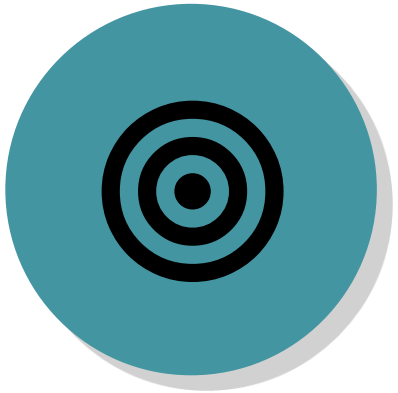
Group exercise Progettazione Data Warehouse 2023/2024

- Valerio Morelli
- Federica Paganica
- Federico Staffolani
- Enrico Maria Sardellini
- Micol Zazzarini
- Mattia Sbattella

The solution



ERD reverse engineering

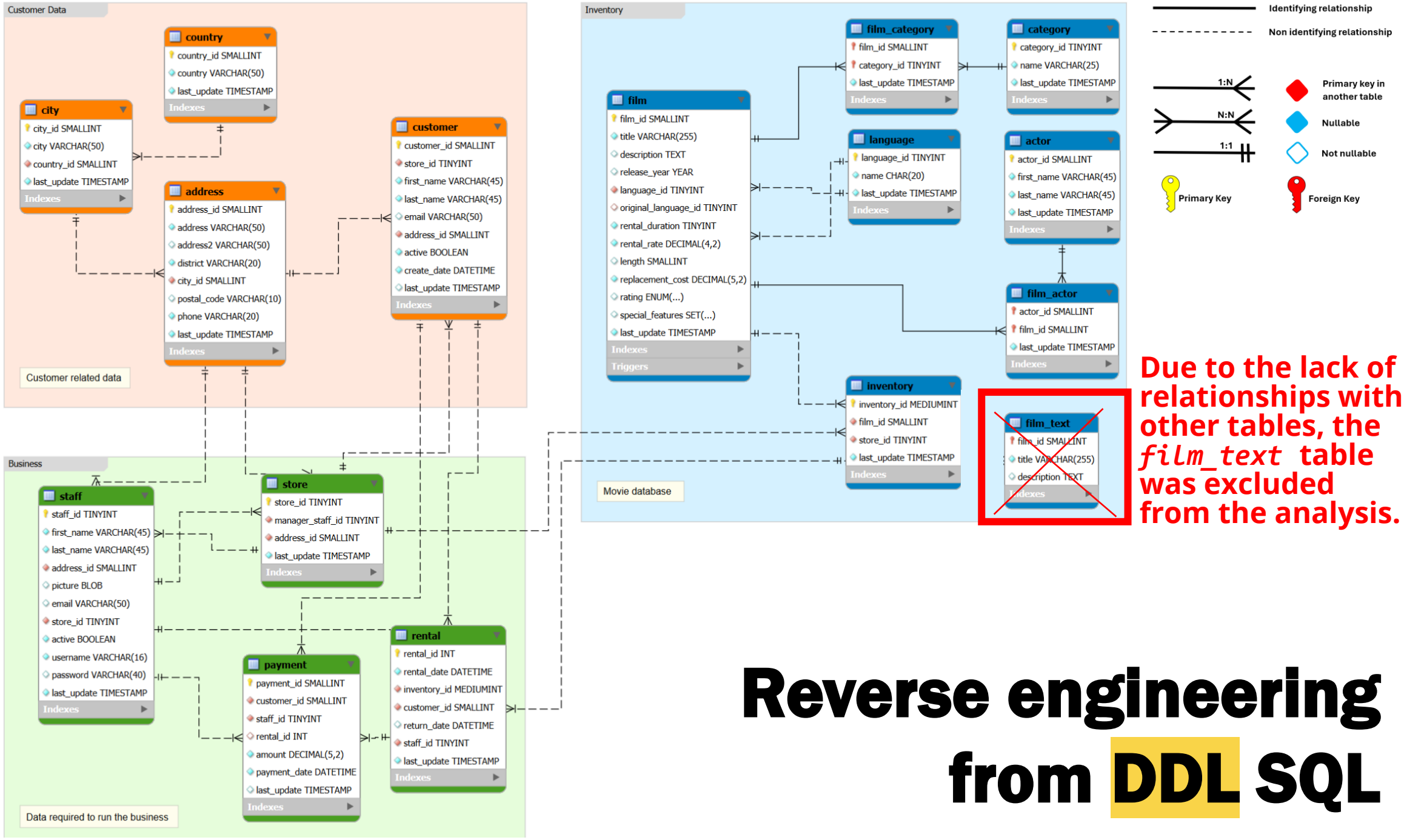
How did we obtain the ER schema from the logical schema?

Data Warehouse design

How was the multidimensional scheme defined?

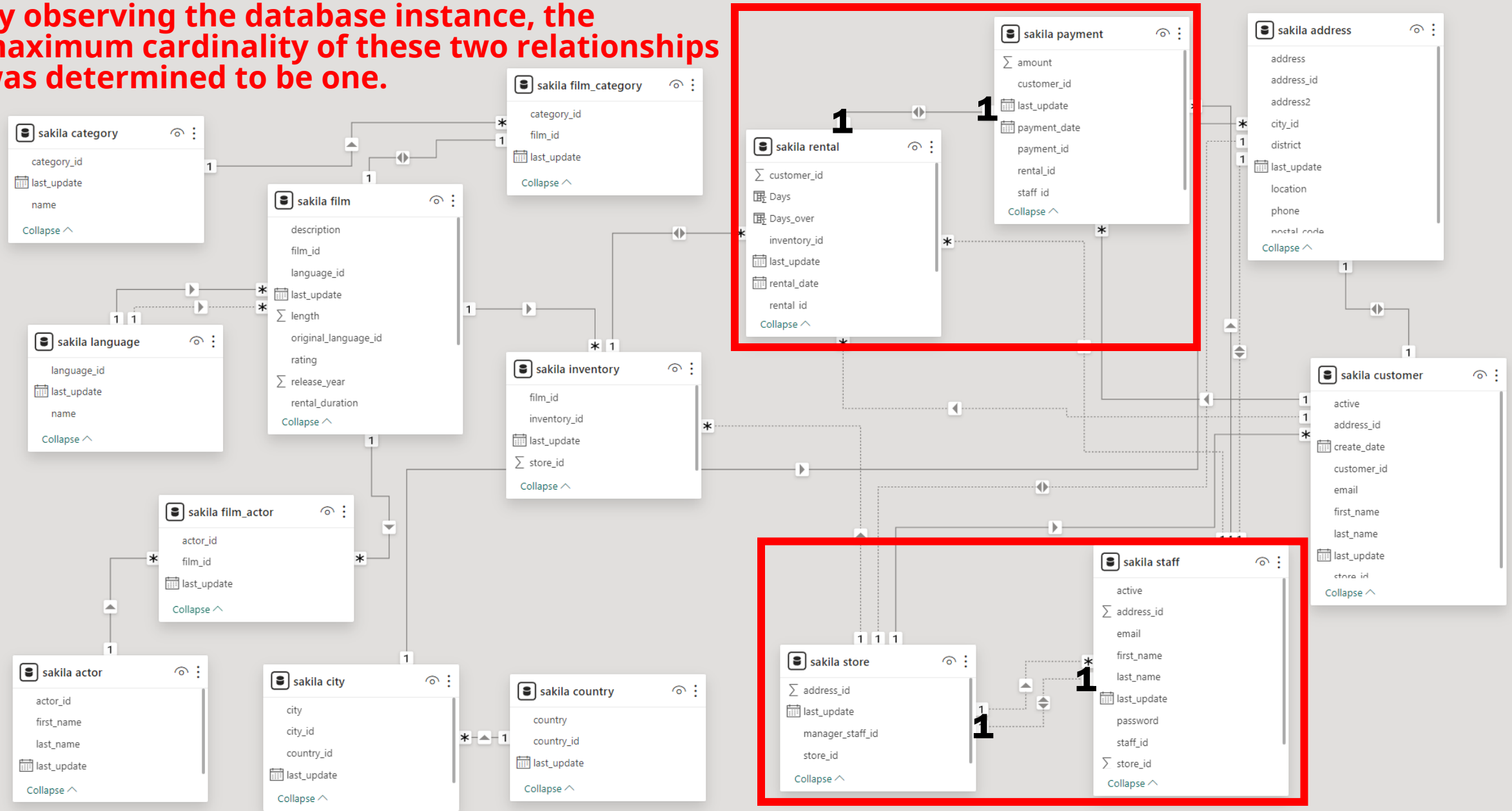
Analyses implementation

What were the results of the analyses?

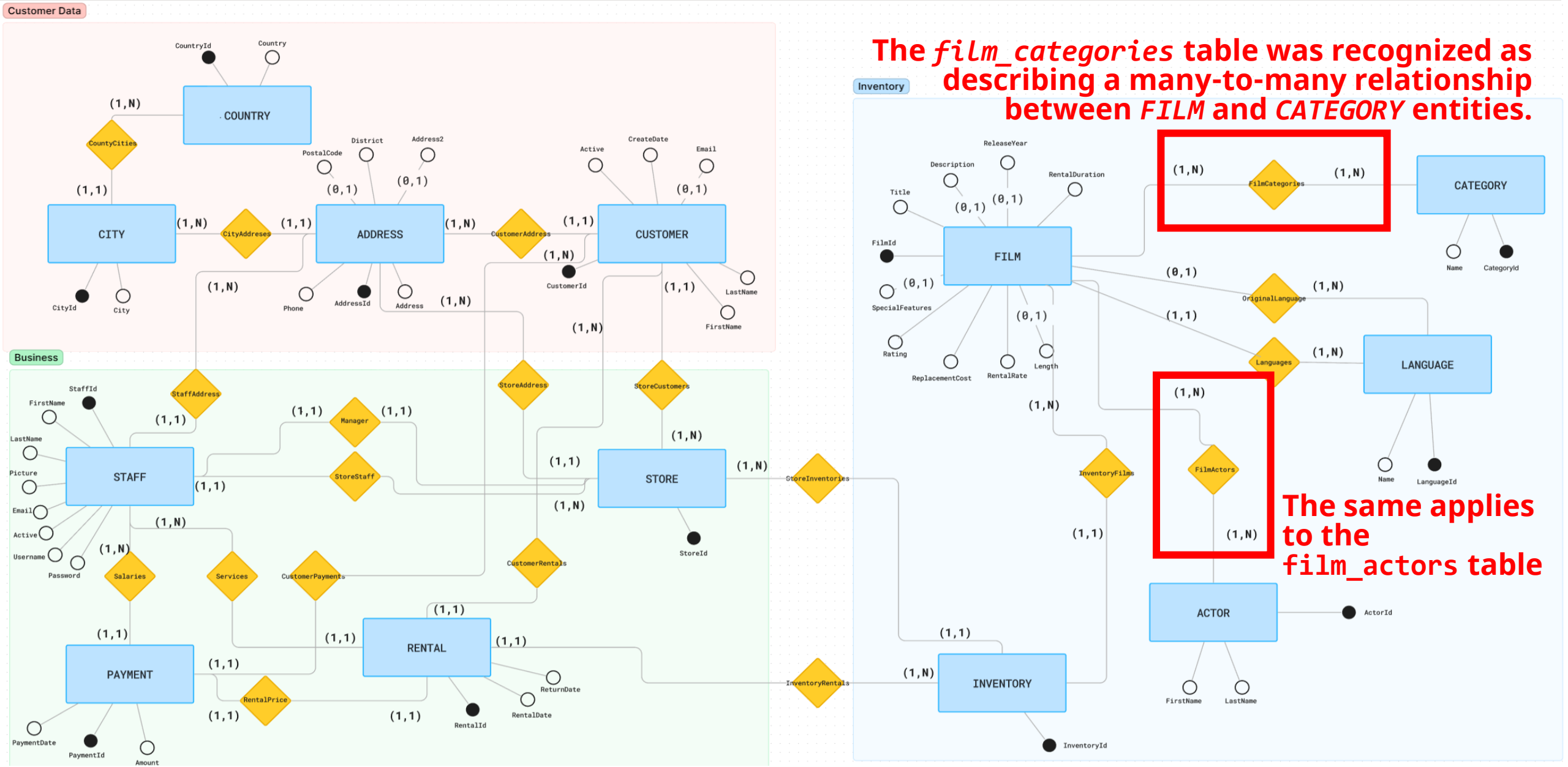

Customer Data
country
country_id SMALLINT
country VARCHAR(50)
last_update TIMESTAMP
Indexes
city
city_id SMALLINT
city VARCHAR(50)
country_id SMALLINT
last_update TIMESTAMP
Indexes
address
address_id SMALLINT
address VARCHAR(50)
address2 VARCHAR(50)
district VARCHAR(20)
city_id SMALLINT
postal_code VARCHAR(10)
phone VARCHAR(20)
last_update TIMESTAMP
Indexes
customer
customer_id SMALLINT
store_id TINYINT
first_name VARCHAR(45)
last_name VARCHAR(45)
email VARCHAR(50)
address_id SMALLINT
active BOOLEAN
create_date DATETIME
last_update TIMESTAMP
Indexes
Customer related data
Business
staff
staff_id TINYINT
first_name VARCHAR(45)
last_name VARCHAR(45)
address_id SMALLINT
picture BLOB
email VARCHAR(50)
store_id TINYINT
active BOOLEAN
username VARCHAR(16)
password VARCHAR(40)
last_update TIMESTAMP
Indexes
store
store_id TINYINT
manager_staff_id TINYINT
address_id SMALLINT
last_update TIMESTAMP
Indexes
payment
payment_id SMALLINT
customer_id SMALLINT
staff_id TINYINT
rental_id INT
amount DECIMAL(5,2)
payment_date DATETIME
last_update TIMESTAMP
Indexes
rental
rental_id INT
rental_date DATETIME
inventory_id MEDIUMINT
customer_id SMALLINT
return_date DATETIME
staff_id TINYINT
last_update TIMESTAMP
Indexes
Data required to run the business
Inventory
film_category
film_id SMALLINT
category_id TINYINT
last_update TIMESTAMP
Indexes
category
category_id TINYINT
name VARCHAR(25)
last_update TIMESTAMP
Indexes
film
film_id SMALLINT
title VARCHAR(255)
description TEXT
release_year YEAR
language_id TINYINT
original_language_id TINYINT
rental_duration TINYINT
rental_rate DECIMAL(4,2)
length SMALLINT
replacement_cost DECIMAL(5,2)
rating ENUM(...)
special_features SET(...)
last_update TIMESTAMP
Indexes
Triggers
language
language_id TINYINT
name CHAR(20)
last_update TIMESTAMP
Indexes
actor
actor_id SMALLINT
first_name VARCHAR(45)
last_name VARCHAR(45)
last_update TIMESTAMP
Indexes
film_actor
actor_id SMALLINT
film_id SMALLINT
last_update TIMESTAMP
Indexes
inventory
inventory_id MEDIUMINT
film_id SMALLINT
store_id TINYINT
last_update TIMESTAMP
Indexes
film_text
film_id SMALLINT
title VARCHAR(255)
description TEXT
Indexes
Movie database
Identifying relationship
Non identifying relationship
1:N
N:N
1:1
Primary key in another table
Nullable
Not nullable
Primary Key
Foreign Key
Due to the lack of relationships with other tables, the film_text table was excluded from the analysis.
Reverse engineering from DDL SQL

Reverse engineering from DML SQL

By observing the database instance, the maximum cardinality of these two relationships was determined to be one.



The overall E-R diagram



The solution

ERD reverse engineering

How did we obtain the ER schema from the logical schema?

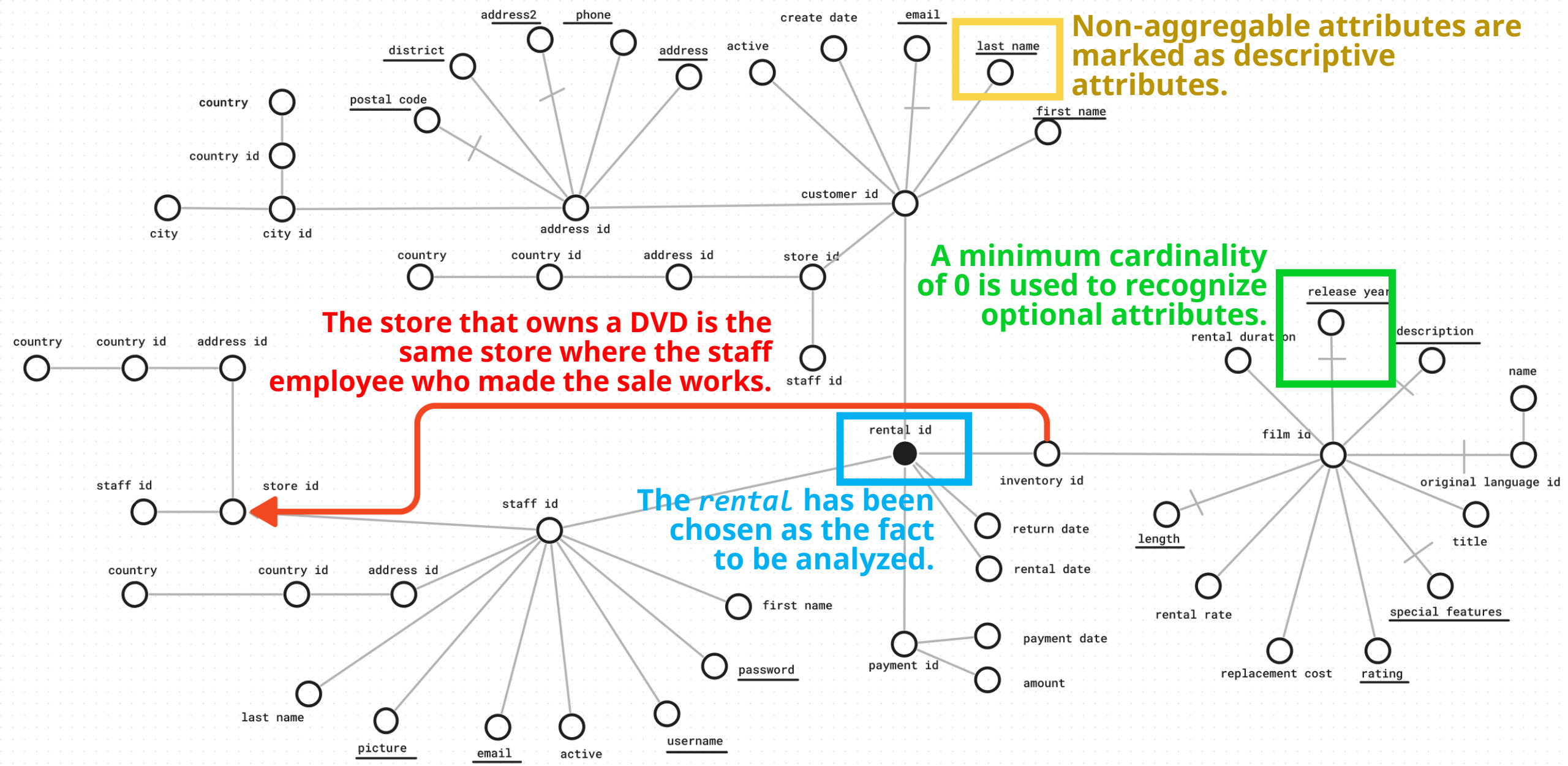
Data Warehouse design

How was the multidimensional scheme defined?

Analyses implementation

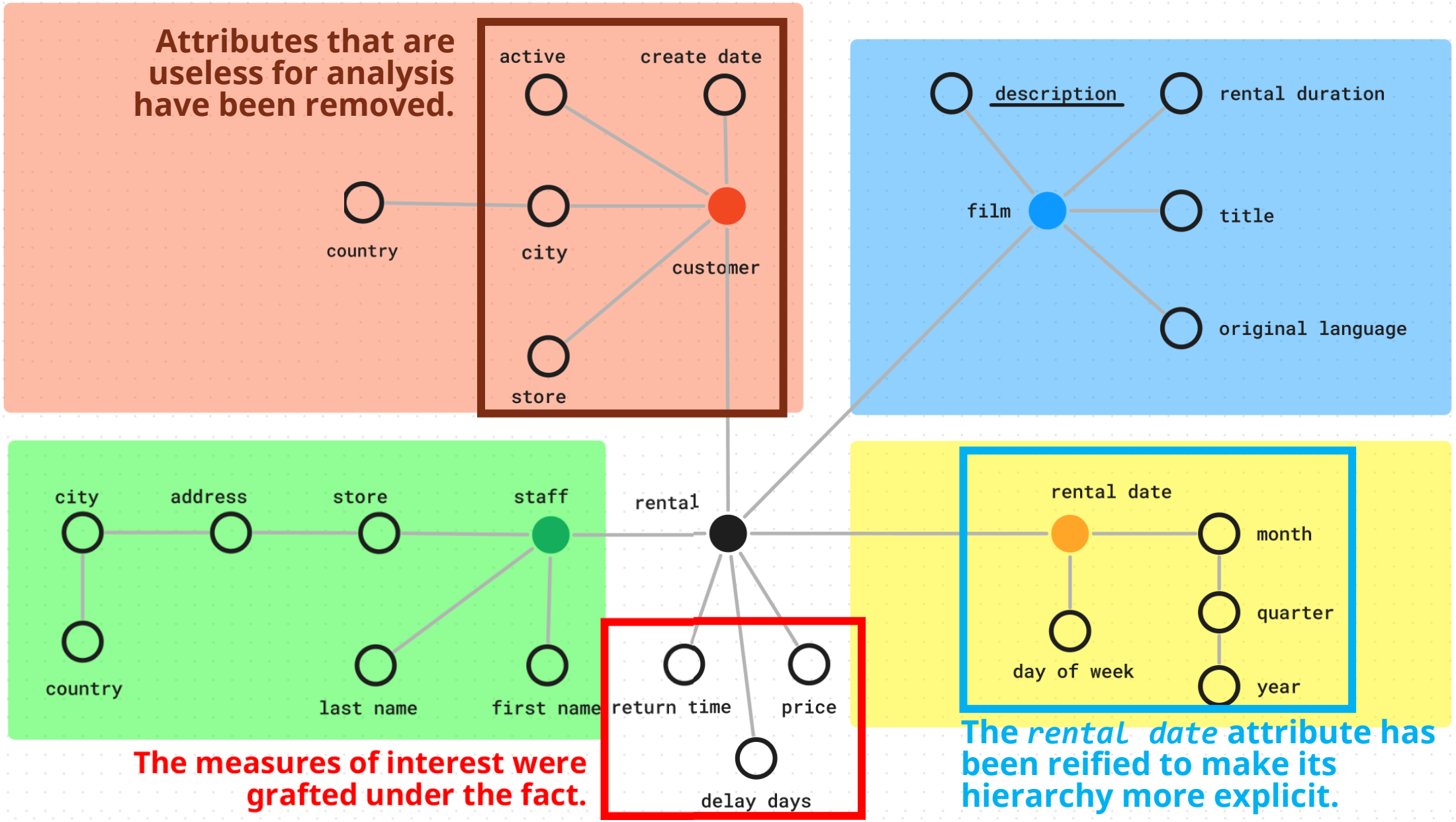
What were the results of the analyses?

Attribute tree



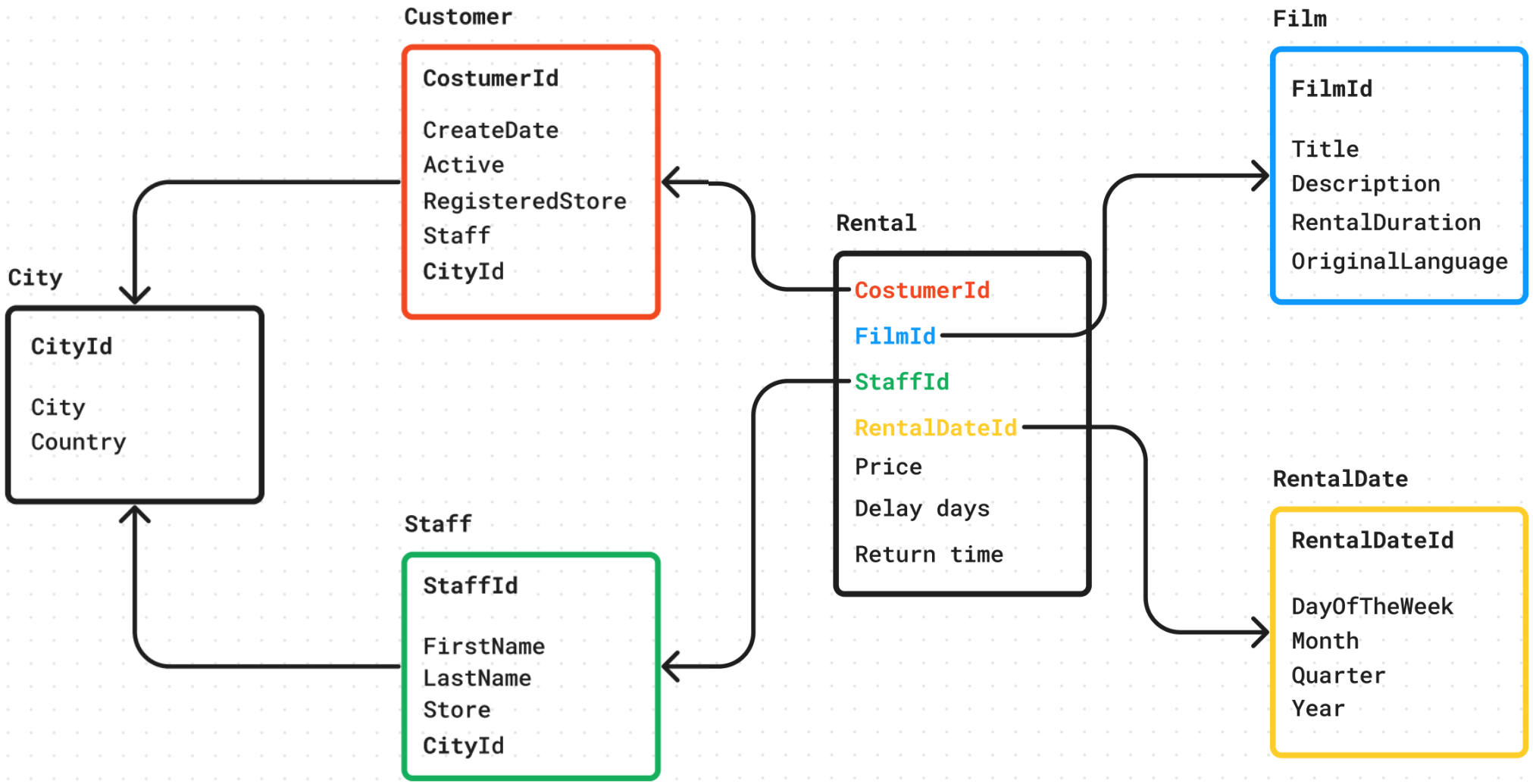
The *rental* has been chosen as the fact to be analyzed.

Refined attribute tree



Hybrid ROLAP Representation

To reduce redundancy at the cost of slower data access, the address hierarchy was extracted from both the customer and employee dimensions and a foreign key constraint was created between them. This resulted in a hybrid multidimensional scheme **using both the star and snowflake schemes**.



Measurements analysis

The following measurements were chosen to perform the analyses:

- Average cost of rentals = `avg(payment.amount)`
- Average return days = `avg(datediff(rental.return_date, rental.rental_date))`
- Average days of delay = `avg(greatest(0, datediff(rental.return_date, date_add(rental.rental_date, interval film.rental_duration day))))`
- Number of rentals = `count(*)`

To determine when it is possible to **do a pre-aggregation** of each of them, **their additivity** along each dimension should be proven.

	Film	**Customer**	**Staff**	**Rental date**
Average cost of rentals	No	No	No	No
Average return days	No	No	No	No
Average days of delay	No	No	No	No
Number of rentals	Yes	Yes	Yes	Yes

The solution

ERD reverse engineering

How did we obtain the ER schema from the logical schema?

Data Warehouse design

How was the multidimensional scheme defined?

Analyses implementation

What were the results of the analyses?

SQL implementation (part 1)

```sql
# Average rental cost, aggregated on STAFF dimension
select s.last_name, round(avg(p.amount), 2) as 'Average cost of rentals'
from rental r
         join payment p on r.rental_id = p.rental_id
         join staff s on p.staff_id = s.staff_id
group by s.last_name;
```

For the first analysis, average rental costs were aggregated under the staff dimension

	staff	`Average cost of rentals`
1	Hillyer	4.16
2	Stephens	4.25

The second measure was aggregated under the customer's country of origin attribute.

```sql
# Average aggregate return time on the CUSTOMER dimension (Country of its address)
select c3.country, round(avg(datediff(r.return_date, r.rental_date)), 2) as 'Average return days'
from rental r
         join customer c on c.customer_id = r.customer_id
         join address a on c.address_id = a.address_id
         join city c2 on a.city_id = c2.city_id
         join country c3 on c2.country_id = c3.country_id
group by c3.country;
```

	country	`Average return days`
1	Afghanistan	4.67
2	Algeria	5.25
3	American Samoa	3.85
4	Angola	5.25
5	Anguilla	4.40
6	Argentina	5.15
7	Armenia	5.17

SQL implementation (part 2)

The delay measure was aggregated under the film title dimension

```sql
# Average days beyond the maximum duration of an aggregate rental on the FILM dimension
select film, round(avg(days_late), 2) as 'Average days of delay'
from (select f.title                                                    as film,
             greatest(0, datediff(r.return_date,
                                  date_add(r.rental_date, interval f.rental_duration day)))  as days_late
      from rental r
               join inventory i on r.inventory_id = i.inventory_id
               join film f on i.film_id = f.film_id) as fgr
group by film;
```

	film	`Average days of delay`
1	ACADEMY DINOSAUR	0.50
2	ACE GOLDFINGER	3.00
3	ADAPTATION HOLES	0.17
4	AFFAIR PREJUDICE	1.05
5	AFRICAN EGG	1.45
6	AGENT TRUMAN	3.05
7	AIRPLANE SIERRA	0.53

The number of rentals were aggregated on the film rental duration in days

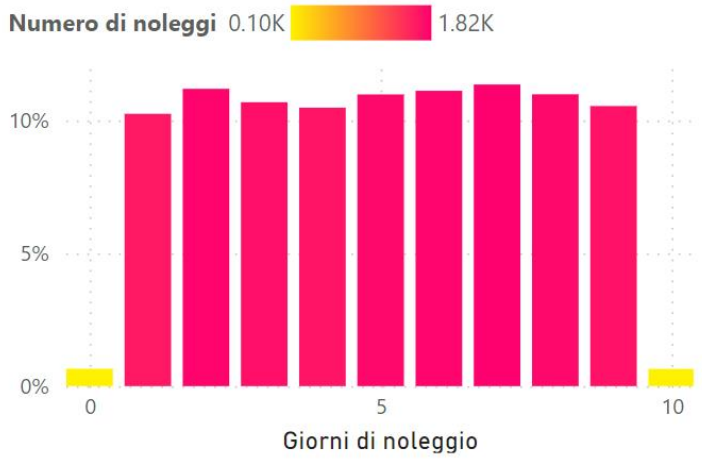
```sql
# Number of Rentals by film rental duration
select f.rental_duration as 'Rental duration', count(*) as 'Number of rentals'
from rental r
         join inventory i on r.inventory_id = i.inventory_id
         join film f on i.film_id = f.film_id
group by f.rental_duration
```

	`Rental duration` 1	`Number of rentals`
1	3	3412
2	4	3251
3	5	3165
4	6	3392
5	7	2824

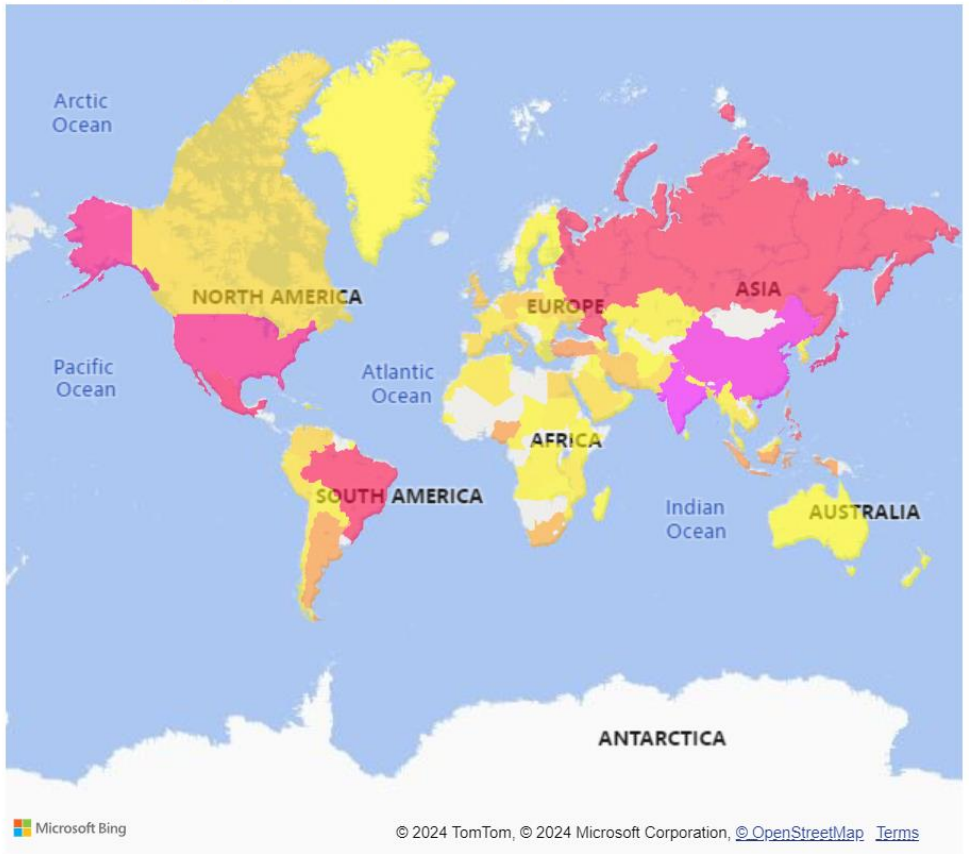
Power BI dashboard (part 1)

4.20	5.03	1.26	16.04K
Costo medio dei noleggi	Giorni medi di restituzione	Media dei giorni di ritardo	Numero di noleggi effettuati

Tempo di restituzione dei noleggi



Numero noleggi per Nazionalità del cliente



Impiegato

- [ ] Hillyer
- [ ] Stephens

Durata del noleggio

- [ ] 3
- [ ] 4
- [ ] 5
- [ ] 6
- [ ] 7

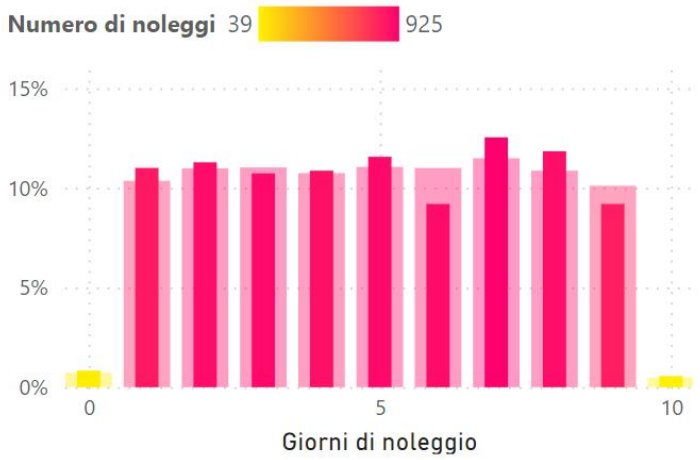
Film

- [ ] ACADEMY DINOSAUR
- [ ] ACE GOLDFINGER
- [ ] ADAPTATION HOLES
- [ ] AFFAIR PREJUDICE
- [ ] AFRICAN EGG
- [ ] AGENT TRUMAN
- [ ] AIRPLANE SIERRA
- [ ] AIRPORT POLLOCK
- [ ] ALABAMA DEVIL
- [ ] ALADDIN CALENDAR
- [ ] ALAMO VIDEOTAPE
- [ ] ALASKA PHANTOM
- [ ] ALI FOREVER
- [ ] ALICE FANTASIA
- [ ] ALIEN CENTER
- [ ] ALLEY EVOLUTION
- [ ] ALONE TRIP
- [ ] ALTER VICTORY
- [ ] AMADEUS HOLY
- [ ] AMELIE HELLFIGHTERS
- [ ] AMERICAN CIRCUS
- [ ] AMISTAD MIDSUMMER
- [ ] ANACONDA CONFESSIONS

Power BI dashboard (part 2)

4.11	4.95	1.23	718
Costo medio dei noleggi	Giorni medi di restituzione	Media dei giorni di ritardo	Numero di noleggi effettuati

Tempo di restituzione dei noleggi



Impiegato

- Hillyer
- Stephens

Durata del noleggio

- 3
- 4
- 5
- 6
- 7

Numero noleggi per Nazionalità del cliente

Film

- ACADEMY DINOSAUR
- ACE GOLDFINGER
- ADAPTATION HOLES
- AFFAIR PREJUDICE
- AFRICAN EGG
- AGENT TRUMAN
- AIRPLANE SIERRA
- AIRPORT POLLOCK
- ALABAMA DEVIL
- ALADDIN CALENDAR
- ALAMO VIDEOTAPE
- ALASKA PHANTOM
- ALI FOREVER
- ALICE FANTASIA
- ALIEN CENTER
- ALLEY EVOLUTION
- ALONE TRIP
- ALTER VICTORY
- AMADEUS HOLY
- AMELIE HELLFIGHTERS
- AMERICAN CIRCUS
- AMISTAD MIDSUMMER
- ANACONDA CONFESSIONS