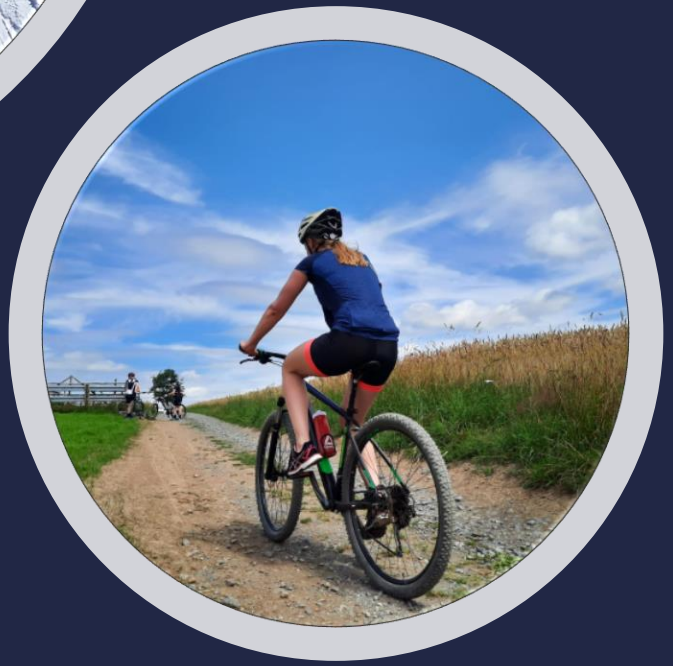
Dimensions Demystified

Paulien van Eijk
02-06-2022

Demystify several types of dimensions and discuss which problems they can solve.

Paulien van Eijk



Who am I?

Before we start..

Focused on Power BI

Intermediate level

First time speaker

Q&A

Star Schema

Slowly Changing Dimension

Degenerate Dimension

Role-playing Dimension

Junk Dimension

Star Schema

Facts

- Store observations and events and contain numeric data about business processes
  - Orders
  - Units
  - Revenue

Dimensions

- Contains descriptive attributes which can determine the aggregation level of the fact:
  - Date
  - Product
  - Customer

Relationships

- Dimension contains a unique row identifier for each row
- Fact table contains a column with corresponding keys

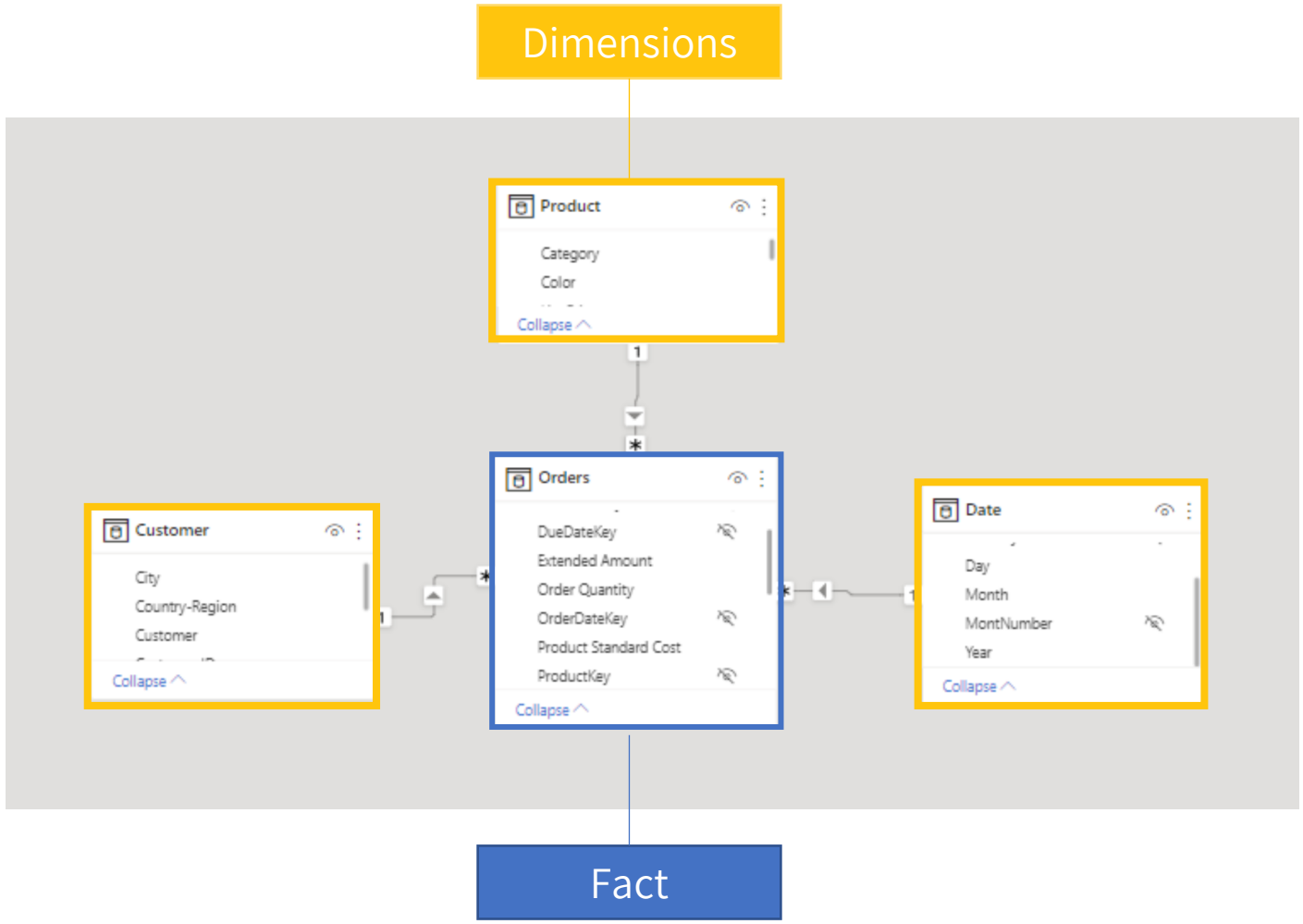
Dim Date

DateKey	Date
01042022	01-04-2022
02042022	02-04-2022
03042022	03-04-2022

Fact Orders

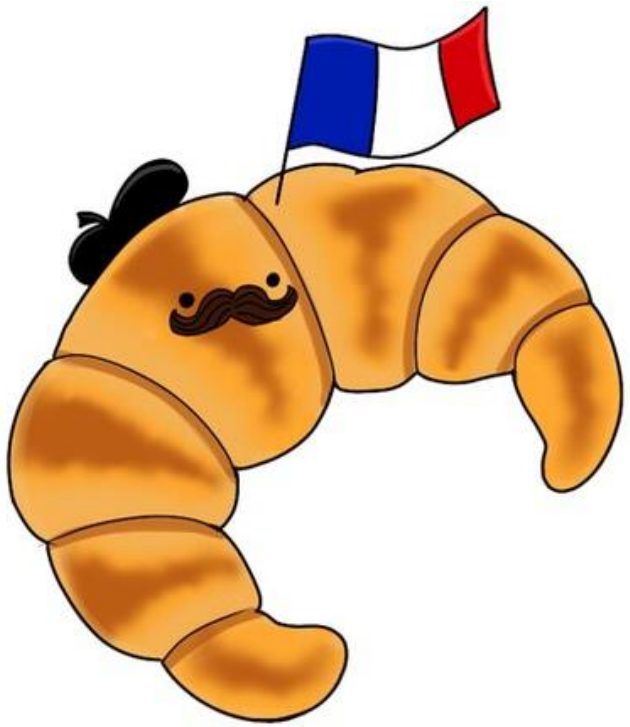
DateKey	Quantity	..
01042022	100	..
01042022	200	..
01042022	100	..

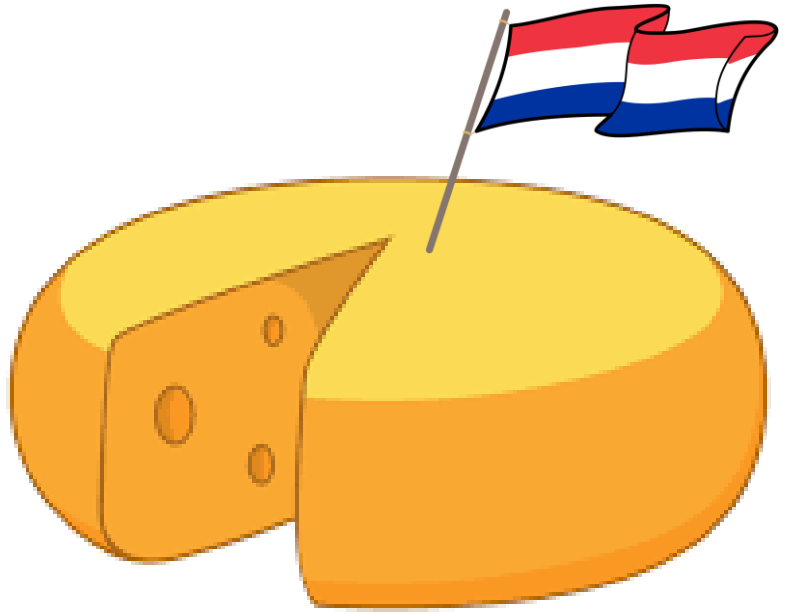
Star schema


Dimensions
Product
Category
Color
Collapse
1
*
Orders
DueDateKey
Extended Amount
Order Quantity
OrderDateKey
Product Standard Cost
ProductKey
Collapse
Customer
City
Country-Region
Customer
Collapse
Date
Day
Month
MontNumber
Year
Collapse
Fact

Slowly Changing Dimensions

Paulien van Eijk





Paulien van Eijk

What should we do?

- Replace old address with new address

- Store old address and new address

Slowly Changing Dimension

- Dimension that manages the changes of the attributes over time
  - Phone number
  - Email address
  - Address
- Type 1 and Type 2

Slowly Changing Dimension Type 1

- Overwrite the last value in the dimension with the new value
  - Email adresses
  - Phone numbers
- Not an issue for supplemantary values

Slowly Changing Dimension Type 2

- Keep te past and current values in the dimension
  - In this way we do not lose any information

- The business key stays the same, but the surrogate key changes.
  - **Surrogate key:** unique identifier of the row without any business meaning
  - **Business key / Alternate key:** might have a meaning..

Demo Time

Type 1 vs. Type 2

- Keep only the latest value: Type 1
- Keep past and current value: Type 2

But there is even a third option

- And it’s not called Slowly Changing Dimension Type 3
- Create a new table
  - Dim Address

But there is even a third option

Fact Orders

CustomerKey	AddressKey	DateKey	..
12301	100000	10042021	..
12301	100003	11052021	..
12301	100003	12062021	..

Dim Customer

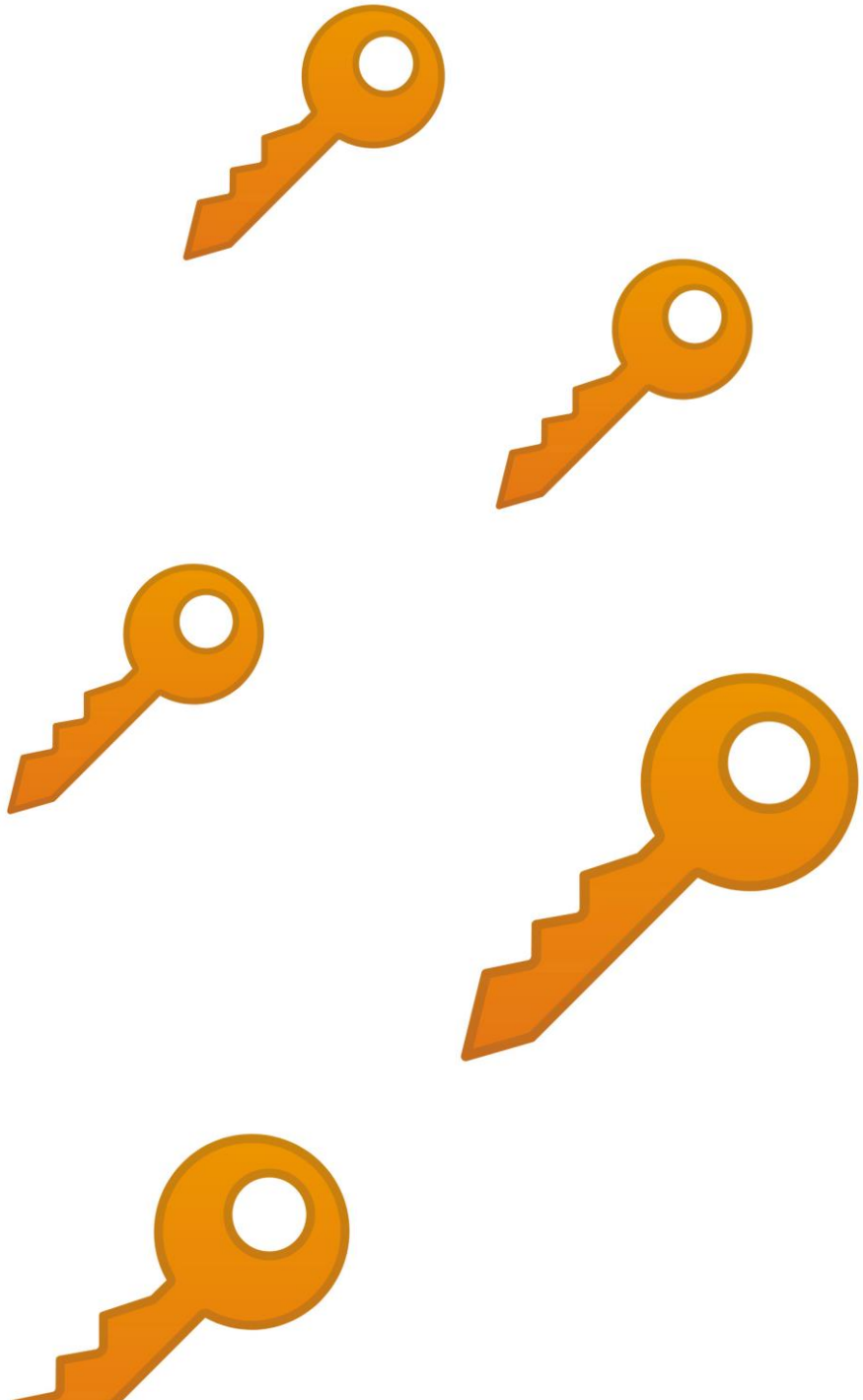
CustomerKey	Name
12301	Sam Turner
12302	Isa Cox
12303	Rosa Hu

Dim Address

AddressKey	Country	Postal Code
100000	France	95003
100002	Netherlands	5352 CZ
100003	Netherlands	1027 AB

But there is even a third option

- The address will always be assigned correctly
- But more difficult to find out where the customer actually lives.
  - There we could store his current address back in dim Customer
  - But then we are in a way multplicating the data
  - Or we create a snowflake schema..
- Many solutions to model one problem



Degenerate Dimensions

Orderlines

- Orderline table with keys and values
- Therefore, the only information left of the Order is the Ordernumber

Orderlines

Ordernumber	DateKey	ProductKey	CustomerKey	Amount
AB3732	20210901	345	12301	200
AB3732	20210901	221	12301	100
AB3733	20210902	343	14221	300
AB3733	20210902	221	14221	200

Following the standard ★

- Dim Order with :
  - OrderKey (surrogate key)
  - Ordernumber (business key)

Dim Order

OrderKey	Ordernumber
3732	AB3732
3733	AB3733

Following the standard ★

Dim Order

OrderKey	Ordernumber
3732	AB3732
3733	AB3733

1 .. *

Fact Orderlines

OrderKey	DateKey	ProductKey	CustomerKey	Amount
3732	20210901	345	12301	200
3732	20210901	221	12301	100
3733	20210902	343	14221	300
3733	20210902	221	14221	200

Following the standard ★

- Very clear to the end user
- A lot of new unique values → model size might blow up

Degenerate Dimension 🔑

- Dimension key is placed in the fact table, but there is no associated dimension table with it.
- All necessary attributes are stored in other dimensions

Degenerate Dimension 

- Dimension key placed in the fact table, but there is no associated dimension table.
- All necessary attributes are stored in other dimensions

Fact Orderlines

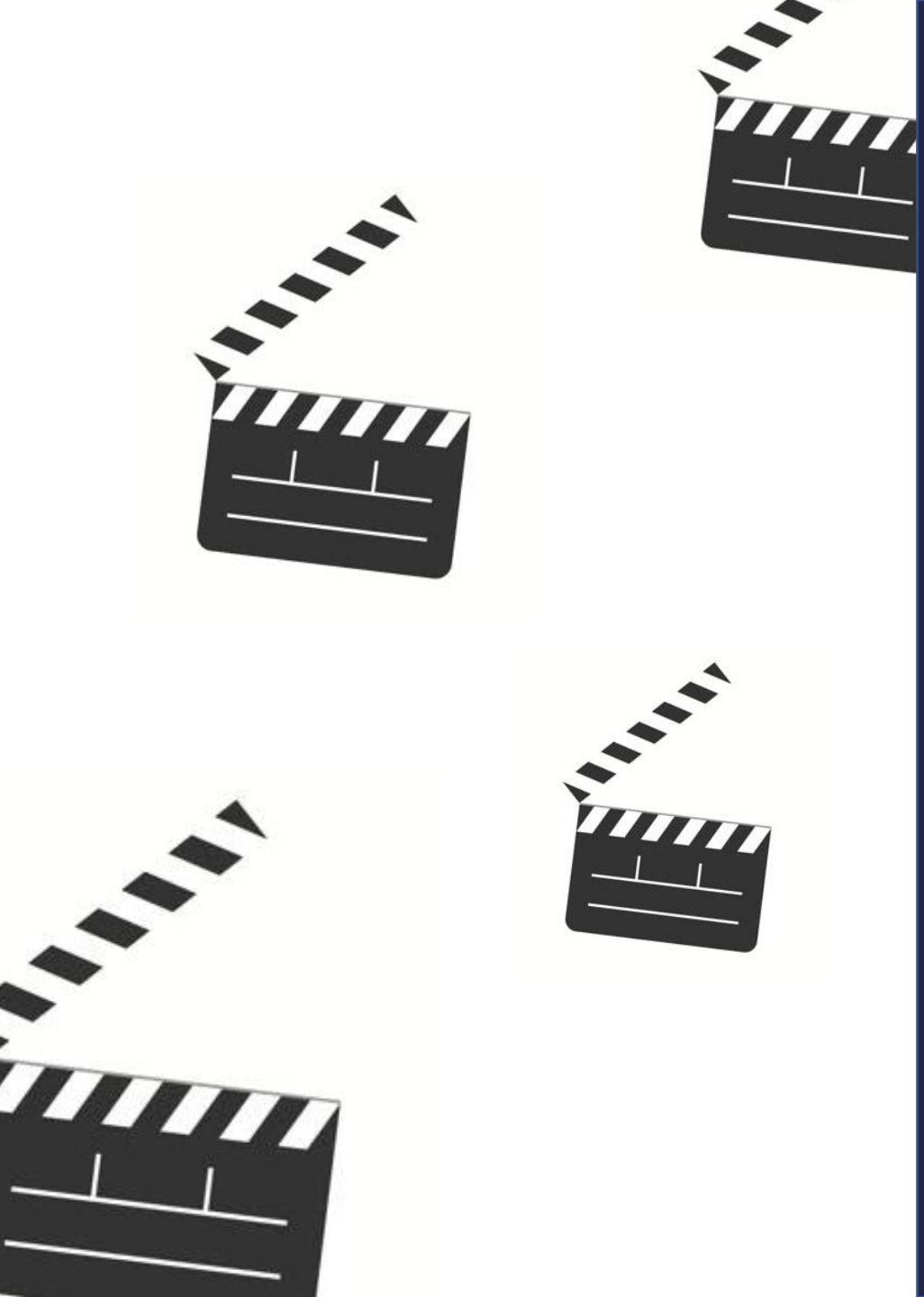
Ordernumber	DateKey	ProductKey	CustomerKey	Amount
AB3732	20210901	345	12301	200
AB3732	20210901	221	12301	100
AB3733	20210902	343	14221	300
AB3733	20210902	221	14221	200

Degenerate Dimension 🔑

- Model size will be smaller
- Performance of might decrease when filtering on ordernumber (depending on the calculations)
- Unclear for the end user

Star schema vs. Degenerate Dimension

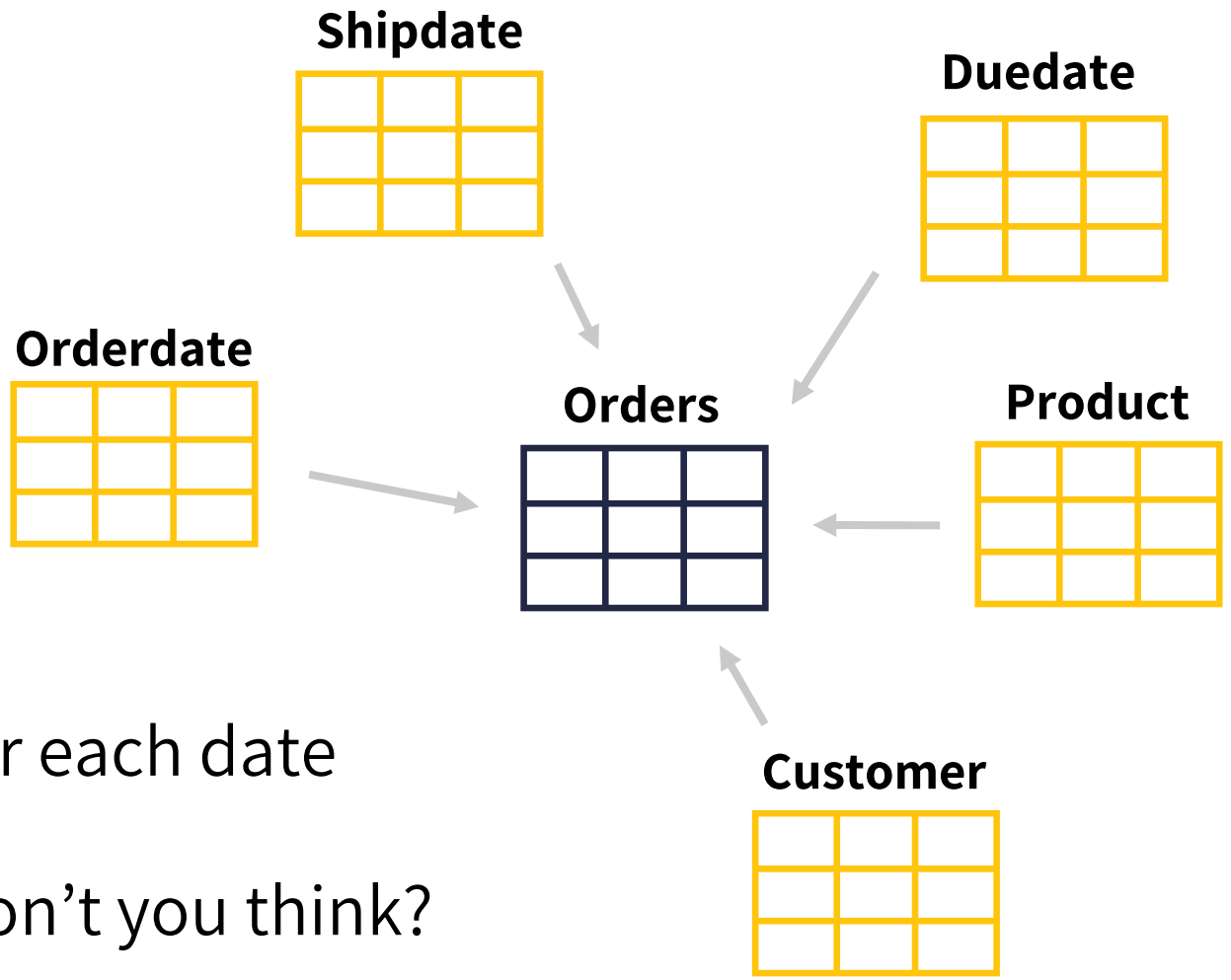
- Filter report on Ordernumber 
- Clear model for the end user 
- Model size 



Role Playing Dimensions

Following the standard ★

- Three dates in fact table:
  - Order Date
  - Ship Date
  - Due Date
- Import a date dimension for each date
- It’s getting a bit crowded don’t you think?

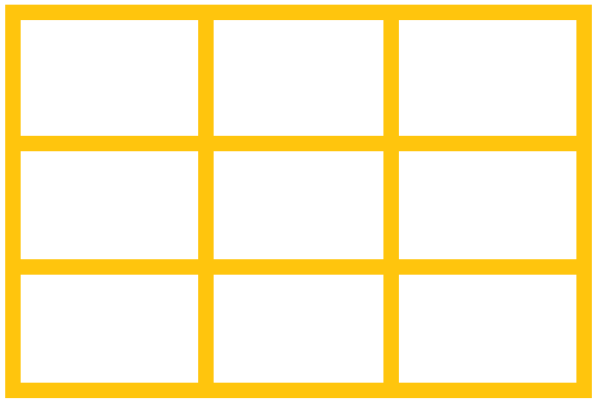


Role Playing Dimension 

- Dimension that is used to filter the related fact table over multiple different relationships.

Role Playing Dimension

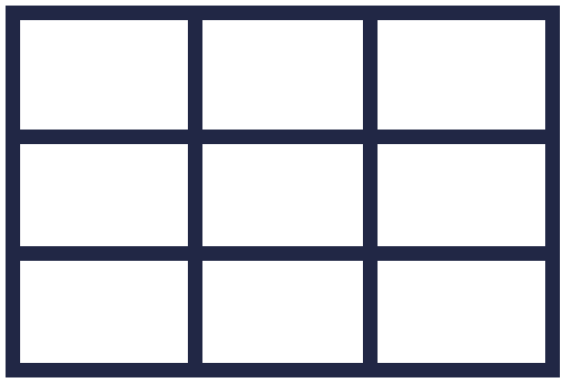
Date



Date[DateKey]

1 .. *

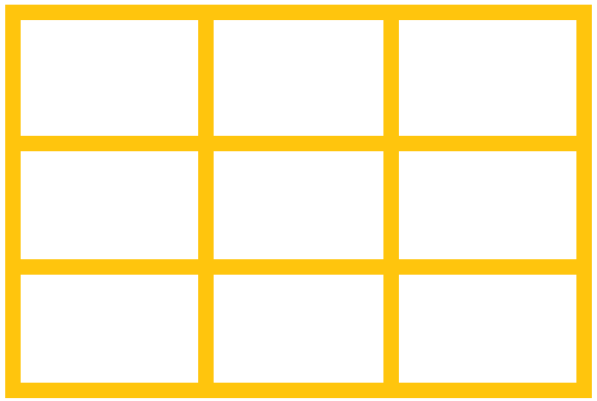
Orders



Orders[OrderDateKey]

Role Playing Dimension

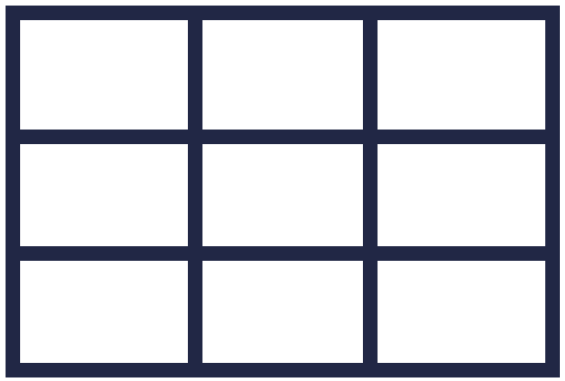
Date



Date[DateKey]

1 .. *

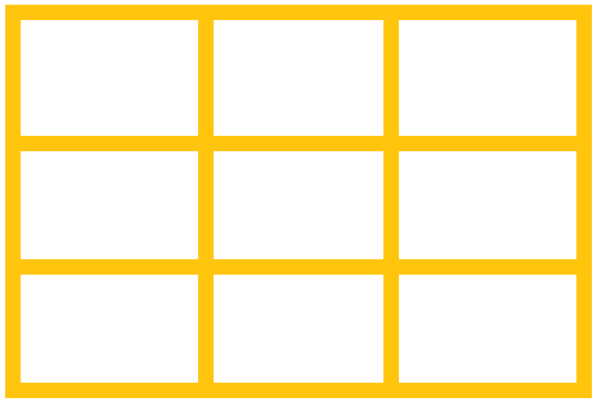
Orders



Orders[ShipDateKey]

Role Playing Dimension

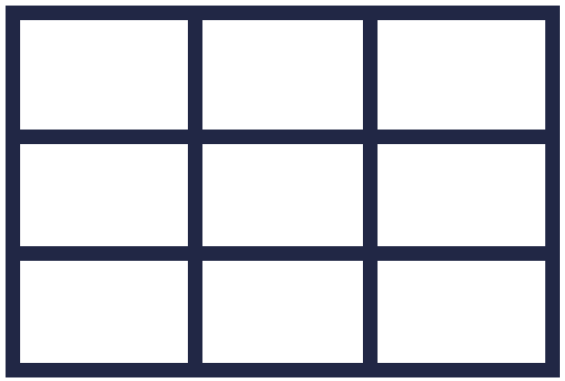
Date



Date[DateKey]

1 .. *

Orders



Orders[DueDateKey]

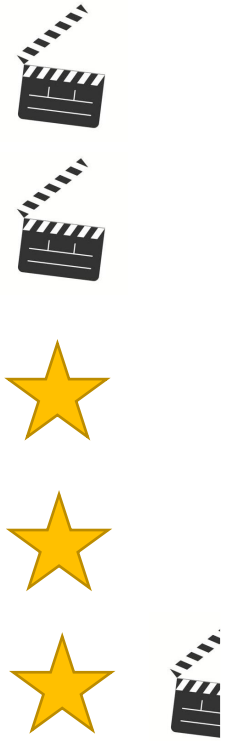
Demo Time

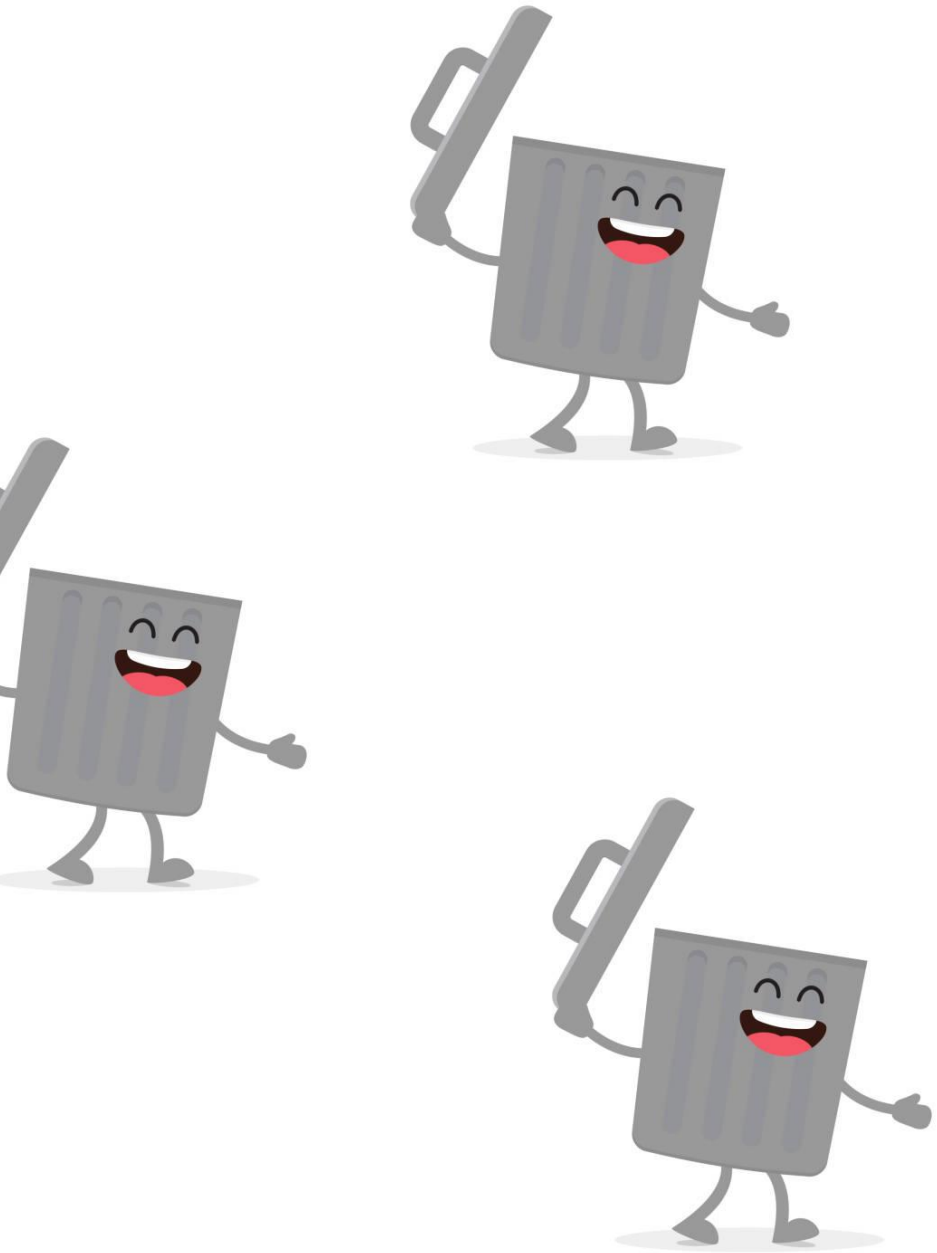
Role Playing Dimension

- USERELATIONSHIP()

Star schema vs. Role Playing Dimension

- Only a few date tables (instead of 3.. Or 6.. Or 21)
- Easily show multiple measures in one graph
- Filter sales simultaneously by different types of dates
- # Measures
- Clear for the end user





Junk Dimensions

So many flags?

- Flags on lowest granularity
- Add them to the fact table?

Fact Ordelines

Ordernumber	ProductKey	IsShipped	IsPaid	IsPackaged	IsTransferred
AB3732	345	1	1	1	1
AB3732	221	0	1	0	1
AB3733	343	1	0	1	1
AB3733	221	1	0	1	1

Following the standard ★

- Dim Orderline with:
  - Orderline Key & Ordernumber
  - Add the flags

Dim Orderline

OrderLineKey	Ordernumber	IsShipped	IsPaid	IsPackaged	IsTransferred
1	AB3732	1	1	1	1
2	AB3732	0	1	0	1
3	AB3733	1	0	1	1
4	AB3733	1	0	1	1

Dim Orderline

OrderLineKey	Ordernumber	IsShipped	IsPaid	IsPackaged	IsTransferred
1	AB3732	1	1	1	1
2	AB3732	0	1	0	1
3	AB3733	1	0	1	1
4	AB3733	1	0	1	1

1 .. * (but actually 1 .. 1)

Fact Ordelines

OrderLineKey	Quantity
1	100
2	200
3	300
4	200

Paulien van Eijk

Following the standard ★

- Very clear for the end user
- A lot of new unique values → model size might blow up

Junk Dimension

- Dimension that stores indicators and flags with low cardinality.
- **Low cardinality:** A column that has a lot of duplicated data (0 / 1, yes / no, low / medium / high) has low cardinality

Junk Dimensions 

Fact Orderlines

Ordernumber	ProductKey	IsShipped	IsPaid	IsPackaged	IsTransferred
AB3732	345	1	1	1	1
AB3732	221	0	1	0	1
AB3733	343	1	0	1	1
AB3733	221	1	0	1	1

Junk Dimensions

- Dimension that stores indicators and flags with low cardinality.
- Create one row for each of the flag combinations

Dim Ordeline Status

OrderLineStatusKey	IsShipped	IsPaid	IsPackaged	IsTransferred
1	1	1	1	1
2	0	1	0	1
3	1	0	1	1

Junk Dimensions

Dim Orderline Status

OrderLineStatusKey	IsShipped	IsPaid	IsPackaged	IsTransferred
1	1	1	1	1
2	0	1	0	1
3	1	0	1	1

1 .. *

Fact Orderlines

Ordernumber	ProductKey	OrderlineStatusKey
AB3732	345	1
AB3732	221	2
AB3733	343	3
AB3733	221	3

Paulien van Eijk

Star schema vs. Junk Dimension

- Model size
- Clear for the end user

There isn’t always one good answer

- And that’s why the following phrase is our favorite phrase in our field of work..

“Lunch break”

“It depends”

Paulien van Eijk

There isn’t always one good answer

- But make sure, you always document your model very well so the end user knows what to expect!

How?

- Add descriptions to measures
- Add descriptions to (ambigious) columns
- Create content page on e.g. sharepoint
- Add a report with all the definitions

Wrap up

- Four implementations of a dimension
- Each their pros and cons
- Document, document, document..

Q&A

Feedback

https://evals.datagrillen.com