Data management and visualization

Domanda 1
Risposta non data
Punteggio max.: 1,50
Contrassegna domanda
Modifica domanda

The query

db.students.find({ "addresses": { $elemMatch: { "streetNumber": { $lt: 123 }, "types": "home" } } })

returns:

Scegli un'alternativa:

- a. all the documents where the "addresses" array has at least one embedded document that contains the "streetNumber" field lower than 123 or "types" contains "home"
- b. all the documents where at least one embedded document in the "addresses" array has the "streetNumber" field lower than 123 and at least one embedded document (but not necessarily the same one) in the "addresses" array has the "types" field equal to "home"
- c. all the documents where the "addresses" array has at least one embedded document containing both the field "streetNumber" lower than 123 and the field "types" with an element being "home"
- d. one document where the "addresses" array has at least one embedded document that contains the field "streetNumber" lower than 123
- e. none of the other answers is correct.

Risposta errata.

La risposta corretta è: all the documents where the "addresses" array has at least one embedded document containing both the field "streetNumber" lower than 123 and the field "types" with an element being "home"

Domanda 2
Risposta corretta
Punteggio ottenuto 1,50 su 1,50
Contrassegna domanda
Modifica domanda

In MongoDB, which is the best modeling pattern to manage changes to an insurance company's contractual conditions while also maintaining the history of changes?

Scegli un'alternativa:

- a. Schema versioning
- b. Attribute pattern
- c. none of the other answers is correct.
- d. Document versioning ✔
- e. Bucket pattern

Risposta corretta.

La risposta corretta è: Document versioning

Domanda 3
Risposta corretta
Punteggio ottenuto 1,00 su 1,00
Contrassegna domanda
Modifica domanda

How do the concepts of data accuracy and the Open or Closed World Assumption relate to data quality in information systems?

Scegli un'alternativa:

- a. The Open World Assumption suggests that absence of data implies falsity, which contradicts the principles of data accuracy.
- b. Data accuracy focuses on the precision of data, whereas the Open or Closed World Assumption relates to data privacy and security.
- c. Data accuracy is concerned with the correctness of data, while the Open World Assumption implies all data is accurate unless proven otherwise.
- d. Both data accuracy and the Open/Closed World Assumptions are primarily concerned with the speed and efficiency of data processing.
- e. The Closed World Assumption operates on the belief that what is not known to be true is false, which directly influences the strategies for ensuring data accuracy. ✔

Risposta corretta.

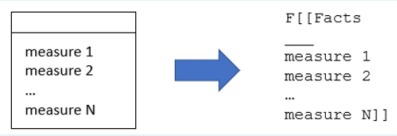
La risposta corretta è: The Closed World Assumption operates on the belief that what is not known to be true is false, which directly influences the strategies for ensuring data accuracy.

Domanda 4
Completo
Punteggio ottenuto 0,00 su 4,00
Contrassegna domanda
Modifica domanda

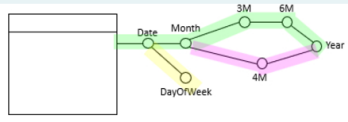
The question starts below.

Conceptual Schema: You must use the formalism discussed during the lectures to solve the conceptual schema.

Declare the fact table as



The attributes as arcs starting from the fact table (**F**), using multiple lines for each dimension or path, for instance:



becomes

```
F --- date --- month --- 3M --- 6M --- year
Date --- DayOftheWeek
Month --- 4M --- Year
```

If you need multiple arcs and/or texts for the arcs:

```
F --- book --- genre
book === author --(author location)-> region
F --(library location)-> region --- state
```

Data analysts at an international book sharing organization are required to design a data warehouse to gain deeper insights into their readers and the book exchanges in terms of total number of books, and average number of pages per book.
Books are shared from a person or library, the lender, to a person, the borrower. Each book is characterized by one main category, one language, a known number of pages, and many target audiences. Each book is shared for a period, e.g., for 2 weeks, or 1 month, etc, then it has to be returned.
The data warehouse must allow the analysis to be performed according to the following requirements:

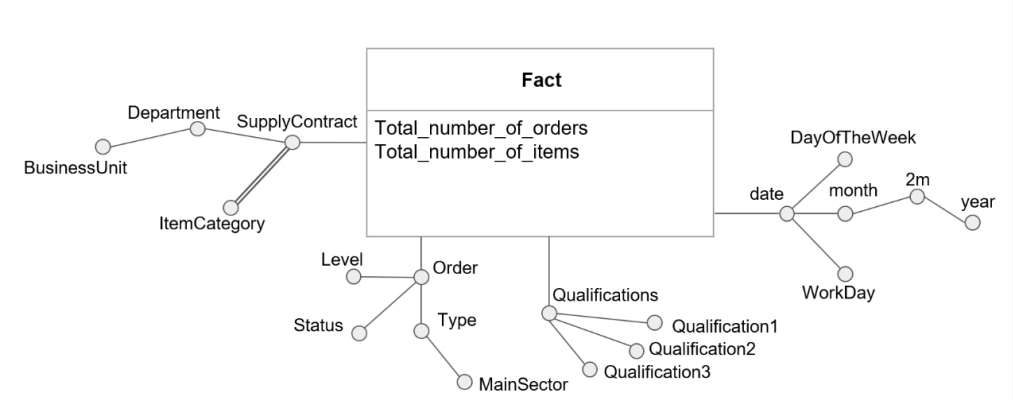
- The timing of the sharing, i.e., when the book exchange was initiated: monthly, every 2 months, 3 months, 4 months, 6 months, annually, and according to the reading season. Reading seasons are as follows: Spring includes March, April, and May; Summer includes June, July, and August; Autumn includes September, October, and November; Winter includes December, January, and February.
- The sharing period can be short, medium, or long. Short periods are shorter than 14 days, medium periods are between 14 and 30 days, long periods are longer than 30 days.
- The lender city, region (e.g. Piedmont), country (e.g., Italy), continent (e.g., Europe), and the main language of the country.
- The borrower's city, region, country, continent, and the country's main language.
- The book's main category (e.g., fantasy, historical, etc.), the book's language, and the book's target audience. Each book can target many audiences, such as kids, teenagers, adults, couples & families, professionals, etc.. The list of target audiences is limited and known.

Write the textual formalism to describe the described conceptual schema.

```
graph LR
F[[
  Book sharing
  ___________
  Total_Book_Count
  Total_Book_Pages
]]

F -- LENDER --> City
F -- BORROWER --> City
subgraph geographical_dimension
City --> Region --> Country --> Continent
Country --> Main_Language
end

F --> Target_audience
subgraph audience_configuration_Yes_No
Target_audience --- Kids
Target_audience --- Couples_and_families
Target_audience --- Professionals_...
end

F --- Month
subgraph time_dimension
Month --> 3M --> 6M --> Year
Month --> 2M --> 4M --> Year
2M --> 6M
Month --> Season
end

F --> Sharing_period
F --> Category
F --> Language
```

Domanda 5
Completo
Punteggio ottenuto 0,00 su 1,00
Contrassegna domanda
Modifica domanda

Given the following conceptual schema:



- The data warehouse collects orders of items.

Provide the logical design of the conceptual DW schema indicated in the picture.

- Example: TABLE_NAME (PrimaryKey, Attribute1, Attribute2)
- Write each table on a new line.
- Use the bold or the underline for identifying primary-key attributes.

Fact (SupplyContract_ID, Order_ID, Time_ID, Qualifications_ID,
Total_number_of_orders, Total_number_of_items)
SupplyContract (SupplyContract_ID, Department, BusinessUnit)
SupplyContractHasItemCategory (SupplyContract_ID, ItemCategory)
Order (Order_ID, Level, Status, Type, MainSector)
Time (Time_ID, Date, DayOfTheWeek, WorkDay, Month, 2m, Year)

Domanda 6
Completo
Punteggio ottenuto 0,00 su 4,00
Contrassegna domanda
Modifica domanda

```
Exams(CourseID, TeacherID, StudentID, TimeID, Grade)
Courses(CourseID, Name, numberOfCredits, Category)
Time(TimeID, Date, Month, DayOfWeek, 2m, 3m, 4m, 6m, year, monthOfTheYear)
Teacher(TeacherID, Name, Role, Department,
University)
Student(StudentID, Name, EnrollmentYear, Status)
```

●The datawarehouse collects exam grades given by teachers to students for specific courses.
●A teacher's Department belongs to a unique University.

Separately for each student and semester, select:

1. the total number of exams
2. the percentage of exams of each student in each semester with respect to the total exams of each student in the whole year
3. assign a rank to each student in each semester, separately for each student status, based on the number of exams passed in each semester (rank 1st the highest)

Write the corresponding SQL query.

```
SELECT student, status, 6m, year,
       COUNT(*) as A,
       COUNT(*) / SUM(COUNT(*)) OVER (PARTITION BY student, year) as B,
       RANK() OVER (PARTITION BY status, 6m,
  ORDER BY COUNT(*) DESC) as C
FROM Exams E, Student S, Time Ti
WHERE E.TimeID = Ti.TimeID AND E.StudentID = S.StudentID
GROUP BY student, status, 6m, year
```

Domanda 7
Completo
Punteggio ottenuto 0,00 su 4,00
Contrassegna domanda
Modifica domanda

```
Exams(CourseID, TeacherID, StudentID, TimeID, Grade)
Courses(CourseID, Name, numberOfCredits, Category)
Time(TimeID, Date, Month, DayOfWeek, 2m, 3m, 4m, 6m, year, monthOfTheYear)
Teacher(TeacherID, Name, Role, Department, University)
Student(StudentID, Name, EnrollmentYear, Status)
```

- The datawarehouse collects exam grades given by teachers to students for specific courses.
- A teacher's Department belongs to a unique University.

Separately for each course and trimester, select:

1. the average exam grade
2. the percentage of exams of the course in the trimester with respect to all the exams in the same period by all courses of the same category
3. the cumulative number of exams since the beginning of the year

Write the corresponding SQL query.

```
SELECT CourseID, category, 3m, year,
       AVG(Grade) as A,
       100 * COUNT(*) / SUM(COUNT(*)) OVER (PARTITION BY category, 3m) as B,
       SUM(COUNT(*)) OVER (
            PARTITION BY CourseID, year
            ORDER BY 3m
            ROWS UNBOUNDED PRECEDING) as C,
FROM Exams E, Courses C, Time Ti
WHERE E.TimeID = Ti.TimeID AND E.CourseID = C.CourseID
GROUP BY CourseID, category, 3m, year
```

Domanda 8
Completo
Punteggio ottenuto 0,00 su 4,00
Contrassegna domanda
Modifica domanda

Design a MongoDB database to store vehicle transits through highway toll gates.

Highways are identified by a predefined alphanumeric ID and characterized by a name, the year of construction, and the management company's name. Highways have one or more toll gates. Each highway toll gate is identified by a unique number and characterized by its name and the highway to which it belongs. The list of towns surrounding the toll gate are also known. Please note that the highway's name is always shown when a toll gate is displayed.

Vehicles traveling on highways are characterized by their number plate and the category to which they belong. Each vehicle is registered to one or more people, i.e., the vehicle owners. Each owner must be retrieved simultaneously as its vehicle information. The owner's information includes first name, surname, social security number, and city of residence.

The daily transits for each toll gate are to be stored. Daily transits are characterized by the list of vehicles transited, the number of vehicles transited, and the total amount of tolls collected. The transit of a vehicle is characterized by the vehicle number plate, the timestamp of the transit, and the toll amount paid. Since a toll gate can record many transits within the same day, the daily transits of each toll gate are stored in groups containing a maximum of 1000 transits.

Write a sample document for each collection of the database.
Important: Besides sample documents, explicitly indicate the design patterns used

```
Highway
{
_id: <string>,
name: <string>,
year_of_construction: <date>,
management_company_name: <string>
}

GATES
{
_id: <number>,
name: <string>,
```

```
highway: {
  _id: <string>,
  name: <string>
},
cities: [<string>]
}

alternative solution:

Highway
{
_id: <string>,
name: <string>,
year_of_construction: <date>,
management_company_name: <string>
gates: [
{
  _id: <number>,
  name: <string>,
  cities: [<string>]
},
{ … }
]
}


VEHICLE

{
_id: ObjectId(),
plate: <string>,
category: <string>,
owners: [{
  name: <string>,
  surname:<string>,
  SSN: <string>
  city: <string>
}, {...} ]
}


Daily_transits

{_id: ObjectId(),
gate: {
  _id: <number>,
  name: <string> },
date: <date>,
vehicles: [
  {timestamp: <datetime>,
   vehicle: {
     _id: ObjectId(),
     plate: <string>}
   toll_amount: <number>
],
total_amount: <number>,
tot_transits: <number>
}


Patterns:
Attribute
Extended reference
Bucket
Computed
```

Domanda 9

Risposta non data

Punteggio max.: 2,00

Contrassegna domanda

Modifica domanda

The following document structure represents an insurance policy characterized by the policy number, insurance type (e.g. car, home, etc.), date of signature, expiration date, price, insurance holder, and the list of coverages. Each coverage is described by its category, the maximum amount insured, the deductible amount, and the premium amount charged. The insurance holder is characterized by first name, surname, date of birth, and city of residence.

```
{
    "_id" : ObjectId("61fa5b8f6f631bb5339dc4b7"),
    “policy_number”: “2024-1034”,
    “type”: “car”,
        “date_of_signature”: “2024-06-01”,
        “expiration_date”: “2025-05-31”,
        “holder”: {
         “name”: “Mario”,
         “surname”: “Rossi”
         “birth_date”: “1980-05-26”,
 “city”: “Torino”},
    “price”: 653,
    “coverages” : [
      {
       “category” : “theft”,
         “maximum”: 2000,
 “deductible”: 100,
 “premium”: 50
},
        {
       “category” : “crystals”,
         “maximum”: 3000,
 “deductible”: 0,
         “premium”: 150
      }]
}
```

Compute the total daily income and the daily number of insurance policies signed for the “car” category. The date to be considered is the date of signature. The sum of the prices gives the income to be considered.

```
db.policy.aggregate (
{$match: {
“type”: “car”
},
{ $group: {
' id': “$date_of_signature”
```

```
n: {$sum: 1}
income: {$sum: "$price"}
}
)
```

Domanda 10

Risposta non data

Punteggio max.: 3,00

Contrassegna domanda

Modifica domanda

The following document structure represents an insurance policy characterized by the policy number, insurance type (e.g. car, home, etc.), date of signature, expiration date, price, insurance holder, and the list of coverages. Each coverage is described by its category, the maximum amount insured, the deductible amount, and the premium amount charged. The insurance holder is characterized by first name, surname, date of birth, and city of residence.

```
{
    "_id" : ObjectId("61fa5b8f6f631bb5339dc4b7"),
    "policy_number": "2024-1034",
    "type": "car",
        "date_of_signature": "2024-06-01",
        "expiration_date": "2025-05-31",
        "holder": {
         "name": "Mario",
         "surname": "Rossi"
         "birth_date": "1980-05-26",
"city": "Torino"},
    "price": 653,
    "coverages" : [
      {
       "category" : "theft",
         "maximum": 2000,
"deductible": 100,
"premium": 50
},
        {
       "category" : "crystals",
         "maximum": 3000,
"deductible": 0,
         "premium": 150
       }]
}
```

Considering only insurance policies signed in the year 2023 by people in Turin and covering a maximum amount lower than 1,000 euros, separately for each type of insurance and category of coverage, compute the average value of the premium amount. Show only the results whose average premium value is greater than 100 euros, and sort them in descending order according to the average premium value.

```
db.policy.aggregate([
{$match: {
"date_of_signature":{$gte: "2023-01-01", $lte: "2023-12-31"}
"holder.city": "Turin"
},
{ $unwind: '$coverages},
{$match: {
"coverages.maximum": {$lt: 1000}
},
{ $group: {
'_id': {
policy_type: '$type,
category: '$coverages.category''},
avg_premium:  {$avg: '$coverages.premium'}
},
{$match: {
"avg_premium": {$gt: 100}
},
{ $sort: { avg_premium: -1}
])
```

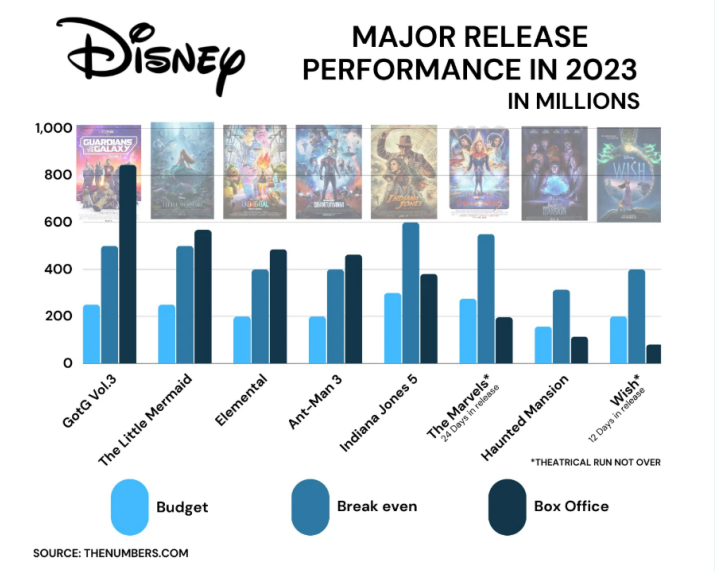
Domanda 11

Risposta corretta

Punteggio ottenuto 0,25 su 0,25

Contrassegna domanda

Modifica domanda



Question

Which one of the following questions represents the purpose of this visualization?

Scegli un'alternativa:

a. How popular were Disney movies in 2023 based on social media metrics?

b. How did the month-by-month box office performance of Disney movies unfold in 2023?

c. How do the budget, break-even point, and box office performance of Disney's major releases in 2023 compare? ✔

d. How profitable were Disney's theme parks in 2023?

e. How were Disney movies released in 2023 received in terms of critical acclaim and audience ratings?

Risposta corretta.

La risposta corretta è:
How do the budget, break-even point, and box office performance of Disney's major releases in 2023 compare?

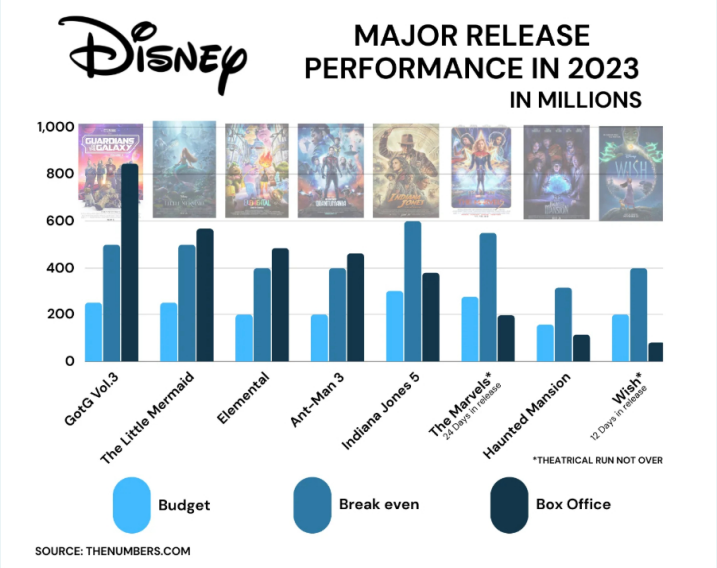
Domanda 12

Parzialmente corretta

Punteggio ottenuto 0,94 su 1,25

Contrassegna domanda

Modifica domanda



Data

Is the data quality appropriate? Select true answers only.

Scegli una o più alternative:

a. The data is not current because it includes movies that have not completed their theatrical run.

b. The credibility of the data may be questionable because it is unclear whether 'THENUMBERS.COM' is an authoritative source.

c. The chart lacks consistency because some movies have asterisks while others do not.

d. The chart is not credible because it does not source its data.

e. The data appears to be complete as it covers a range of Disney movies released in 2023 with no obvious gaps. ✔ Correct, because all the major releases are listed, and there are no blank spaces where data for a particular movie should be.

f. The data is consistent as it uses the same unit of measurement (millions) for all the financial figures. ✔ Correct, because using the same unit across all figures allows for easy comparison and understanding.

g. The visualization demonstrates accuracy as it provides reasonable figures for the budget and box office returns. ✔ Correct, assuming the source is reliable, the use of reasonable figures indicates a high level of accuracy.

h. The data lacks precision since it rounds figures to the nearest million.

i. The data is not understandable because it uses technical financial terms.

j. The data is not accurate because the budget figures are estimates.

Risposta parzialmente esatta.

Hai selezionato correttamente 3.
Le risposte corrette sono: The data appears to be complete as it covers a range of Disney movies released in 2023 with no obvious gaps.,
The data is consistent as it uses the same unit of measurement (millions) for all the financial figures.,

The visualization demonstrates accuracy as it provides reasonable figures for the budget and box office returns.,

The credibility of the data may be questionable because it is unclear whether 'THENUMBERS.COM' is an authoritative source.

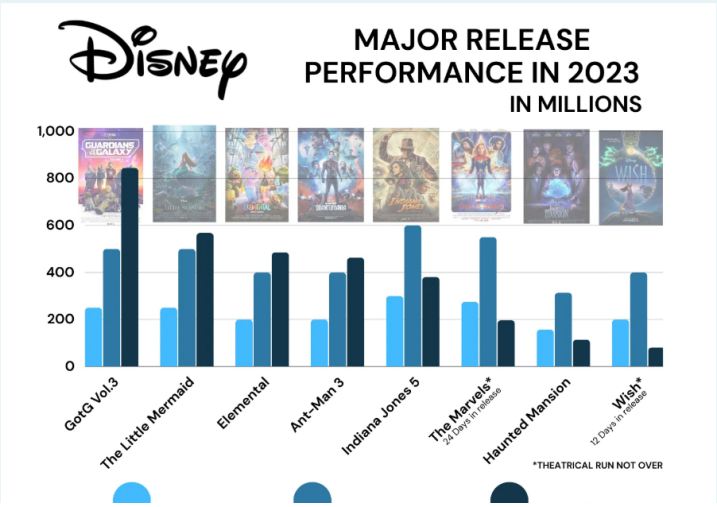
Domanda 13

Completo

Punteggio ottenuto 0,50 su 0,75

Contrassegna domanda

Modifica domanda



SOURCE: THENUMBERS.COM

Visual Proportionality

Are the values encoded in a uniformly proportional way?

Yes, the values on the graph are encoded in a uniformly proportional way. The axis starts at zero, which means that the visual representation of the data is proportionate to the actual values, allowing for an accurate comparison between the different movies' financial metrics.

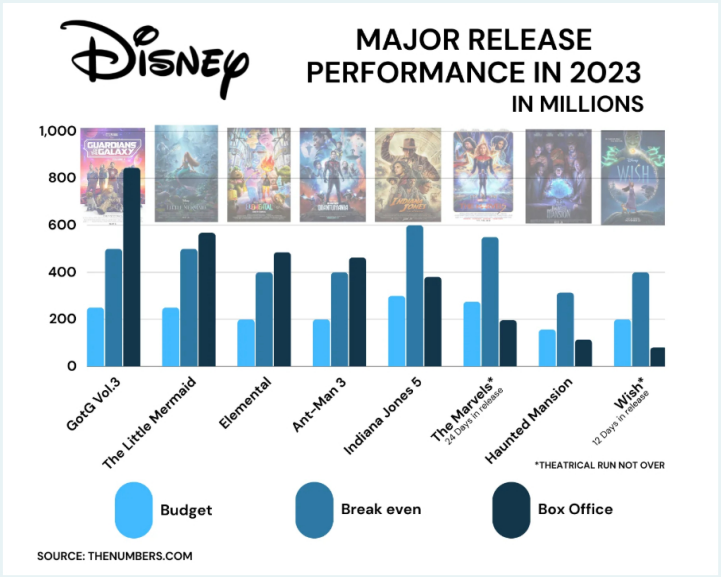
Domanda 14

Risposta non data

Punteggio max.: 0,75

Contrassegna domanda

Modifica domanda



SOURCE: THENUMBERS.COM

Visual Utility

All the elements in the graph convey useful information?

No, not all elements in the graph convey useful information in terms of data representation. The Disney logo and movie posters, while visually appealing and providing context, do not contribute to the actual data analysis and could be removed without losing any critical information about the movie performances.

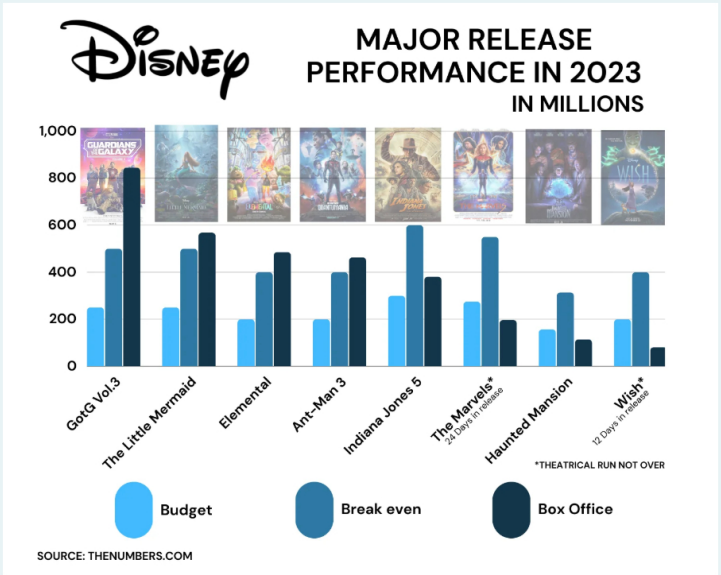
Domanda 15

Risposta non data

Punteggio max.: 0,50

Contrassegna domanda

Modifica domanda



SOURCE: THENUMBERS.COM

Visual Clarity

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

Yes, the data in the graph are clearly identifiable and understandable. The graph is properly labeled with the names of the movies, and the metrics are clearly marked as budget, break-even points, and box office performance. The use of consistent and clear labeling makes the data easy to interpret.

Domanda 16

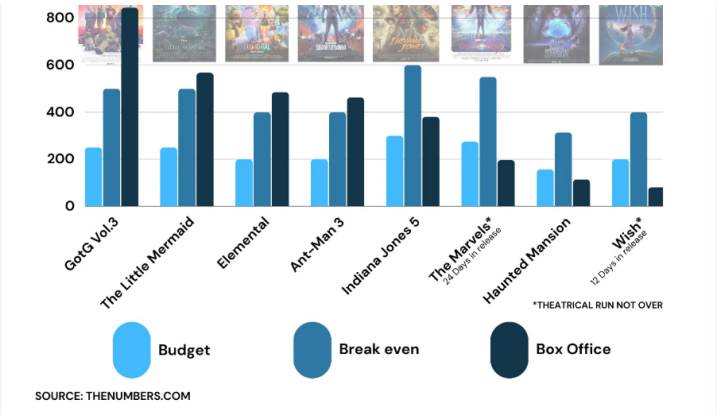
Risposta corretta

Punteggio ottenuto 0,25 su 0,25

Contrassegna domanda

Modifica

domanda



Design data

Design the visualization based on the following data structure.

RUN_STATUS	Dimension ✔
BOX_OFFICE	Measure ✔
BREAK_EVEN	Measure ✔
MOVIE_TITLE	Dimension ✔
BUDGET	Measure ✔

Risposta corretta.

La risposta corretta è:
RUN_STATUS → Dimension,

BOX_OFFICE → Measure,

BREAK_EVEN → Measure, MOVIE_TITLE → Dimension,

BUDGET → Measure

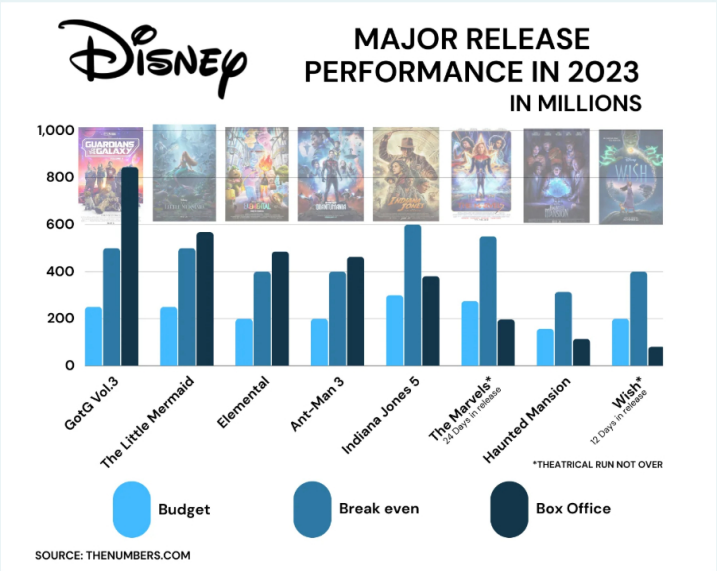
Domanda 17

Completo

Punteggio ottenuto 0.00 su 1,25

Contrassegna domanda

Modifica domanda



Design schema & Sketch

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

Schema	Details
Columns	SUM(BUDGET), SUM(BOX_OFFICE), SUM(BREAK_EVEN)
Rows	MOVIE_TITLE
Graph type	Bar, Bar, Circle
Color	MEASURE_NAMES
Size	Default
Label	Default

Design proposal: The redesigned visualization will be a bar chart with each movie title listed along the vertical axis (Rows) for clarity. The Budget and Box Office Revenue will be represented as bars along the horizontal axis (Columns), with the length of each bar corresponding to the magnitude of the figures. The Break-Even Point will be overlaid on the same axis as the Budget using a dual-axis chart, represented by an orange circle. This will help in understanding how the Box Office Revenue compares to the Budget and Break-Even Point visually. The movies which are still running in theaters will have their bars marked in a different color to indicate that the Box Office data is not final. Labels will be present for all measures to provide exact figures and enhance understandability. To reduce clutter, the Disney logo and movie posters will be removed.