Data Engineering Case - Retviews

**1) Data Model Try to make a normalized data model with the following characteristics:**

* **We gather data about clothes that are sold online**
* **We keep track of the title, brand, material, colors, and price**
* **The price can vary per color. We get daily updates of the price, and need to keep track of the whole price history. The rest of the data (title, brand, material & colors) stay the same.**

**The data model should contain:**

* **the primary keys**
* **the foreign keys**
* **the cardinalities**

This would be the graphical representation of the model. In a data warehousing context, the fact table would be price listing and the remaining tables would be dimensions. In the real world the date field in price\_listing should be implemented as its own dimension.

Graphical user interface, application

Description automatically generated

**2) Queries Write a query for your data model that gives us the following:**

* **Get all brands that have a clothes item with title 'Baggy Trousers'**

**SELECT** **DISTINCT** brand.brand  
**FROM** clothes  
**INNER** **JOIN** title **ON** clothes.title\_id = title.id  
**INNER** **JOIN** brand **ON** clothes.brand\_id = brand.id  
**WHERE** title = 'Baggy Trousers';

* **Get the highest price per color of a clothes item with title 'Sleeveless Shirt' from the brand Zara**

**SELECT** **MAX**(lp.last\_price)  
**FROM** clothes  
**INNER** **JOIN** title **ON** clothes.title\_id = title.id  
**INNER** **JOIN** brand **ON** clothes.brand\_id = brand.id  
**INNER** **JOIN** clothes\_attributes **ON** clothes.id = clothes\_attributes.clothes\_id  
**INNER** **JOIN** color **ON** clothes\_attributes.color\_id = color.id  
**INNER** **JOIN** (  
 **SELECT**   
 price\_listing.clothes\_attributes\_id  
 ,price **AS** last\_price  
 **FROM** price\_listing  
 **INNER** **JOIN** (  
 **SELECT**   
 clothes\_attributes\_id  
 ,**MAX**(DATE) **AS** max\_date  
 **FROM** price\_listing  
 **GROUP** **BY** clothes\_attributes\_id  
 ) md   
 **ON** price\_listing.clothes\_attributes\_id = md.clothes\_attributes\_id  
 **AND** price\_listing.date = md.max\_date  
 ) lp **ON** clothes\_attributes.id = lp.clothes\_attributes\_id  
**WHERE**   
 title = 'Sleeveless Shirt'   
**AND** brand = 'Zara';

**3) Indexes**

* **What indexes would you put, assuming that the 2 queries above are the ones executed mostly?**

clothes.title\_id, clothes.brand\_id and title.title

* **What are the disadvantages of putting an index on every single column?**

Increased disk usage and performance decrease for write operations.

**4) Elaborate the data model**

**What would you change about the model if we want to keep track of the sizes of the clothes? The sizes can be different per color, and the prices can be different per size as well**

I would add a new ‘size’ dimension under clothes\_attributes, as in the following diagram.

Graphical user interface, application

Description automatically generated