POLITECNICO
DI TORINO

Data Science and Database Technology

Exam 2023-01-23

	PEGAH YARAHMADI 315908
Iniziato	lunedì, 23 gennaio 2023, 11:14
Terminato	lunedì, 23 gennaio 2023, 12:46
Tempo impiegato	1 ora 31 min.
Valutazione	**0,84** su un massimo di 32,00 (**3**%)

Domanda 1

Risposta errata

Punteggio ottenuto -0,15 su 1,00

1 point (15% penalty for a wrong answer)

The following is a document taken from a MongoDB collection named "entries" that maintains information about gym customer entries.

```
{
  "first": "DOROTHY",
  "last": "MILLER",
  "age": 20,
  "gender": "F",
  "date": {
    "day": 16,
    "month": 8,
    "year": 2022
  },
}
```

Based on the structure inferred from this document, which of the following MongoDB queries extracts the minimum number of entries made in the years where the average duration of entries is greater than 60 minutes?

○ (a)

```
db.collection.aggregate([
 {
   $group: {
     _id: null,
     "avg_duration": {
       $avg: "$duration"
     },
     "count": {
       $sum: 1
     }
   }
 },
 {
   $match: {
     avg_duration: {
       $gte: 60
     }
   }
 },
 {
   $group: {
     _id: "$date.year",
     min_count: {
       $min: "$count"
     }
   }
 }
])
```

○ (b)

```
db.collection.aggregate([
 {
   $match: {
     avg_duration: {
       $gte: 60
     }
   }
 },
 {
   $group: {
     _id: null,
     "avg_duration": {
       $avg: "$duration"
     },
     "count": {
       $sum: 1
     }
   }
 },
 {
   $group: {
     _id: "$date.year",
     min_count: {
       $min: "$count"
     }
   }
 }
])
```

○ (c)

```
db.collection.aggregate([
 {
  $match: {
   avg_duration: {
    $gte: 60
   }
  }
 },
 {
  $group: {
   _id: null,
   "avg_duration": {
    $avg: "$duration"
   }
  }
 },
 {
  $group: {
   _id: "$date.year",
   min_count: {
    $min: "$count"
   }
  }
 }
])
```

○ (d)

```
db.collection.aggregate([
 {
  $group: {
   _id: null,
   "avg_duration": {
    $avg: "$duration"
   }
  }
 },
 {
  $match: {
   avg_duration: {
    $gte: 60
   }
  }
 },
 {
  $group: {
   _id: "$date.year",
   min_count: {
    $min: "$count"
   }
  }
 }
])
```

◉ (e)

```
db.collection.aggregate([
  {
    $match: {
      avg_duration: {
        $gte: 60
      }
    }
  },
  {
    $group: {
      _id: "$date.year",
      "avg_duration": {
        $avg: "$duration"
      }
    }
  },
  {
    $group: {
      _id: null,
      min_count: {
        $min: "$count"
      }
    }
  }
])
```

✗

○ (f)

```
db.collection.aggregate([
 {
   $group: {
     _id: "$date.year",
     "avg_duration": {
       $avg: "$duration"
     },
     "count": {
       $sum: 1
     }
   }
 },
 {
   $match: {
     avg_duration: {
       $gte: 60
     }
   }
 },
 {
   $group: {
     _id: null,
     min_count: {
       $min: "$count"
     }
   }
 }
])
```

○ (g)

```
db.collection.aggregate([
 {
  $group: {
   _id: "$date.year",
   "avg_duration": {
    $avg: "$duration"
   }
  }
 },
 {
  $match: {
   avg_duration: {
    $gte: 60
   }
  }
 },
 {
  $group: {
   _id: null,
   min_count: {
    $min: "$count"
   }
  }
 }
])
```

○ (h)

```
db.collection.aggregate([
  {
    $match: {
      avg_duration: {
        $gte: 60
      }
    }
  },
  {
    $group: {
      _id: "$date.year",
      "avg_duration": {
        $avg: "$duration"
      },
      "count": {
        $sum: 1
      }
    }
  },
  {
    $group: {
      _id: null,
      min_count: {
        $min: "$count"
      }
    }
  }
])
```

Risposta errata.

La risposta corretta è:

```
db.collection.aggregate([
  {
    $group: {
      _id: "$date.year",
      "avg_duration": {
        $avg: "$duration"
      },
      "count": {
        $sum: 1
      }
    }
  },
  {
    $match: {
      avg_duration: {
        $gte: 60
      }
    }
  },
  {
    $group: {
      _id: null,
      min_count: {
        $min: "$count"
      }
    }
  }
])
```

Domanda 2

Parzialmente corretta

Punteggio ottenuto -0,06 su 5,00

5 points overall (penalty -15% for a wrong answer)

The following tables are provided:

Film(CodF, Title, CodD, CodC, Genre, Year, AverageRating)
Director(CodD, Name, Surname, Nationality, BirthDate)
DistributionCompany(CodC, CompanyName, FoundationDate, HeadQuarterArea)
User(CodU, Name, Surname, Subscription)
Viewed(CodF, CodU, DateLastViewed, Rating)

Assume the following cardinalities:

- card(FILM) = $6\cdot10^{6}$ tuples,
  - Distinct values of Genre = 20
  - Reduction factor of AverageRating>4.5 1/20
- card(DIRECTOR) = $3\cdot10^{4}$ tuples,
  - Distinct values of Nationality = 100
  - MIN(BirthDate) = 1/1/1910, MAX(BirthDate) = 31/12/1999
- card(DISTRIBUTIONCOMPANY) = 10^{2} tuples,
  - Distinct values of HeadQuarterArea = 10
- card(USER) = $4\cdot10^{7}$ tuples,
  - MIN(BirthDate) = 1/1/1940, MAX(BirthDate) = 31/12/2009
  - Distinct values of Subscription = 4
- card(VIEWED)= $12\cdot10^{8}$ tuples

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

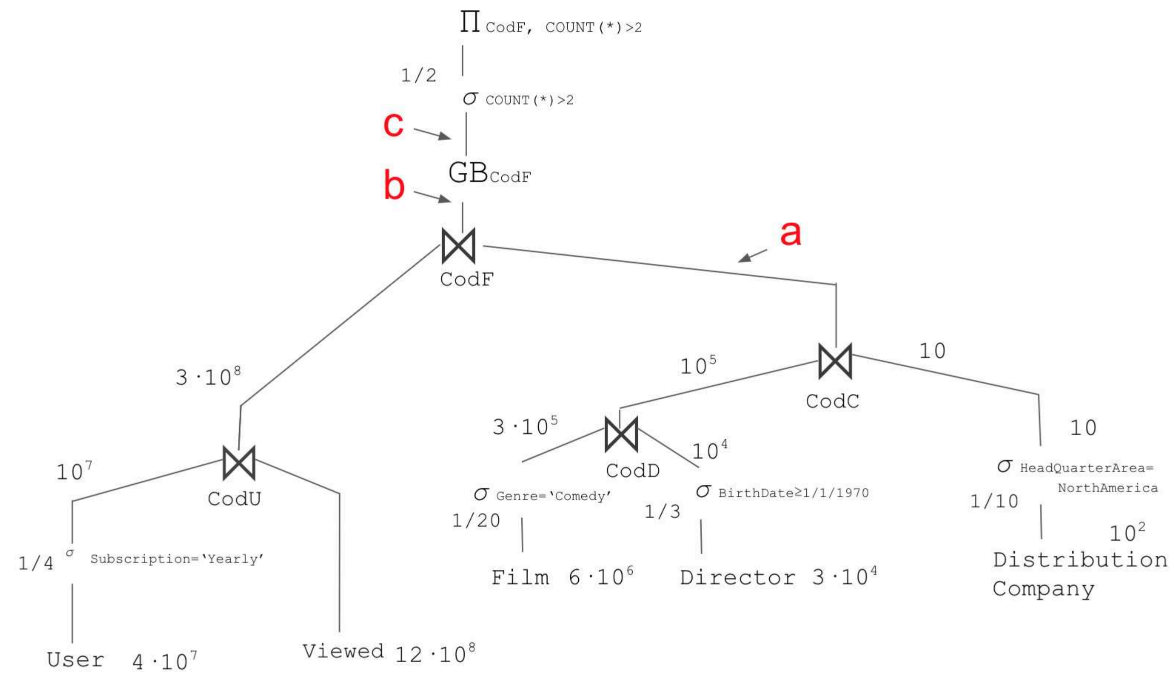
Having COUNT(*)>2 = 1/5

Consider the following query:

```
select F.CodF, COUNT(*)
from  User U, Viewed V, Film F, Director D, DistributionCompany C
where V.CodU=U.CodU and F.CodD=D.CodD
   and F.CodC=C.CodC and V.CodF=F.CodF
   and U.Subscription=‘Yearly’
   and C.HeadQuarterArea=‘NorthAmerica’
   and D.BirthDate≥1/1/1970
   and F.Genre=‘Comedy’
group by F.CodF
having COUNT(*)>2
```

Cardinalities

(1.5 points, penalty -15% for a wrong answer)
The figure below represents the query tree for the query above.



Specify the correct cardinality for **(a)**:

○ 10^4 ● $2 \cdot 10^5$ ✗ ○ $5 \cdot 10^3$ ○ 10^5

Punteggio ottenuto -0,75 su 5,00

La risposta corretta è: 10^4

Specify the correct cardinality for **(b)**:

○ $5 \cdot 10^4$ ○ $4 \cdot 10^3$ ● $2 \cdot 10^4$ ✗ ○ $5 \cdot 10^5$

Punteggio ottenuto -0,75 su 5,00

La risposta corretta è: $5 \cdot 10^5$

Specify the correct cardinality for **(c)**:

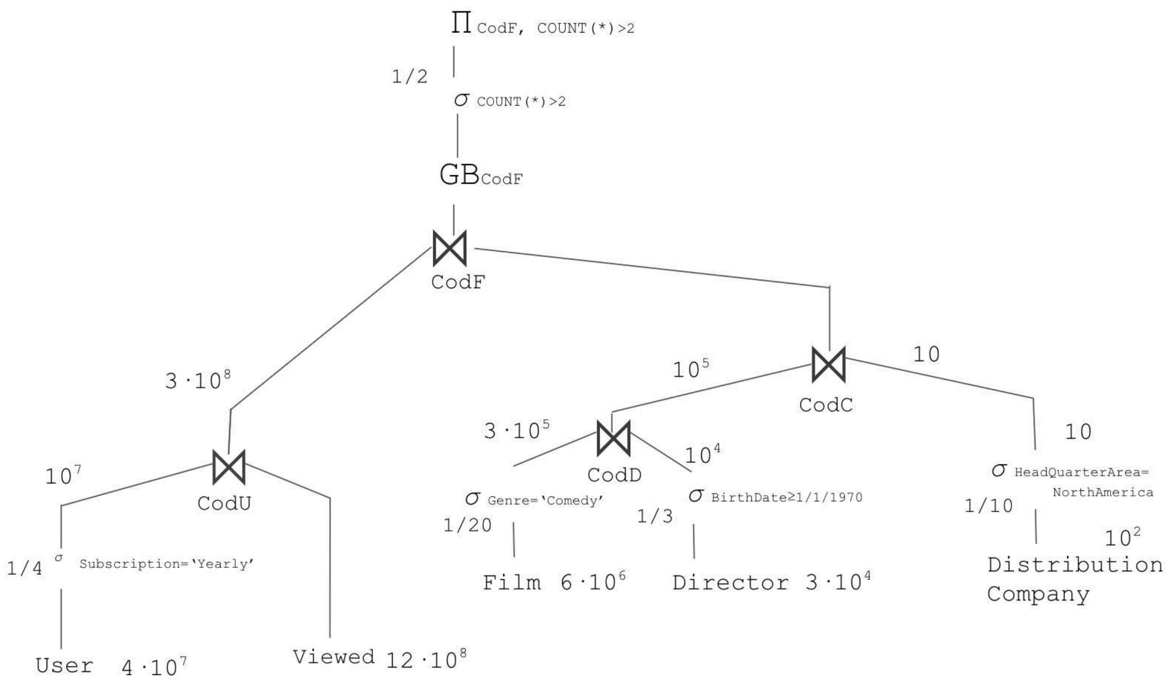
○ 10^5 ○ 10^4 ○ $4 \cdot 10^4$ ● $2 \cdot 10^5$ ✗

Punteggio ottenuto -0,75 su 5,00

La risposta corretta è: 10^4

Indexes

(1.5 points, penalty -15% for a wrong answer)
The figure below represents the query tree for the query above.



Select, for each table, one or more secondary physical structures to increase query performance (if possible) among the options below.

Table VIEWED

○ CREATE INDEX IndexA ON VIEWED(DateLastViewed) - HASH

○ None of the proposed secondary physical structures on this table would increase query performance

● CREATE INDEX IndexB ON VIEWED(DateLastViewed) - B+- Tree ❌

Punteggio ottenuto -0,45 su 3,00

La risposta corretta è: None of the proposed secondary physical structures on this table would increase query performance

Table USER

○ None of the proposed secondary physical structures on this table would increase query performance

○ CREATE INDEX IndexC ON USER(Subscription) - HASH

● CREATE INDEX IndexD ON USER(Subscription) - B+- Tree ✗

Punteggio ottenuto -0,45 su 3,00

La risposta corretta è: None of the proposed secondary physical structures on this table would increase query performance

Table FILM

○ None of the proposed secondary physical structures on this table would increase query performance

○ CREATE INDEX IndexF ON FILM(Genre) - B+- Tree

● CREATE INDEX IndexE ON FILM(Genre) - HASH ✓

Punteggio ottenuto 3,00 su 3,00

La risposta corretta è: CREATE INDEX IndexE ON FILM(Genre) - HASH

Table DIRECTOR

○ None of the proposed secondary physical structures on this table would increase query performance

● CREATE INDEX IndexH ON DIRECTOR(BirthDate) - B+- Tree ✗

○ CREATE INDEX IndexG ON DIRECTOR(BirthDate) - HASH

Punteggio ottenuto -0,45 su 3,00

La risposta corretta è: None of the proposed secondary physical structures on this table would increase query performance

Table DISTRIBUTIONCOMPANY

○ CREATE INDEX IndexJ ON DISTRIBUTIONCOMPANY(HeadQuartedArea) - B+- Tree

○ CREATE INDEX IndexI ON DISTRIBUTIONCOMPANY(HeadQuartedArea) - HASH

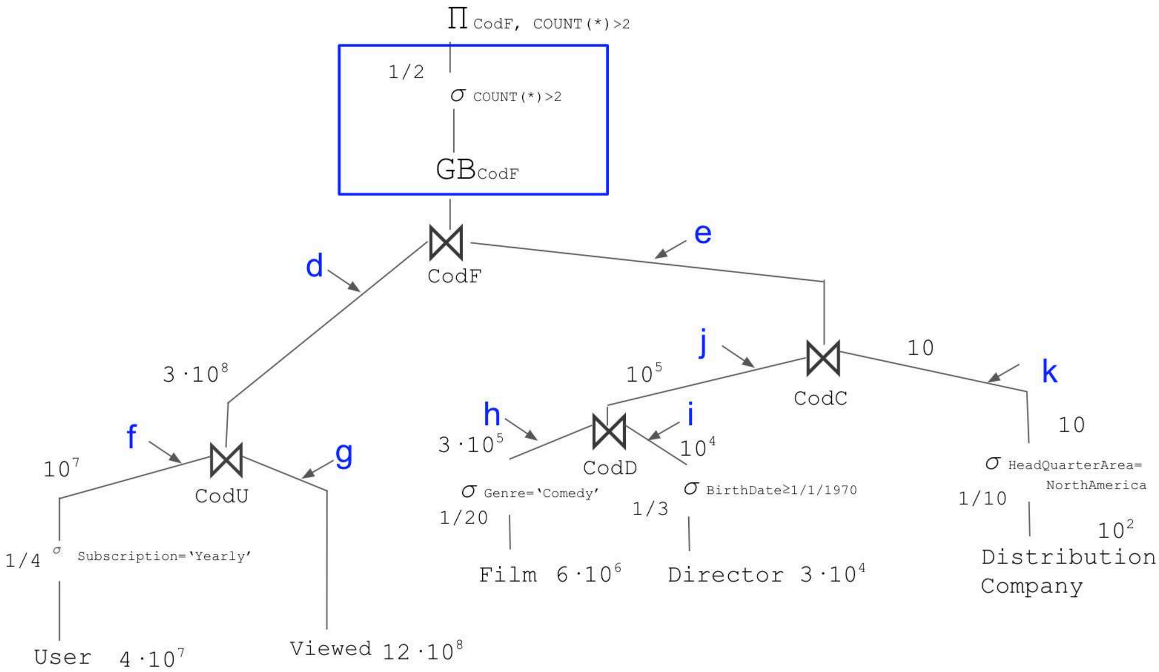
● None of the proposed secondary physical structures on this table would increase query performance ✓

Punteggio ottenuto 3,00 su 3,00

La risposta corretta è: None of the proposed secondary physical structures on this table would increase query performance

Group By Anticipation

(2 points, -15% penalty for a wrong answer)
The figure below represents the query tree for the query above.



Analyze the group by anticipation of **group by F.CodF having COUNT(*)>2** represented in the box. Select the solution that **allows maximum efficiency** in executing the query (if any).

○ It is possible to anticipate it in branch g

○ It is possible to anticipate it in branch d

○ It is possible to anticipate it in branch i

● It is possible to anticipate it in branch j ✗

○ It is possible to anticipate it in branch e

○ It is not possible to anticipate the Group BY **group by F.CodF having COUNT(*)>2)**

○ It is possible to anticipate it in branch f

○ It is possible to anticipate it in branch h

○ It is possible to anticipate it in branch k

Punteggio ottenuto -3,00 su 20,00

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

1) La risposta corretta è : 10^4
2) La risposta corretta è : $5 \cdot 10^5$
3) La risposta corretta è : 10^4
4) La risposta corretta è : None of the proposed secondary physical structures on this table would increase query performance
5) La risposta corretta è : None of the proposed secondary physical structures on this table would increase query performance
6) La risposta corretta è : CREATE INDEX IndexE ON FILM(Genre) - HASH
7) La risposta corretta è : None of the proposed secondary physical structures on this table would increase query performance
8) La risposta corretta è : None of the proposed secondary physical structures on this table would increase query performance
9) La risposta corretta è : It is possible to anticipate it in branch d

Domanda 3

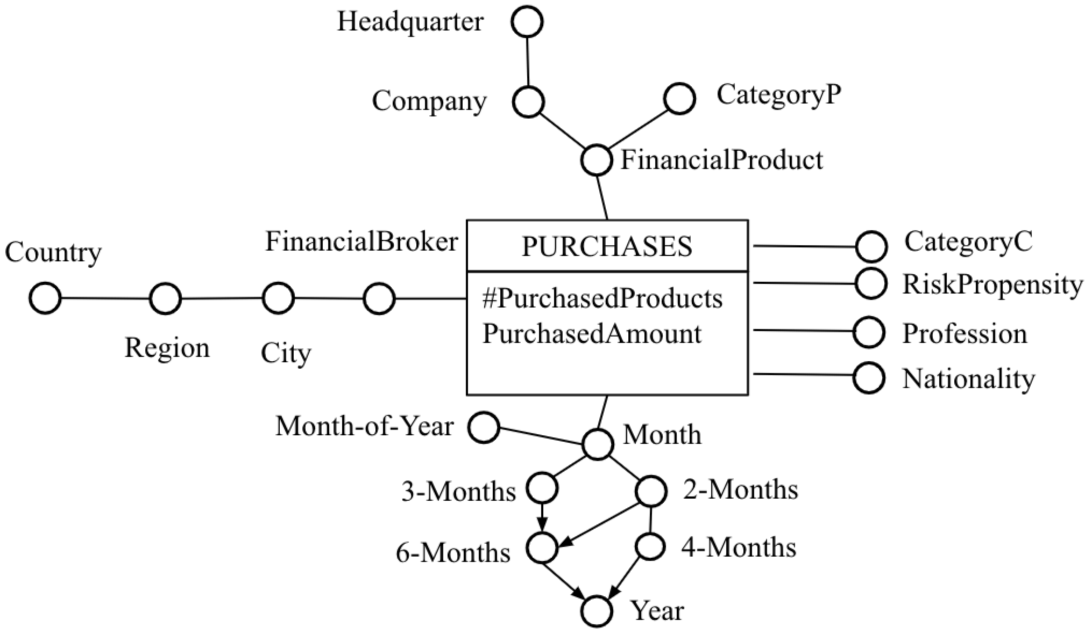
Completo

Punteggio ottenuto 0,25 su 11,00

11 points overall (no penalties for a wrong answer)

The following data warehouse describes the purchasing performance of financial products (stocks, securities, investment funds, etc.) offered by different companies through financial brokers (bank, broker, agencies, etc.). Each financial product is issued by a specific company and pertains to only one category (bulk securities, derivative contracts, etc.). For each financial broker, the city, region and state in which it operates are known. The data warehouse stores the customer category (senior, medium, junior), risk propensity (high, medium and low), profession and nationality of customers who have purchased financial products through financial brokers over time.

The metrics to be analyzed are the number of products purchased (#PurchasedProducts) and the corresponding value (PurchasedAmount). The metrics are to be analyzed for each month, 2-month, 3-month, 4-month, 6-month, and year period.



FINANCIAL-BROKER(**IDFinancialBroker**, FinancialBroker, City, Region, Country)
FINANCIAL-PRODUCT (**IDFinancialProduct**, FinancialProduct, Company, Headquarter, CategoryP)
CUSTOMER-FEATURES (**IDCF**, CategoryC, RiskPropensity, Profession, Nationality)
TIME (**IDTime**, Month, Month-of-Year, 2-Months, 3-Months, 4-Months, 6-Months, Year)
PURCHASES (**IDFinancialBroker**, **IDFinancialProduct**, **IDCF**, **IDTime**, #PurchasedProducts, PurchasedAmount)

Given the previous logical schema, write the following queries in extended SQL in the box below, separated by a space:

- **(3 points) Query 1**
  Show for each financial broker and four-month period the total number of purchased products and the corresponding amount. Associate each record with a rank:
  - which identifies the position of the record according to the total amount of purchases (1 for the lowest total value)
  - which identifies the position of the record in descending order of the number of purchased products, separately by financial broker country

- **(4 points) Query 2**
  For financial products belonging to the derivative contracts category (CategoryP='Derivative Contracts'), show separately for each financial broker and four-month period
  - the total purchased amount,

- the cumulative purchased amount from the beginning of the year as the 4-months periods pass, separately by financial broker,
- the total purchased amount independent of the financial broker

- **(4 points) Query 3**
  Show for each financial product, customer profession, and three-month period
  - the number of purchased products,
  - the average amount per purchased product,
  - the average monthly amount,
  - the percentage of the purchased amount with respect to the total purchased amount of financial products in the same Category, separately by customer profession and three-month period

---

Query 1:

```
SELECT PurchasedProducts, PurchasedAmount, FinancialBroker, 4-Months
FROM PURCHASES
WHERE
JOIN FINANCIAL-BROKER, TIME
GROUP BY FinancialBroker, 4-Months;
```

Query 1

```
SELECT FinancialBroker, 4-Months,
SUM(#PurchasedProducts), SUM(PurchasedAmount),
RANK() OVER (ORDER BY SUM(PurchasedAmount)),
RANK() OVER (PARTITION BY Country ORDER BY SUM(#PurchasedProducts) DESC)
FROM FINANCIAL-BROKER F, PURCHASES P, TIME T
WHERE F.IDFinancialBroker=P.IDFinancialBroker AND P.IDTime=T.IDTime
GROUP BY FinancialBroker, 4-Months, Country
```

Query 2

```
SELECT 4-Months, FinancialBroker, SUM(PurchasedAmount)
     SUM(SUM(PurchasedAmount)) over (PARTITION BY FinancialBroker, Year
  ORDER BY 4-Months
  rows unbounded preceding)
SUM(SUM(PurchasedAmount)) over (PARTITION BY 4-Months)
FROM FINANCIAL-BROKER FB, PURCHASES P, TIME T, FINANCIAL-PRODUCT FP
WHERE FB.IDFinancialBroker=P.IDFinancialBroker AND P.IDTime=T.IDTime AND
FP.IDFinancialProduct=P.IDFinancialProduct AND CategoryP=’Derivative Contracts’
GROUP BY 4-Months, FinancialBroker, Year;
```

Query 3

```
SELECT FinancialProduct, Profession, 3-Months,
 SUM(#PurchasedProducts), SUM(PurchasedAmount)/SUM(#PurchasedProducts),
SUM(PurchasedAmount)/ COUNT(DISTINCT Month)
100*SUM(PurchasedAmount)/SUM(SUM(PurchasedAmount)) OVER (PARTITION BY Profession, 3-Months, CategoryP)
FROM PURCHASES P, FINANCIAL-PRODUCT FP, CUSTOMER-FEATURES C, TIME T
WHERE P.IDFinancialProduct=FP.IDFinancialProduct AND
C.IDCF=P.IDCF AND T.IDTime=P.IDTime
GROUP BY FinancialProduct, Profession, 3-Months, CategoryP
```

Commento:
Query 1: **0.25 points**

SELECT PurchasedProducts, PurchasedAmount, FinancialBroker, 4-Months

SUM(#PurchasedProducts), SUM(PurchasedAmount),

RANK() OVER (ORDER BY SUM(PurchasedAmount)),

RANK() OVER (PARTITION BY Country ORDER BY SUM(#PurchasedProducts) DESC)

FROM PURCHASES

WHERE

~~JOIN~~ FINANCIAL-BROKER, TIME

GROUP BY FinancialBroker, 4-Months;**, Country**

Query 2) **0 points**

Query 3) **0 points**

Domanda 4

Risposta errata

Punteggio ottenuto -0,38 su 2,50

2.5 points (15% penalty for a wrong answer)

The following list of transactions is given.

abcd
acde
abc
abd
cde
acde
b
d
abde
acd

The FP-growth algorithm is applied using minsup = 2 (an itemset is considered frequent if it appears in at least two transactions).

What is the BC-CPB (BC-Conditional Pattern Base)?

○ (a) ae: 1, a: 2

○ (b) ad: 3, a: 4

○ (c) ad: 1, a: 1

○ (d) ae: 1, a: 1

○ (e) ade: 2, a: 1

○ (f) ae: 1, a: 2

● (g) ade: 1, a: 2 ✗

○ (h) abc: 1, a: 1

Risposta errata.

La risposta corretta è: ad: 1, a: 1

Domanda 5

Risposta errata

Punteggio ottenuto -0,15 su 1,00

1 point (15% penalty for a wrong answer)

Notation:

- rN(V): read of object V by transaction N
- wN(V): write of object V by transaction N

The following schedule of 3 transactions is given:

S = w0(z) w0(y) r2(y) r0(y) w2(z) r1(z) w2(x) r2(x) w1(x) w1(z)

S is conflict serializable because it is conflict equivalent to the serial schedule:

○ (a) r2(y) w2(z) w2(x) r2(x) w0(z) w0(y) r0(y) r1(z) w1(x) w1(z)

○ (b) w0(z) w0(y) w2(z) r1(z) w2(x) r2(x) w1(x) w1(z) r2(y) r0(y)

○ (c) S is not conflict serializable because it is not conflict equivalent to any serial schedule.

○ (d) r2(y) w2(z) w2(x) r2(x) r1(z) w1(x) w1(z) w0(z) w0(y) r0(y)

○ (e) w0(z) w0(y) r0(y) r1(z) w1(x) w1(z) r2(y) w2(z) w2(x) r2(x)

○ (f) w0(z) w0(y) r0(y) r2(y) w2(z) w2(x) r2(x) r1(z) w1(x) w1(z)

○ (g) r1(z) w1(x) w1(z) w0(z) w0(y) r0(y) r2(y) w2(z) w2(x) r2(x)

● (h) r1(z) w1(x) w1(z) r2(y) w2(z) w2(x) r2(x) w0(z) w0(y) r0(y) ✗

Risposta errata.

La risposta corretta è: w0(z) w0(y) r0(y) r2(y) w2(z) w2(x) r2(x) r1(z) w1(x) w1(z)

Domanda 6

Parzialmente corretta

Punteggio ottenuto 1,85 su 3,00

3 points overall (penalty 15% for a wrong answer)

We want to analyze the enrollment information of schools in the Italian territory. Each school offers some laboratories (e.g., for theater or musical activities) and some services to support students (e.g., cafeteria service or afternoon study support).

Students may have a certification for a special educational need (e.g., DSA or BSE). The number of students enrolled in the schools may change during the year due to possible school changes or arrivals from other schools.

We want to analyze (1) the percentage of students in attendance relative to the number of students enrolled and (2) the ratio of the number of students in attendance to the number of teachers lecturing based on:

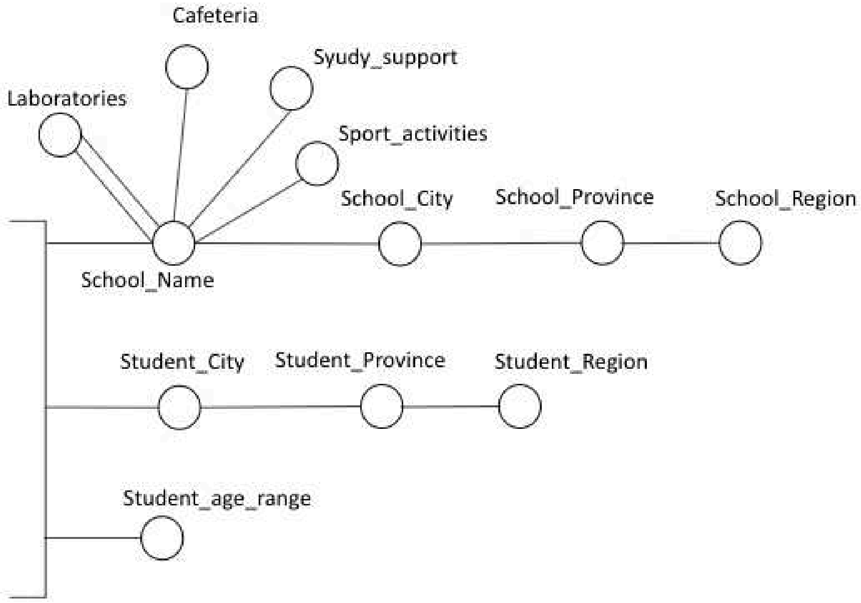
- gender of student
- date, week, month, month of the year, four-month period, semester, year in which the student was present at school, and time slot of attendance at school (one value from the following: 8:00-9:59 a.m., 10:00-12:59 p.m., 1:00-5:00 p.m.)
- school name, characterized by its geographic location (expressed in terms of city, province, region) and all services offered by the school (e.g., cafeteria, study support, afternoon course for sports activities, etc.). Each school is also characterized by the laboratories offered by the school (one or more of the following values: STEM lab, drama lab, music lab)
- city, province and region of the student
- certification of educational need possibly presented by the student (one value from the following: DSA, BSE, none)
- age group of the student (one value from the following: between 6 and 9 years, between 10 and 14 years, more than 14 years)

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

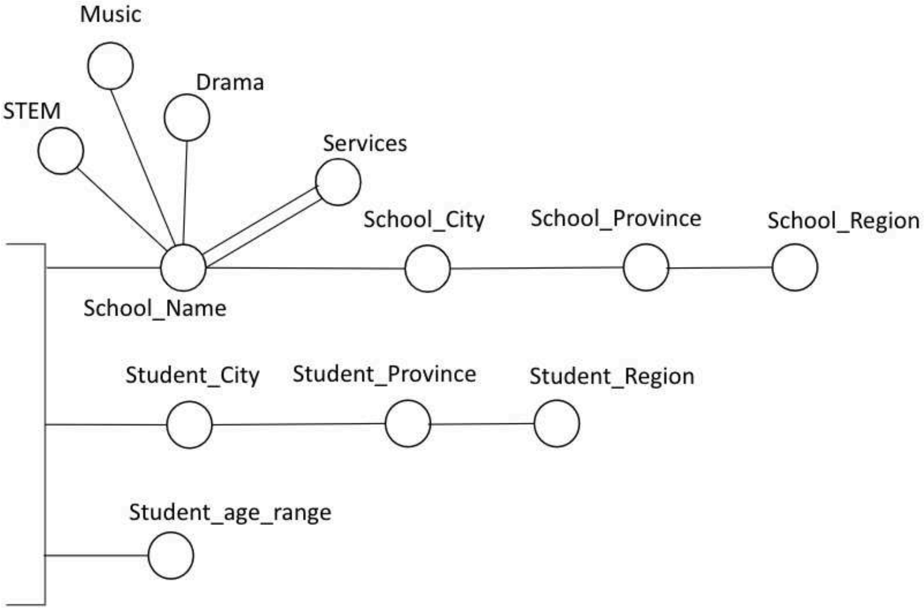
Conceptual Schema 1

1 point (penalty 15% for a wrong answer)

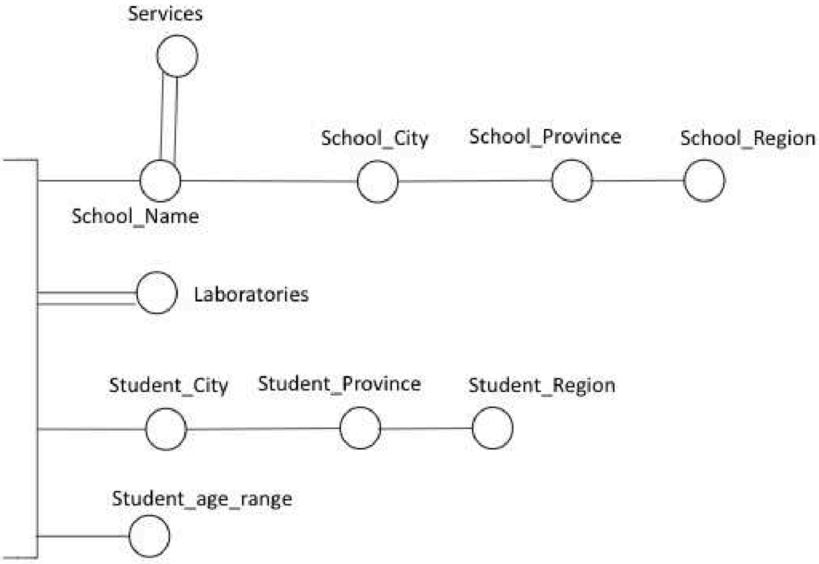
a)


Cafeteria
Syudy_support
Laboratories
Sport_activities
School_City
School_Province
School_Region
School_Name
Student_City
Student_Province
Student_Region
Student_age_range

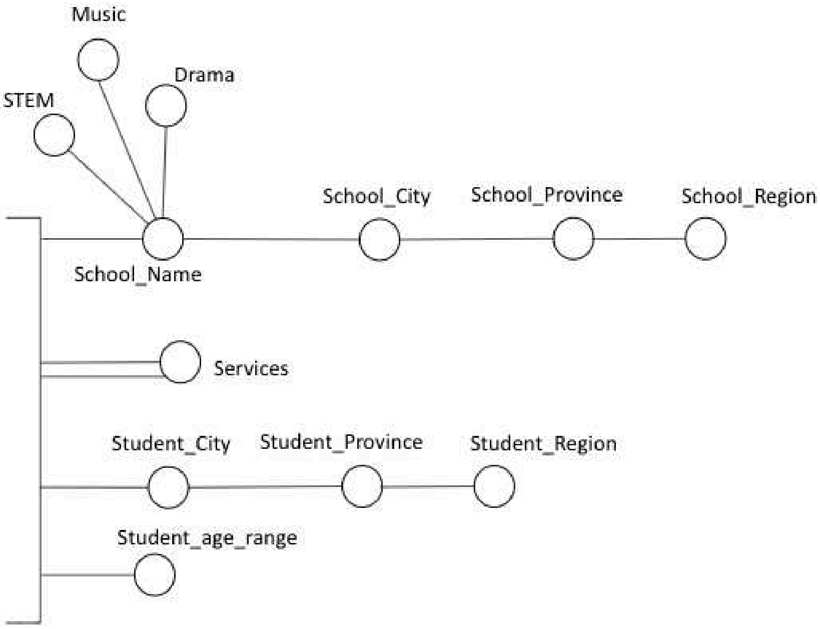
b)


Music
Drama
STEM
Services
School_City
School_Province
School_Region
School_Name
Student_City
Student_Province
Student_Region
Student_age_range

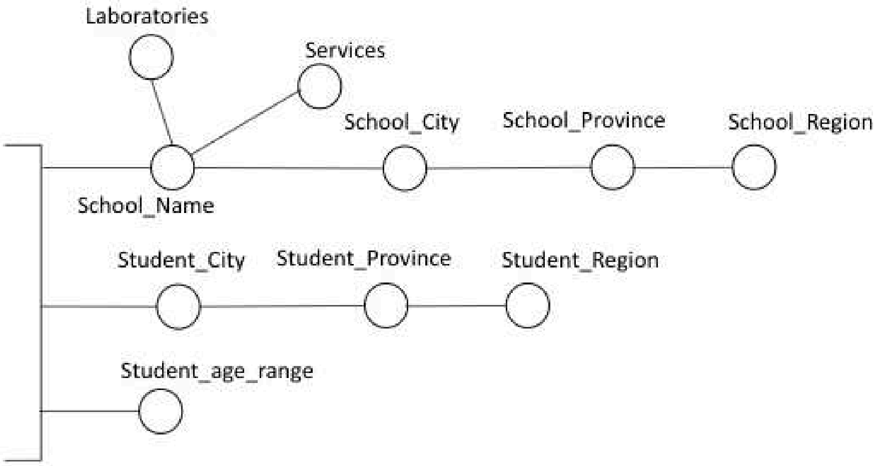
c)


Services
School_City
School_Province
School_Region
School_Name
Laboratories
Student_City
Student_Province
Student_Region
Student_age_range

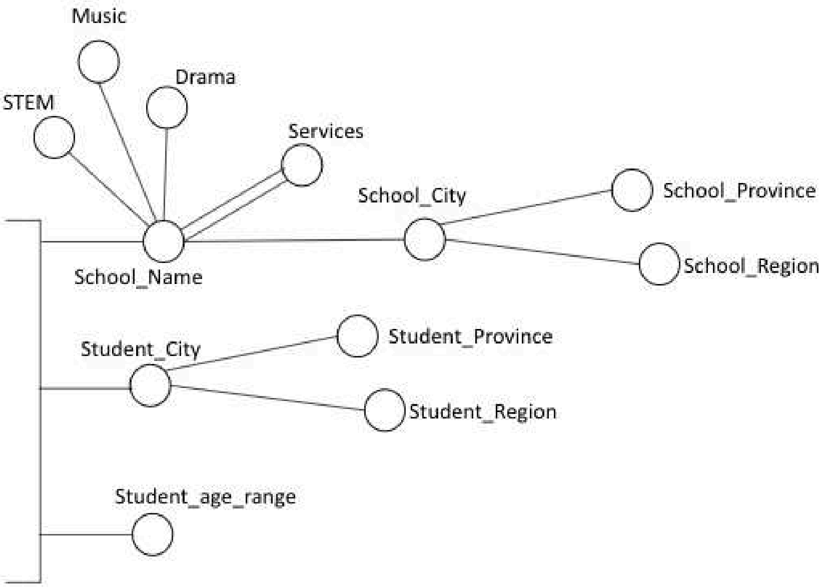
d)


Music
Drama
STEM
School_City
School_Province
School_Region
School_Name
Services
Student_City
Student_Province
Student_Region
Student_age_range

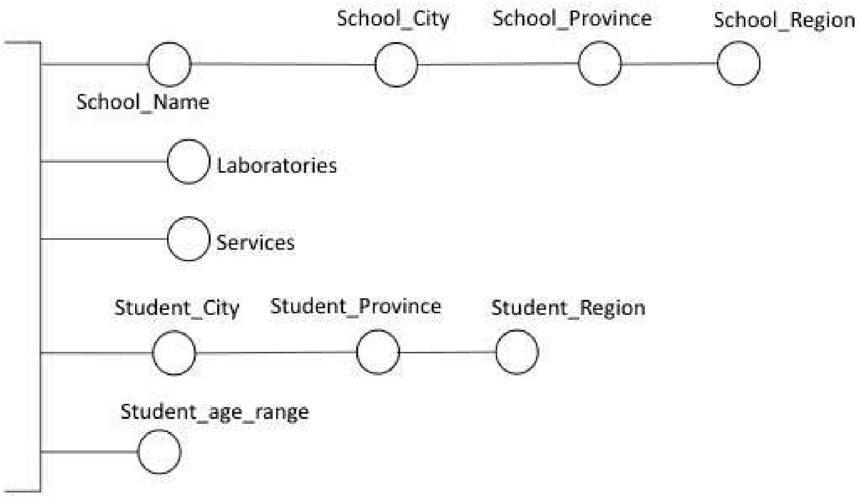
e)


Laboratories
Services
School_City
School_Province
School_Region
School_Name
Student_City
Student_Province
Student_Region
Student_age_range

f)


Music
Drama
STEM
Services
School_City
School_Province
School_Name
School_Region
Student_City
Student_Province
Student_Region
Student_age_range

g)



○ a ● e ✗ ○ g ○ b ○ f ○ d ○ c

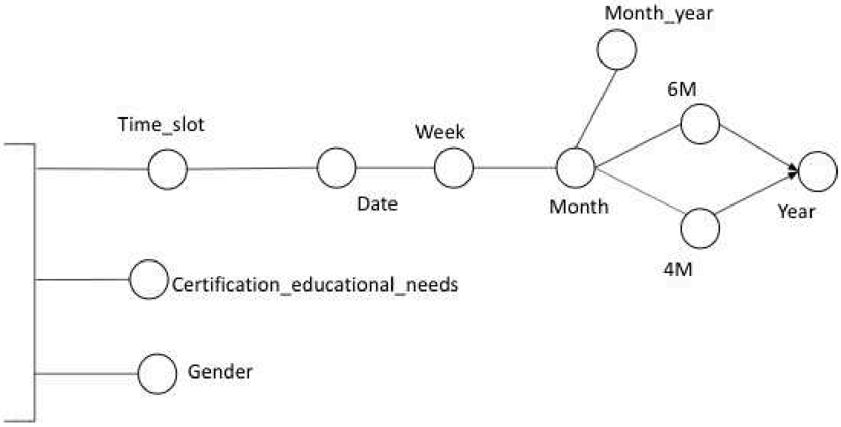
Punteggio ottenuto -0,15 su 1,00

La risposta corretta è: b

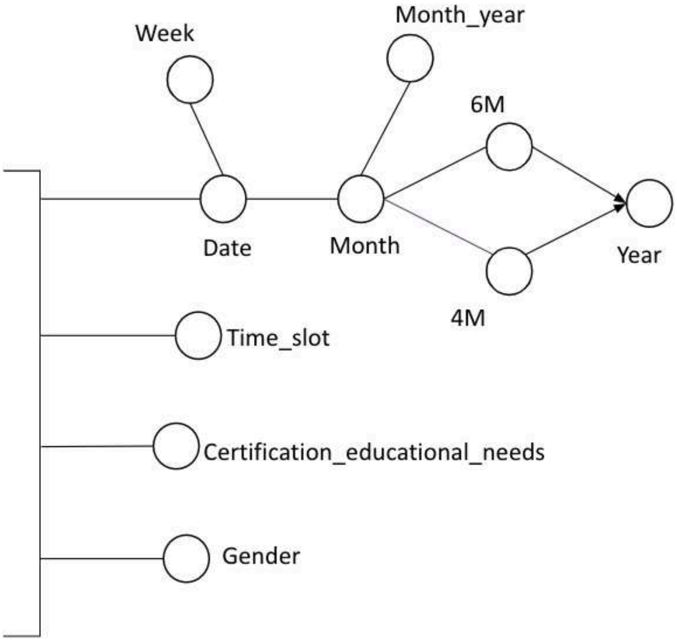
Conceptual Schema 2

1 point (penalty 15% for a wrong answer)

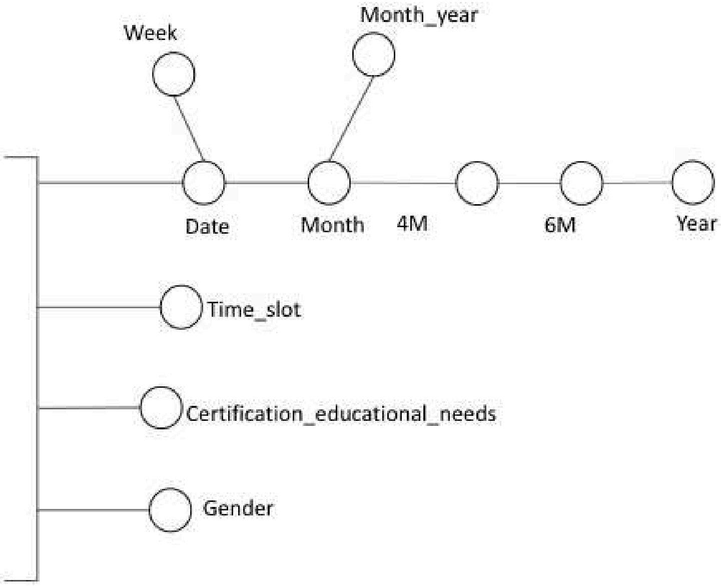
a)



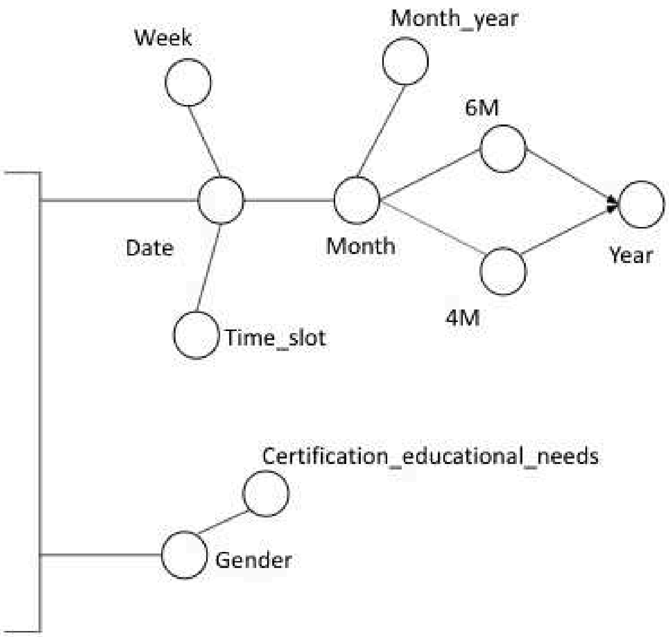
b)


Week
Month_year
6M
Date
Month
Year
4M
Time_slot
Certification_educational_needs
Gender

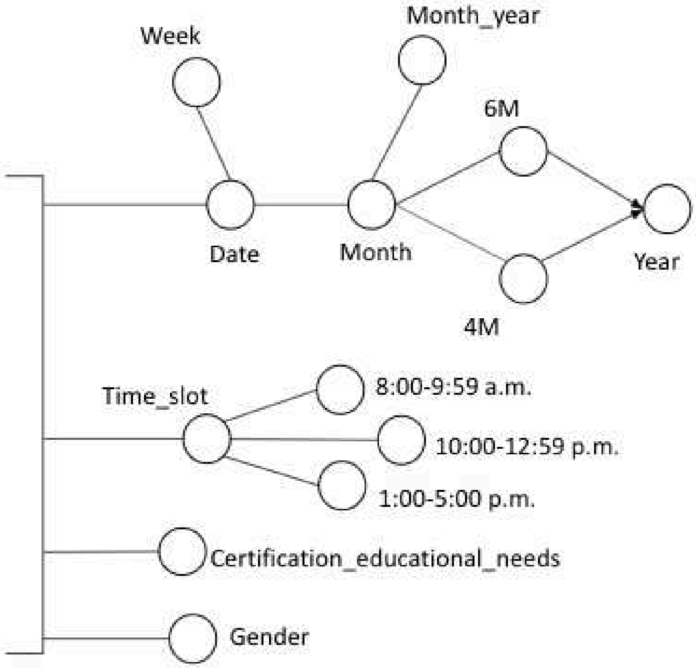
c)


Week
Month_year
Date
Month
4M
6M
Year
Time_slot
Certification_educational_needs
Gender

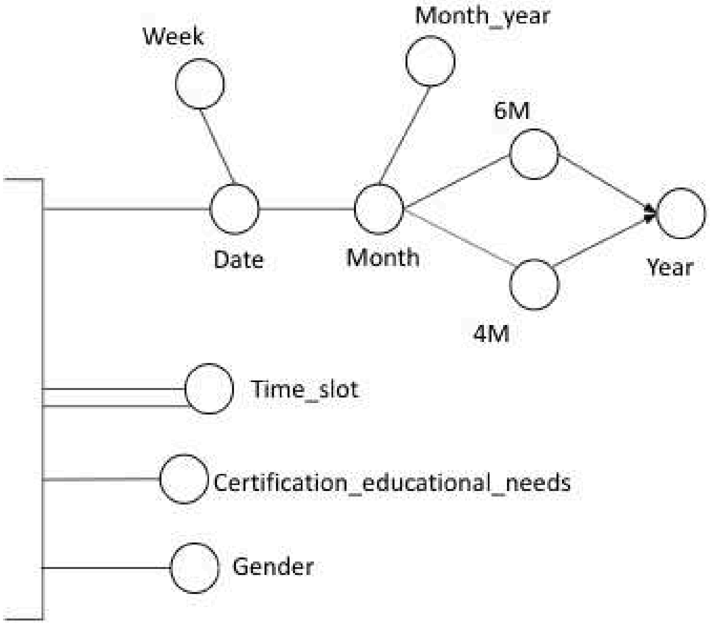
d)


Week
Month_year
6M
Date
Month
Year
4M
Time_slot
Certification_educational_needs
Gender

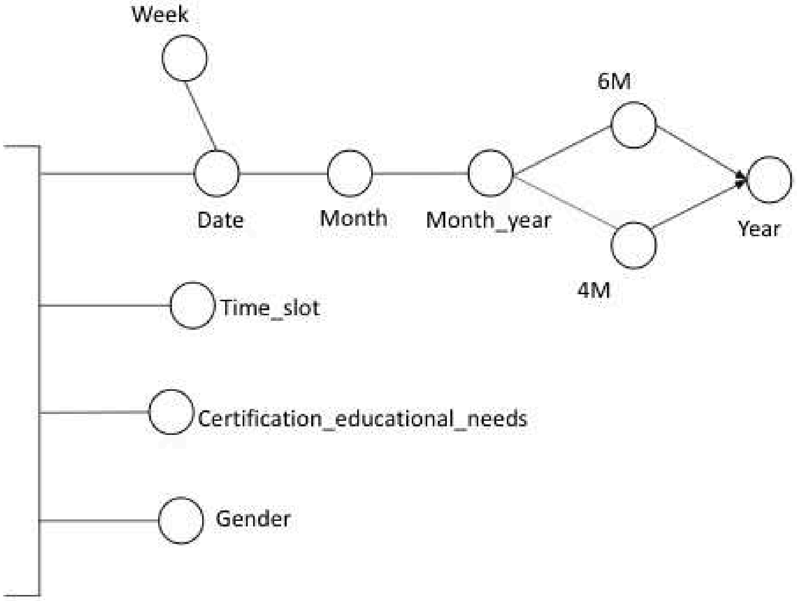
e)


Week
Month_year
6M
Date
Month
Year
4M
Time_slot
8:00-9:59 a.m.
10:00-12:59 p.m.
1:00-5:00 p.m.
Certification_educational_needs
Gender

f)



g)



○ c ○ g ○ e ○ d ○ a ● b✓ ○ f

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: b

Measures

1 point (penalty 15% for a wrong answer)

○ Total number of students in attendance, Total number of class hours, Total number of schools

● Total number of students in attendance, Total number of enrolled students, Total number of teachers lecturing✓

○ Total number of schools, Total number of classes, otal number of class hours

○ Total number of certifications, Total number of enrolled students, Total number of teachers lecturing

○ Total number of students in attendance, Total number of teachers lecturing, Total number of classes

○ Total number of enrolled students,Total number of schools, Average number of students in attendance

○ Total number of class hours, Total number of certifications, Total number of teachers lecturing

○ Total number of students in attendance, Total number of enrolled students, Average number of students in attendance

Punteggio ottenuto 1,00 su 1,00

La risposta corretta è: Total number of students in attendance, Total number of enrolled students, Total number of teachers lecturing

1) La risposta corretta è : b
2) La risposta corretta è : b
3) La risposta corretta è : Total number of students in attendance, Total number of enrolled students, Total number of teachers lecturing

Domanda 7

Risposta errata

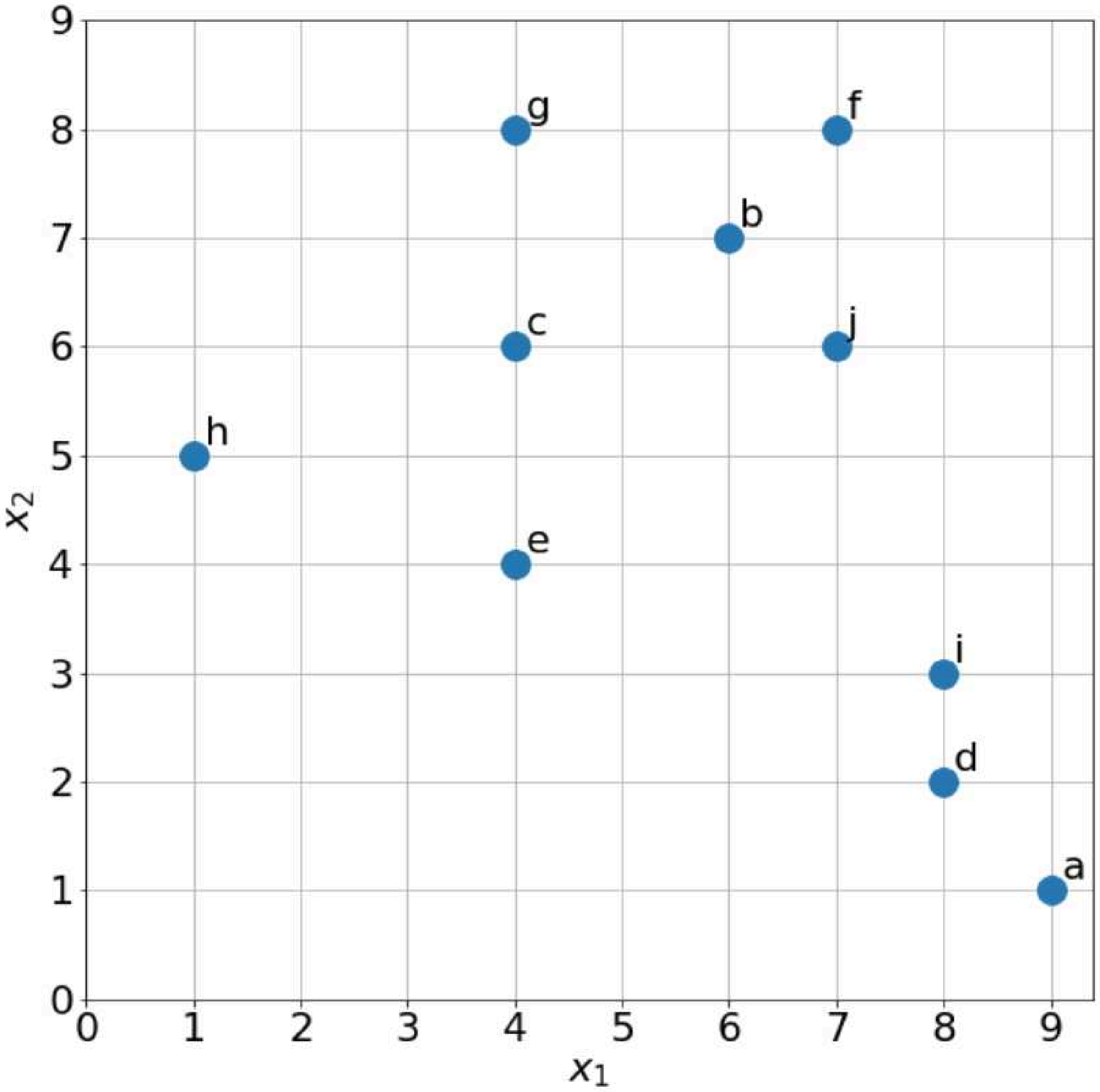
Punteggio ottenuto -0,15 su 1,00

1 point (15% penalty for a wrong answer)

The Manhattan distance (or cityblock) between two points a and b $\in R^n$ is defined as:

$d(a, b) = \sum_{k=1}^{n} |a_k - b_k|$
(k represents the k-th dimension)

DBSCAN is applied to the following two-dimensional dataset. The hyperparameters used are ϵ = 2.5, minpoints = 2 (a point is considered a "core" point if it has at least minpoints points within a neighborhood of radius ϵ. The Manhattan distance is used to calculate the distance between pairs of points.



Each point is assigned a category between noise, border, and core.

What categories are assigned to points h, d, j?

○ (a) h noise, d core, j border

○ (b) h core, d border, j core

○ (c) h noise, d border, j core

○ (d) h noise, d core, j core

○ (e) h core, d core, j core

○ (f) h border, d core, j border

● (g) h core, d core, j border ✗

○ (h) h border, d border, j core

○ (i) None of the other answers is correct.

○ (j) h border, d core, j core

Risposta errata.

La risposta corretta è: h noise, d core, j core

Domanda 8

Completo

Non valutata

This is not an exam question.

You can use the text box below for notes or drafts (for example, to write the intermediate steps of an exercise).

Any comments/feedback for the teacher can be written here.

The text entered in this exercise will not be considered in the exam correction phase.

Comment about class: Most of us really needed the real practice of Query/optimization in the real world. I mean if all students could participate and do it in the real IDE, that would be a huge benefit for us but unfortunatley, there was just the prof fiori and the practice was in the VScode!!

Or the theory topics usage in the real worl projects. I mean doing group projects in the topics related to the course. The level of stress and the homesickness for my family and my country is so high but all in all, the prof and assistants were amazing and I appreciate all their efforts and patience.

Domanda 9

Risposta errata

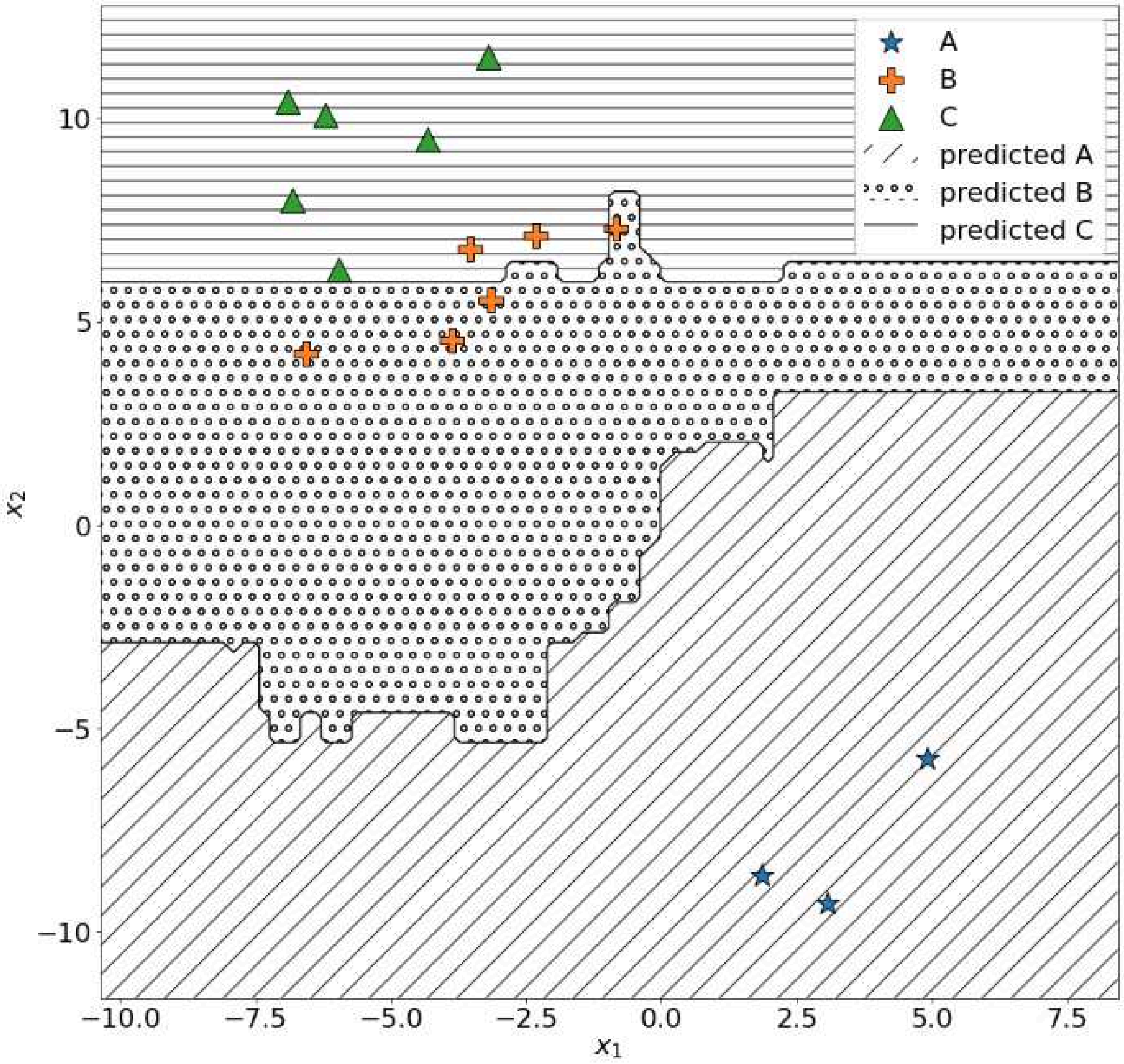
Punteggio ottenuto -0,23 su 1,50

1.5 points (15% penalty for a wrong answer)

Precision(X) is the fraction of correct predictions made for class X out of all predictions made for class X.
Recall(X) is the fraction of correct predictions made for class X out of all points that belong to class X.

A random forest classifier has been trained with a two-dimensional dataset (features x1, x2). Each point in the dataset is labeled with a class label of A, B, or C (represented respectively as star, cross, triangle).

The following figure represents a test set that has been used to validate the classifier.



The labels assigned by the model are shown in the figure:

- The areas with diagonal lines represent zones where the model predicts class A
- The areas with dots represent zones where the model predicts class B
- The areas with horizontal lines represent zones where the model predicts class C

Based on this information, what is the precision for class C and recall for class B?

○ (a) precision C: 1, recall B: 1

○ (b) precision C: 2/3, recall B: 3/4

○ (c) precision C: 2/3 recall B: 1

● (d) precision C: 3/4, recall B: 1 ✗

○ (e) precision C: 1, recall B: 2/3

○ (f) None of the other answers is correct.

○ (g) precision C: 1, recall B: 3/4

Risposta errata.

La risposta corretta è: None of the other answers is correct.

Domanda 10

Risposta errata

Punteggio ottenuto -0,15 su 1,00

1 point (15% penalty for a wrong answer)

The following sequence of operations is given in a log file:

```
B(T0) I0(o4) CK(T0) I0(o0) B(T1) Commit(T1) B(T2) B(T3) U0(o2) D0(o4) U0(o2) U0(o4) I0(o2) D3(o0)
U3(o4) FAILURE
```

Notation:

- Tn: Id of transaction n
- B(Tn): Begin of transaction Tn
- CK(Ta,Tb,...): checkpoint with transactions Ta, Tb, … not completed
- Commit(Tn): commit of transaction Tn
- Abort(Tn): abort (rollback) of transaction Tn
- Un(ox): update executed by transaction Tn on object ox
- In(ox): insert executed by transaction Tn on object ox
- Dn(ox): delete executed by transaction Tn on object ox

What are the **final** UNDO and REDO sets for the warm restart?

● (a) undo: [T0, T1,T3], redo: [T2] ✗

○ (b) undo: [T2], redo[T3, T1]

○ (c) undo: [T0, T2, T3], redo: [T1]

○ (d) undo: [T0, T1, T2, T3], redo: []

○ (e) undo: [T1,T3], redo: [T2]

○ (f) undo: [T2, T3], redo: [T1]

○ (g) undo: [T0, T2], redo[T3, T1]

○ (h) undo: [T1, T2, T3], redo: []

Risposta errata.

La risposta corretta è: undo: [T0, T2, T3], redo: [T1]

Domanda 11

Completo

Punteggio ottenuto 0,00 su 5,00

5 points (no penalty for a wrong answer)

The following data warehouse describes the subscription trends of insurance policies offered by different companies through distribution networks (branches, brokers, agencies, etc.). Each insurance product is issued by a specific company and pertains to only one branch of insurance (life, vandalism, MTPL, etc.). For each distribution network, the city, province, region, and corresponding geographic area are known. The data warehouse stores the gender, civil status, profession, and age group of customers who have subscribed to insurance products over time through the different distribution networks.

The metrics to be analyzed are the number of subscriptions, total duration, and the corresponding premium amount. Metrics should be analyzed for each month, two-month, three-month, six-month, year, three-year period, and five-year period.

DISTRIBUTION-NETWORK (**IDDistributionNetwork**, DistributionNetwork, City, Province, Region, GeographicArea)
INSURANCE-PRODUCT (**IDInsuranceProduct**, InsuranceProduct, Company, InsuranceBranch)
CUSTOMER-FEATURES (**IDCF**, AgeRange, Gender, CivilStatus, Profession)
TIME (**IDTime**, Month, 2-Months, 3-Months, 6-Months, Year, 3-Years, 5-Years)
SUBSCRIPTIONS (**IDDistributionNetwork**, **IDInsuranceProduct**, **IDCF**, **IDTime**, #Subscriptions, PremiumsAmount, Duration)

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

a. Separately by quarter (3-Months attribute) and profession, show the total amount of premiums and the average duration of subscriptions.

b. Considering female clients (value 'F' of the Gender attribute), separately by year and distribution geographic area, show the total amount and the average 3-month amount of premiums.

c. For subscriptions made in the geographical area 'North,' show the total number of subscriptions and the corresponding average duration, separately by distribution region and year.

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 ViewSubscriptions

BUILD IMMEDIATE

REFRESH FAST ON COMMIT

AS
```

Block A

2. Define the **minimal set** of attributes that allows identification of the tuples belonging to the materialized ViewSubscriptions view.

3. Assume that the management of the materialized view (derived table) is carried out by means of triggers. Write the trigger to propagate to the ViewSubscriptions materialized view the changes due to the insertion of a new record into the SUBSCRIPTIONS table.

1-a)

```
SELECT SUM(PremiumsAmount), AVG(Duration)
JOIN TIME, CUSTOMER-FEATURES
FROM TIME, CUSTOMER-FEATURES
GROUP BY 3-Months, Profession
```

1. Block A – Query for materialized view

```
SELECT Gender, Profession, 3-Months, Year, Region, GeographicArea, SUM(PremiumsAmount) AS TotPremiumsAmount, SUM(Duration) AS TotDuration, SUM(#Subscriptions) AS NumTotSubscriptions

FROM CUSTOMER-FEATURES C, DISTRIBUTION-NETWORK D, TIME T, SUBSCRIPTIONS S

WHERE C.IDCF = S.IDCF AND D.IDDistributionNetwork = S.IDDistributionNetwork AND T.IDTime = S.IDTime

GROUP BY Gender, Profession, 3-Months, Year, Region, GeographicArea
```

2. Identifier

Gender, Profession, 3-Months, Region

3. Trigger

```
CREATE OR REPLACE TRIGGER MaintenanceViewSubscriptions

AFTER INSERT ON SUBSCRIPTIONS

FOR EACH ROW

DECLARE
```

```
VarYear, Var3M DATE;
VarGender, VarProfession, VarRegion, VarGeoArea  VARCHAR(10);
N INTEGER;
BEGIN
 SELECT 3-Months, Year INTO Var3M, VarYear
 FROM TIME
 WHERE IDTime = :NEW.IDTime;

 SELECT Gender, Profession INTO VarGender, VarProfession
 FROM CUSTOMER-FEATURES
 WHERE IDCF = :NEW.IDCF;

SELECT Region, GeographicArea INTO VarRegion,  VarGeoArea
 FROM DISTRIBUTION-NETWORK
 WHERE IDDistributionNetwork = :NEW.IDDistributionNetwork;

SELECT COUNT(*) INTO N
FROM ViewSubscriptions
WHERE 3-Months = Var3M AND Gender = VarGender AND Profession = VarProfession AND
Region = VarRegion;

IF N>0 THEN
  UPDATE ViewSubscriptions
  SET TotPremiumsAmount = TotPremiumsAmount + :NEW.PremiumsAmount, TotDuration =
TotDuration + :NEW.Duration,
NumTotSubscriptions = NumTotSubscriptions + :NEW.#Subscriptions
WHERE 3-Months = Var3M AND Gender = VarGender AND Profession = VarProfession AND
Region = VarRegion;
ELSE
  INSERT INTO ViewSubscriptions(…) VALUES (VarGender, VarProfession, Var3M, varYear,
VarRegion,  VarGeoArea,
                  :NEW.PremiumsAmount, :NEW.Duration, :NEW.#Subscriptions);
END IF;
END;
```

Commento:

Solutions for points 1-3 are missing