Data management and visualization

Iniziato	giovedì, 25 febbraio 2021, 07:07
Stato	Completato
Terminato	giovedì, 25 febbraio 2021, 07:08
Tempo impiegato	18 secondi
Valutazione	**0,00** su un massimo di 31,00 (**0**%)

Domanda 1

Risposta non data

Punteggio max.: 1,50

During the ETL process, the correct order of data warehouse loading is:

- (a) update indices, then dimensions, tables, and finally materialized views
- (b) update materialized views, then indices, tables, and finally dimensions
- (c) update dimensions, then tables, finally materialized views and indices
- (d) update tables, then dimensions, and finally materialized views and indices
- (e) update materialized views, then indices, dimensions, and finally tables

Risposta errata.

La risposta corretta è: update dimensions, then tables, finally materialized views and indices

Domanda 2

Risposta non data

Punteggio max.: 1,00

The approximation pattern has the advantage of:

- (a) reduction in the overall number of documents in a collection
- (b) improvement of performance when there are a lot of join operations
- (c) fewer writes to the database
- (d) reduction in the overall size of the working set

Risposta errata.

La risposta corretta è: fewer writes to the database

Domanda 3

Risposta non data

Punteggio max.: 1,00

In the aggregation pipeline, which stage operator is used to execute a recursive search on a collection:

- (a) $group
- (b) $graphLookup
- (c) $project
- (d) $match

Risposta errata.

La risposta corretta è: $graphLookup

Domanda 4

Risposta non data

Punteggio max.: 1,50

Which one of the following visualizations is the most appropriate one for representing a measure as a statistical distribution, a dimension with a high cardinality and another dimension with a low cardinality? For example, think about a visualization representing incomes (measure), level of education (dimension with high cardinality), and gender (dimension with low cardinality).

- (a) Heatmaps
- (b) Multiple box plots
- (c) Gauges
- (d) Stacked bars
- (e) Pie charts

Risposta errata.

La risposta corretta è: Multiple box plots

Domanda 5

Risposta non data

Punteggio max.: 0,50

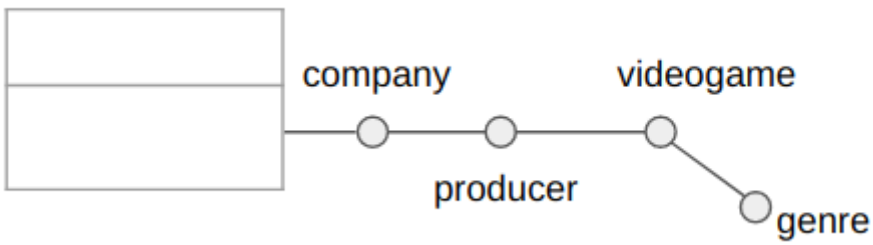
Data analysts of the video game industry are interested in analyzing some metrics for different video games.
Their original system records the video games sold in all stores: they know how many video games are sold, in which store, when, at which price, and some customer information.
The data warehouse must be designed to efficiently analyze the **average revenue for each video game purchase,** according to the following dimensions.

- A video game has a unique name and is produced by a **producer**. A video game producer may design and produce many video games.
- Each video game producer belongs to a video game **company**. A video game company can have different producers and video games.
- Each video game has a specific **genre**. There are 5 possible types of genres: “action”, “adventure” , “simulation”, “fantasy”, and “educational”.
- **Stores** are identified by a unique name. They are analyzed according to their **city** and **country**. A store is located in a specific city. In a city there can be different stores.
- Each store may sell some additional **articles**. There are 4 possible types of additional articles: “collectable”, “toys”, “manga” and “accessories”. The systems records which types of additional articles are sold by each store. For instance, store X can sell “toys” and “manga” only, whereas store Y sells “collectable”, “toys”, and “manga”.

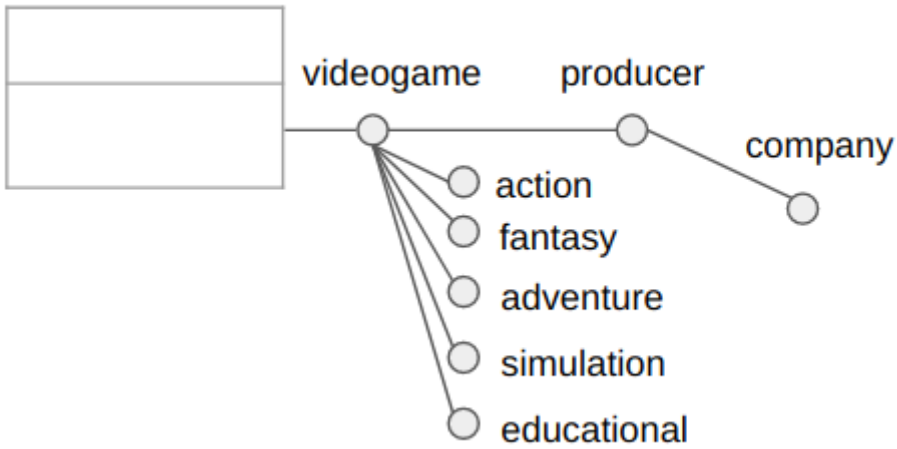
The customer **age group** (18-30, 31-50, >50 years old) is also required.

Select, among the following proposed dimensions, those that meet the requirements described in the problem specification (at most one answer is correct).

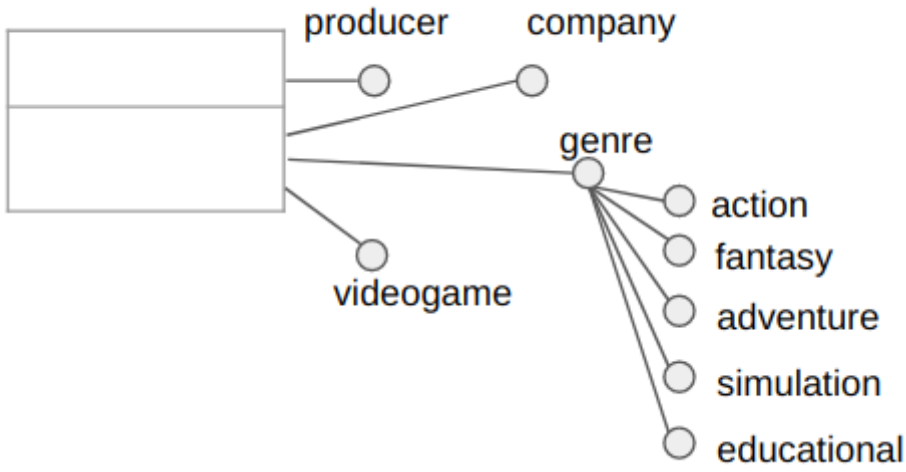
(a)



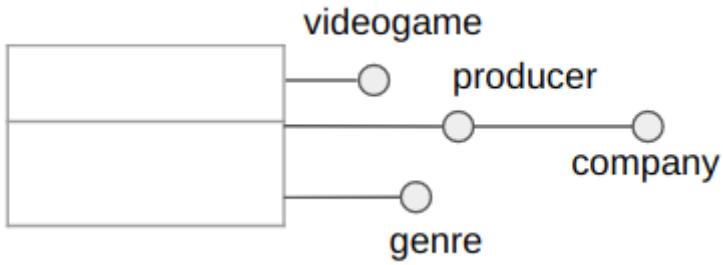
(b)



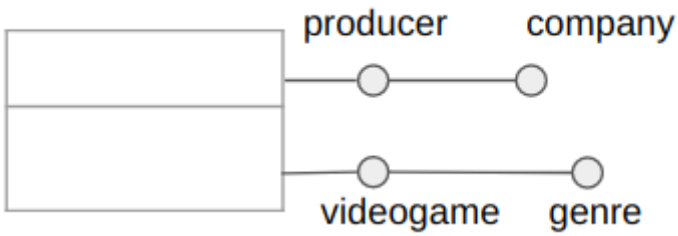
(c)


producer
company
genre
action
fantasy
adventure
simulation
educational
videogame

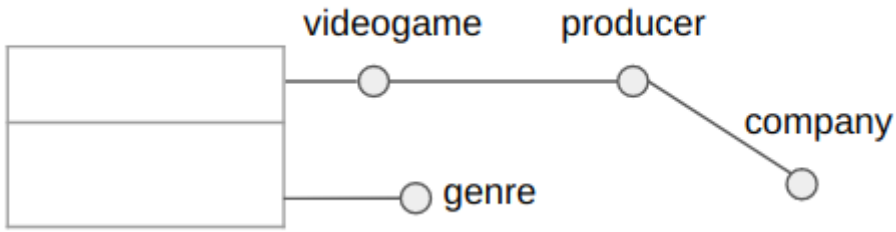
(d)


videogame
producer
company
genre

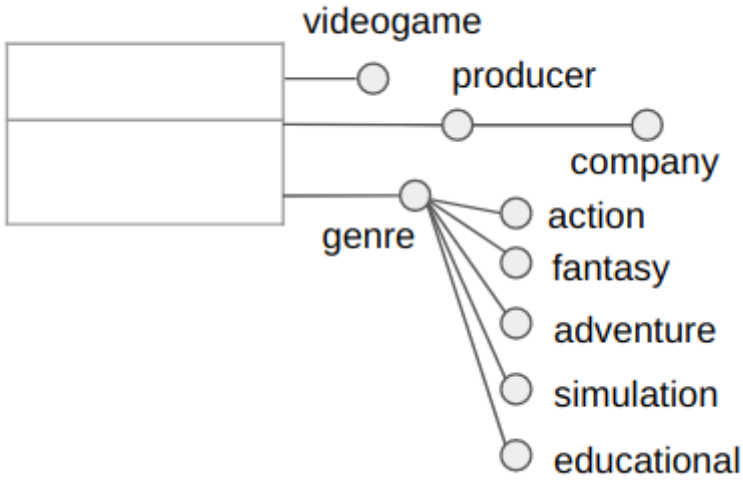
(e)


producer
company
videogame
genre

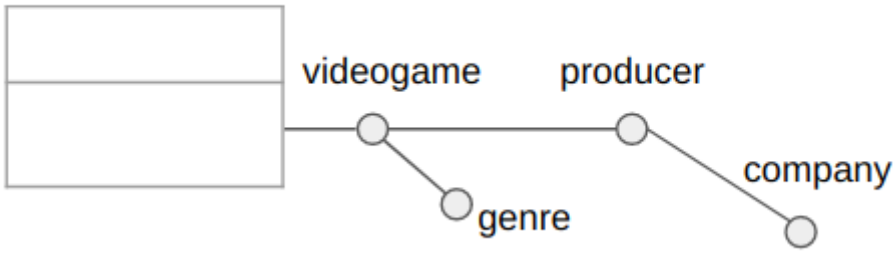
(f)


videogame
producer
company
genre

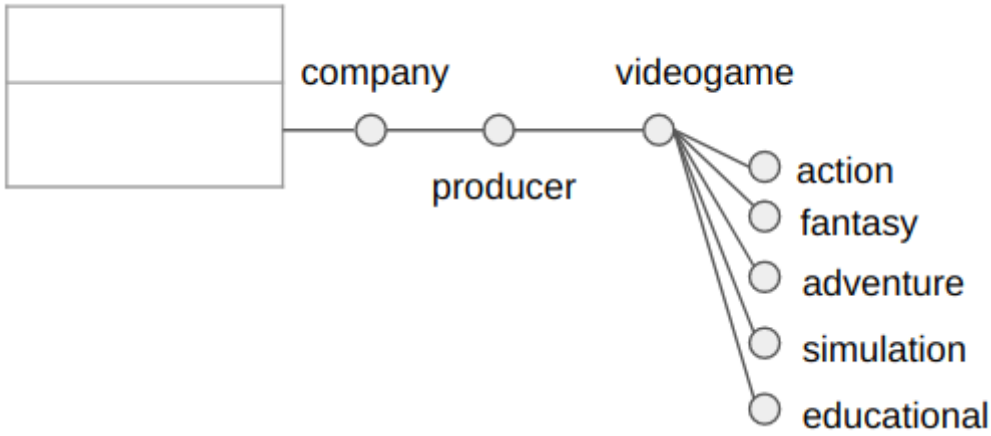
(g)


videogame
producer
company
genre
action
fantasy
adventure
simulation
educational

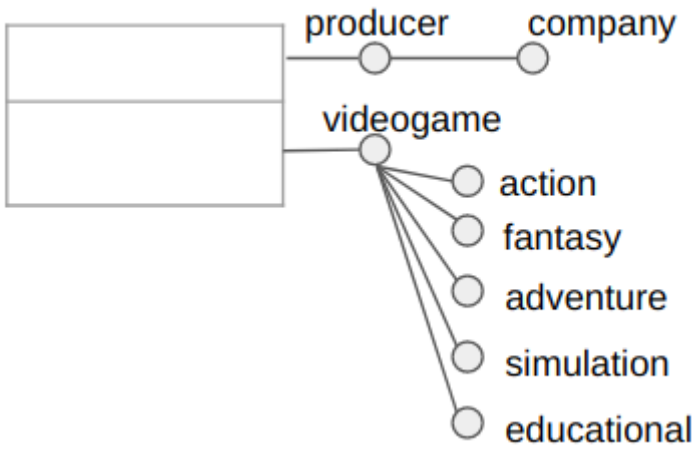
(h)


videogame
producer
company
genre

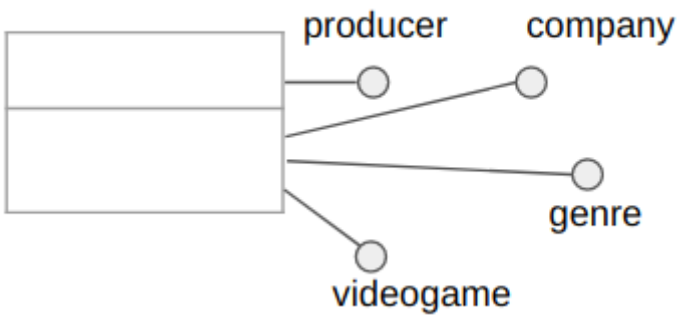
(i)


company
videogame
producer
action
fantasy
adventure
simulation
educational

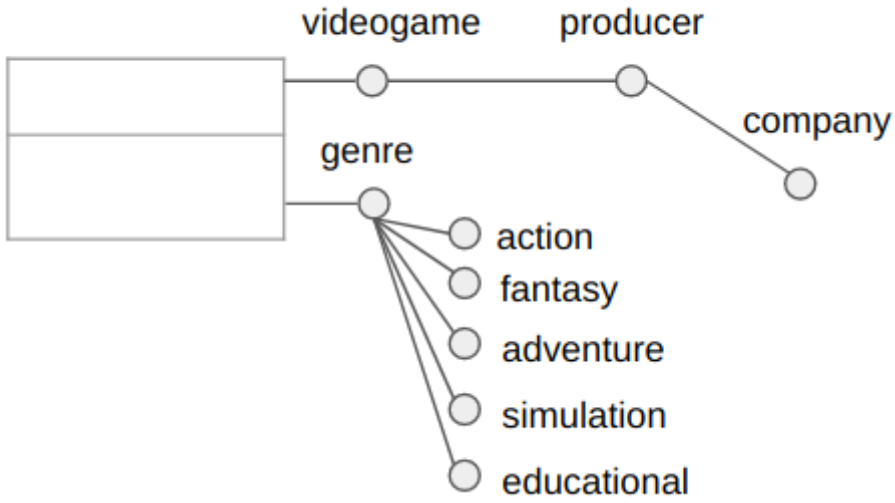
(j)


producer
company
videogame
action
fantasy
adventure
simulation
educational

(k)

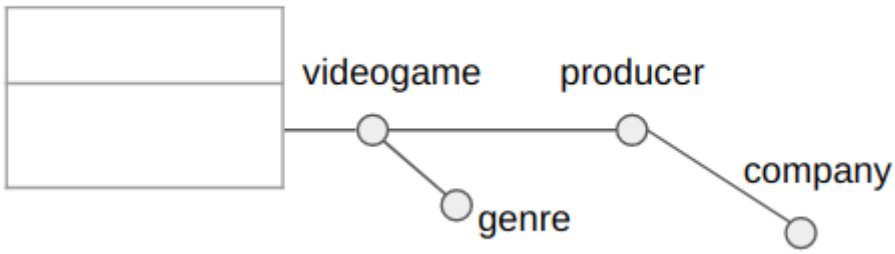

producer
company
genre
videogame

(I)



Risposta errata.

La risposta corretta è:



Domanda 6

Risposta non data

Punteggio max.: 0,50

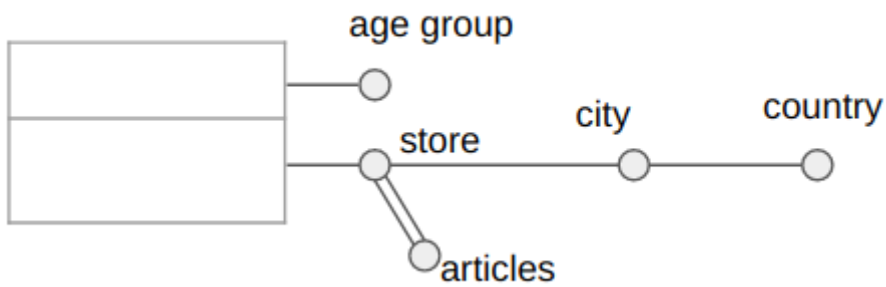
Data analysts of the video game industry are interested in analyzing some metrics for different video games.
Their original system records the video games sold in all stores: they know how many video games are sold, in which store, when, at which price, and some customer information.
The data warehouse must be designed to efficiently analyze the **average revenue for each video game purchase,** according to the following dimensions.

- A video game has a unique name and is produced by a **producer**. A video game producer may design and produce many video games.
- Each video game producer belongs to a video game **company**. A video game company can have different producers and video games.
- Each video game has a specific **genre**. There are 5 possible types of genres: “action”, “adventure” , “simulation”, “fantasy”, and “educational”.
- **Stores** are identified by a unique name. They are analyzed according to their **city** and **country**. A store is located in a specific city. In a city there can be different stores.
- Each store may sell some additional **articles**. There are 4 possible types of additional articles: “collectable”, “toys”, “manga” and “accessories”. The systems records which types of additional articles are sold by each store. For instance, store X can sell “toys” and “manga” only, whereas store Y sells “collectable”, “toys”, and “manga”.

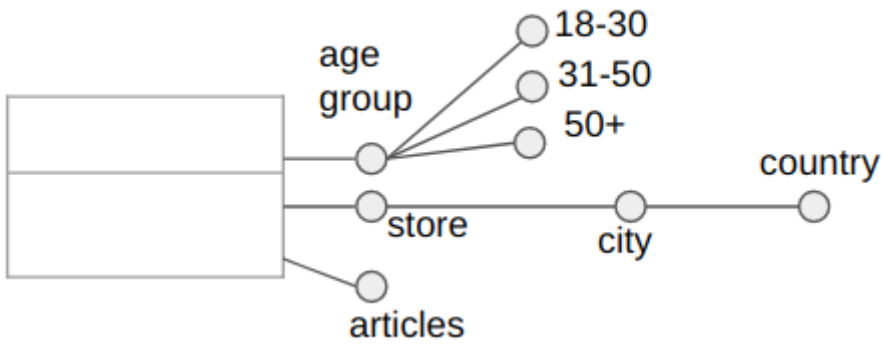
The customer **age group** (18-30, 31-50, >50 years old) is also required.

Select, among the following proposed dimensions, those that meet the requirements described in the problem specification (at most one answer is correct).

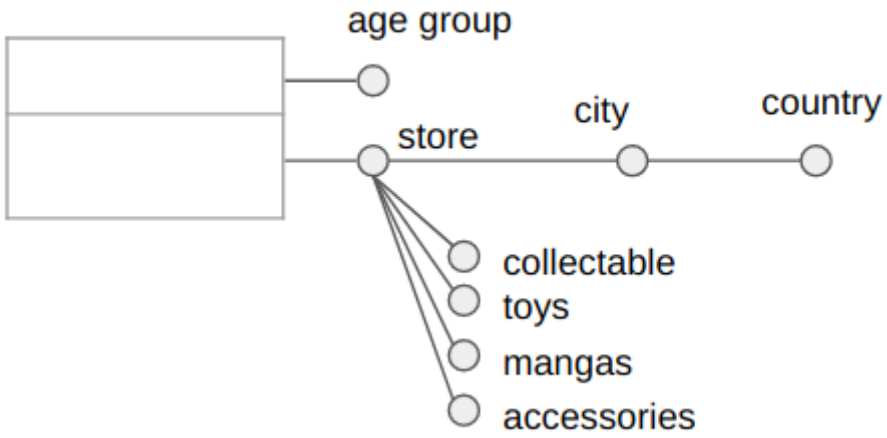
(a)



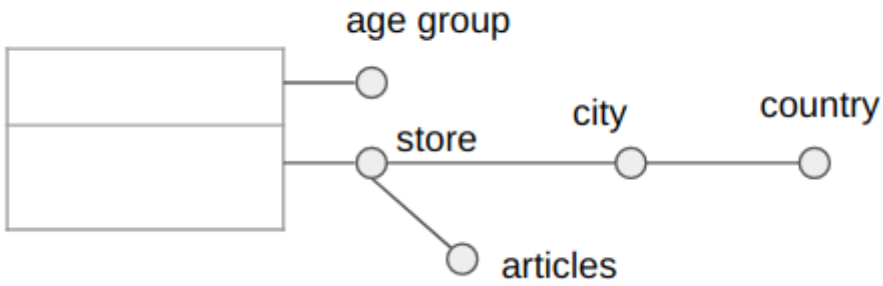
(b)



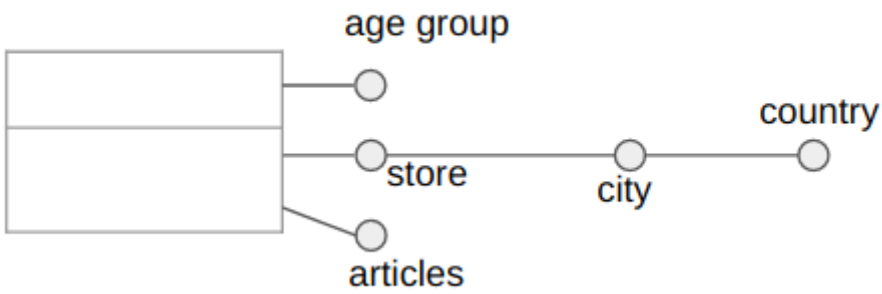
(c)


age group
store
city
country
collectable
toys
mangas
accessories

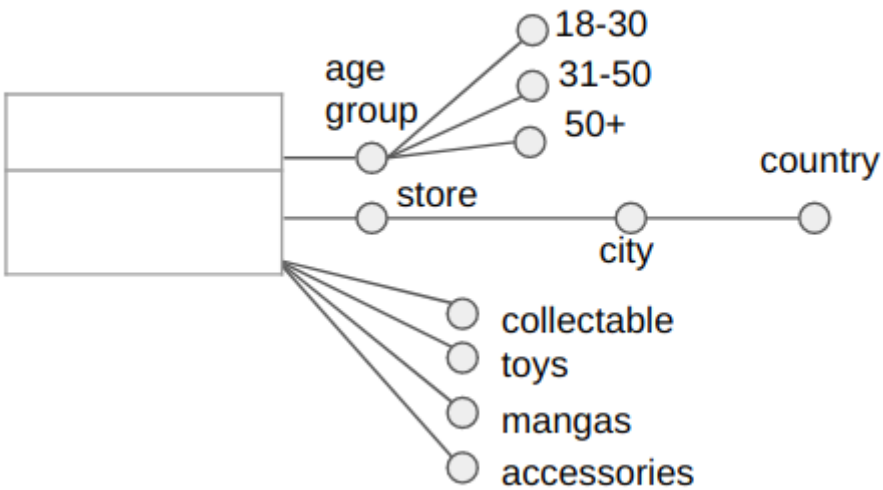
(d)


age group
store
city
country
articles

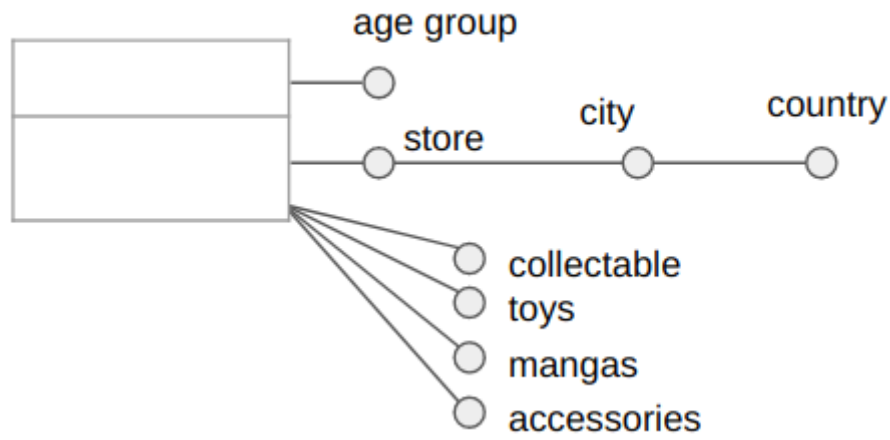
(e)


age group
store
city
country
articles

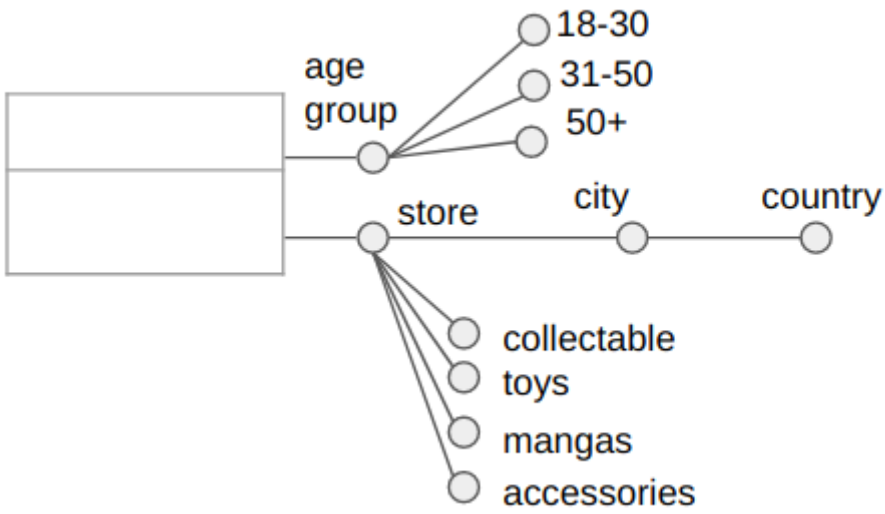
(f)


18-30
age
group
31-50
50+
store
city
country
collectable
toys
mangas
accessories

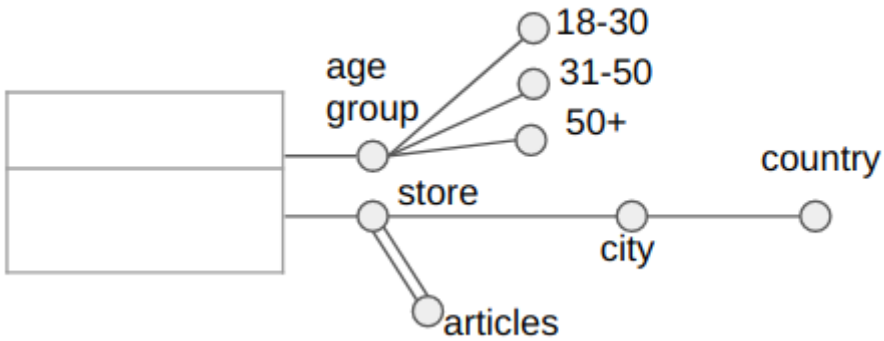
(g)



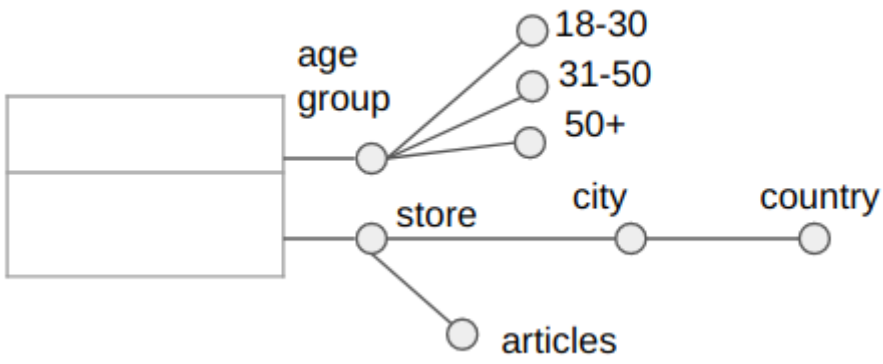
(h)



(i)

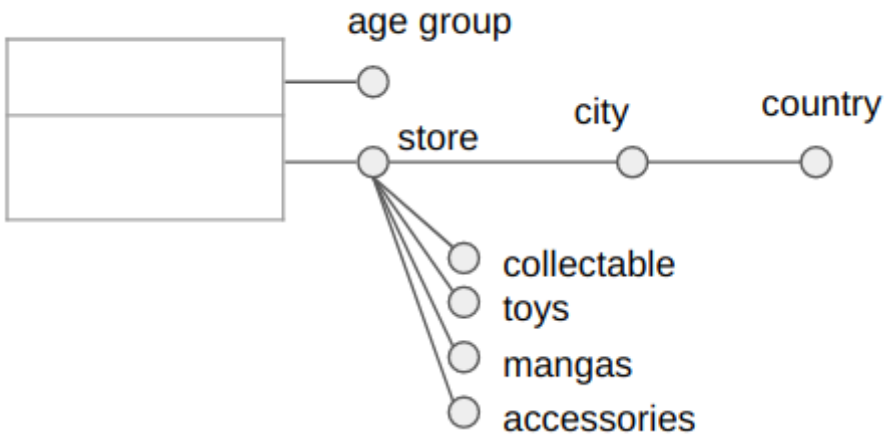


(j)



Risposta errata.

La risposta corretta è:



Domanda 7

Risposta non data

Punteggio max.: 1,00

Data analysts of the video game industry are interested in analyzing some metrics for different video games.
Their original system records the video games sold in all stores: they know how many video games are sold, in which store, when, at which price, and some customer information.
The data warehouse must be designed to efficiently analyze the **average revenue for each video game purchase,** according to the following dimensions.

- A video game has a unique name and is produced by a **producer**. A video game producer may design and produce many video games.
- Each video game producer belongs to a video game **company**. A video game company can have different producers and video games.
- Each video game has a specific **genre**. There are 5 possible types of genres: “action”, “adventure” , “simulation”, “fantasy”, and “educational”.
- **Stores** are identified by a unique name. They are analyzed according to their **city** and **country**. A store is located in a specific city. In a city there can be different stores.
- Each store may sell some additional **articles**. There are 4 possible types of additional articles: “collectable”, “toys”, “manga” and “accessories”. The systems records which types of additional articles are sold by each store. For instance, store X can sell “toys” and “manga” only, whereas store Y sells “collectable”, “toys”, and “manga”.
- The customer **age group** (18-30, 31-50, >50 years old) is also required.

Select all and only the required measures of the fact table in the conceptual schema design among the following (multiple choice question). Hint: do consider the dimensions defined by the previous answers.

Scegli una o più alternative:

- [ ] (a) Average revenues per video game
- [ ] (b) Average number of video games
- [ ] (c) Total number of customers
- [ ] (d) Total number of stores
- [ ] (e) Average revenue for each video game purchase
- [ ] (f) Total number of video game purchases
- [ ] (g) Total revenues of the producer
- [ ] (h) Average price of the video game
- [ ] (i) Total number of producers
- [ ] (j) Total age of the customers
- [ ] (k) Total number of different video games
- [ ] (l) Total revenues
- [ ] (m) Total revenues of the store

Risposta errata.

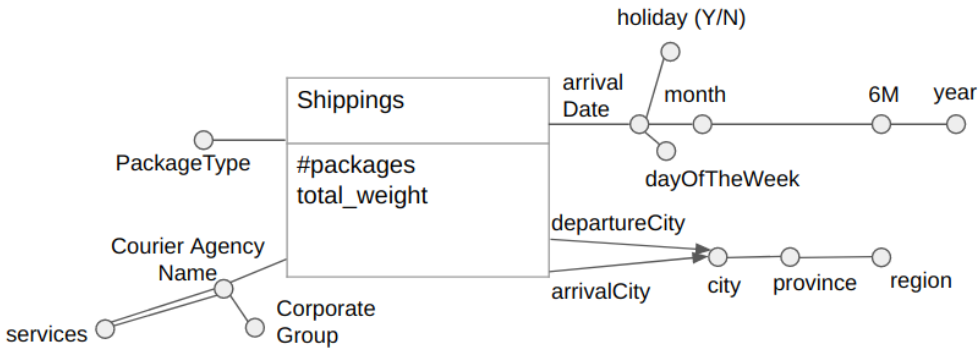
La risposta corretta è: Total revenues, Total number of video game purchases

Domanda 8

Risposta non data

Punteggio max.: 2,00

Given the following conceptual schema:



- For each shipping, the departure and arrival cities, provinces and regions are recorded.
- The cardinality of “PackageType” is 3, and it can be “1” for “Small”, “2” for “Medium” and “3” for “Large”.
- For the shipping, the courier agency is recorded. The courier agency has a unique name. Each agency belongs to a Corporate group.
- An agency may have some additional services available. Examples of additional services are “package tracking”, “package insurance” and “notification by SMS”. The systems records which services are available for each agency. The number of possible additional services is large and growing, hence the full list is not known a prior.
- The system stores the arrival date, the day of the week and if the day was an holiday or not. It also stores the month, year and semester.

Write the logical design of the conceptual DW schema indicated in the picture.
Write each table on a new line.
Use the **bold** or the underline for identifying primary-key attributes.

```
CourierAgency(CourierAgencyId, CourierAgencyName,
CorporateGroup)
Services(ServiceId, ServiceName)
Agency-HAS-Services(CourierAgencyId, ServiceId)
Location(LocationId, city, province, region)
Time(TimeId, arrivalDate, dayOfTheWeek, holiday, month,
6M, year)
Shippings(CourierAgencyId, TimeId, ArrivalLocationId,
DepartureLocationId, PackageType, #packages,
total_weight)
```

Domanda 9

Risposta non data

Punteggio max.: 4,00

```
Gardens(TimeId, GardenCenterId, PlantId, numberOfPlants,
revenue)
Time(TimeId, date, month, 2M, 3M, 4M, 6M, year, dayOfTheWeek,
holiday)
GardenCenter(GardenCenterId, GardenCenter, city, province,
region, greenhouse, accessoryShop, gardenShop, parking)
Plant(PlantId, plantSpecies, genus, family, indoor)
```

A plant species can be either an indoor or an outdoor plant. A plant species belongs to only one genus and one genus belong to only one family.
A garden center can have 0 or more services. There are 4 available services: “parking”, “accessories shop”, “gardening shop” and “greenhouse”.
`Indoor, greenhouse, accessoryShop, gardenShop,` and `parking` attributes can be `“True”` or `“False”`.

Separately for each plant species and month, compute the following metrics:

A. the daily average number of plants

B. the monthly percentage of the number of plants of the species with respect to the number of plants of the genus

C. the cumulative total number of plants since the beginning of the year

Write the requested SQL query.

```
SELECT plantSpecies, month,
    SUM(numberOfPlants)/COUNT(DISTINCT date) as A,
    100*SUM(numberOfPlants)/SUM(SUM(numberOfPlants)) OVER
(PARTITION BY genus, month) as B,
      SUM(SUM(numberOfPlants)) OVER (
            PARTITION BY plantId, plantSpecies, year
            ORDER BY month
            ROWS UNBOUNDED PRECEDING) as C
FROM Plant P, Time T, Gardens G
WHERE P.PlantId=G.PlantId AND T.Timeid=G.Timeid
GROUP BY plantId, plantSpecies, month, year, genus
```

Domanda 10

Risposta non data

Punteggio max.: 4,00

```
Gardens(TimeId, GardenCenterId, PlantId, numberOfPlants,
revenue)
Time(TimeId, date, month, 2M, 3M, 4M, 6M, year, dayOfTheWeek,
holiday)
GardenCenter(GardenCenterId, GardenCenter, city, province,
region, greenhouse, accessoryShop, gardenShop, parking)
Plant(PlantId, plantSpecies, genus, family, indoor)
```

A plant species can be either an indoor or an outdoor plant. A plant species belongs to only one genus and one genus belong to only one family.
A garden center can have 0 or more services. There are 4 available services: “parking”, “accessories shop”, “gardening shop” and “greenhouse”.
`Indoor, greenhouse, accessoryShop, gardenShop,` and `parking` attributes can be `“True”` or `“False”`.

Consider only the garden centers having the “parking” service.
Separately for each garden center and plant genus, compute the following metrics:

A. the average revenue per plant
B. the total revenues of the plant family, for each garden center
C. assign a rank to each garden center within its province, based on its total revenues (rank 1st the garden center with the highest revenue in its province for each plant genus)

Write the requested SQL query.

```
SELECT gardenCenter, genus,
    SUM(revenue)/SUM(numberOfPlants) as A,
      SUM(SUM(revenue)) OVER (PARTITION BY family,
gardenCenter) as B,
      RANK() OVER (PARTITION BY province, genus
                   ORDER BY SUM(revenue) DESC) as C
FROM Plant P, Gardens G, GardenCenter GC
WHERE P.PlantId=G.PlantId AND
G.GardenCenterId=GC.GardenCenterId AND parking=True
GROUP BY gardenCenter, genus, family, province
```

Domanda 11

Risposta non data

Punteggio max.: 2,00

Given the following document structure, representing the measurements received by sensors, where each document collects the measures received in one day:

```
{"_id":ObjectId("5553a998e4b02cf7151190b8"),
"start": Date("2021-02-01T00:00:00.000Z"),
"end": Date("2021-02-01T23:00:00.000Z"),
“sensor”:{
        “_id”: 1000,
        "position":{"type":"Point","coordinates":[-47.9,47.6]},
        "elevation":200,
        “city”: “Turin”,
        “country”: “Italy”
},
"temperature":[
        {ts: Date("2021-02-01T00:00:00.000Z"), value: 12},
        {ts: Date("2021-02-01T01:00:00.000Z"), value: 11},
        …
        {ts: Date("2021-02-01T23:00:00.000Z"), value: 9}
],
nTemp: 24,    // total number of elements in the temperature list
sumTemp: 372    // sum of the values of all elements in the temperature list
}
```

Update the document of the sensor with "_id" equal to 1000 by adding a new "temperature" measurement with "value" 16 received at the timestamp "ts" 2021-02-02T01:10:00.000Z.

Also concurrently update the corresponding statistics (i.e., "nTemp" and "sumTemp").

Suppose that the document with "start" attribute equal to "2021-02-02" exists.

N.B. Use the syntax new Date (string) to manage date attributes, e.g., "start": new Date("2021-02-02")

```
db.measures.updateOne(
{
    'sensor._id': 1000,
    'start': new Date("2021-02-02"),
},
{
    $inc:{ nTemp: 1, sumTemp: 16},
    $push: { temperature: { ts: new Date(“2021-02-20 10:00”), value: 16}}
})
```

Domanda 12

Risposta non data

Punteggio max.: 3,00

Given the following document structure, representing the measurements received by sensors, where each document collects the measures received in one day:

```
{"_id":ObjectId("5553a998e4b02cf7151190b8"),
"start": Date("2021-02-01T00:00:00.000Z"),
"end": Date("2021-02-01T23:00:00.000Z"),
“sensor”:{
        “_id”: 1000,
        "position":{"type":"Point","coordinates":[-47.9,47.6]},
        "elevation":200,
        “city”: “Turin”,
        “country”: “Italy”
},
"temperature":[
        {ts: Date("2021-02-01T00:00:00.000Z"), value: 12},
        {ts: Date("2021-02-01T01:00:00.000Z"), value: 11},
        …
        {ts: Date("2021-02-01T23:00:00.000Z"), value: 9}
],
nMeasure: 24,    // total number of elements in the temperature list
totMeasure: 372    // sum of the values of each element in the temperature list
}
```

Considering the sensor located in Italy and the measures received in the month of January 2021, show the sensor id, sensor city and the date in which the average measure of the sensor was greater than or equal to 15.

N.B. Use the syntax new Date (string) to manage date attributes, e.g., new Date("2021-02-01")

```
db.measures.aggregate([
   {$match: {
        “sensor.country”: "Italy",
        “start”: {$gte: new Date (“2021-01-01”)},
        “start”: {$lte: new Date (“2021-01-31”)},
   {$addFields: {
        avg: { $divide: [“$tot”, “$n” ]}
   }},
   { $match : {avg: {$gte: 15}}},
   {$project: { “sensor._id”: 1, “sensor.city”: 1, start: 1}}
])
```

Domanda 13

Risposta non data

Punteggio max.: 4,00

Design a MongoDB database to manage museum exhibitions according to the following requirements.

Museums are characterized by their name, address, a telephone number, and a website (if available). The address consists of geographical coordinates, street name and number, postal code, and city.

The items exhibited in the museums are identified by a progressive number and characterized by a title, a description and the list of author names. The items are categorized as either archaeological finds, or paintings, or sculptures. The database must record all the main features of each item, such as its dimensions (i.e., width, height, weight, etc.). Each feature has at least a name and a value, and possibly a unit of measure. For instance, the main material is a feature of an archaeological find, the geometrical sizes are features of a painting. For each item, the museum to which it belongs must be recorded, with the museum name frequently accessed together with the item itself.

Several exhibitions are hosted in each museum. The exhibition is characterized by a title, a description, the list of curator names. You must record all the items associated with each exhibition, they can be in the order of hundreds. An item can be part of different exhibitions. Moreover, each exhibition can be hosted by several museums in different periods. You must record the start and end dates of each exhibition in each museum.

Given an item, the database must be designed to efficiently provide the name of the museum that owns it.

Given an exhibition, the database must be designed to efficiently provide the name of the museum and the geographical coordinates where it has been hosted.

Furthermore, given an exhibition, the list of items included in the exhibition and their number must be efficiently returned.

Write a sample document for each collection of the database.

Explicitly indicate the design patterns used.

Museum

```
{
_id: ObjectId(),
name: <string>,
address: {
        street: <string>,
        number: <string|number>,
        postal_code: <number>,
        city: <string>,
        province: <string>,
        geo_ref: {type: <string>, coordinates: [ <number> ] }
}
tel: <string>,
website: <string> // optional
}
```

Items

```
{
_id: <number>,
title: <string>,
description:<string>,
authors: [ <string> ],
category: <string>,
features: [ {k: <string>, v: <string>, u: <string>} ],
museum: {
      _id: itemId(),
      name: <string>
      }
}
```

Exhibition

```
{
_id: ObjectId(),
title: <string>,
description:<string>,
curators: [ <string> ],
events: [
        {start: <date>,
        end: <date>,
        museum:{
                _id: itemId(),
                name: <string>,
                geo_ref: {type: <string>, coordinates: [ <number> ] }
        }
],
items: [<number> ] // _id of items
}
```

Pattern used:

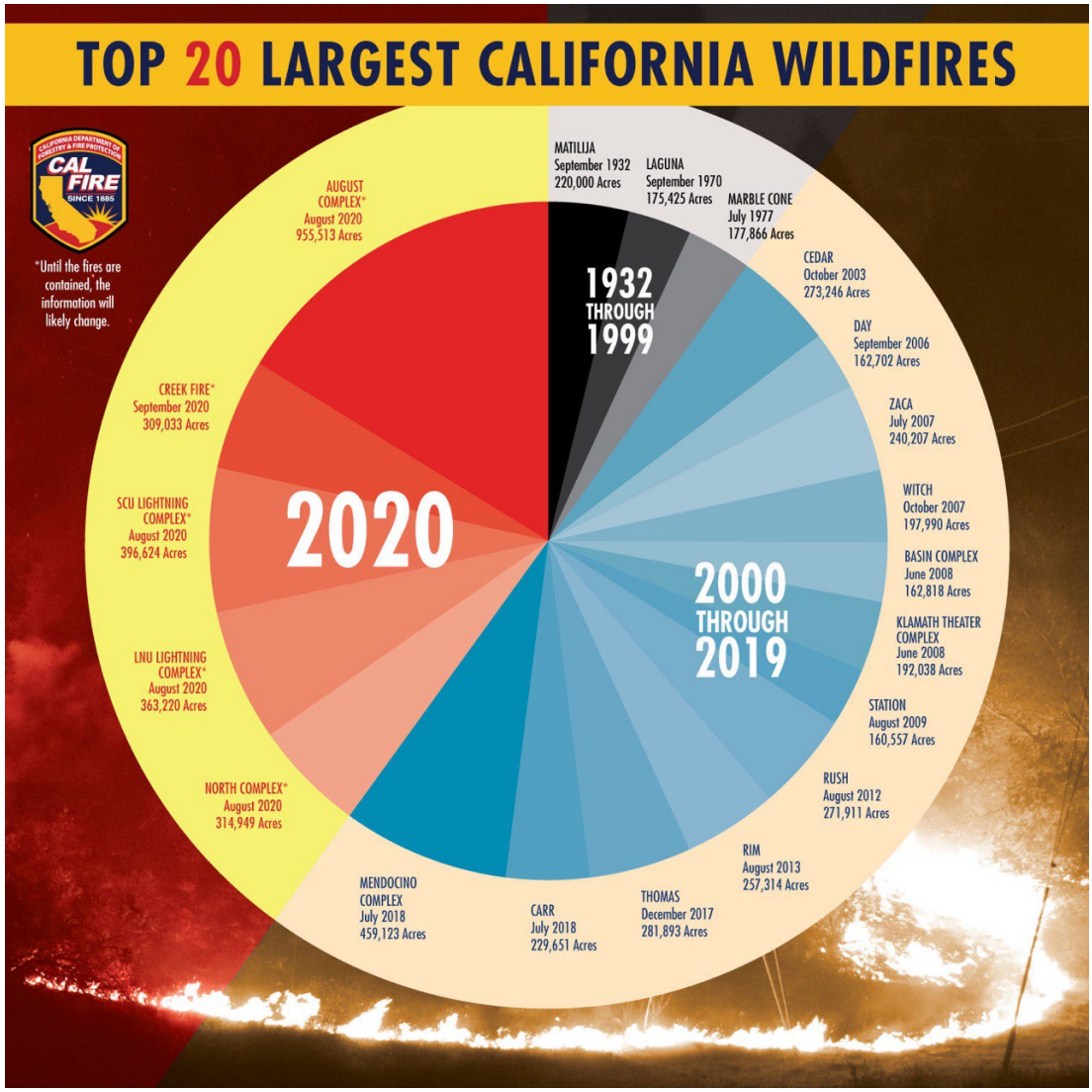
Polymorphic pattern to track the website information in the museum collection.

Attribute pattern (with polymorphic pattern) to track the features of each item.

Extended reference pattern to track the museum information associated with each item.

Bucket pattern with extended reference pattern to track when an exhibition is hosted in a museum.

Informazione



Analyze the above graph illustrating the top 20 largest California wildfires. The visualization was created by CAL FIRE, that is the California Department of Forestry and Fire Protection.

Domanda 14

Risposta non data

Punteggio max.: 0,25

Question

Is there a clearly defined question addressed by the visualization? Write it down.

Domanda 15

Risposta non data

Punteggio max.: 1,25

Data

Is the data quality appropriate? Identify the inadequate characteristics and explain.

Domanda 16

Risposta non data

Punteggio max.: 0,75

Visual Proportionality

Are the values encoded in a uniformly proportional way?

Domanda 17

Risposta non data

Punteggio max.: 0,75

Visual Utility

All the elements in the graph convey useful information?

Domanda 18

Risposta non data

Punteggio max.: 0,50

Visual Clarity

Are the data in the graph clearly identifiable and understandable (properly described)?

Domanda 19

Risposta non data

Punteggio max.: 0,25

Design data

Design the visualization based on the following data structure (to be completed).

Domanda 20

Risposta non data

Punteggio max.: 1,25

Design schema & Sketch

Fill in the required schema elements; formulas can be used if required. Then describe in words the design proposal.

Domanda 21

Risposta non data

Non valutata

This is a blank question to be used as your personal notepad during the exam.

Anything written here will NOT be evaluated.

/