Intro to Dimensional Modelling

Koen Verbeeck

MINECRAFT
STAR SCHEMA EDITION

Contact

Koen Verbeeck

@Ko_Ver

@koenv.bsky.social

SQLkover.com

Meet Tom!

Business analyst at fictional online retailer

Tasked with finding their top customers

Knows a bit of SQL and Power BI

dalanzo

Tom needs data

Tom asks IT if he can access the source databases

They obviously say "no"

Tom talks to the CFO, and he "convinces" IT to give Tom read access

But the SQL queries are slow and complex

One of Tom's queries accidentally contains a cross join and takes the production server down

Poor Tom

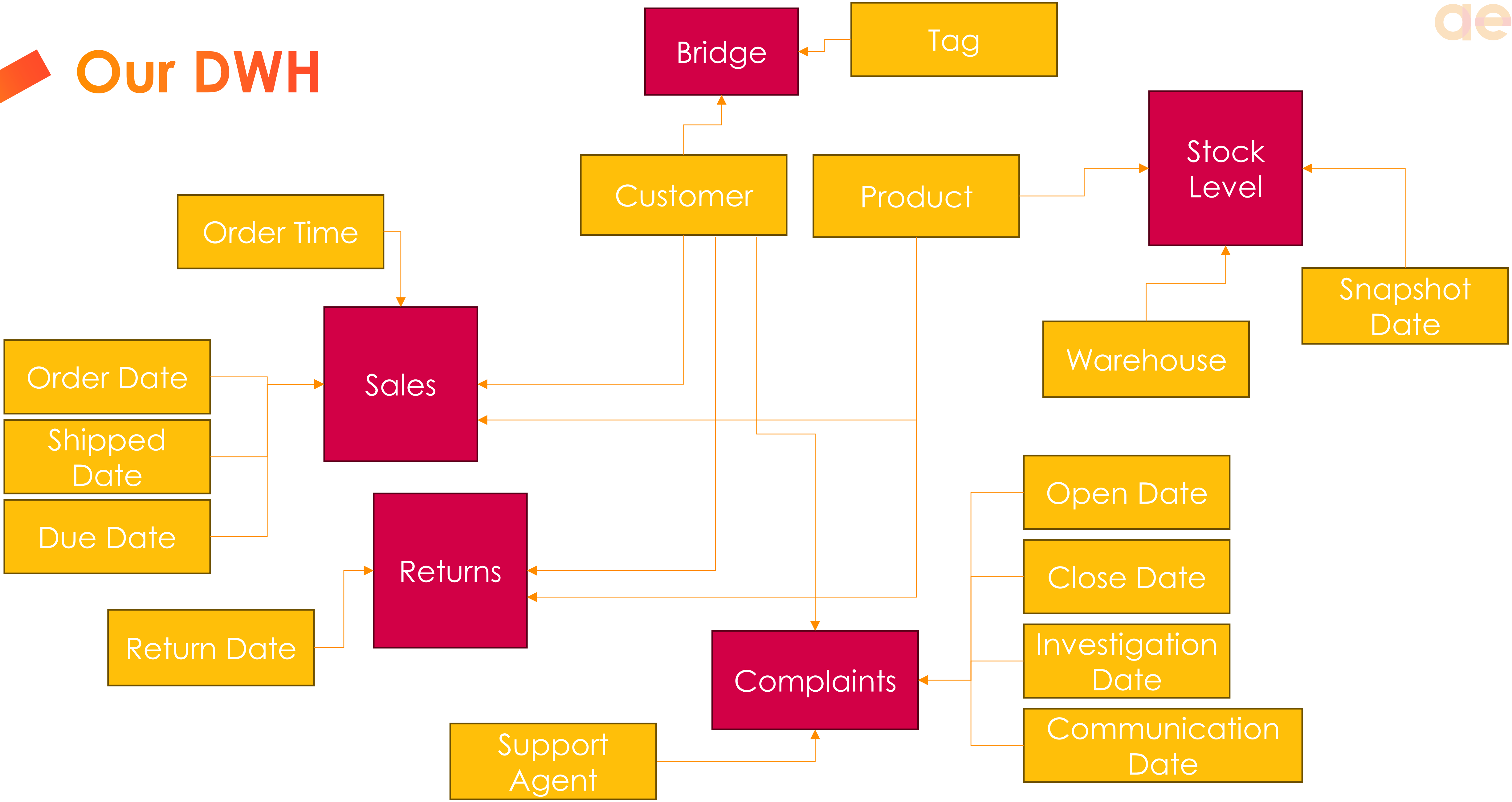
The data is in multiple source systems and he doesn't know how to combine them

Only the latest customer information is kept, making trend analysis hard

He doesn't get the data or results that he wants

Maybe he should talk with those "analytics engineers" from the data team?

Our DWH


Bridge
Tag
Customer
Product
Stock Level
Order Time
Snapshot Date
Warehouse
Sales
Order Date
Shipped Date
Due Date
Returns
Return Date
Open Date
Close Date
Investigation Date
Complaints
Communication Date
Support Agent

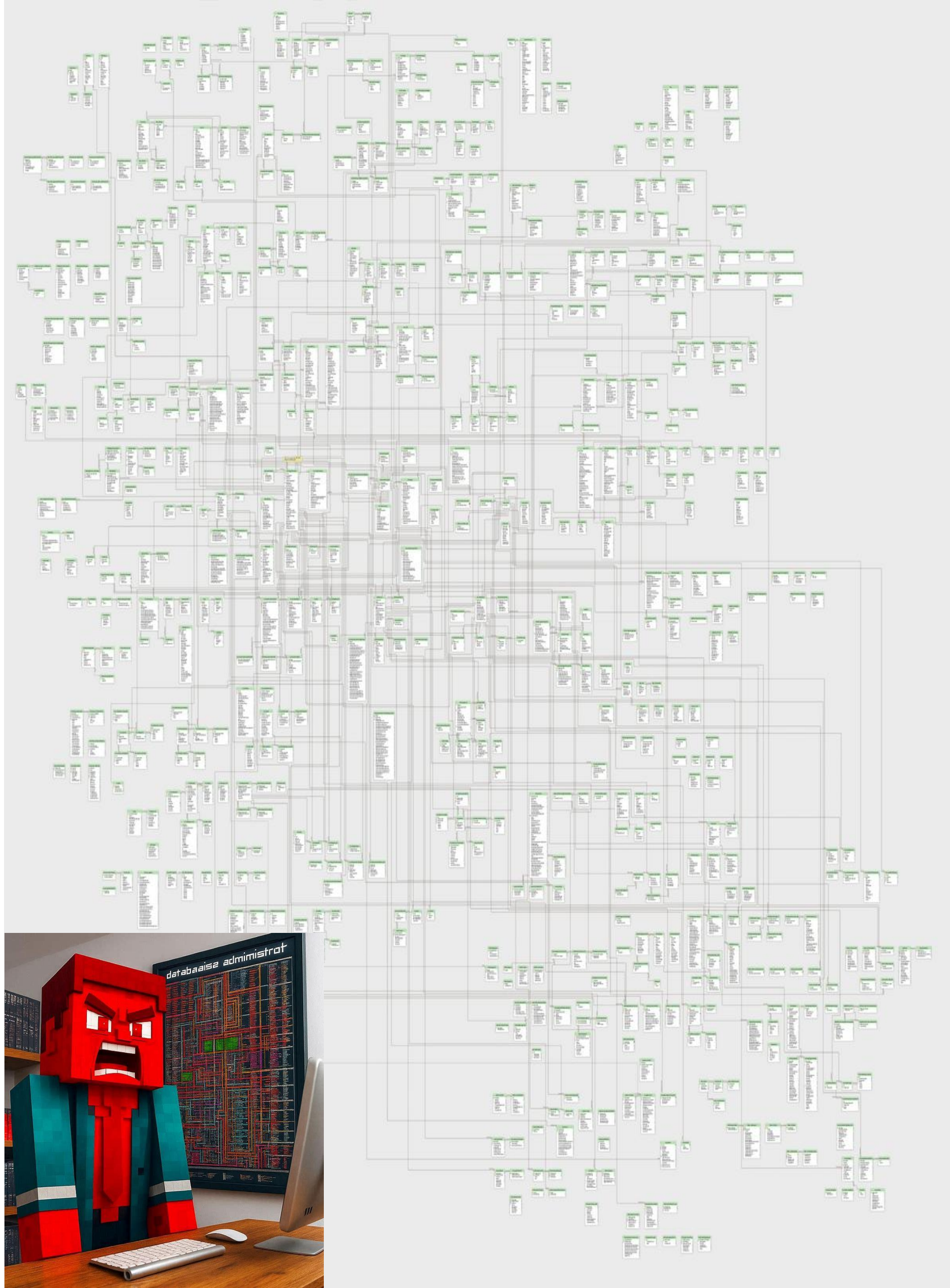
Outline

Why dimensional modelling?

Operational Systems

- Typically highly normalized
  - eliminate redundancy
- Optimized for writes
- Supports **execution** of business process

- Multiple source systems in one company
- Usually no history (or it’s hard to query)
- Analytical queries become very complex
  - These put a burden on the critical DB
- No user access

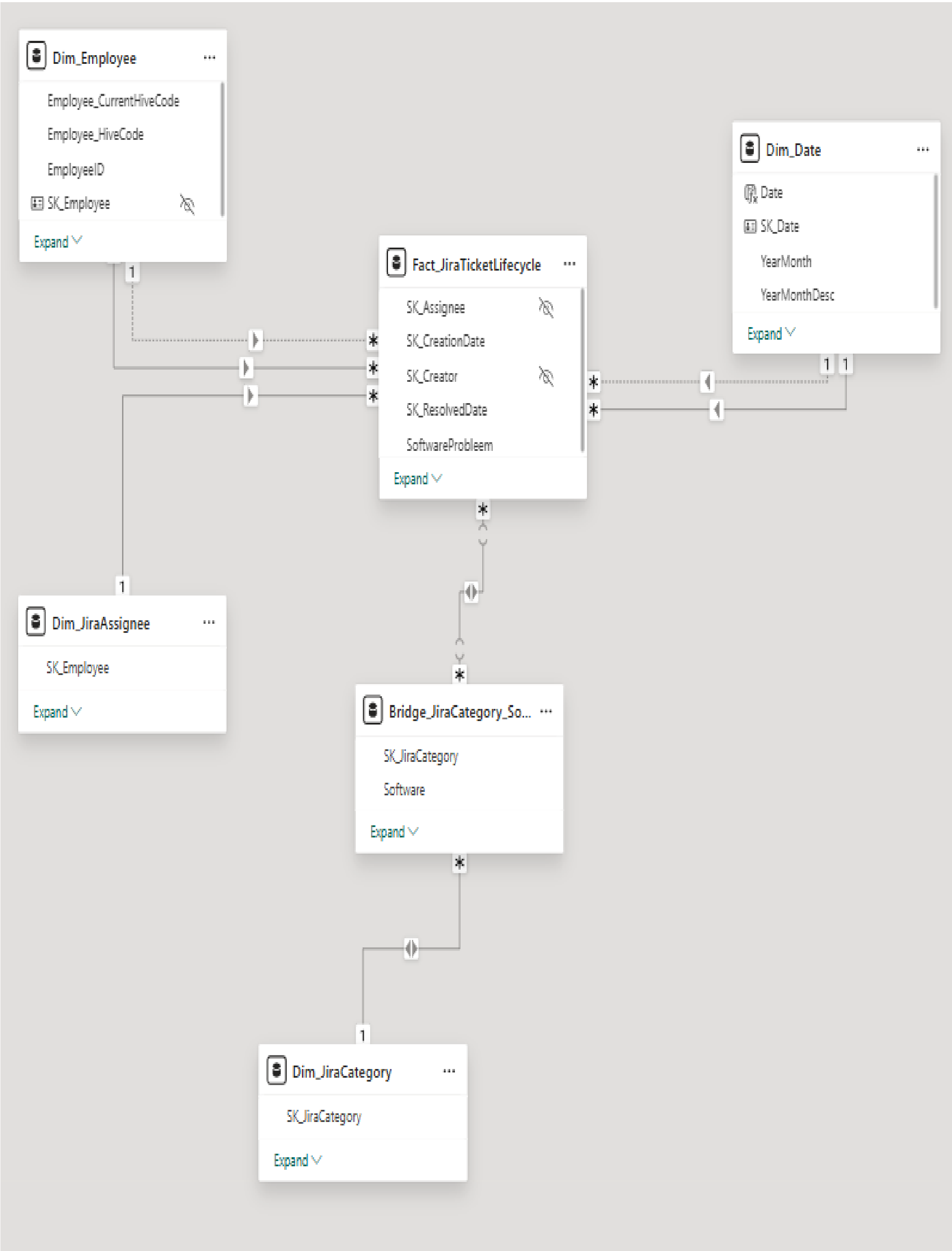


Data Warehouse

- Centralizes data from different systems
- Optimized for reads
  - Redundancy is allowed again

- Tracks history of business entities
- Easier analytical queries
- Open to anyone who uses data for analysis

- Supports **analysis** or **evaluation** of business process



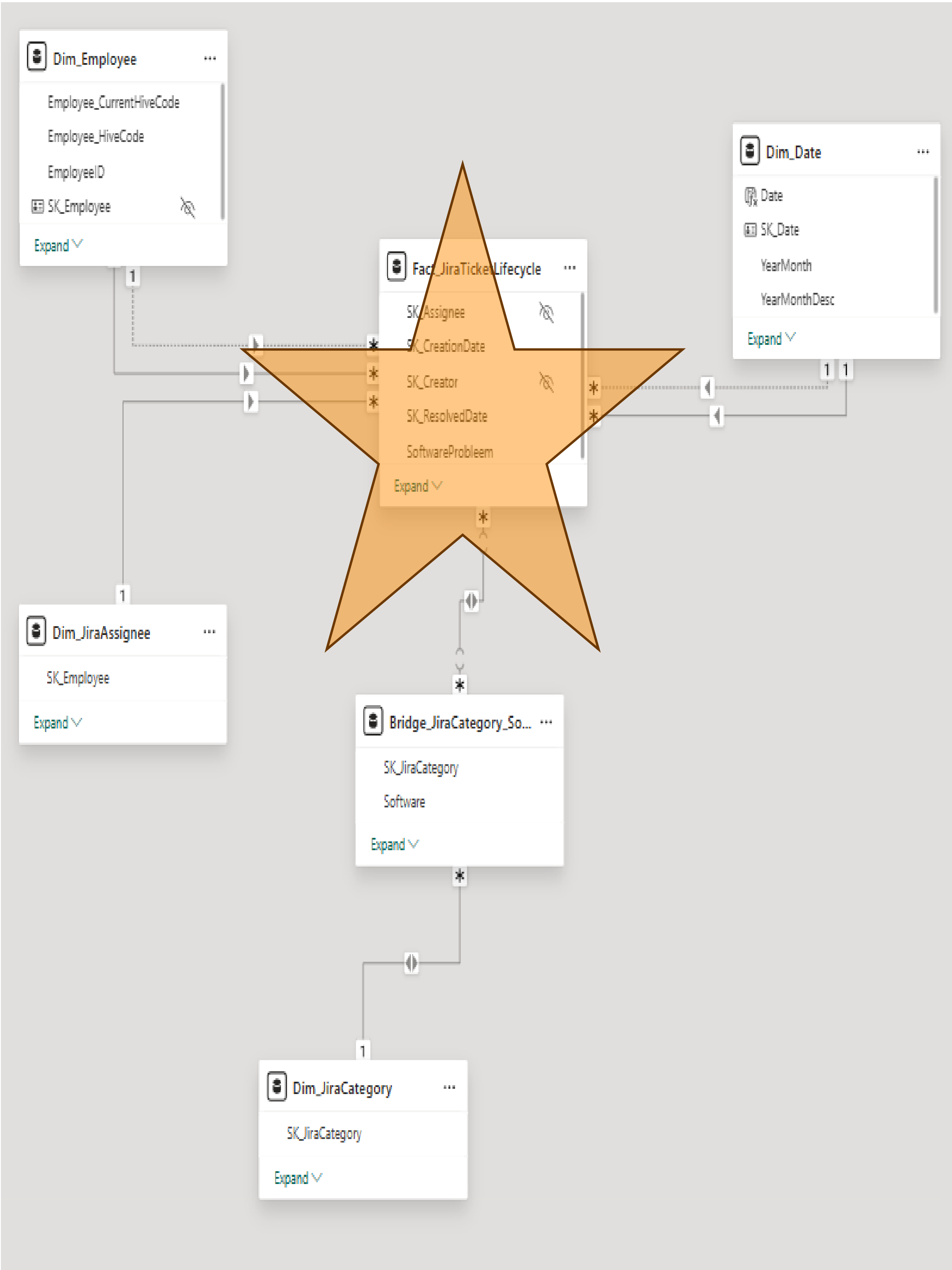
Star Schema

- Two main building blocks

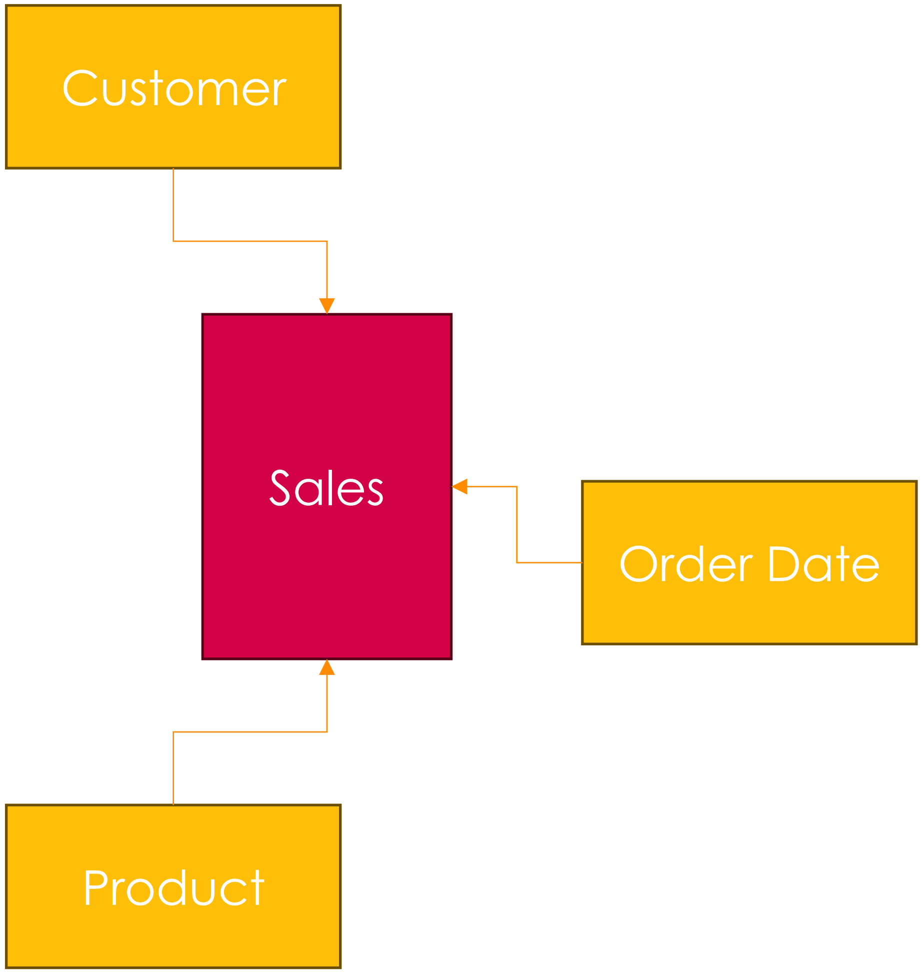
- **Dimensions**
  - descriptive information about business entities
  - employee, product, date, order, customer …

- **Facts**
  - measurements of a business process
  - sales, returns, stock level, temperature …

- Easy to interpret by business users
- Allows for easier and faster analytical queries



Sales Star Schema



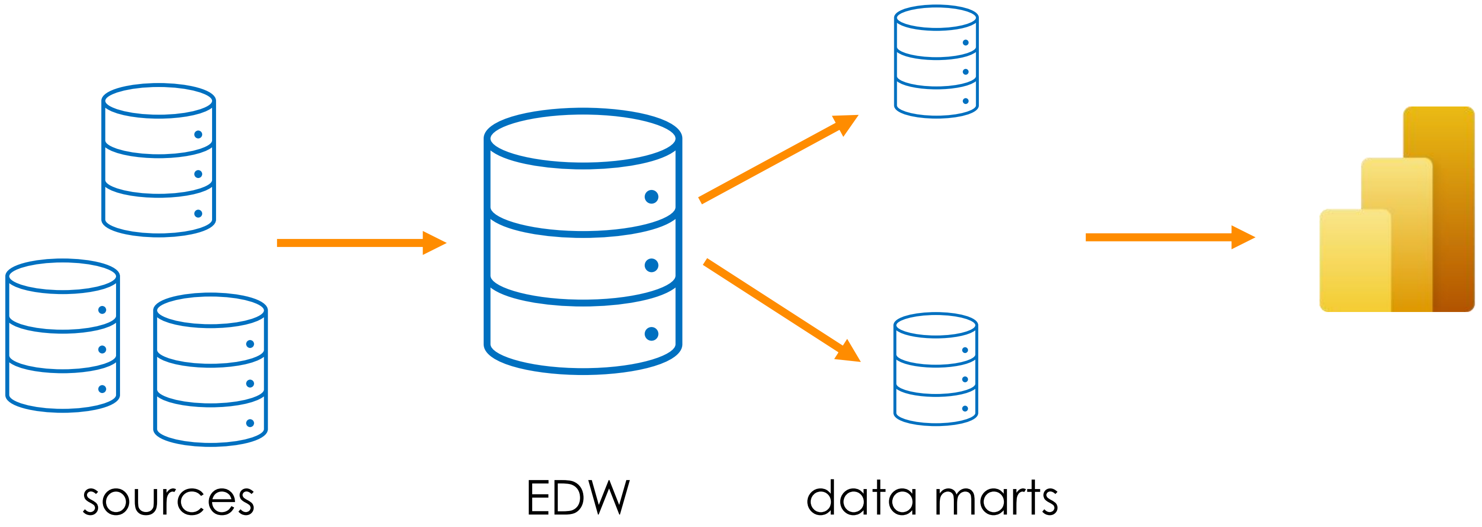
Data modelling for data warehouse

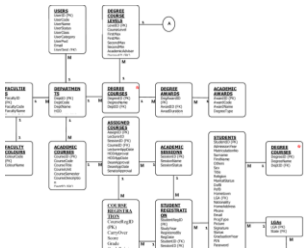
Different methods exist for modelling data warehouses

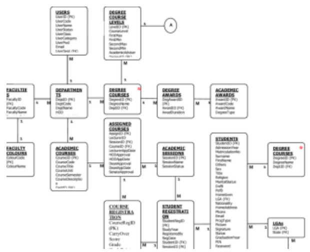
They all have their strengths & weaknesses

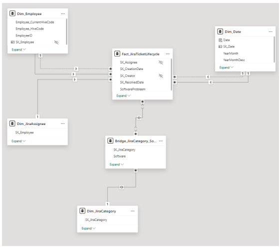
Let’s take a tour!

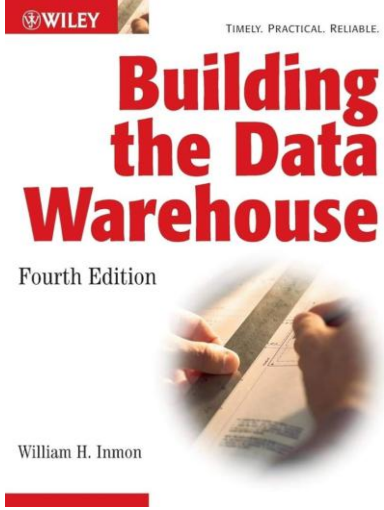
Inmon


sources
EDW
data marts

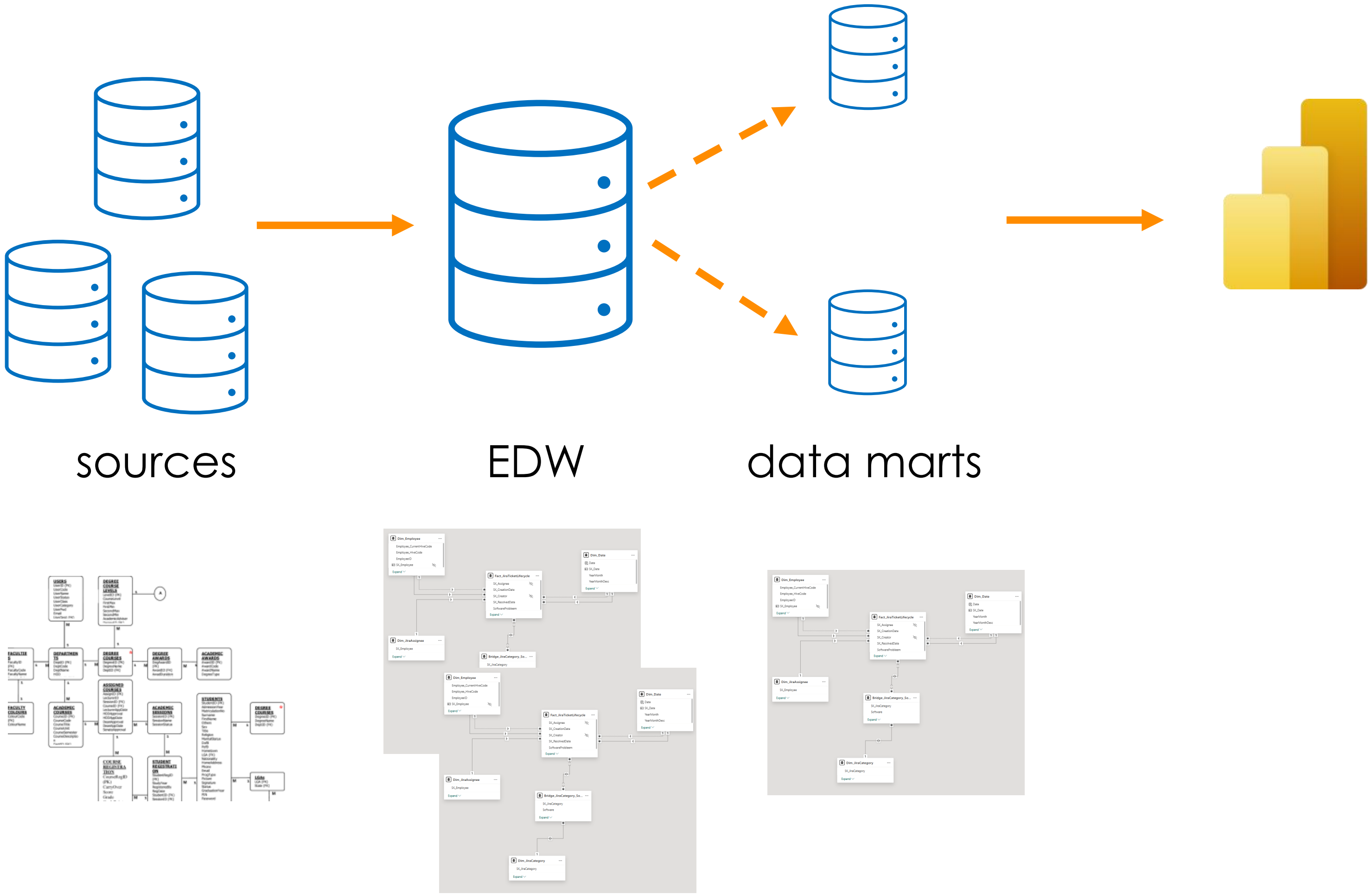


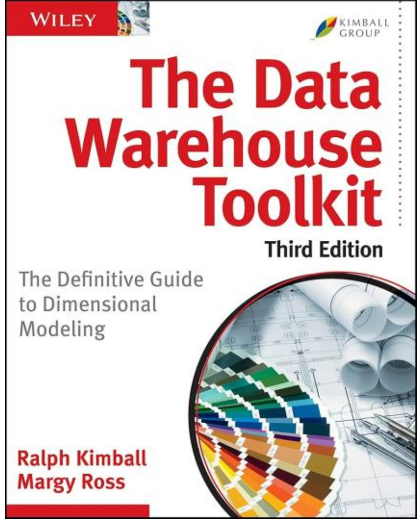


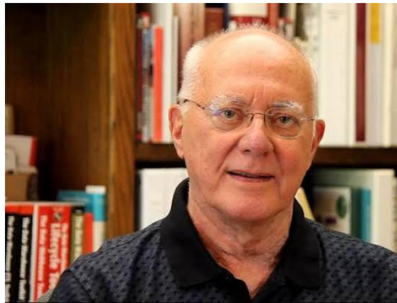



WILEY
TIMELY. PRACTICAL. RELIABLE.
Building the Data Warehouse
Fourth Edition
William H. Inmon

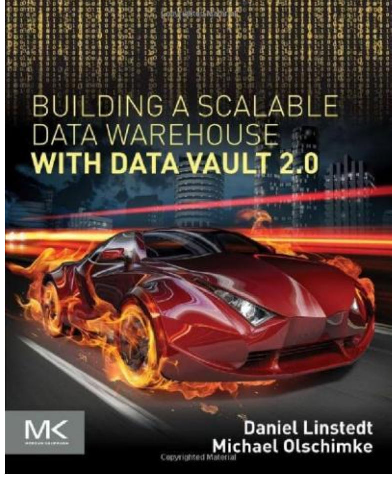
Kimball

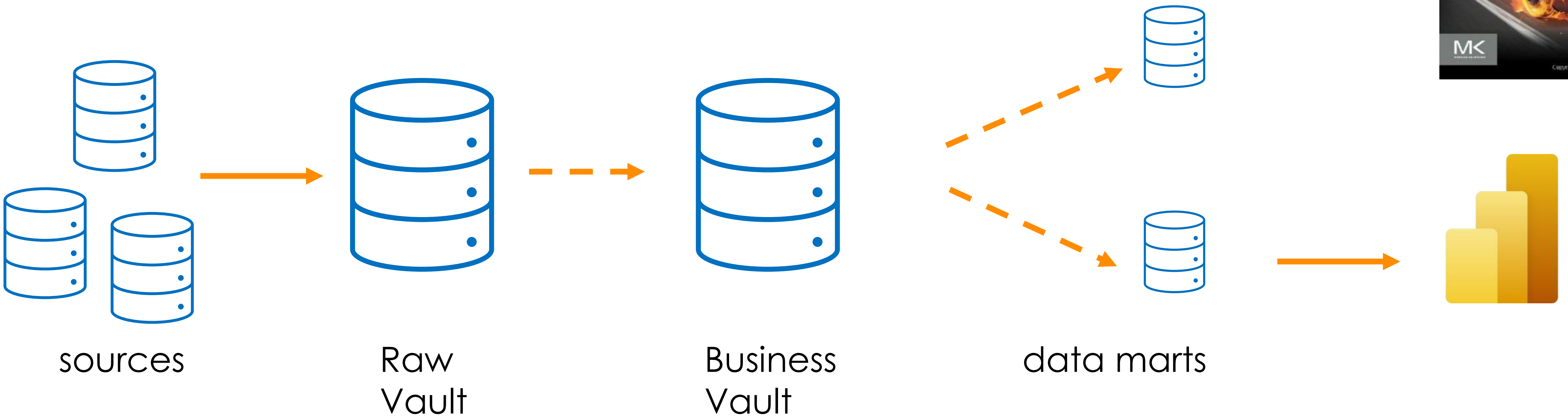

sources
EDW
data marts

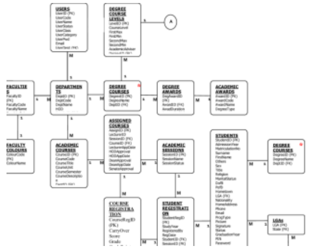

WILEY
KIMBALL GROUP
The Data Warehouse Toolkit
Third Edition
The Definitive Guide to Dimensional Modeling
Ralph Kimball
Margy Ross

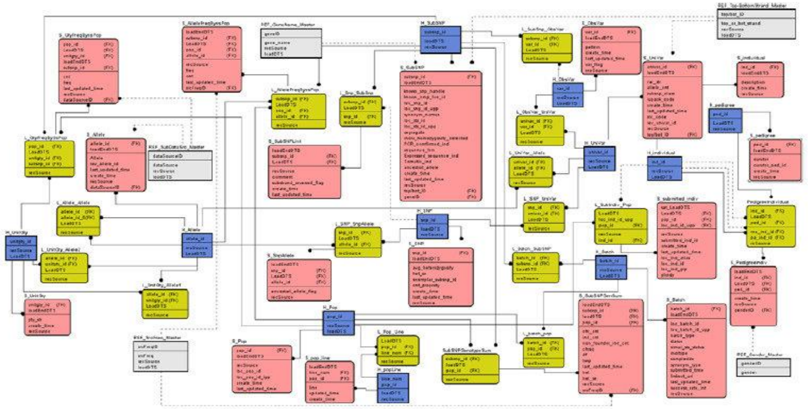


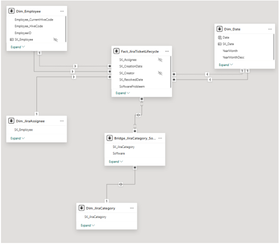
Data Vault


BUILDING A SCALABLE
DATA WAREHOUSE
WITH DATA VAULT 2.0
MK
Daniel Linstedt
Michael Olschimke


sources
Raw Vault
Business Vault
data marts







One Big Table

One Table To Rule Them All!

Typically created by data scientists, citizen developers or by data engineers who think per use case

OK for **proof of concepts** or **machine learning**

Many drawbacks for analytical use cases

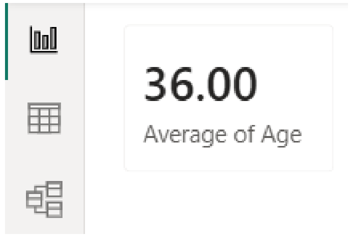
One Big Table

Tom used Power Query to pull in data from the sales system into Power BI

He created one query where he merges different queries together

Tom now wants to calculate the average age of their customers

Customer	Age	Item	Amount
Sarah	30	Pants	100
David	70	T-Shirt M	25
Sarah	30	Blouse XXS	30
Sarah	30	Blouse XS	30
Lucie	20	Shoes 38	85



One Big Table – what went wrong?

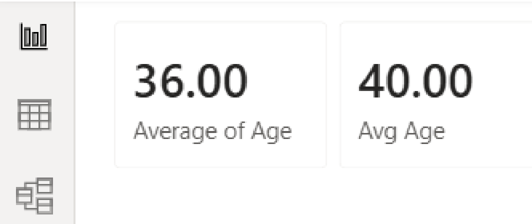
When averaging age over all rows, you get a **weighted** average

The correct average customer age needs to be calculated using a DAX formula:

```
AVERAGEX(
    SUMMARIZE(Sales,Sales[Customer],Sales[Age])
    ,Sales[Age])
```

If age was stored in a separate customer table, there wouldn't be a need for complex DAX

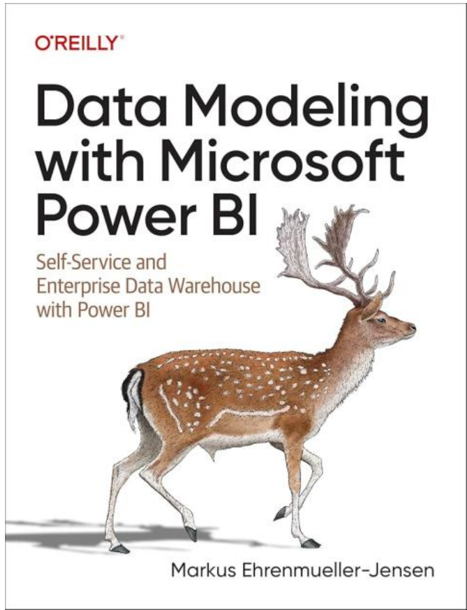
Customer	Age	Item	Amount
Sarah	30	Pants	100
David	70	T-Shirt M	25
Sarah	30	Blouse XXS	30
Sarah	30	Blouse XS	30
Lucie	20	Shoes 38	85



Why dimensional modelling?

- **Easy to understand** for business users
- Minimizes joins, but still with great flexibility
  - SQL Server has specific optimizations for star joins
- Acts as a **semantic layer**

- Most **optimal model for Power BI**
  - The documentation says so
  - Dimensions are for *filtering* and *grouping*
  - Facts are for *summarization*
  - Best performance is achieved with star schemas (presentation by Fabric CAT Team member Benni)
  - It simplifies DAX formulas



What are dimensions?

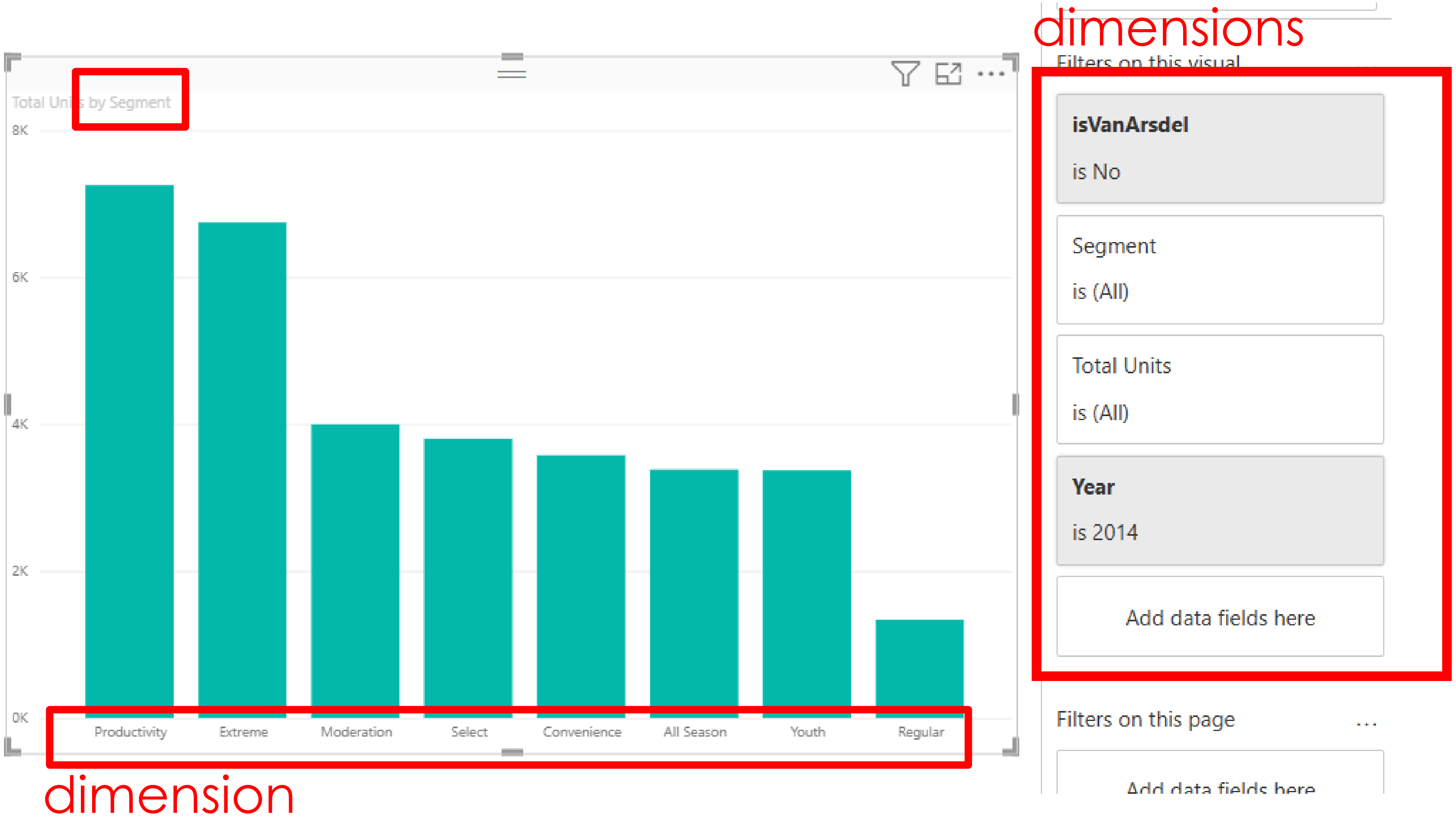
What are dimensions?

Ask the question: what do you want to analyze?

"I want to analyze *sales amounts*,
by customer,
by order date,
by product,
by store,
by promotion,
…
"

Or ask **the 7Ws**
(who, what, where, when, how many, why, how)

What are dimensions?



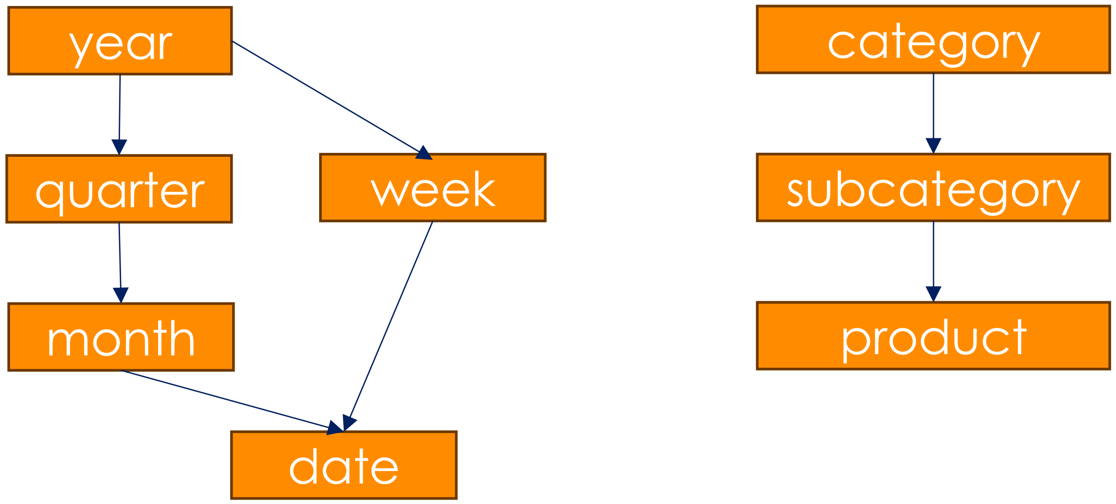
Dimension characteristics

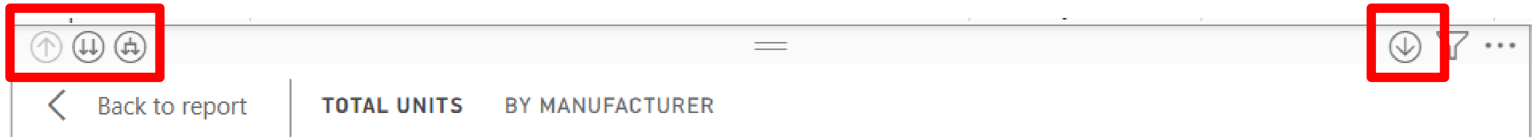
- Many (textual) attributes
  - Cleaned and with no NULL values
  - the data warehouse is a **user interface!**
- Typically lots of columns, but not that many rows → Wide tables
- Content doesn't change or changes slowly
- Data is **denormalized**
  - e.g. product, product subcategory and product category are all in the same table

Code	Name	Age	Location	First Purchase	Tier	Married Flag	…
C000457	Sarah	30	Leuven	2014-04-11	Gold	Y	…
C000057	David	70	Antwerp	2003-12-24	Bronze	Y	…
C003785	Lucie	20	Brussels	2024-08-23	Silver	N	…

Hierarchies

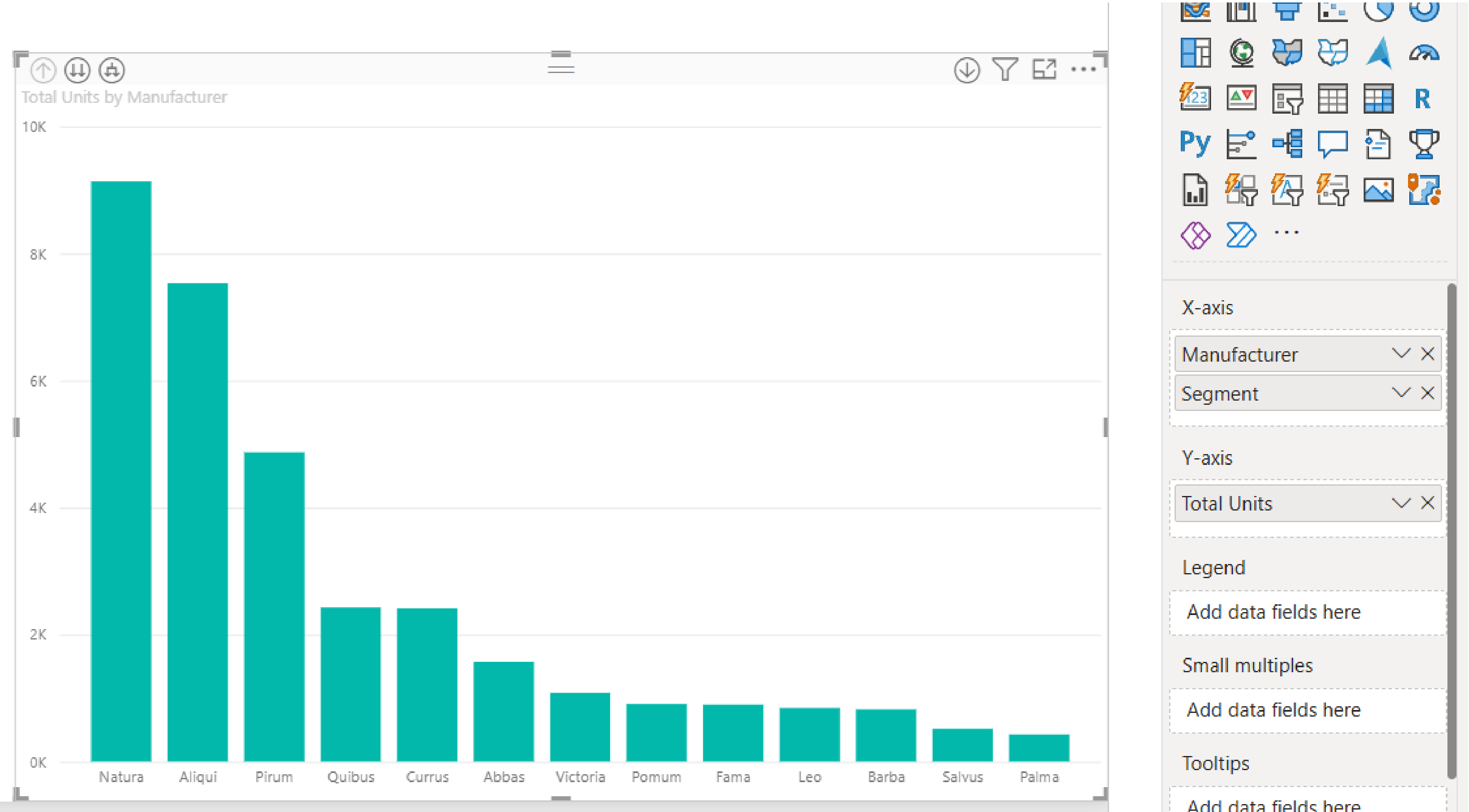
- Hierarchy describes a relationship between the attributes of a dimension
- Often used for “drilling down” in reports





Hierarchies

Power BI can create custom hierarchies on the fly by combining attributes (even from different dims) in a visual



Keys

- Business key
  - *natural key*
  - what makes a row unique in the source system?

- Surrogate key
  - **meaningless keys** used in the relationships between facts & dimensions
  - only exception is in the date dimension: 20260221
  - decouples dimension from source systems
  - allows for the tracking of history

SK customer	Code	Name	Age	Location	First Purchase	Tier	Married Flag	…
1	C000457	Sarah	30	Leuven	2014-04-11	Gold	Y	…
2	C000057	David	70	Antwerp	2003-12-24	Bronze	Y	…
3	C003785	Lucie	20	Brussels	2024-08-23	Silver	N	…

Surrogate keys

- More efficient joins in data warehouse when integers are used
- Power BI supports only single-column relationships
- Less space required for fact tables (SKs are the foreign keys to the dimensions)

- Allows for **dummy records**, like 'N/A'

SK customer	Code	Name	Age	Location	First Purchase	Tier	Married Flag	…
-1	N/A	N/A	NULL	N/A	NULL	N/A	N	…
1	C000457	Sarah	30	Leuven	2014-04-11	Gold	Y	…
2	C000057	David	70	Antwerp	2003-12-24	Bronze	Y	…
3	C003785	Lucie	20	Brussels	2024-08-23	Silver	N	…

Date dimension

- Every dimensional model has a date table
- Usually built upfront

- Don’t include timestamps!
  - create a separate time dimension

Date	Year	Month Name	Month Nbr	Year Month	Quarter	Day	Working Day	…
2020-01-01	2020	January	1	2020 Jan	1	Wednesday	Y	…
2020-01-02	2020	January	1	2020 Jan	1	Thursday	Y	…
2020-01-03	2020	January	1	2020 Jan	1	Friday	Y	…
2020-01-04	2020	January	1	2020 Jan	1	Saturday	N	…

SK_Date	DayOfWeek	DayOfMonth	DayOfQuarter	DayOfSemester	DayOfYear	DaysAgo	WeeksAgo	MonthsAgo	QuartersAgo	SemestersAgo	YearsAgo	IsCurrentDate	IsCurrentWeek	IsCurrentMonth	IsCurrentQuarter	IsCurrentSemester	IsCurrentYear	Last12Months	PreviousMonth	PreviousWeek	eekDa
2024-09-01	7	1	63	63	245	12	1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-02	1	2	64	64	246	11	1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	yes	
2024-09-03	2	3	65	65	247	10	1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	yes	
2024-09-04	3	4	66	66	248	9	1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	yes	
2024-09-05	4	5	67	67	249	8	1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	yes	
2024-09-06	5	6	68	68	250	7	1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	yes	
2024-09-07	6	7	69	69	251	6	1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	yes	
2024-09-08	7	8	70	70	252	5	0	0	0	0	0	no	no	yes	yes	yes	yes	no	no	yes	
2024-09-09	1	9	71	71	253	4	0	0	0	0	0	no	yes	yes	yes	yes	yes	no	no	no	
2024-09-10	2	10	72	72	254	3	0	0	0	0	0	no	yes	yes	yes	yes	yes	no	no	no	
2024-09-11	3	11	73	73	255	2	0	0	0	0	0	no	yes	yes	yes	yes	yes	no	no	no	
2024-09-12	4	12	74	74	256	1	0	0	0	0	0	no	yes	yes	yes	yes	yes	no	no	no	
2024-09-13	5	13	75	75	257	0	0	0	0	0	0	yes	yes	yes	yes	yes	yes	no	no	no	
2024-09-14	6	14	76	76	258	-1	0	0	0	0	0	no	yes	yes	yes	yes	yes	no	no	no	
2024-09-15	7	15	77	77	259	-2	-1	0	0	0	0	no	yes	yes	yes	yes	yes	no	no	no	
2024-09-16	1	16	78	78	260	-3	-1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-17	2	17	79	79	261	-4	-1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-18	3	18	80	80	262	-5	-1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-19	4	19	81	81	263	-6	-1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-20	5	20	82	82	264	-7	-1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-21	6	21	83	83	265	-8	-1	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-22	7	22	84	84	266	-9	-2	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-23	1	23	85	85	267	-10	-2	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-24	2	24	86	86	268	-11	-2	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-25	3	25	87	87	269	-12	-2	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-26	4	26	88	88	270	-13	-2	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-27	5	27	89	89	271	-14	-2	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-28	6	28	90	90	272	-15	-2	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-29	7	29	91	91	273	-16	-3	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	
2024-09-30	1	30	92	92	274	-17	-3	0	0	0	0	no	no	yes	yes	yes	yes	no	no	no	

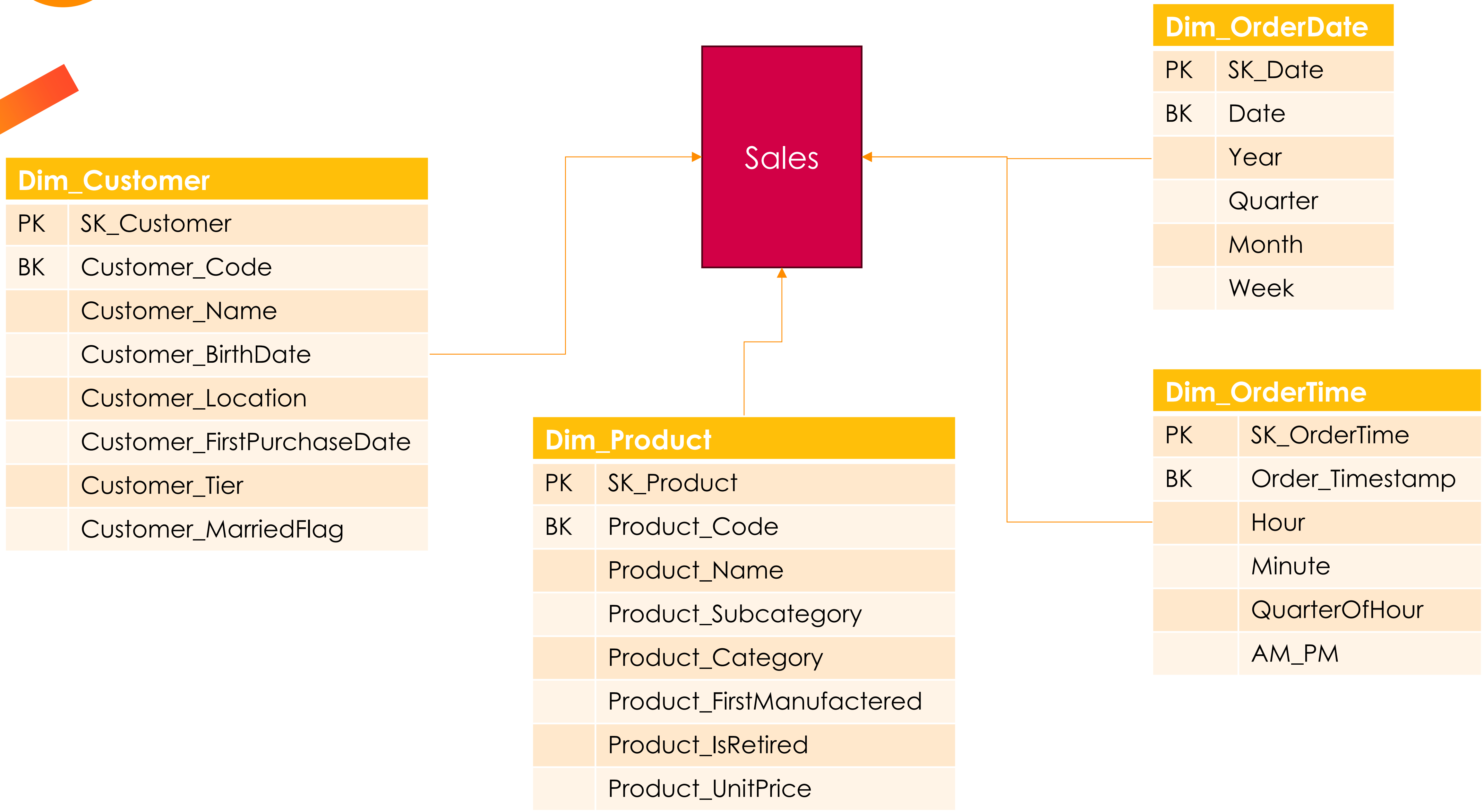
Time dimension

- Date dimension: max 366 rows per year
- Time dimension: 86.400 records (one row for each second)

- Allows for rollups per quarter of hour, hour, AM/PM …
- Allows for certain periods like rush hour, shifts …

Results | Messages

	SK_Time	TimeOfDay	TimeDesc12	TimeDesc24	AMorPM	DayPart	MilitaryTime	HourOfDay	MinuteOfDay	MinuteOfHour	SecondOfDay	SecondOfHour	SecondOfMinute
86...	86383	23:59:43	11:59:43 PM	23:59:43	PM	Night	2359	23	1439	59	86383	3583	43
86...	86384	23:59:44	11:59:44 PM	23:59:44	PM	Night	2359	23	1439	59	86384	3584	44
86...	86385	23:59:45	11:59:45 PM	23:59:45	PM	Night	2359	23	1439	59	86385	3585	45
86...	86386	23:59:46	11:59:46 PM	23:59:46	PM	Night	2359	23	1439	59	86386	3586	46
86...	86387	23:59:47	11:59:47 PM	23:59:47	PM	Night	2359	23	1439	59	86387	3587	47
86...	86388	23:59:48	11:59:48 PM	23:59:48	PM	Night	2359	23	1439	59	86388	3588	48
86...	86389	23:59:49	11:59:49 PM	23:59:49	PM	Night	2359	23	1439	59	86389	3589	49
86...	86390	23:59:50	11:59:50 PM	23:59:50	PM	Night	2359	23	1439	59	86390	3590	50
86...	86391	23:59:51	11:59:51 PM	23:59:51	PM	Night	2359	23	1439	59	86391	3591	51
86...	86392	23:59:52	11:59:52 PM	23:59:52	PM	Night	2359	23	1439	59	86392	3592	52
86...	86393	23:59:53	11:59:53 PM	23:59:53	PM	Night	2359	23	1439	59	86393	3593	53
86...	86394	23:59:54	11:59:54 PM	23:59:54	PM	Night	2359	23	1439	59	86394	3594	54
86...	86395	23:59:55	11:59:55 PM	23:59:55	PM	Night	2359	23	1439	59	86395	3595	55
86...	86396	23:59:56	11:59:56 PM	23:59:56	PM	Night	2359	23	1439	59	86396	3596	56
86...	86397	23:59:57	11:59:57 PM	23:59:57	PM	Night	2359	23	1439	59	86397	3597	57
86...	86398	23:59:58	11:59:58 PM	23:59:58	PM	Night	2359	23	1439	59	86398	3598	58
86...	86399	23:59:59	11:59:59 PM	23:59:59	PM	Night	2359	23	1439	59	86399	3599	59


Sales
Dim_Customer
PK SK_Customer
BK Customer_Code
Customer_Name
Customer_BirthDate
Customer_Location
Customer_FirstPurchaseDate
Customer_Tier
Customer_MarriedFlag
Dim_Product
PK SK_Product
BK Product_Code
Product_Name
Product_Subcategory
Product_Category
Product_FirstManufactered
Product_IsRetired
Product_UnitPrice
Dim_OrderDate
PK SK_Date
BK Date
Year
Quarter
Month
Week
Dim_OrderTime
PK SK_OrderTime
BK Order_Timestamp
Hour
Minute
QuarterOfHour
AM_PM

What are facts?

What are facts?

Ask the question: what do you want to analyze?

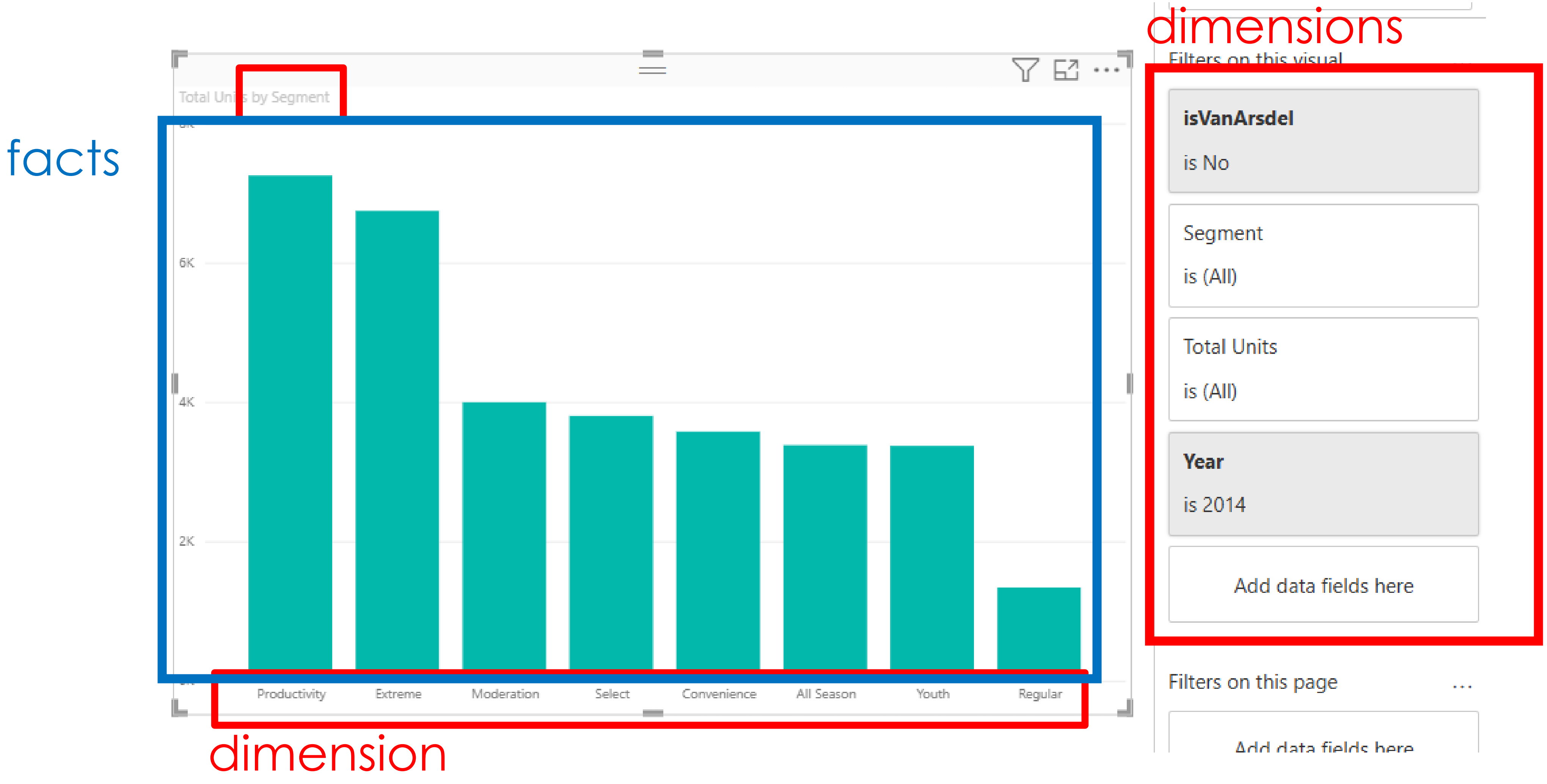
“I want to analyze ***sales amounts***,

by customer,
by order date,
by product,
by store,
by promotion,
…

“

Or ask **the 7Ws**
(who, what, where, when, **how many**, why, how)

What are facts?



Facts characteristics

- Mostly **numerical** attributes
  - Something we want to measure in a business process
  - Can contain NULL values
- Foreign keys relationships (using SKs) to dimension tables
- Typically lots of rows, but not that many columns → Long tables
- Sometimes text attributes are kept in the fact because they have the same *granularity* as the fact table → **Degenerate Dimension**

SK_Customer	SK_Product	SK_OrderDate	OrderID	Amount	Quantity
1	8	2026-01-15	D20260001	100	1
2	5	2026-01-30	D20260002	25	2
1	1	2026-02-05	D20260003	30	1
1	1	2026-02-05	D20260003	30	1
3	3	2026-02-18	D20260004	85	3

Granularity

- **= level of detail** at which data is kept inside the fact table
- Defined by the number of dimensions in the fact table

- Too high → many aggregations needed
  - e.g. sensor data at second level
- Too low → loss of precision
  - e.g. sales at month level

- Rule of thumb: always go as high as you technically can, while it still makes sense

Granularity

- Each row in the fact table must have the **same grain**
  - for example, you cannot mix sales numbers of the month and the day level in the same table
  - or budget numbers on the month level and actuals at the day level

- Store measures/events from the same business process in a single fact
- Store different business processes in different facts

- Sometimes difficult, for example invoice header vs invoice line

Kimball Tip #95 (+ book)

Types of measures

Additive

Add across **all dimensions**

e.g. revenue, quantity, # titles streamed

Semi-additive

Add across **some dimensions** (except time)

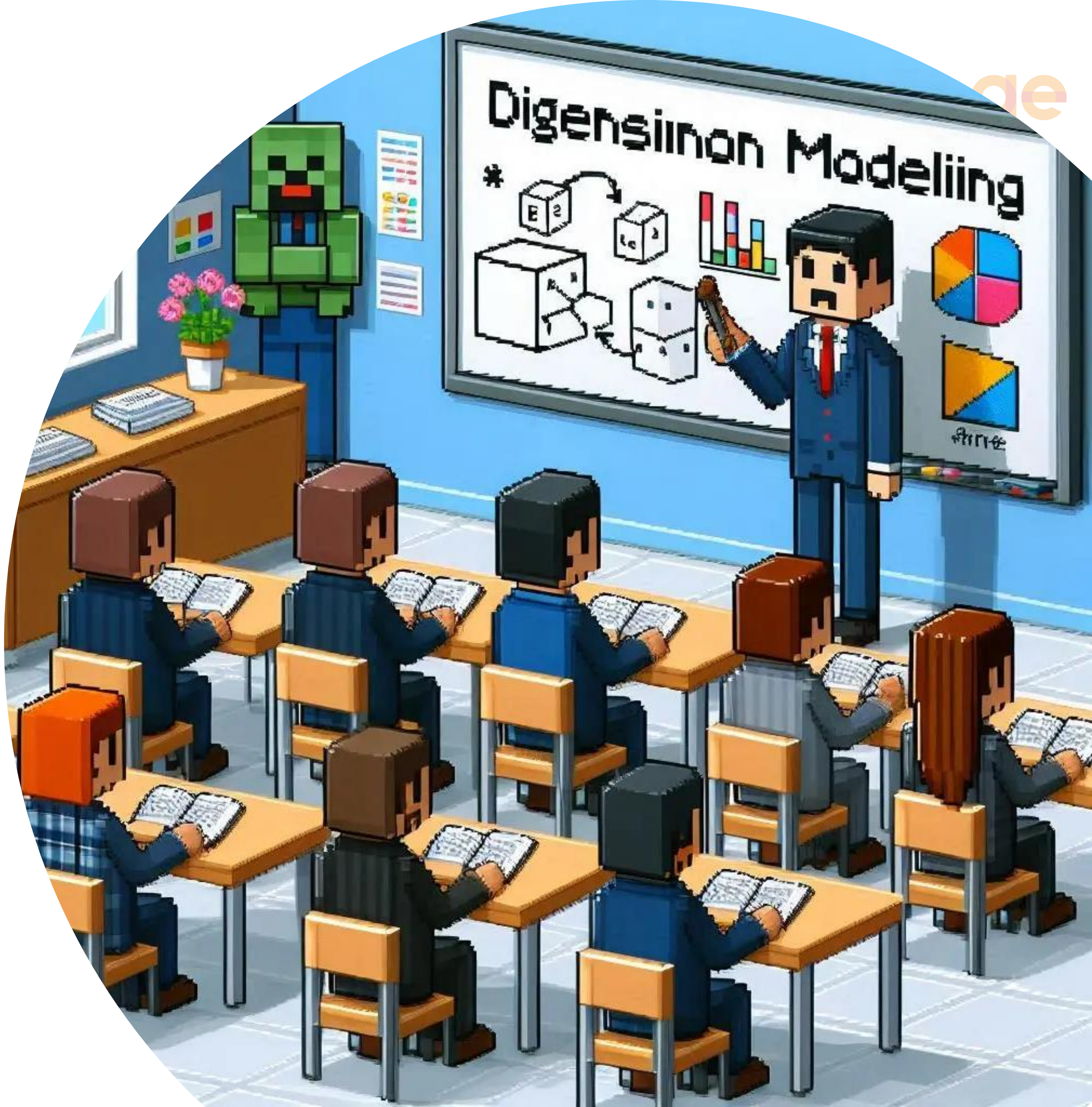
e.g. cash balance, stock level

Non-additive

Cannot be added across any **dimension**

e.g. price, temperature, ratio

Types of Fact Tables



Transaction Fact Table

- Easiest and most often used

- Transactions are point in time events
  - Facts exist when an event occurs
  - Anti-pattern: creating fact rows for something that didn't happen e.g. student didn't attend class, customer did not use promotion

- Most measures are fully additive

Fact_Sales	
PK,FK	SK_Product
PK,FK	SK_Customer
PK,FK	SK_OrderDate
DD	OrderID
	Sales Amount
	Quantity
	Discount Amount
	Tax Amount
	Total Sales

Dim_Customer	
PK	SK_Customer
BK	Customer_Code
	Customer_Name
	Customer_BirthDate
	Customer_Location
	Customer_FirstPurchaseDate
	Customer_Tier
	Customer_MarriedFlag

Dim_OrderDate	
PK	SK_Date
BK	Date
	Year
	Quarter
	Month
	Week

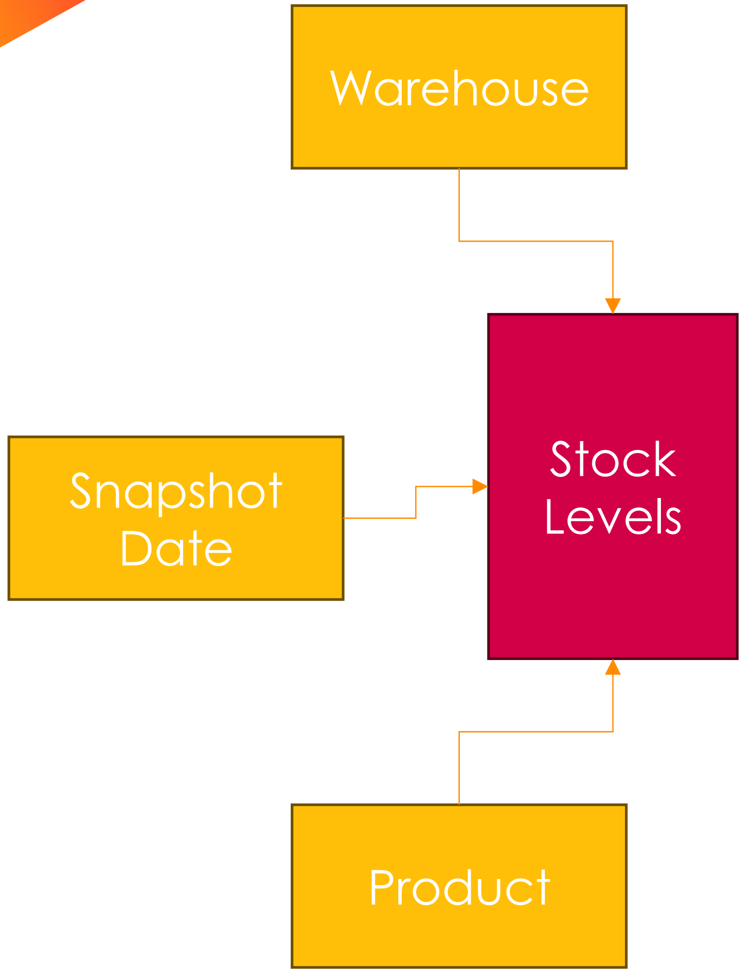
Dim_Product	
PK	SK_Product
BK	Product_Code
	Product_Name
	Product_Subcategory
	Product_Category
	Product_FirstManufactered
	Product_IsRetired
	Product_UnitPrice

Periodic Snapshot Fact Table

- Captures state at a specific regular interval (e.g. day, week, month)

- Simplifies certain queries
  - For example, calculating account balancing using all transactions since the opening of the bank account would be quite ineffective
  - Trend analysis becomes possible over the different snapshots
- Can grow quickly in size
- Zero measurements are possible (for example out-of-stock)

- Measures are **semi-additive**
- Fact table can be a roll-up of a transaction fact table



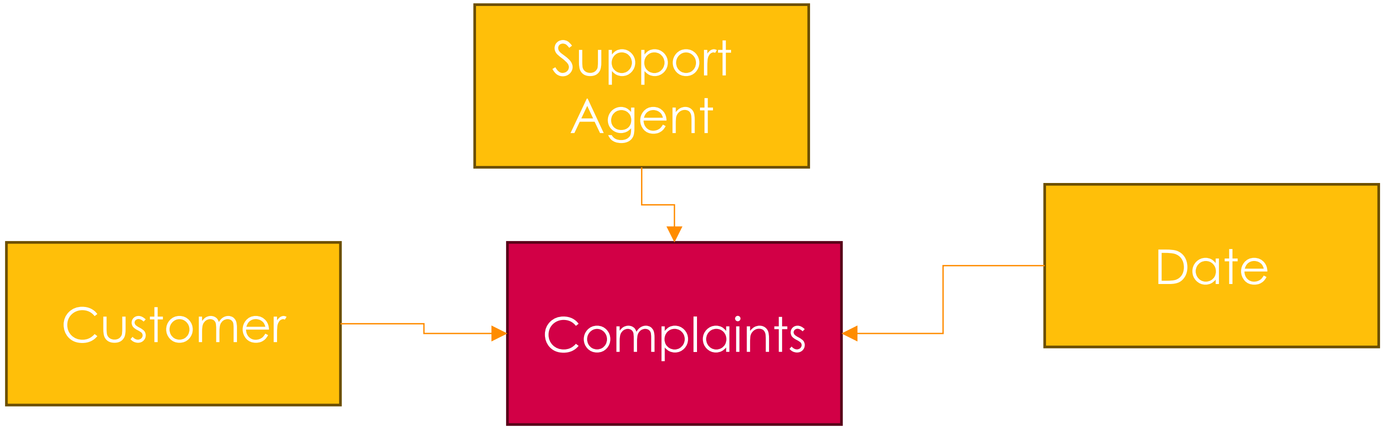
Snapshot Date	Warehouse	Product Code	Stock Level	Delta
2026-02-01	Laakdal	Pants_3032	24	0
2026-02-01	Laakdal	Shirt_M	12	0
2026-02-01	Laakdal	Blouse_XXS	1	0
2026-02-01	Laakdal	Blouse_XS	8	0
2026-02-01	Laakdal	Shoes_38	5	0
2026-02-08	Laakdal	Pants_3032	23	-1
2026-02-08	Laakdal	Shirt_M	25	13
2026-02-08	Laakdal	Blouse_XXS	**0**	-1
2026-02-08	Laakdal	Blouse_XS	**0**	-8
2026-02-08	Laakdal	Shoes_38	3	-2

Accumulating Snapshot Fact Table

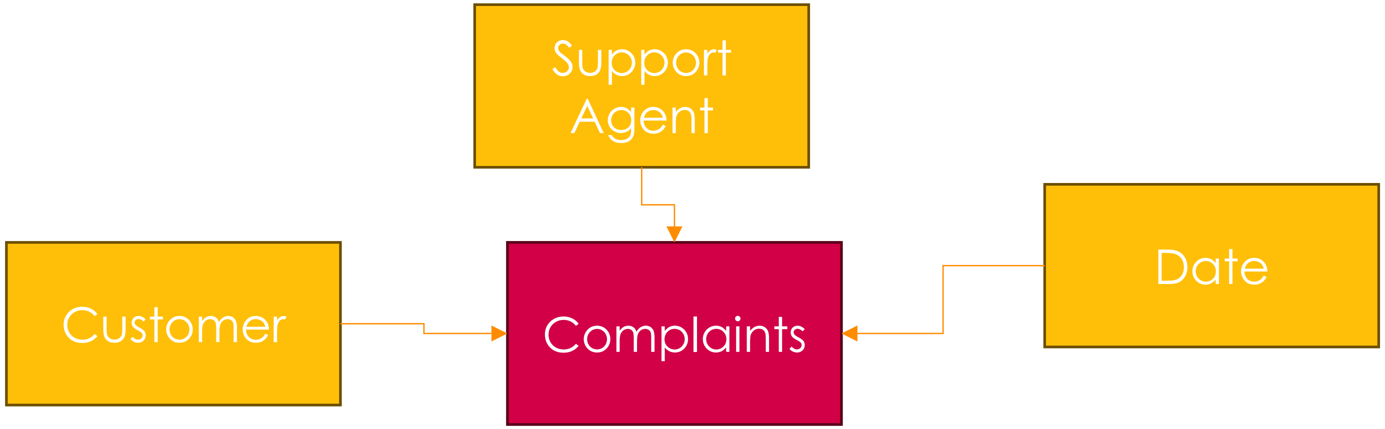
- One line for an entity going through a certain (fixed) lifecycle

- Examples
  - parcel moving through delivery process
  - claims processing in insurance
  - sales funnel

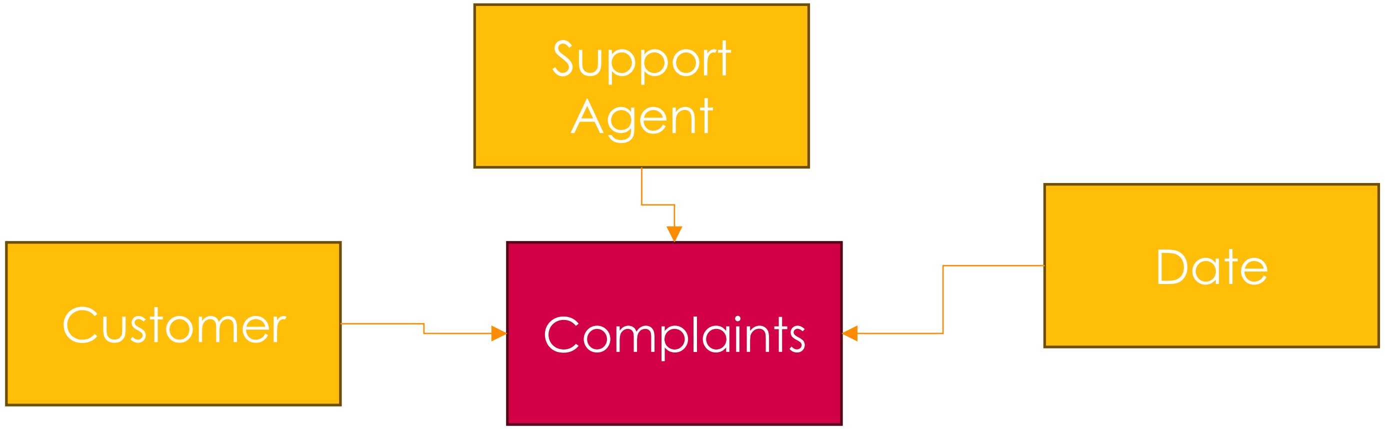
- Major milestones are pre-defined
- Contains many date(time) stamps
- Measures are mostly durations (and counts)
- Only fact table type that is **updated frequently!**



Customer	Agent	Open Date	Close Date	Total Duration	Investigation Date	Delta1	Communication Date	Delta2
C000457	Sophie	2026-02-03						

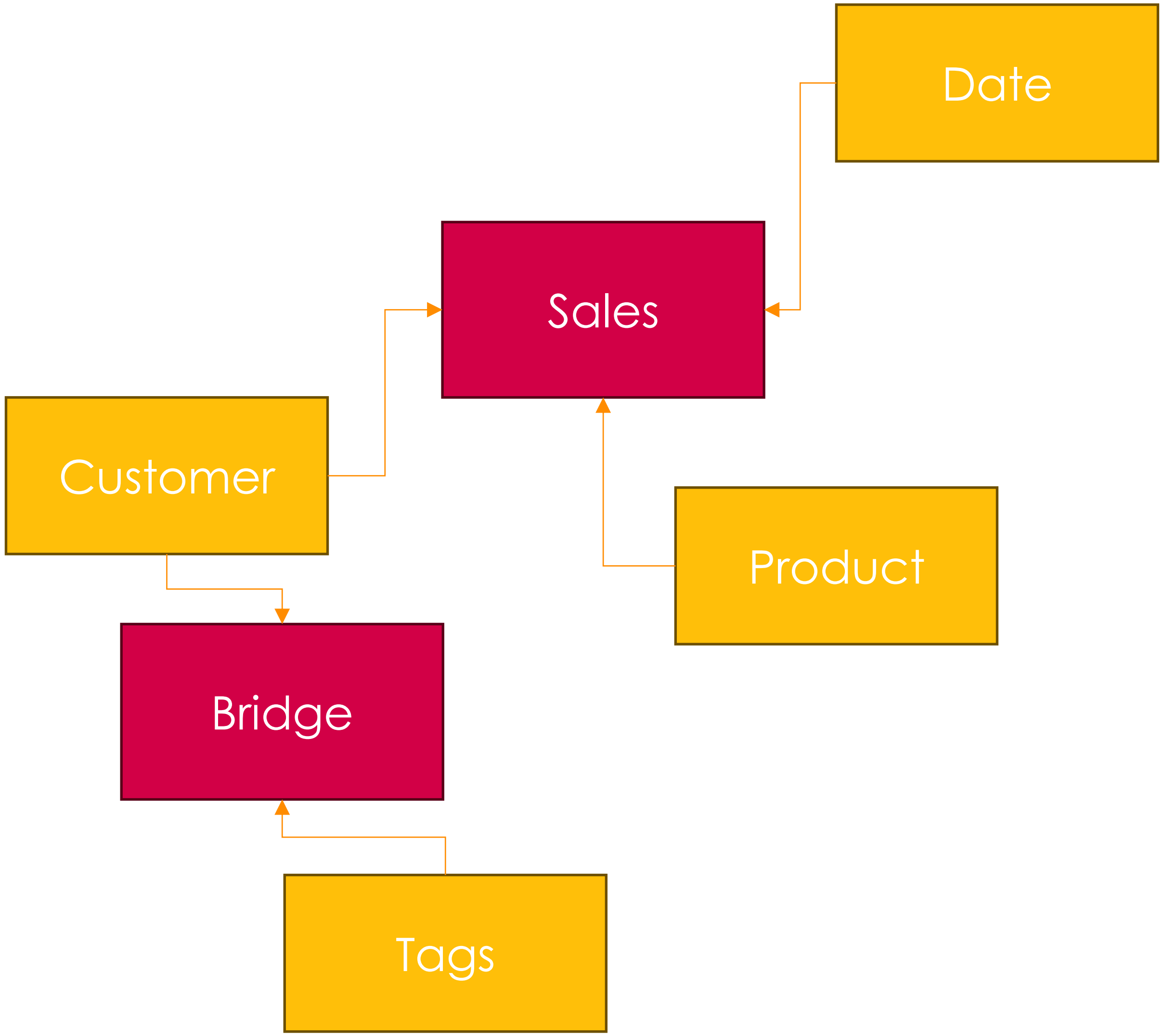


Customer	Agent	Open Date	Close Date	Total Duration	Investigation Date	Delta1	Communication Date	Delta2
C000457	Sophie	2026-02-03			2026-02-03	0		

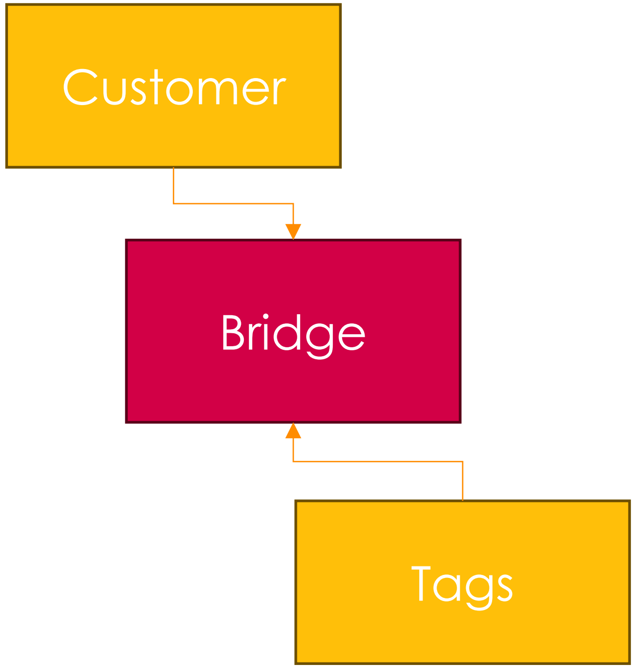


Customer	Agent	Open Date	Close Date	Total Duration	Investigation Date	Delta1	Communication Date	Delta2
C000457	Sophie	2026-02-03	2026-02-05	2	2026-02-03	0	2026-02-05	2

Many 2 Many Relationships



Bridge tables

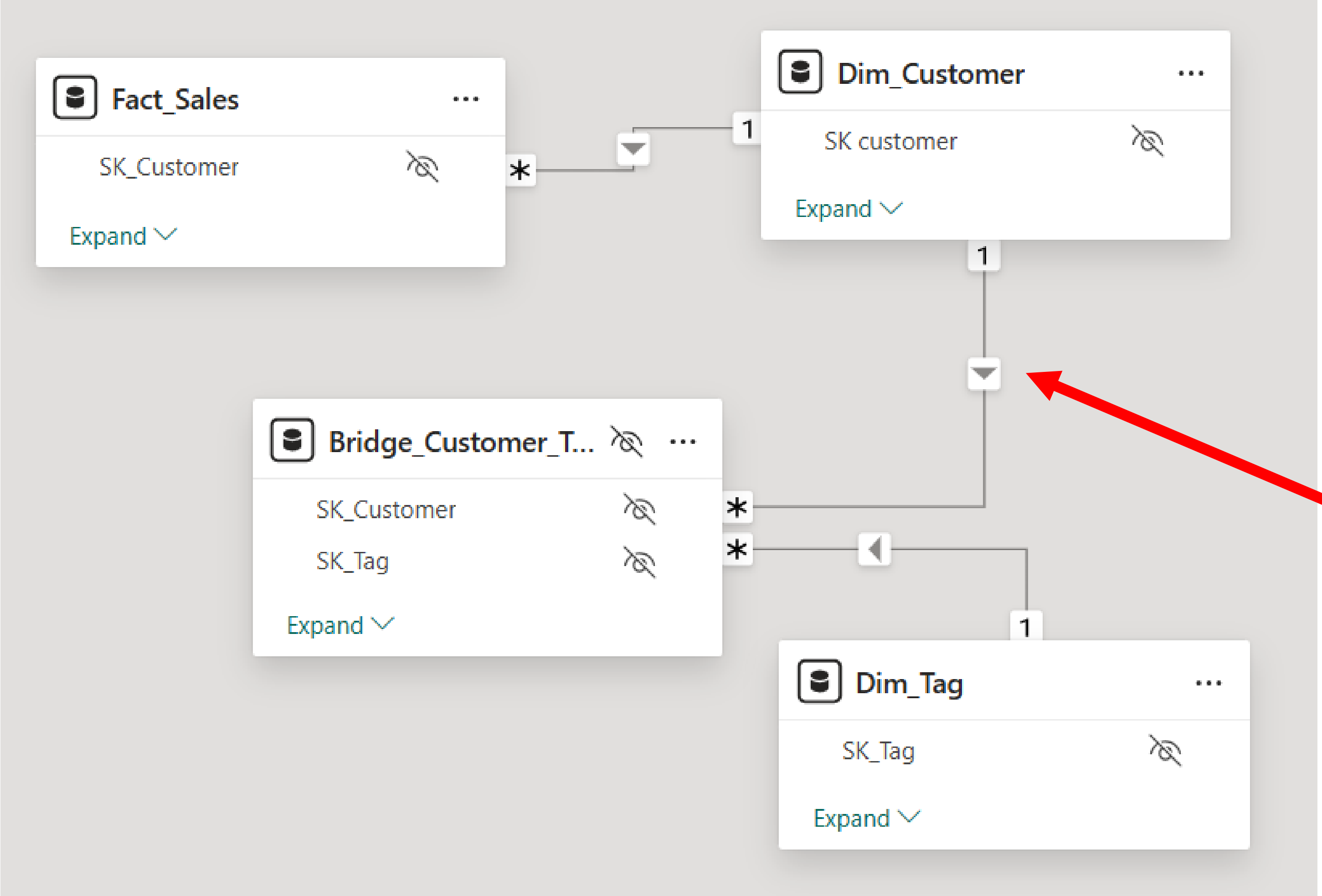


Customer

SK_Customer	SK_Tag
3	1
3	4
3	5
8	2
8	5
10	1
10	6

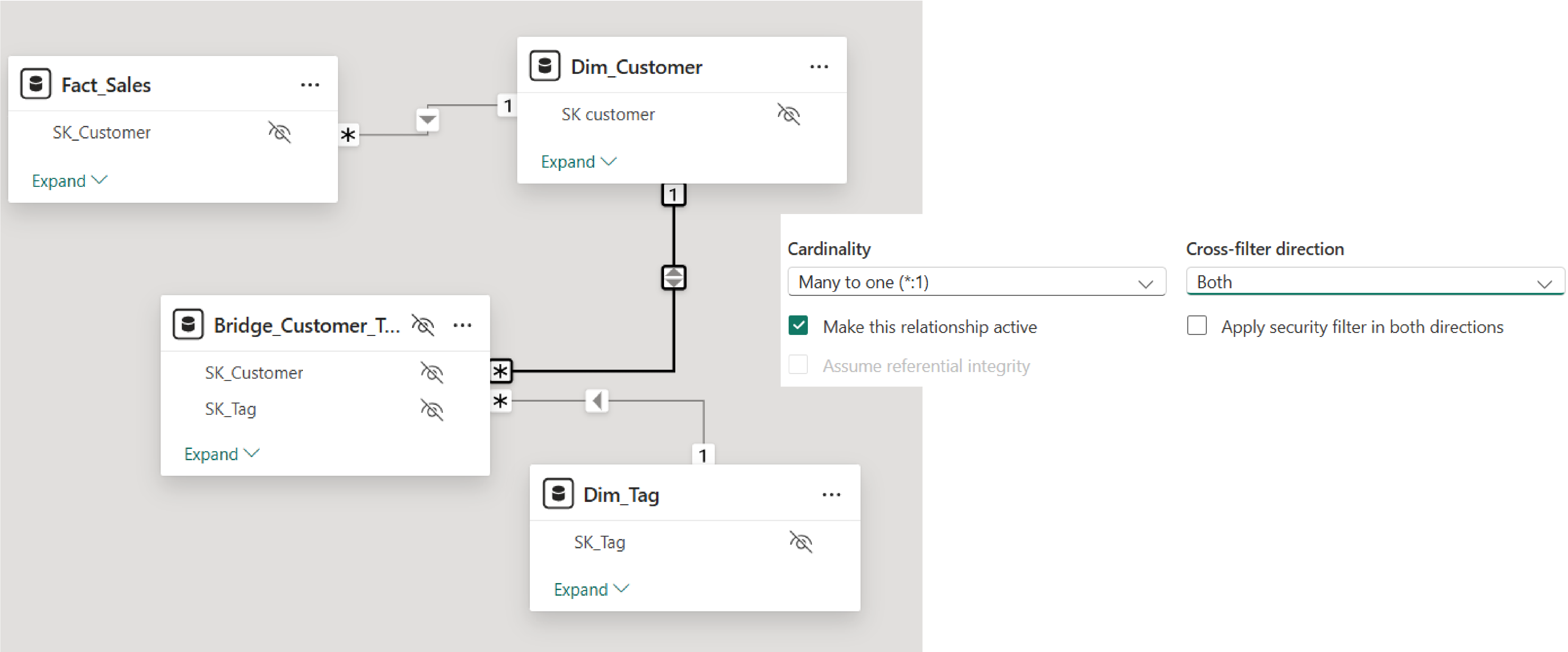
SK_Tag	TagName
1	Acquisitions 2024
2	Acquisitions 2025
3	Acquisitions 2026
4	Active Customers 2025
5	Active Customers 2026
6	Lost Customers 2025

Bridge tables in Power BI



filtering on Tag doesn't propagate to Customer

Bridge tables in Power BI – Option 1



Bridge tables in Power BI – Option 2

Solve it with DAX function `CROSSFILTER`

```
CALCULATE([Sales Amount])
          ,CROSSFILTER(Dim_Customer[SK customer]
                      ,Bridge_Customer_Tag[SK_Customer]
                      ,Both
                      )
          )
```

TagName	Sum of Amount	Sales Amount Correct
Acquisitions 2024	270	110
Acquisitions 2025	270	160
Acquisitions 2026	270	
Active Customers 2025	270	85
Active Customers 2026	270	245
Lost Customers 2025	270	25
Total	**270**	**270**

Dimensions: the SeQueL

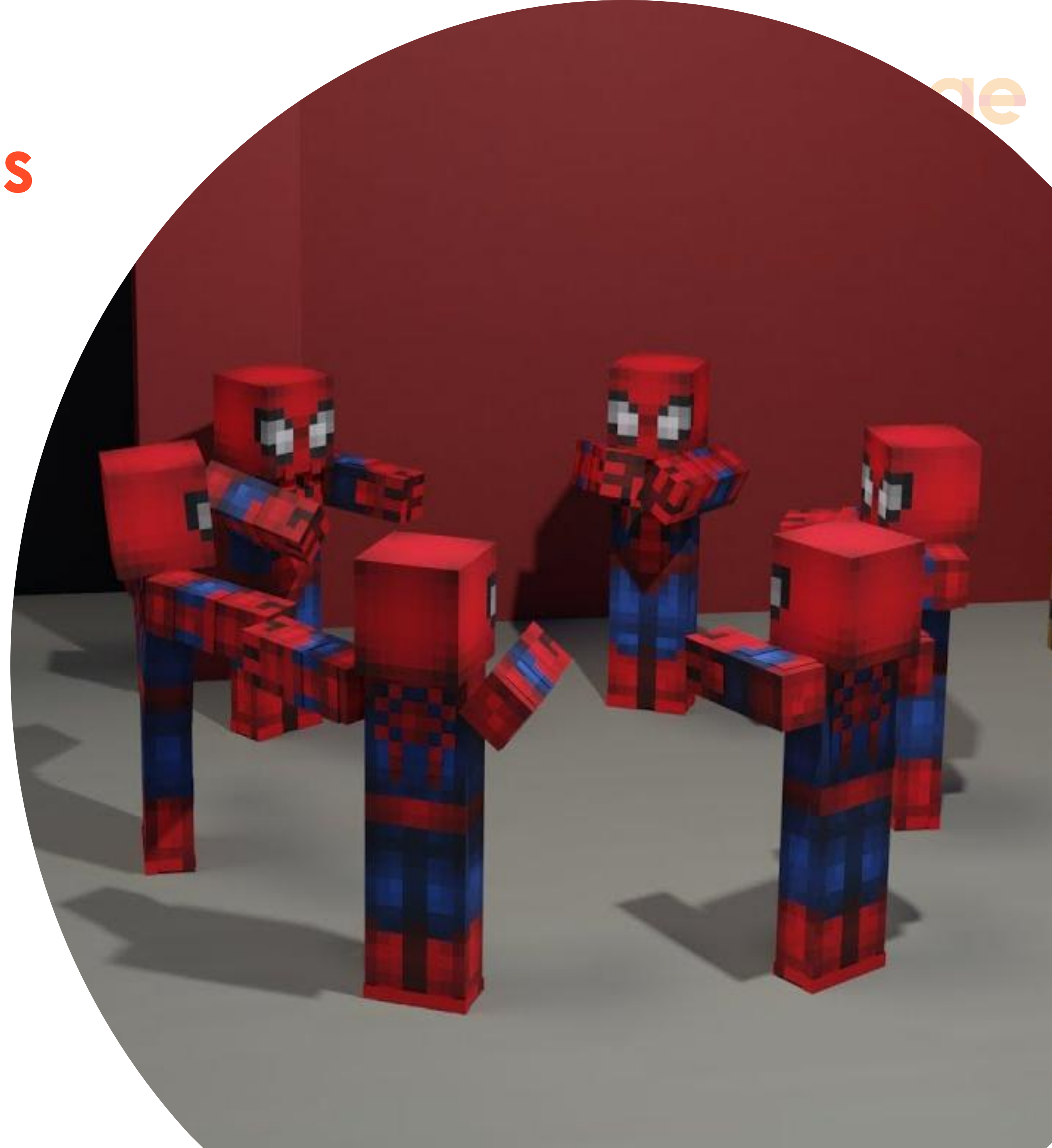
Role Playing Dimensions

The same dimension is linked multiple times to a fact table

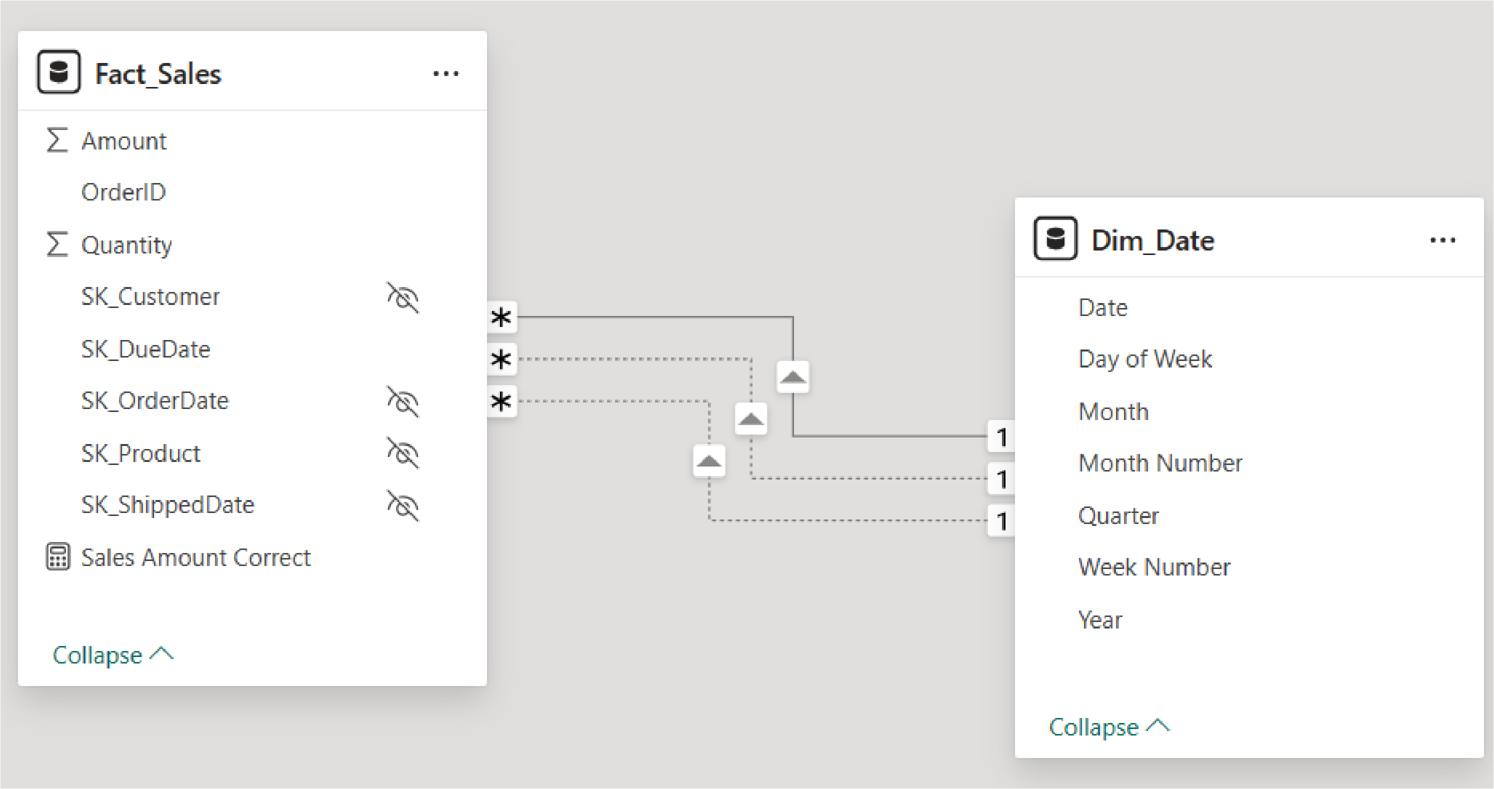
- Order Date, Due Date, Delivery Date etc.
- Employee, Manager, Sales Person etc.

Two options:

- Link the same dimension multiple times
- Create copies of the dimension

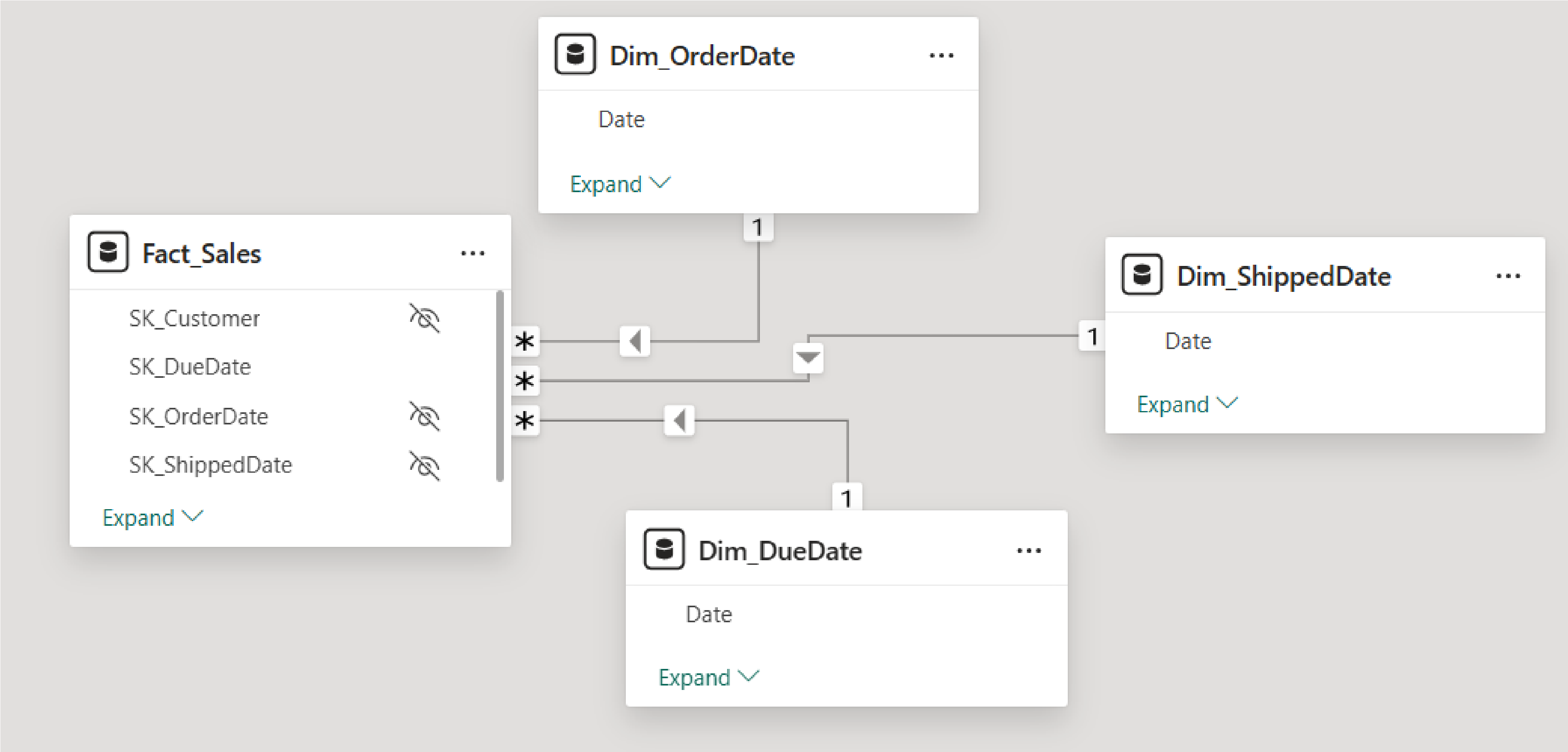


Role Playing Dimensions




```
Sales by ShippedDate =
        CALCULATE(Fact_Sales[Sales Amount]
                ,USERELATIONSHIP(Fact_Sales[SK_ShippedDate], Dim_Date[Date])
                )
```

Role Playing Dimensions



Slowly Changing Dimensions

- Source systems (sometimes) do not track changes
  - A row is overwritten with the new values
- The data warehouse can track those changes instead
- Usually, attributes of a dimension don't change all the time, but rather **slowly**.
- Thanks to surrogate keys, implementing history is straightforward

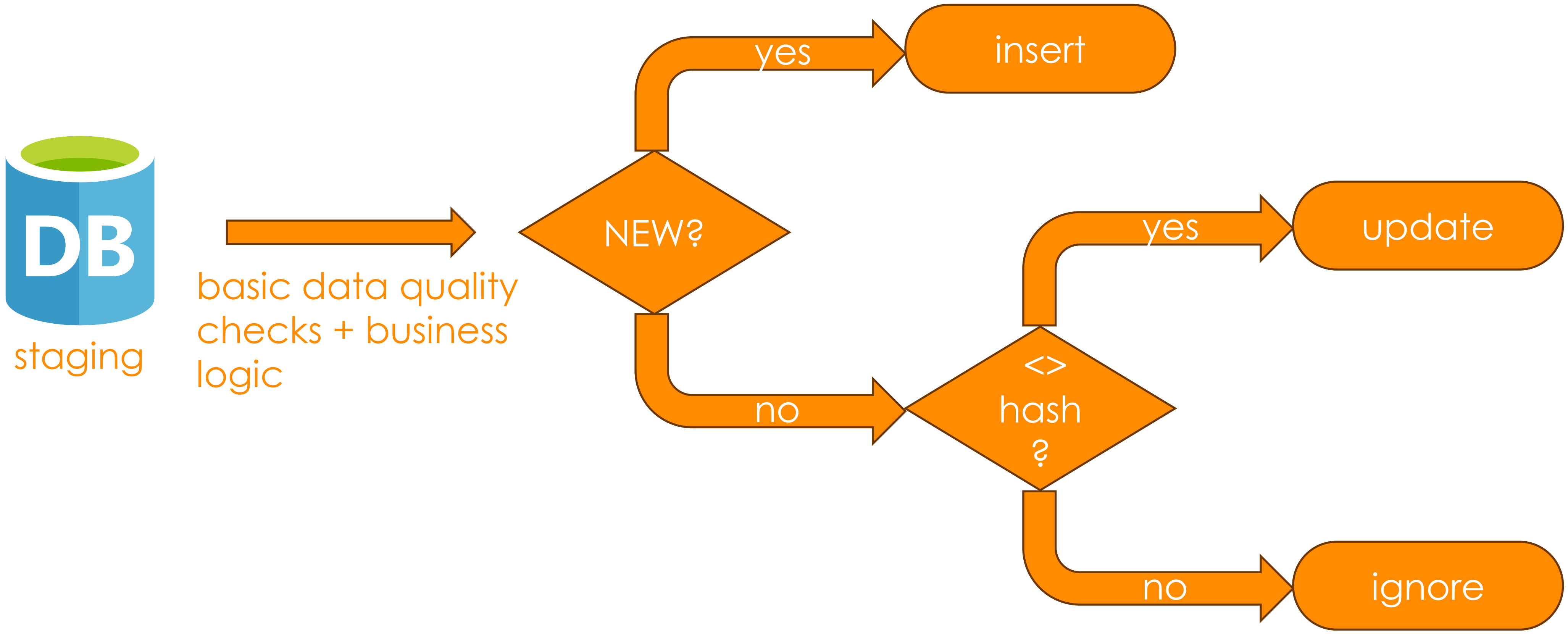
SCD Type 1

- The attribute is overwritten
- Only the most recent values are kept
- Fact records are now associated with these new values

SK customer	Code	Name	Age	Location	First Purchase	Tier	Married Flag
1	C000457	Sarah	30	Leuven	2014-04-11	Gold	Y
2	C000057	David	70	Antwerp	2003-12-24	Bronze	Y
3	C003785	Lucie	20	Brussels	2024-08-23	Silver	N

SK customer	Code	Name	Age	Location	First Purchase	Tier	Married Flag
1	C000457	Sarah	30	Leuven	2014-04-11	Gold	Y
2	C000057	David	70	Antwerp	2003-12-24	**Silver**	Y
3	C003785	Lucie	20	Brussels	2024-08-23	Silver	N

SCD Type 1



DB
staging
basic data quality checks + business logic
NEW?
yes
insert
no
<> hash ?
yes
update
no
ignore

SCD Type 1

dummy record

hash

SK_Customer	CustomerID	CustomerTitle	CustomerFirstName	CustomerMiddleName	CustomerLastName	CustomerSuffix	CustomerFullName	InsertDate	UpdateDate	Hash_SCD1
-1	-1	N/A	N/A	N/A	N/A	N/A	N/A	2024-04-09 09:31:47	NULL	0x8B5F58F7085BC8A155D9A475A71E4F162C85
1	1	Mr.	Orlando	N.	Gee	N/A	Mr. Orlando N. Gee	2024-03-30 15:27:32	NULL	0xDA552154A22203DF4493AC947116867A52DB
2	2	Mr.	Keith	N/A	Harris	N/A	Mr. Keith Harris	2024-03-30 15:27:32	NULL	0xABC12DFC977EC4F45E4A113421A2168FEE20
3	3	Ms.	Donna	F.	Carreras	N/A	Ms. Donna F. Carreras	2024-03-30 15:27:32	NULL	0xEF1AF96F3E368A2EB8945AFD392E4BDD7F20
4	4	Ms.	Janet	M.	Gates	N/A	Ms. Janet M. Gates	2024-03-30 15:27:32	NULL	0x1631D609EEDC38AB757244EFA2EF67BE6E4
5	5	Mr.	Lucy	N/A	Harrington	N/A	Mr. Lucy Harrington	2024-03-30 15:27:32	NULL	0xE1137F7C2078AAA419B7A314DB6F3E6C779A
6	6	Ms.	Rosmarie	J.	Carroll	N/A	Ms. Rosmarie J. Carroll	2024-03-30 15:27:32	NULL	0x4C2C20F550E05628BD868E182B14DE02DE3
7	7	Mr.	Dominic	P.	Gash	N/A	Mr. Dominic P. Gash	2024-03-30 15:27:32	NULL	0x348BC2BB79F465058D376E10ACDDE8D8816
8	10	Ms.	Kathleen	M.	Garza	N/A	Ms. Kathleen M. Garza	2024-03-30 15:27:32	NULL	0x97536ABD7C1BE70BD6EE9FF4B5F73FE5CA5
9	11	Ms.	Katherine	N/A	Harding	N/A	Ms. Katherine Harding	2024-03-30 15:27:32	NULL	0xFFD0AE20EB50DFC2D9D5E54FF3560A49A3C
10	12	Mr.	Johnny	A.	Caprio	Jr.	Mr. Johnny A. Caprio Jr.	2024-03-30 15:27:32	NULL	0x457A6D3A9B83B8251E262FC23DBB4934FA32
11	16	Mr.	Christopher	R.	Beck	Jr.	Mr. Christopher R. Beck Jr.	2024-03-30 15:27:32	NULL	0x93E0036C66B0ACB81748332E3EB283586A8E
12	18	Mr.	David	J.	Liu	N/A	Mr. David J. Liu	2024-03-30 15:27:32	NULL	0xE7E68D20F27A76322C33A3CEC9FC76527E55
13	19	Mr.	John	A.	Beaver	N/A	Mr. John A. Beaver	2024-03-30 15:27:32	NULL	0x3945D48429EA6E4F50609A83C5764EE14345
14	20	Ms.	Jean	P.	Handley	N/A	Ms. Jean P. Handley	2024-03-30 15:27:32	NULL	0xF786973C0280A9A22233E2DB9E2BC953C4AD

SK = identity(1,1) = PK

BK = unique index

audit columns

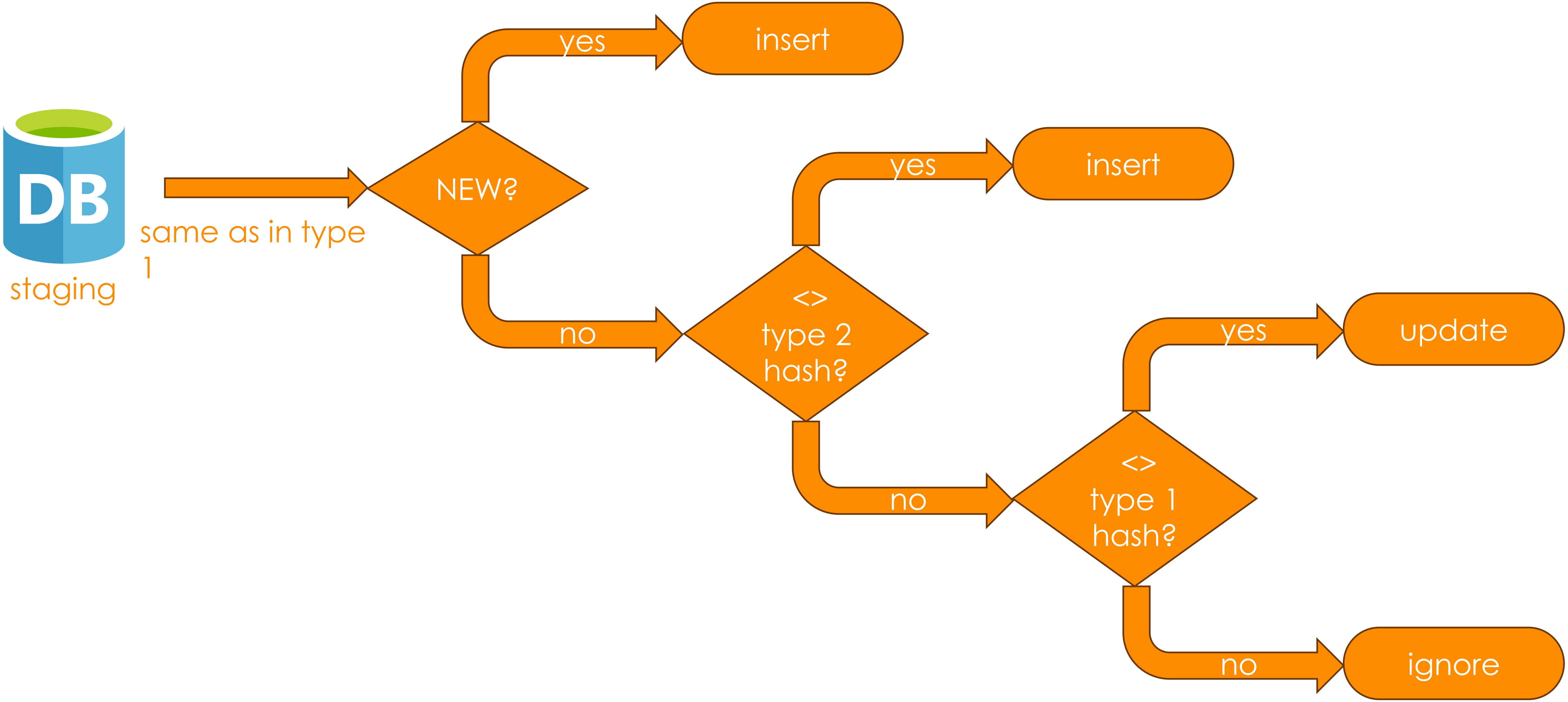
SCD Type 2

- History is preserved by inserting a new row
- The BK remains the same, but a new SK is generated
- Old facts remain associated with the old values, new facts with the new values

SK customer	Code	Name	Age	Location	Valid From	Valid To	Current Flag
1	C000457	Sarah	30	Leuven	2014-04-11		Y
2	C000057	David	70	Antwerp	2003-12-24		Y
3	C003785	Lucie	20	Brussels	2024-08-23		Y

SK customer	Code	Name	Age	Location	Valid From	Valid To	Current Flag
1	C000457	Sarah	30	Leuven	2014-04-11	2026-02-09	N
2	C000057	David	70	Antwerp	2003-12-24		Y
3	C003785	Lucie	20	Brussels	2024-08-23		Y
4	C000457	Sarah	30	Gent	2026-02-10		Y

SCD Type 2


DB
staging
same as in type 1
NEW?
yes
insert
no
<> type 2 hash?
yes
insert
no
<> type 1 hash?
yes
update
no
ignore

SCD Type 2

hashes

SK_Employee	EmployeeID	Employee_FullName	Employee_CurrentHiveName	Employee_HiveName	InsertDate	UpdateDate	ValidFrom	ValidTo	Hash_SCD2	Hash_SCD1
236	000000444	Koen Verbeeck	Data & AI L2	Information Management	2018-06-05 14:52:01	2024-04-10 04:37:49	2000-01-01	2018-09-15	0x904CE4C0394D...	0x0BBFDDCB7...
806	000000444	Koen Verbeeck	Data & AI L2	Information Management	2018-09-17 22:12:15	2024-04-10 04:37:49	2018-09-16	2018-10-22	0xDBFA27A2138E...	0x0BBFDDCB7...
872	000000444	Koen Verbeeck	Data & AI L2	Information Management	2018-10-24 22:13:11	2024-04-10 04:37:49	2018-10-23	2019-03-20	0x28BC68704106...	0x0BBFDDCB7...
1146	000000444	Koen Verbeeck	Data & AI L2	Information Management	2019-03-22 04:31:49	2024-04-10 04:37:49	2019-03-21	2019-03-27	0x28BC68704106...	0x0BBFDDCB7...
1169	000000444	Koen Verbeeck	Data & AI L2	Information Management	2019-03-29 04:29:58	2024-04-10 04:37:49	2019-03-28	2020-10-15	0x852ED155224B...	0x0BBFDDCB7...
2555	000000444	Koen Verbeeck	Data & AI L2	Information Management	2020-10-17 05:19:06	2024-04-10 04:37:49	2020-10-16	2022-01-31	0xE9CCEAF67A27...	0x0BBFDDCB7...
3645	000000444	Koen Verbeeck	Data & AI L2	Information Management	2022-02-24 15:25:03	2024-04-10 04:37:49	2022-02-01	2022-11-30	0xE9CCEAF67A27...	0x0BBFDDCB7...
4196	000000444	Koen Verbeeck	Data & AI L2	Information Management	2022-12-10 05:38:08	2024-04-10 04:37:49	2022-12-01	2022-12-31	0xBDEBE03DDB7...	0x0BBFDDCB7...
4234	000000444	Koen Verbeeck	Data & AI L2	Data	2023-01-03 05:35:32	2024-04-10 04:37:49	2023-01-01	2023-09-30	0xE17FA12D8597...	0x0BBFDDCB7...
4935	000000444	Koen Verbeeck	Data & AI L2	Data	2023-10-27 04:36:30	2024-04-10 04:37:49	2023-10-01	2023-12-31	0xA24B7715C374...	0x0BBFDDCB7...
5191	000000444	Koen Verbeeck	Data & AI L2	Data & AI L2	2024-01-16 05:37:27	2024-04-10 04:37:49	2024-01-01	2024-01-31	0x83DB83E126CC...	0x0BBFDDCB7...
5509	000000444	Koen Verbeeck	Data & AI L2	Data & AI L2	2024-02-01 05:36:26	2024-04-10 04:37:49	2024-02-01	NULL	0xEE5CF2745FC0...	0x0BBFDDCB7...

SK = identity(1,1) = PK

BK + valid from = unique index

audit columns

SCD Type 3

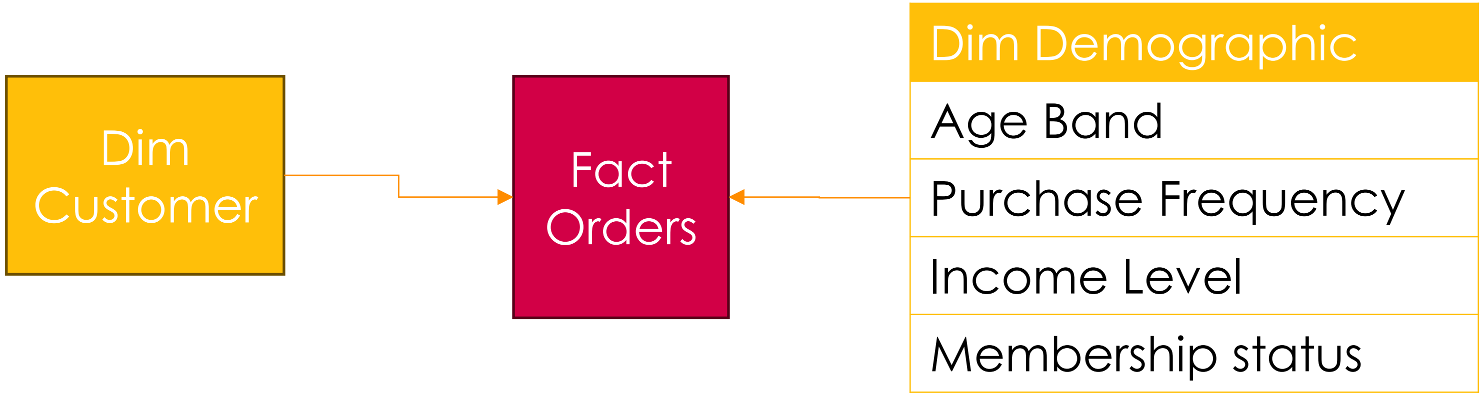
- The old value is kept next to the new value, by adding a column
- Often used in migrations, or change in business logic

SK customer	Code	Name	Age	Location	First Purchase	Tier	Tier <2026
1	C000457	Sarah	30	Leuven	2014-04-11	Gold	Gold
2	C000057	David	70	Antwerp	2003-12-24	Bronze	Silver
3	C003785	Lucie	20	Brussels	2024-08-23	Silver	Silver

Other types of SCD

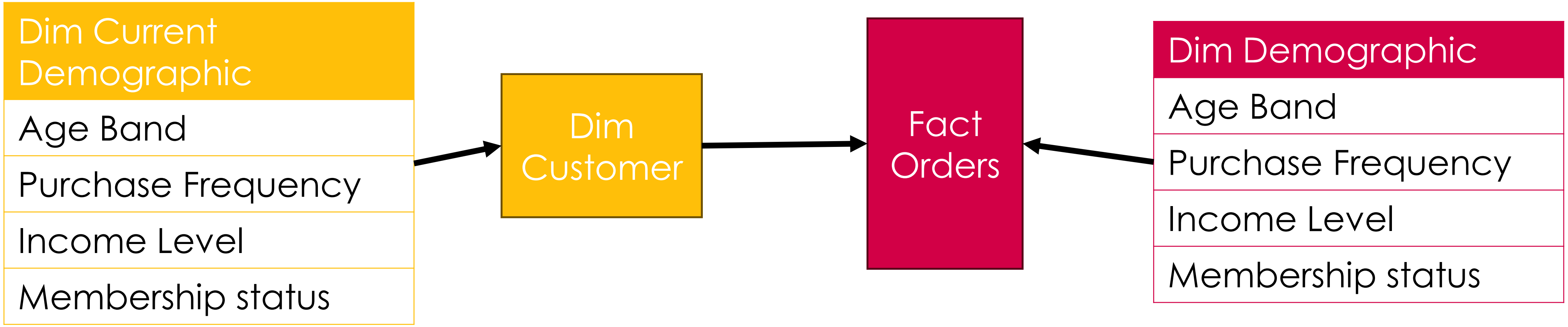
- Type 0
  - Don’t do anything
  - “original” fields

- Type 4
  - Add **mini-dimension**
  - For *rapidly changing monster dimensions*



Other types of SCD

- Type 5 (4 + 1)
  - **Mini-dimension + type 1 outrigger**
  - To keep track of "current values"



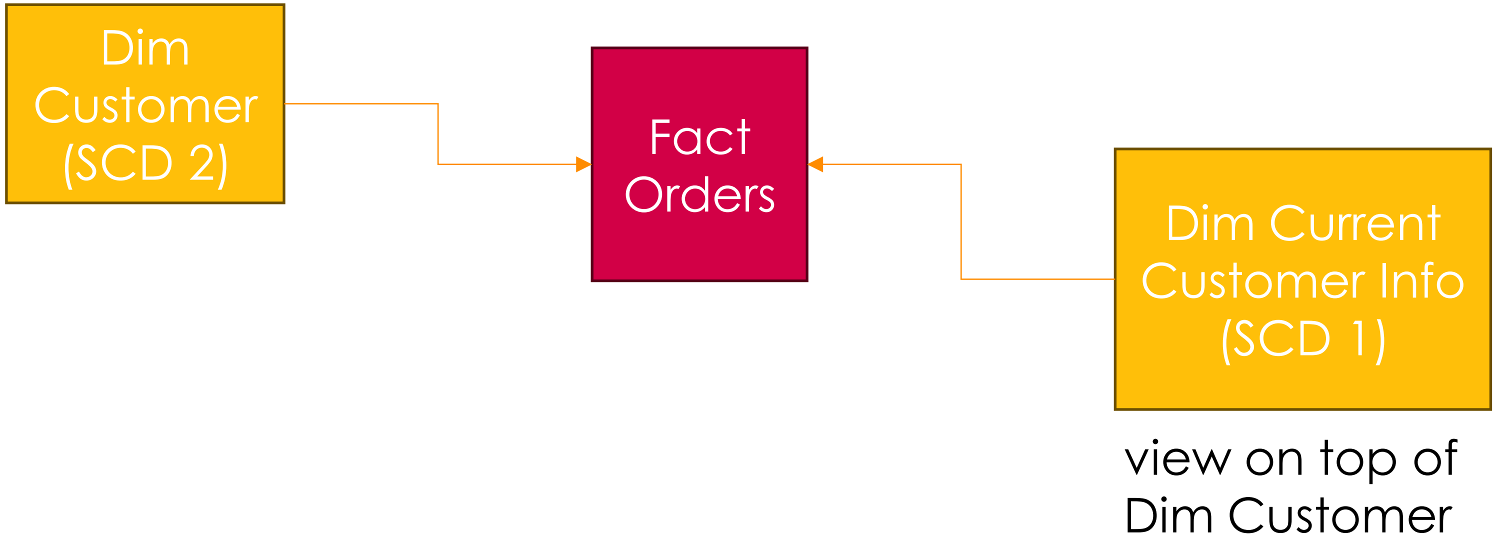
Other types of SCD

- Type 6 (2 + 1 + 3)
  - **Type 1 attributes in type 2** dimension
  - Keep track of current and historical attributes

SK_Employee	EmployeeID	Employee Name	Historical Coach	Current Coach
1	000123	Gert	Bart	Lode
2	000123	Gert	An	Lode
3	000123	Gert	Joke	Lode
4	000123	Gert	Lode	Lode

Other types of SCD

- Type 7
  - **Dual Type 1 and Type 2** dimension
  - When there are too many Type 6 attributes

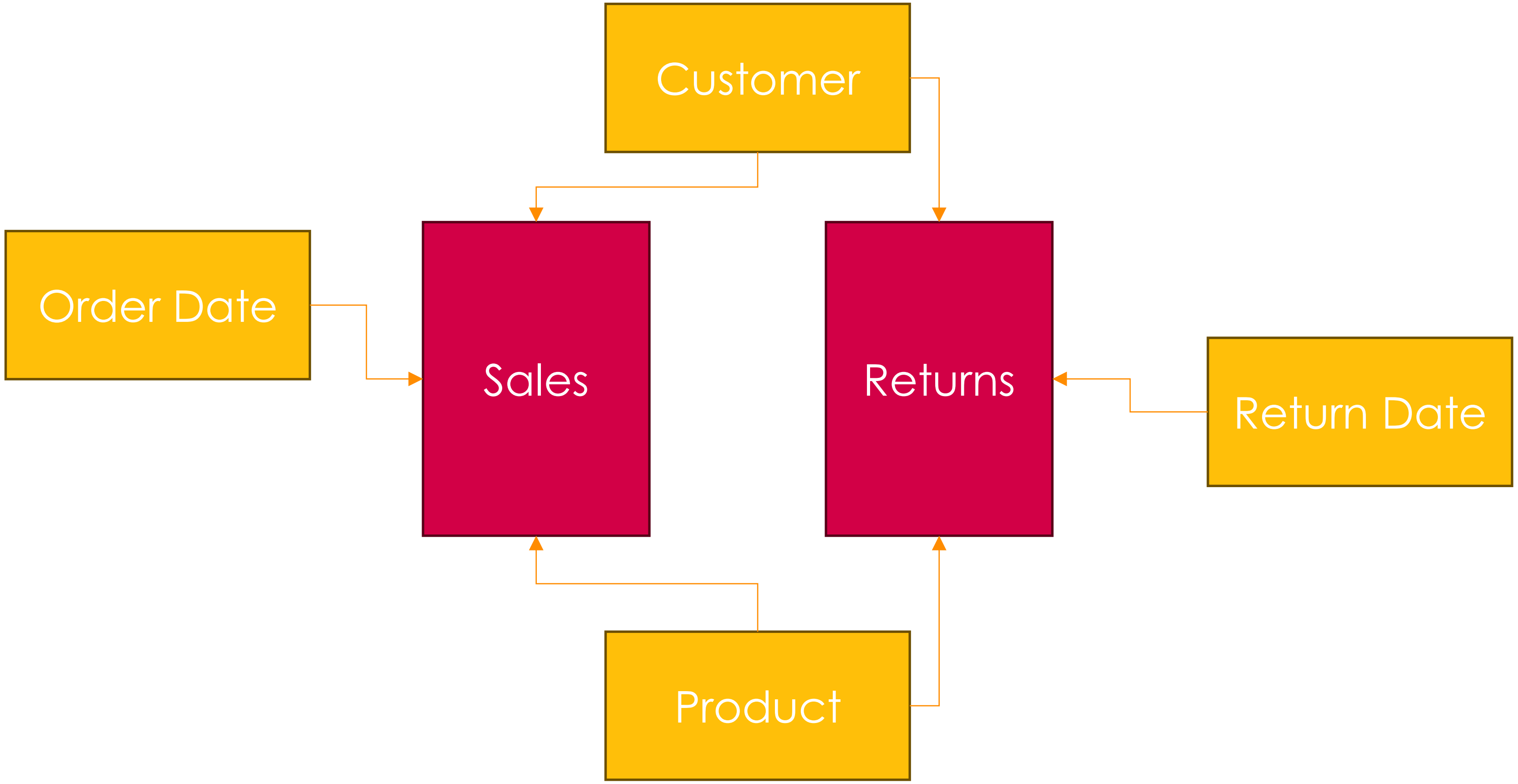


Top Customer?

SK_Customer	SK_Product	SK_OrderDate	OrderID	Amount
1	8	2026-01-15	D20260001	100
2	5	2026-01-30	D20260002	25
1	1	2026-02-05	D20260003	30
1	1	2026-02-05	D20260003	30
3	3	2026-02-18	D20260004	50

Customer_Name	Sum of Amount
Sarah	160
Lucie	50
David	25
Total	**235**

What if we include returns?



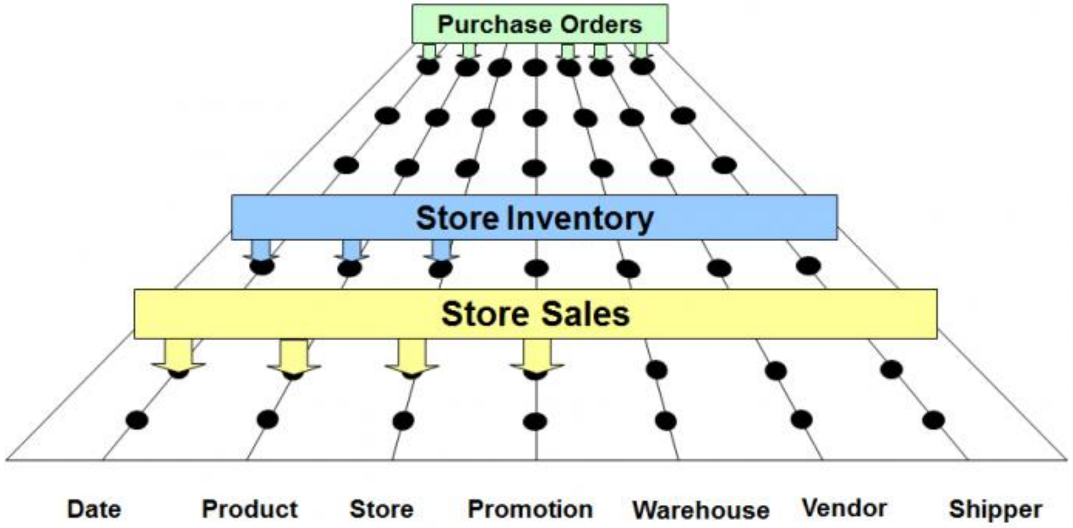
Top Customer?

SK_Customer	SK_Product	SK_ReturnDate	OrderID	Amount
1	8	2026-01-17	D20260001	100
1	1	2026-02-07	D20260003	30

Customer_Name	Sum of Amount	Final Sales
Lucie	50	50
Sarah	160	30
David	25	25
Total	**235**	**105**

Conformed Dimensions

- Dimensions shared between multiple fact tables
- **Essential** for cross-analysis of different business processes
  - also called "drilling-across"
- Kimball calls this the "Bus Architecture"
- Advantages
  - only one dimension to build
  - data and UI are consistent
  - no misinterpretation

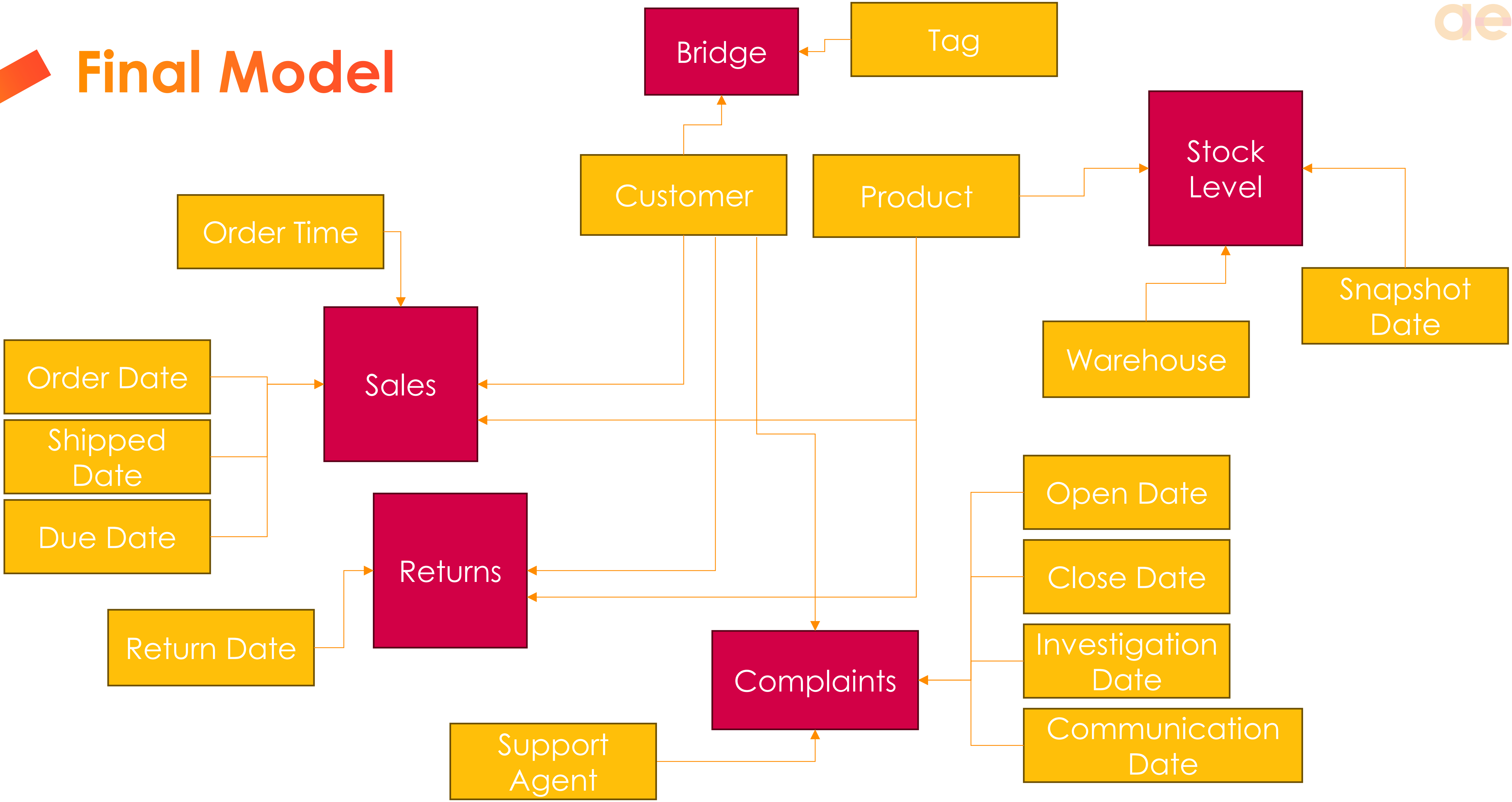


Source: [Kimball Group]

Information Matrix

	Sales	Returns	Complaints	Stock Level	*Bridge Tag Customer*
Customer	X	X	X		X
Product	X	X		X	
Date	X (Order/Shipped/ Due)	X	X (Open, Close, Inv., Comm.)	X (Snapshot)	
Time	X				
Employee			X (Support Agent)		
Tag					X
Complaint			X		
Warehouse				X	

Final Model



Bridge
Tag
Customer
Product
Stock Level
Order Time
Snapshot Date
Warehouse
Order Date
Sales
Shipped Date
Due Date
Returns
Open Date
Close Date
Investigation Date
Return Date
Complaints
Communication Date
Support Agent

Conclusion

Why Dimensional Modelling?

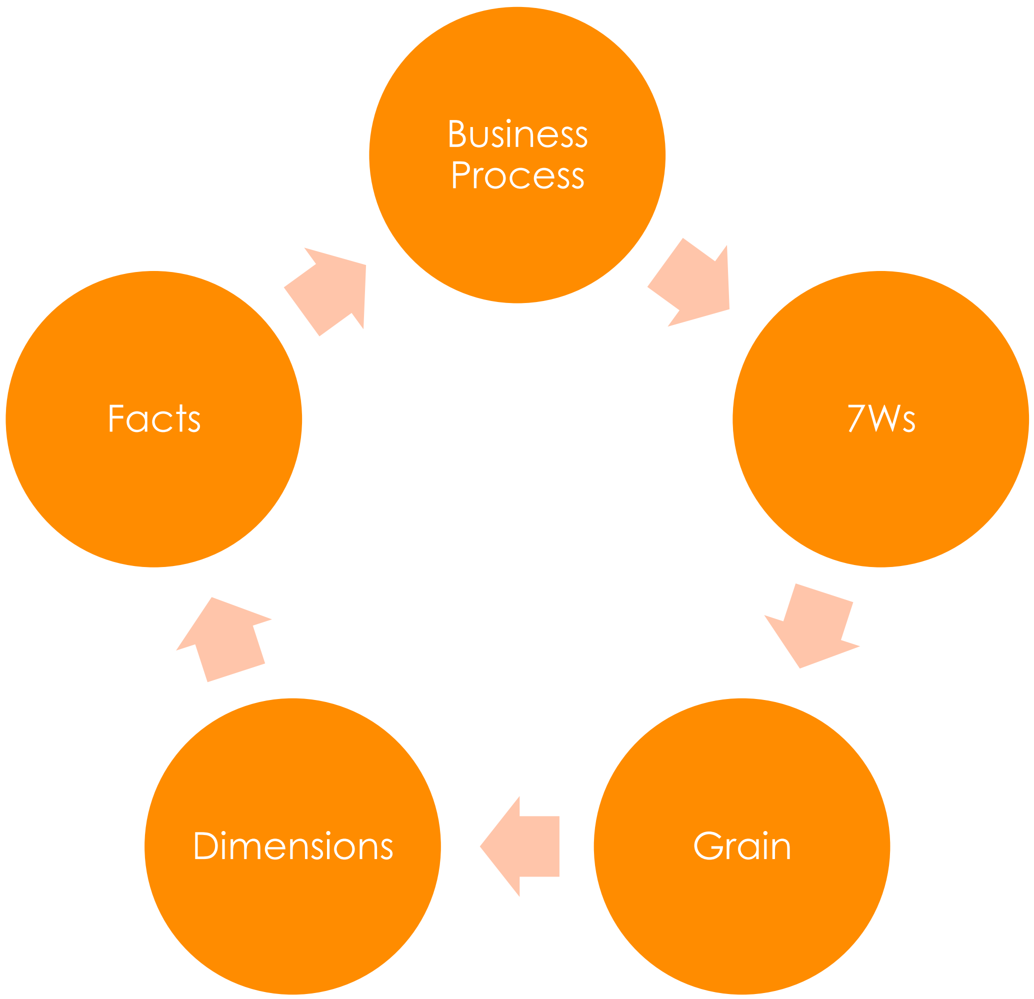
- Easy to understand, intuitive way of working with data
- Well-established in the industry
- Gives best performance in Power BI

- Whatever data warehouse methodology you choose, it all ends with star schema data marts

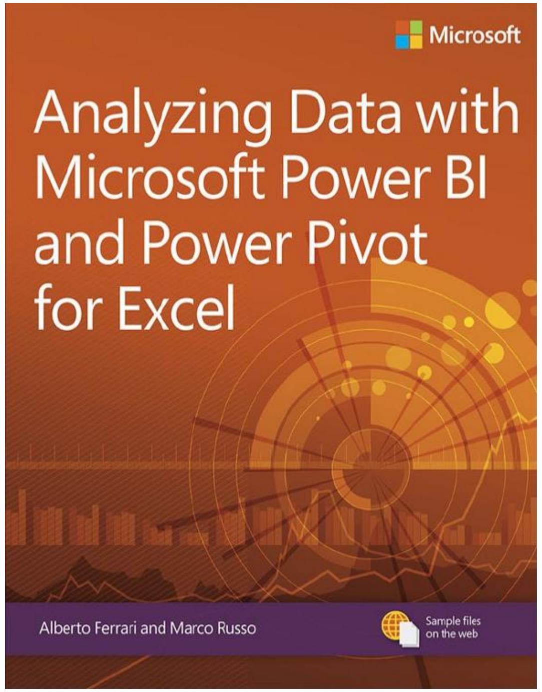
And don't go for the big bang

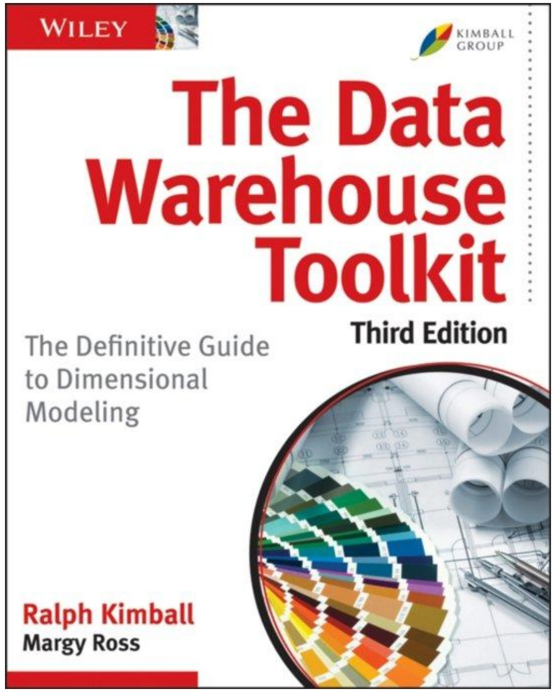
- **Incremental iterations**
- One star schema at a time
- Expect the following issues:
  - terminology definitions
  - data quality issues
  - master data?
  - business process anomalies

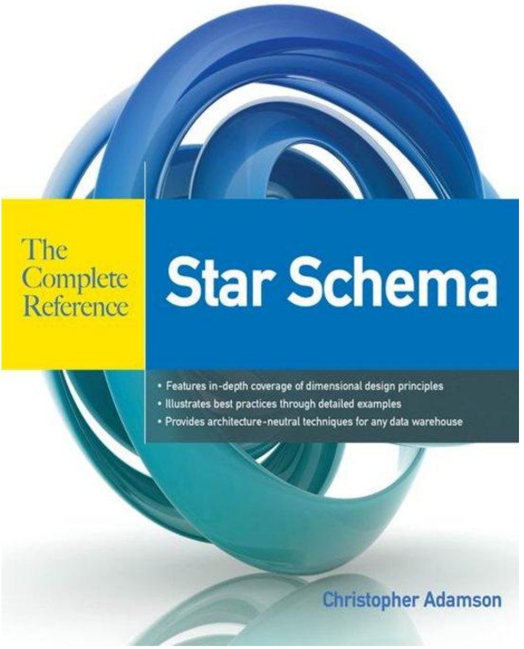
- Include business stakeholders from the start

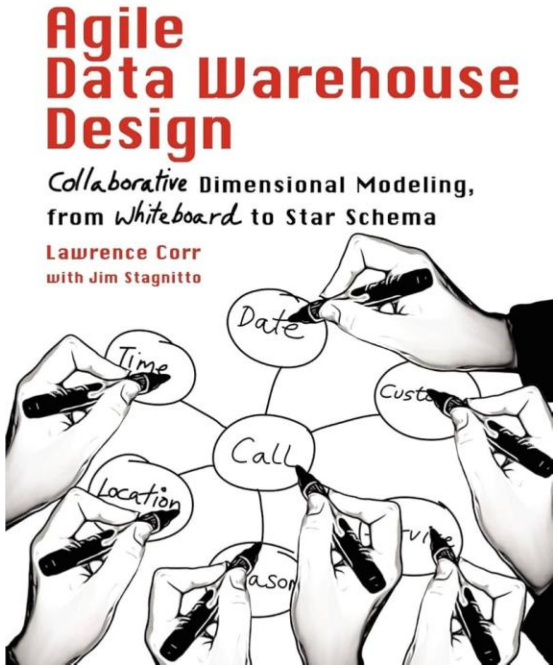


Recommended Reading









Merch

THANK YOU

www.ae.be
inspire@ae.be

+32 16 39 30 60
Interleuvenlaan 27b
3001 Heverlee