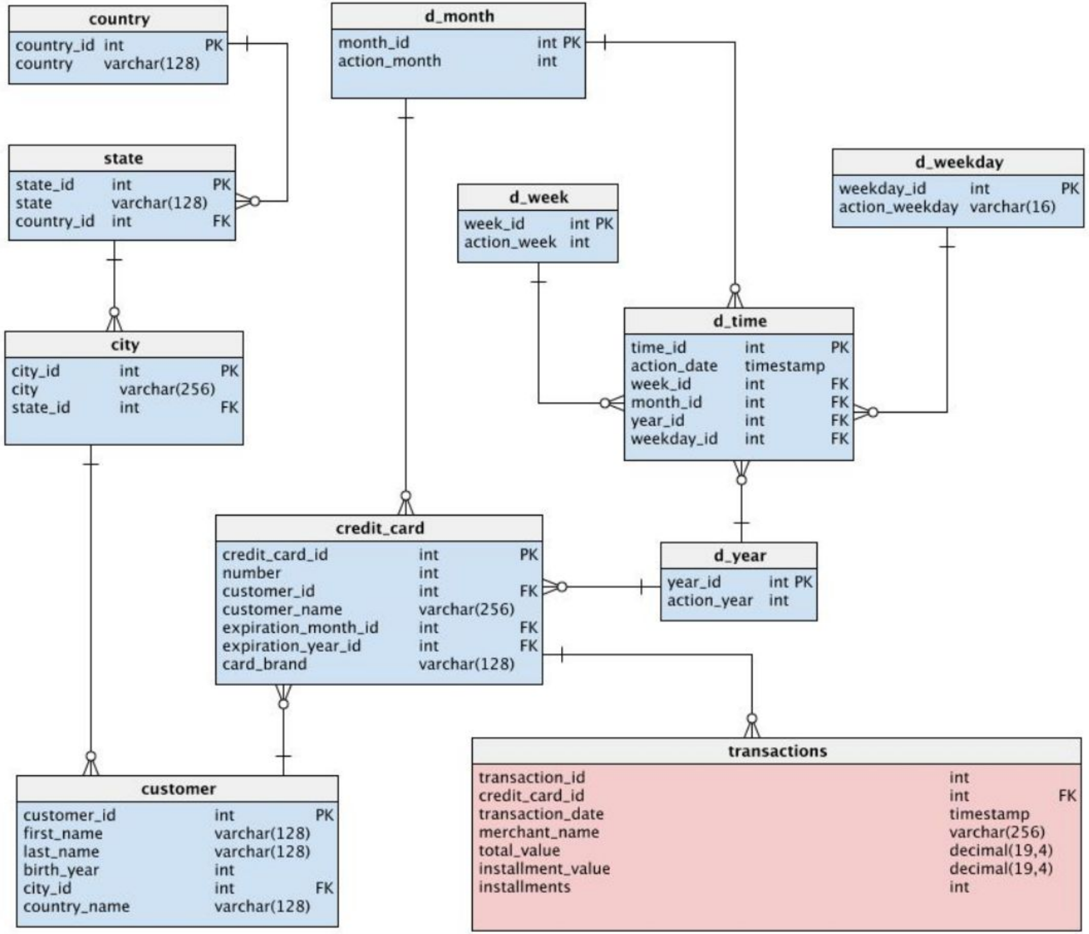
Nubank Data Analyst Challenge

Luiz Eduardo Amaral

Problem Definition

- Legacy warehouse
- *transactions* table
- Hard to query for a monthly *bill.*



country

country_id	int	PK
country	varchar(128)	

d_month

month_id	int	PK
action_month	int	

state

state_id	int	PK
state	varchar(128)	
country_id	int	FK

d_week

week_id	int	PK
action_week	int	

d_weekday

weekday_id	int	PK
action_weekday	varchar(16)	

city

city_id	int	PK
city	varchar(256)	
state_id	int	FK

d_time

time_id	int	PK
action_date	timestamp	
week_id	int	FK
month_id	int	FK
year_id	int	FK
weekday_id	int	FK

credit_card

credit_card_id	int	PK
number	int	
customer_id	int	FK
customer_name	varchar(256)	
expiration_month_id	int	FK
expiration_year_id	int	FK
card_brand	varchar(128)	

d_year

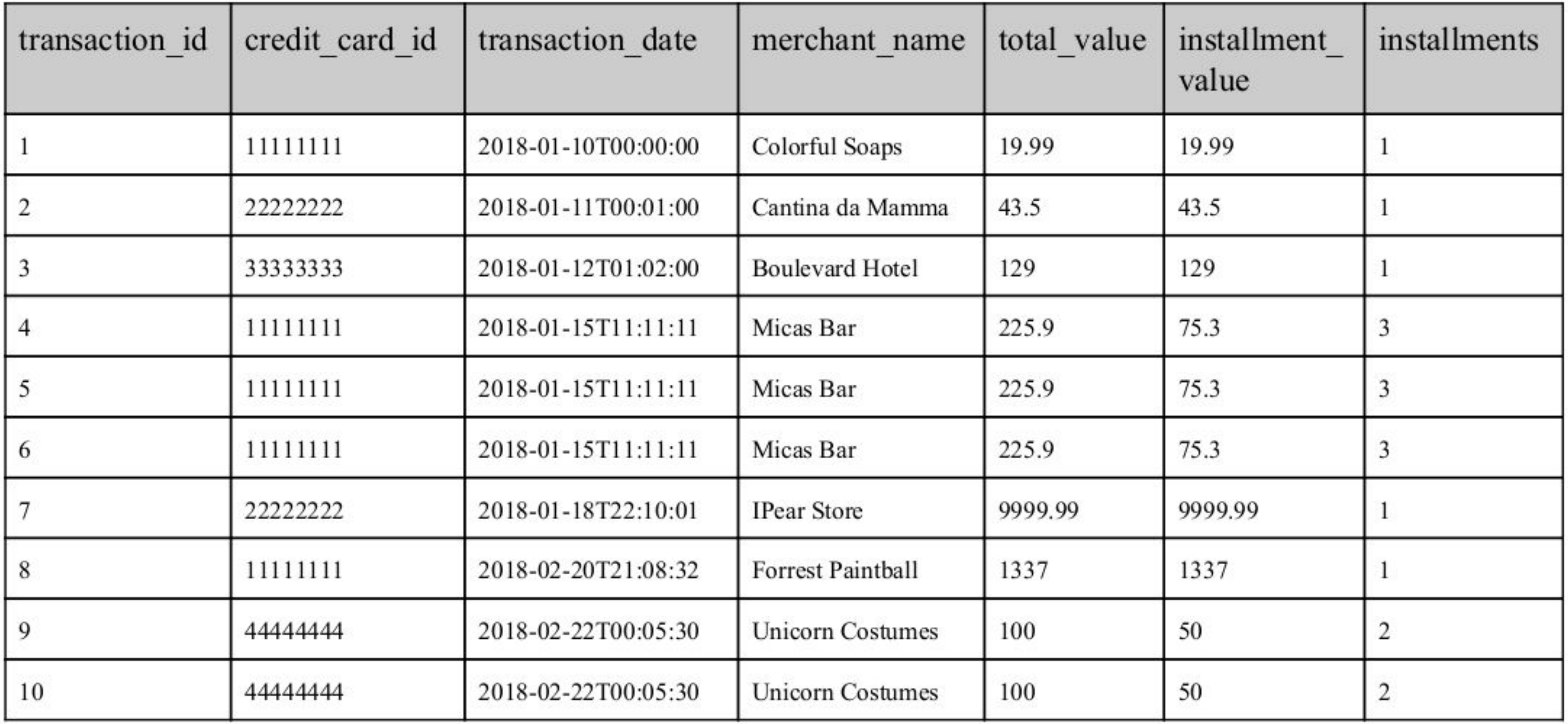
year_id	int	PK
action_year	int	

customer

customer_id	int	PK
first_name	varchar(128)	
last_name	varchar(128)	
birth_year	int	
city_id	int	FK
country_name	varchar(128)	

transactions

transaction_id	int	
credit_card_id	int	FK
transaction_date	timestamp	
merchant_name	varchar(256)	
total_value	decimal(19,4)	
installment_value	decimal(19,4)	
installments	int	



transaction_id	credit_card_id	transaction_date	merchant_name	total_value	installment_value	installments
1	11111111	2018-01-10T00:00:00	Colorful Soaps	19.99	19.99	1
2	22222222	2018-01-11T00:01:00	Cantina da Mamma	43.5	43.5	1
3	33333333	2018-01-12T01:02:00	Boulevard Hotel	129	129	1
4	11111111	2018-01-15T11:11:11	Micas Bar	225.9	75.3	3
5	11111111	2018-01-15T11:11:11	Micas Bar	225.9	75.3	3
6	11111111	2018-01-15T11:11:11	Micas Bar	225.9	75.3	3
7	22222222	2018-01-18T22:10:01	IPear Store	9999.99	9999.99	1
8	11111111	2018-02-20T21:08:32	Forrest Paintball	1337	1337	1
9	44444444	2018-02-22T00:05:30	Unicorn Costumes	100	50	2
10	44444444	2018-02-22T00:05:30	Unicorn Costumes	100	50	2

Problem Definition

1. Redesign the database
2. Build a query for the monthly *bill*
3. How to prevent these mistakes
4. How to better find, understand and consume the data

1. Redesigning the database

Dimensional Modeling Techniques (Kimball/Ross)

- Star Schema
- Fact Tables
- Dimension Tables
- Slowly Changing Dimension

Four-Step Dimensional Design Process (Kimball/Ross)

1. Select the business process.
2. Declare the grain.
3. Identify the dimensions.
4. Identify the facts.

Select the business process

We want to create monthly bills for the customers

Declare the grain

Individual purchases of a customer and individual installments

Identify the dimensions

- Customer
- Credit card
- Merchant
- Date

Identify the facts

- Total value
- Installment value
- installments

Redesigning the database

- One fact table:
    - Installments

- Four dimension tables
    - Customer
    - Credit card
    - Merchant
    - Date

Dimension Tables

Dimension - merchant		
merchant_id	int	PK
merchant_name	varchar(256)	

Dimension - customer		
customer_id	int	PK
first_name	varchar(128)	
last_name	varchar(128)	
birthday	date	
city	varchar(128)	
state	varchar(64)	
country	varchar(64)	
installments_value	decimal(19, 4)	

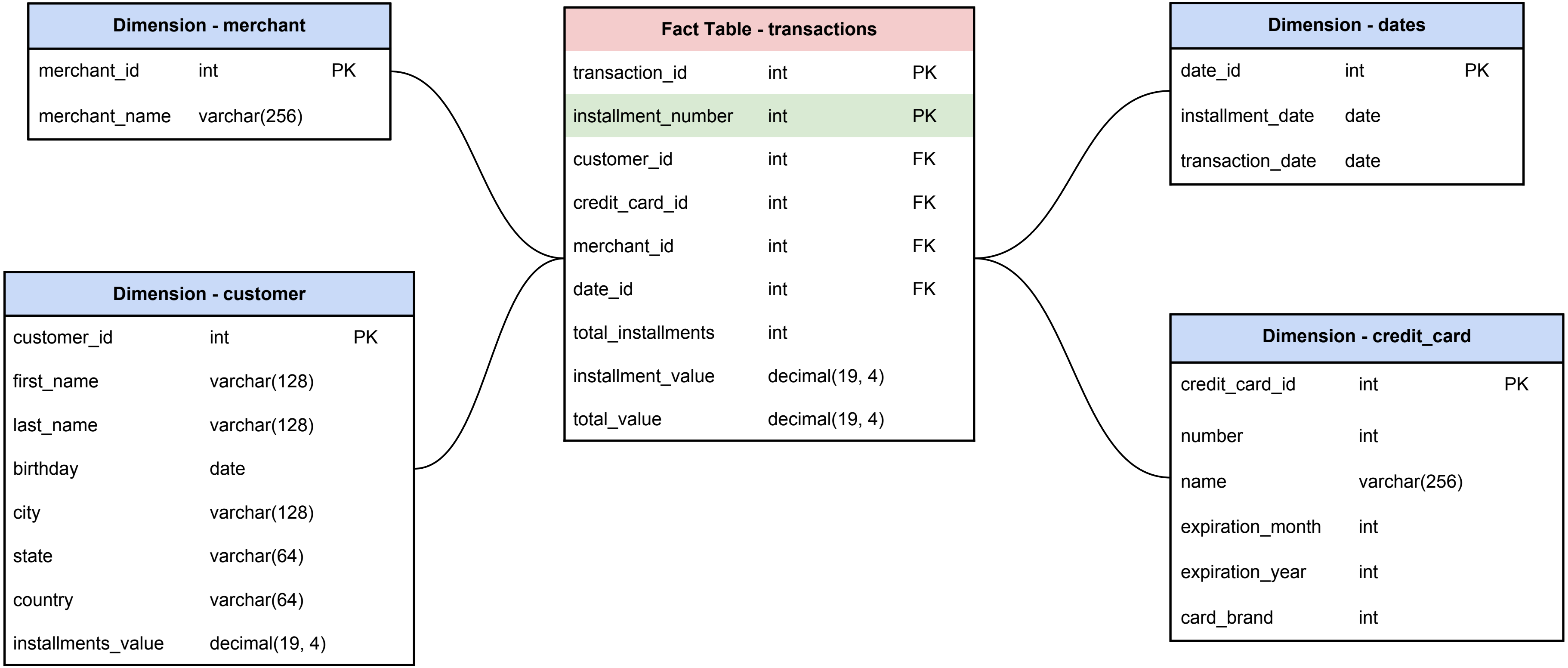
Dimension - dates		
date_id	int	PK
installment_date	date	
transaction_date	date	

Dimension - credit_card		
credit_card_id	int	PK
number	int	
name	varchar(256)	
expiration_month	int	
expiration_year	int	
card_brand	int	

Fact Tables

Fact Table - installments		
transaction_id	int	PK
installment_number	int	PK
customer_id	int	FK
credit_card_id	int	FK
merchant_id	int	FK
date_id	int	FK
total_installments	int	
installment_value	decimal(19, 4)	
total_value	decimal(19, 4)	

Complete Model



Dimension - merchant

merchant_id	int	PK
merchant_name	varchar(256)	

Dimension - customer

customer_id	int	PK
first_name	varchar(128)	
last_name	varchar(128)	
birthday	date	
city	varchar(128)	
state	varchar(64)	
country	varchar(64)	
installments_value	decimal(19, 4)	

Fact Table - transactions

transaction_id	int	PK
installment_number	int	PK
customer_id	int	FK
credit_card_id	int	FK
merchant_id	int	FK
date_id	int	FK
total_installments	int	
installment_value	decimal(19, 4)	
total_value	decimal(19, 4)	

Dimension - dates

date_id	int	PK
installment_date	date	
transaction_date	date	

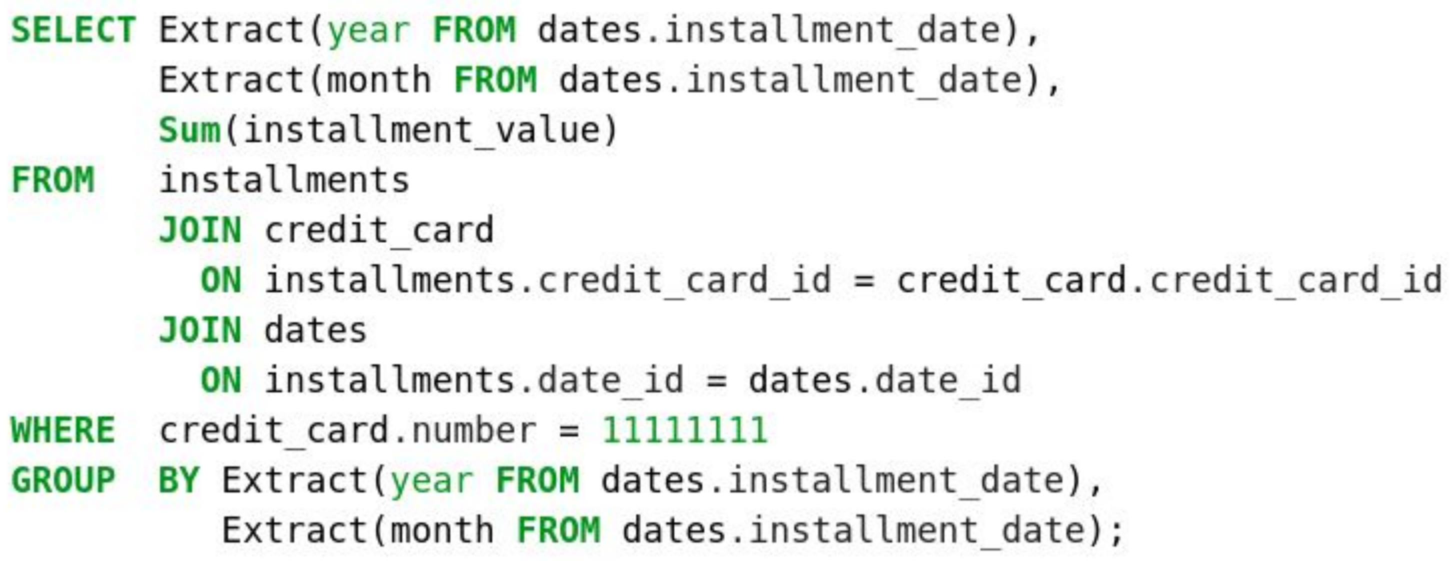
Dimension - credit_card

credit_card_id	int	PK
number	int	
name	varchar(256)	
expiration_month	int	
expiration_year	int	
card_brand	int	

Demo

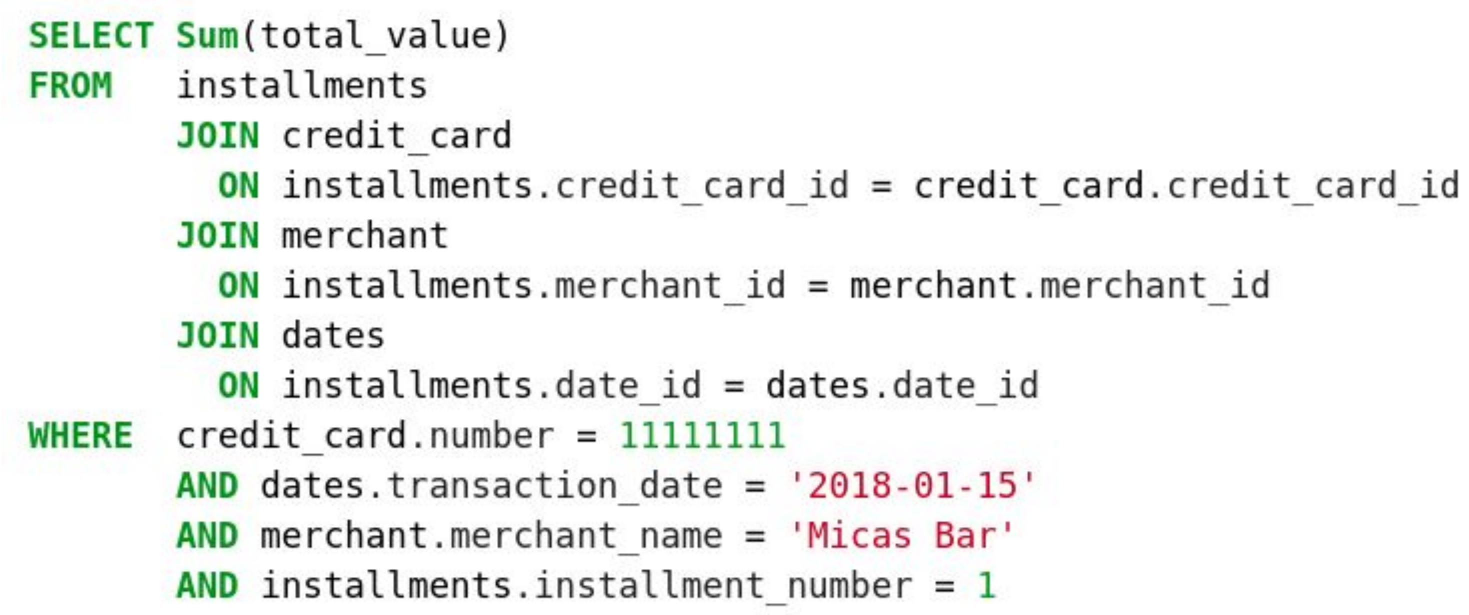
2. Build a query for the monthly bill

Query



```
SELECT Extract(year FROM dates.installment_date),
       Extract(month FROM dates.installment_date),
       Sum(installment_value)
FROM   installments
       JOIN credit_card
         ON installments.credit_card_id = credit_card.credit_card_id
       JOIN dates
         ON installments.date_id = dates.date_id
WHERE  credit_card.number = 11111111
GROUP  BY Extract(year FROM dates.installment_date),
          Extract(month FROM dates.installment_date);
```

Query



```
SELECT Sum(total_value)
FROM   installments
       JOIN credit_card
         ON installments.credit_card_id = credit_card.credit_card_id
       JOIN merchant
         ON installments.merchant_id = merchant.merchant_id
       JOIN dates
         ON installments.date_id = dates.date_id
WHERE  credit_card.number = 11111111
       AND dates.transaction_date = '2018-01-15'
       AND merchant.merchant_name = 'Micas Bar'
       AND installments.installment_number = 1
```

3. How to prevent these mistakes

What happened?

- Implicit Information
- Snowflake Schema ➜ Complex Queries

How to solve it

- Four-Step Dimensional Design Process
- Star Schema
- Slow Changing Dimensions

4. How to better find, understand and consume the data

- Follow Design Patterns
- Have several tools under your belt

Thank You!

Questions?

Nubank Challenge - 09/2018

Luiz Eduardo Amaral