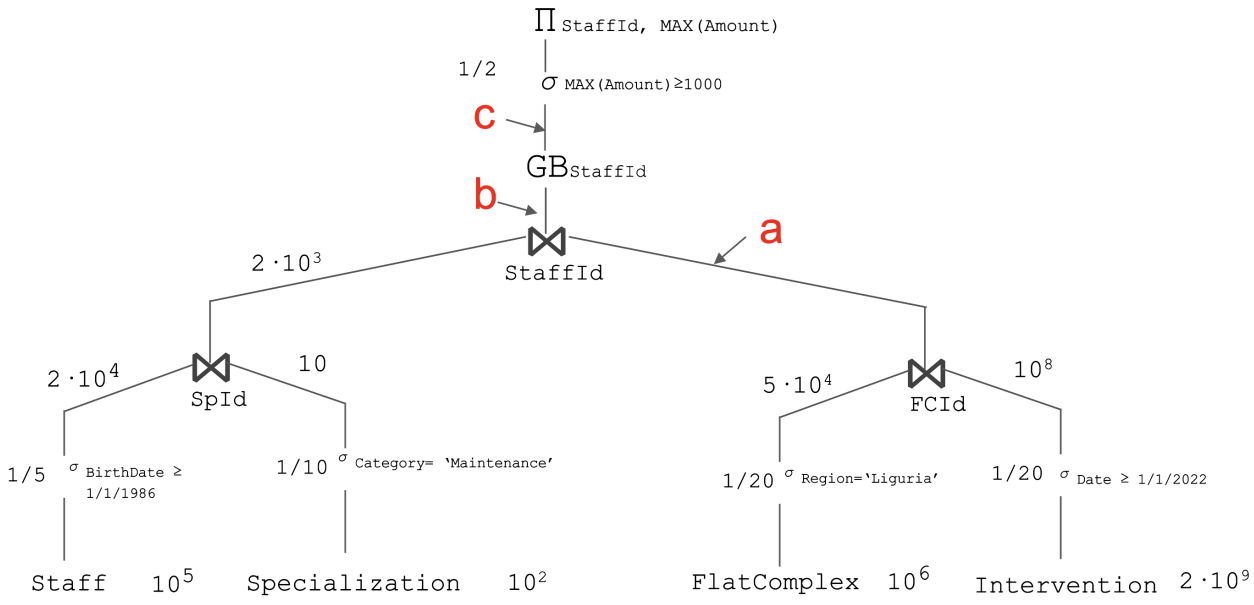
Furthermore, assume the following reduction factor for the having clauses:

Having MAX(Amount)≥1000= 1/2

Consider the following query:

```
SELECT I.StaffId, MAX(Amount)
FROM Intervention I, STAFF ST, SPECIALIZATION S, FLATCOMPLEX FC
WHERE I.StaffId=ST.StaffId and ST.SpId=S.SpId and I.FCId=FC.FCId
and Category='Maintenance' and FC.Region='Liguria' and Date ≥ 1/1/2022
and BirthDate ≥ 1/1/1986)
GROUP BY I.StaffId
HAVING MAX(Amount)≥1000
```

The figure below represents the query tree for the query above.



Specify the cardinality of each node indicated by the red letters (a,b,c) in the figure. There is one right answer for each node a,b,c.

Scegli una o più alternative:

- (a) c: $2 \cdot 10^{2}$
- (b) a: $5 \cdot 10^{4}$
- (c) c: $5 \cdot 10^{4}$

- [ ] (d) a: $2 \cdot 10^5$
- [x] (e) b: 10^5 ✓
- [ ] (f) b: $5 \cdot 10^3$
- [ ] (g) b: 10^4
- [ ] (h) c: 10^5
- [x] (i) a: $5 \cdot 10^6$ ✓
- [ ] (j) a: $2 \cdot 10^7$
- [ ] (k) b: 10^3
- [x] (l) c: $2 \cdot 10^3$ ✓

Risposta corretta.

La risposta corretta è: a: $5 \cdot 10^6$, b: 10^5, c: $2 \cdot 10^3$

Domanda 16

Risposta non data

Punteggio max.: 1,00

1 point (-15% penalty for wrong answer)

- The precision of a class is defined as the fraction of points correctly predicted for that class, compared to all points predicted for that class
- The recall of a class is defined as the fraction of points correctly predicted for that class, compared to all points belonging to that class

Given y_pred and y_true, vectors of predictions and ground truth, respectively, for a test set of 10 points.

y_pred: [B C A A A B C C]
y_true: [C B A A B B C A C]

What are the precision and recall for class A? *All answers are given in the format (precision, recall).*

- (a) (2/3, 1/2)
- (b) (1/3, 1/2)
- (c) (1/2, 1/3)
- (d) None of the other answers are correct
- (e) (1/2, 1/2)
- (f) (1/2, 2/3)
- (g) (2/3, 1/3)
- (h) (1/3, 2/3)

Risposta errata.

La risposta corretta è: (1/2, 2/3)

Domanda 17

Risposta non data

Punteggio max.: 1,00

Theory (1 point, -15% penalty for a wrong answer)

The following is an example of a document taken from a restaurant's guests collection, where information about a restaurant's guests is entered.

```
{
    "firstName": "John",
    "lastName": "Smith",
    "yearOfBirth": 1990,
    "dietaryRequirements": "vegetarian",
    "table": 12
}
```

We want to extract, for each table (table) the number of guests with vegetarian or vegan requests (dietaryRequirements = "vegetarian" or "vegan").

Which of the following queries addresses the request?

- (a)

```
db.guests.aggregate([
  {
    $group: {
      _id: "$table",
      count: { $sum: 1 }
    }
  },
  {
    $match: {
      $and: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  }
])
```

- (b)

```
db.guests.aggregate([
  {
    $match: {
      $or: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  },
  {
    $group: {
      _id: "$table",
      count: { $sum: {} }
    }
  }
])
```

(c)

```
db.guests.aggregate([
  {
    $match: {
      $and: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  },
  {
    $group: {
      _id: "$table",
      count: { $sum: {} }
    }
  }
])
```

(d)

```
db.guests.aggregate([
  {
    $match: {
      $and: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  },
  {
    $group: {
      _id: "$table",
      count: { $sum: 1 }
    }
  }
])
```

(e)

```
db.guests.aggregate([
  {
    $group: {
      _id: "$table",
      count: { $sum: {} }
    }
  },
  {
    $match: {
      $and: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  }
])
```

(f)

```
db.guests.aggregate([
  {
    $match: {
      $or: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  },
  {
    $group: {
      _id: "$table",
      count: { $sum: 1 }
    }
  }
])
```

(g)

```
db.guests.aggregate([
  {
    $group: {
      _id: "$table",
      count: { $sum: {} }
    }
  },
  {
    $match: {
      $or: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  }
])
```

(h)

```
db.guests.aggregate([
  {
    $group: {
      _id: "$table",
      count: { $sum: 1 }
    }
  },
  {
    $match: {
      $or: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  }
])
```

Risposta errata.

La risposta corretta è:

```
db.guests.aggregate([
  {
    $match: {
      $or: [
        { dietaryRequirements: "vegetarian" },
        { dietaryRequirements: "vegan" }
      ]
    }
  },
  {
    $group: {
      _id: "$table",
      count: { $sum: 1 }
    }
  }
])
```