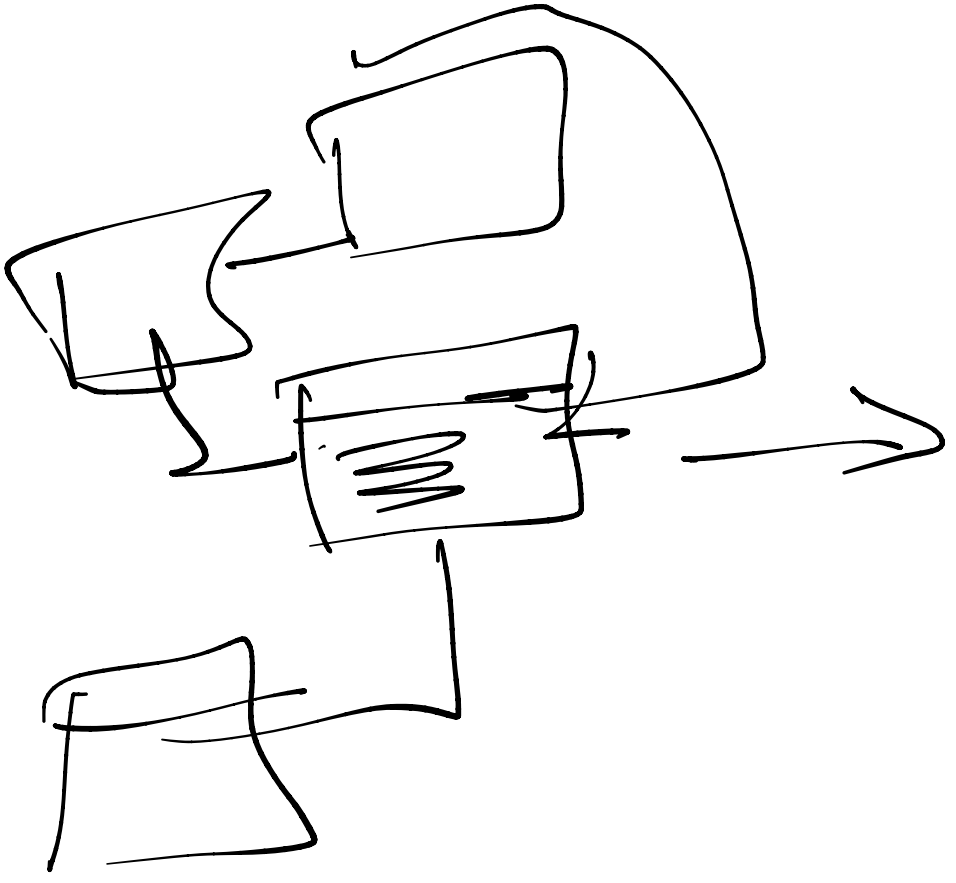


Fact

Dimension

Dimensional modeling

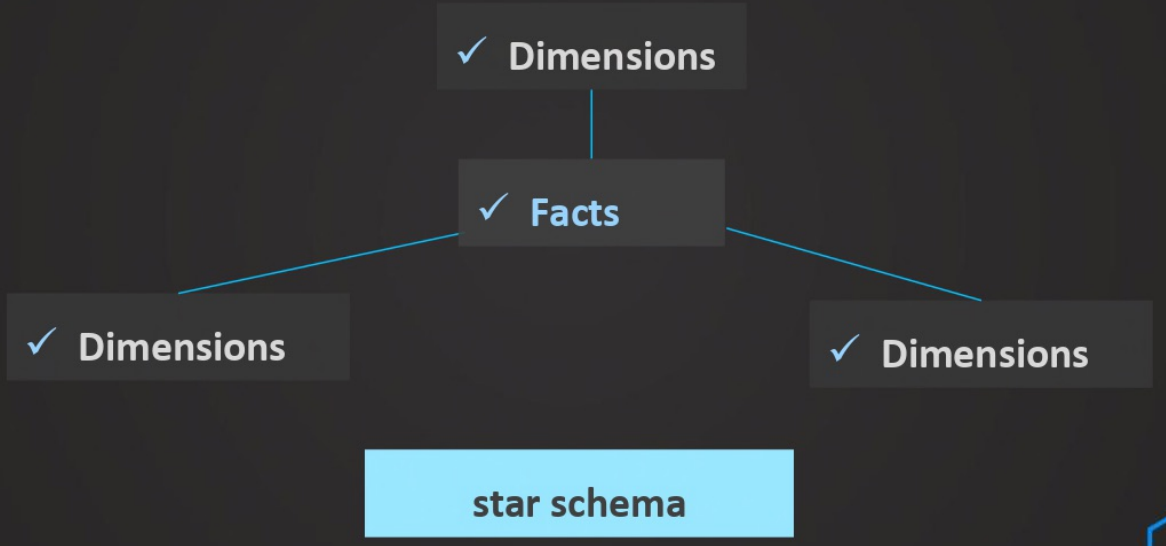
- ✓ Method of organizing data (in a data warehouse)

- ✓ Facts
  - Measurement like profit
- ✓ Dimensions
  - Context like category or period

Profit by year

Profit by category

Dimensional modeling



Dimensional modeling

- ✓ Unique technique of structering data
- ✓ Commonly used in DWH
- ✓ Optimized for faster data retrieval
- ✓ Oriented around performance & usability
- ✓ Designed Reporting / OLAP

Why dimensional modeling?

Dimensional modeling

- ✓ **Goal: Fast data retrieval**
- ✓ **Oriented around performance & usability**

id	date	product	category	customer_id	name	profit
1	1/2/2022	Fulltoss Tangy Tomato	Vegetables	2	Sarah	$23
2	1/2/2022	Chilli - Green, Organically Grown	Snacks	2	Sarah	$12
3	1/2/2022	Masala Powder	Herbs	5	Marc	$93
4	1/2/2022	Cheese Cracker (Mcvities)	Snacks	1	Frank	$23
5	1/2/2022	Centre Filled Chocolate Cake	Snacks	5	Marc	$21

Dimensional modeling

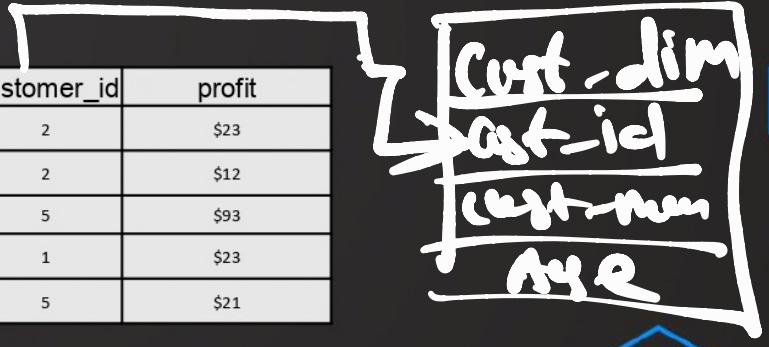
- ✓ **Goal: Fast data retrieval**
- ✓ **Oriented around performance & usability**

id	date	product	category	customer_id	name	profit
1	1/2/2022	Fulltoss Tangy Tomato	Vegetables	2	Sarah	$23
2	1/2/2022	Chilli - Green, Organically Grown	Snacks	2	Sarah	$12
3	1/2/2022	Masala Powder	Herbs	5	Marc	$93
4	1/2/2022	Cheese Cracker (Mcvities)	Snacks	1	Frank	$23
5	1/2/2022	Centre Filled Chocolate Cake	Snacks	5	Marc	$21

Dimensional modeling

- ✓ **Goal: Fast data retrieval**
- ✓ **Oriented around performance & usability**

id	date	product	category	customer_id	profit
1	1/2/2022	Fulltoss Tangy Tomato	Vegetables	2	$23
2	1/2/2022	Chilli - Green, Organically Grown	Snacks	2	$12
3	1/2/2022	Masala Powder	Herbs	5	$93
4	1/2/2022	Cheese Cracker (Mcvities)	Snacks	1	$23
5	1/2/2022	Centre Filled Chocolate Cake	Snacks	5	$21



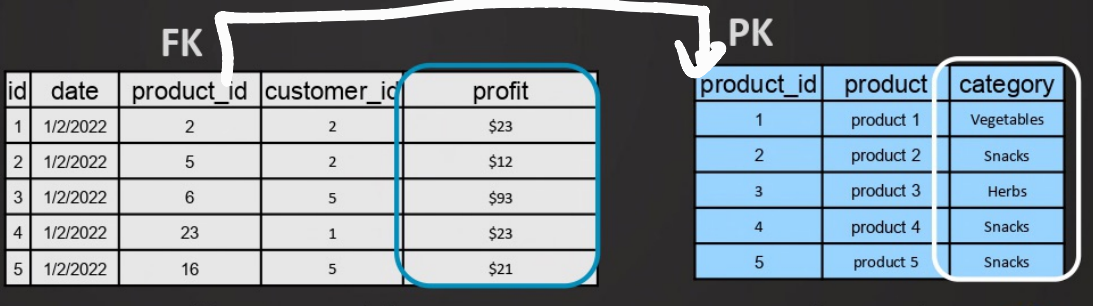
Dimensional modeling

- ✓ **Goal: Fast data retrieval**
- ✓ **Oriented around performance & usability**

id	date	product	category	customer_id	profit
1	1/2/2022	Fulltoss Tangy Tomato	Vegetables	2	$23
2	1/2/2022	Chilli - Green, Organically Grown	Snacks	2	$12
3	1/2/2022	Masala Powder	Herbs	5	$93
4	1/2/2022	Cheese Cracker (Mcvities)	Snacks	1	$23
5	1/2/2022	Centre Filled Chocolate Cake	Snacks	5	$21

Dimensional modeling

- ✓ Goal: Fast data retrieval
- ✓ Oriented around performance & usability



Dimensional modeling

- ✓ **Goal: Fast data retrieval**
- ✓ **Oriented around performance & usability**

id	date_id	product_id	customer_id	profit
1	20220102	2	2	$23
2	20220102	5	2	$12
3	20220102	6	5	$93
4	20220102	23	1	$23
5	20220102	16	5	$21

Dimensional modeling

- ✓ Goal: Fast data retrieval
- ✓ Oriented around performance & usability

id	date_id	product_id	customer_id	profit
1	20220102	2	2	$23
2	20220102	5	2	$12
3	20220102	6	5	$93
4	20220102	23	1	$23
5	20220102	16	5	$21

Profit Fact Table

date_id	weekday	month
20220102	Monday	January
20220103	Tuesday	January
20220104	Wednesday	January
20220105	Friday	January
20220106	Saturday	January

Date Dim

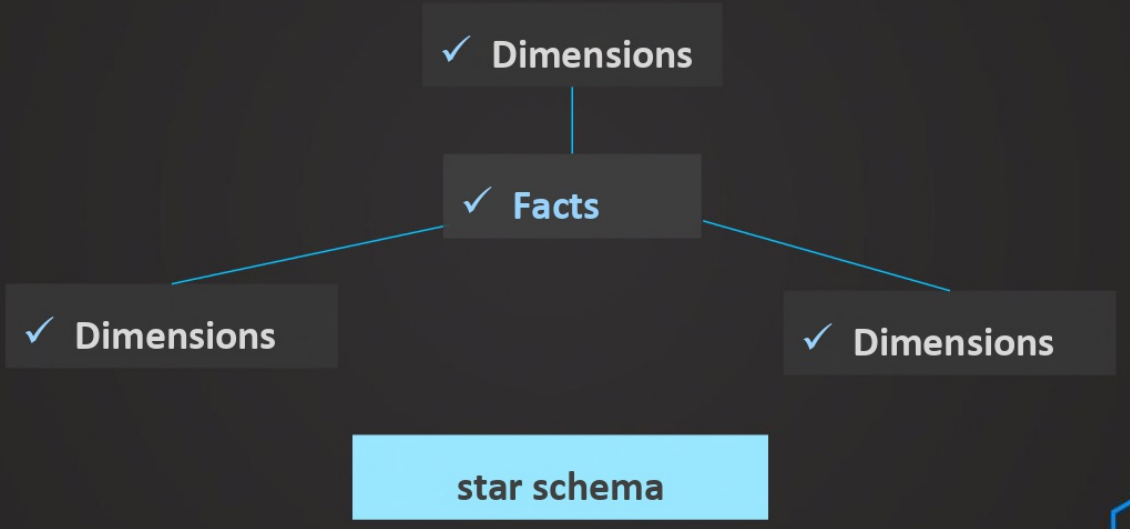
Dimensional modeling

- ✓ Goal: Fast data retrieval
- ✓ Oriented around performance & usability

Prefered technique for data warehouse!

Facts

Facts

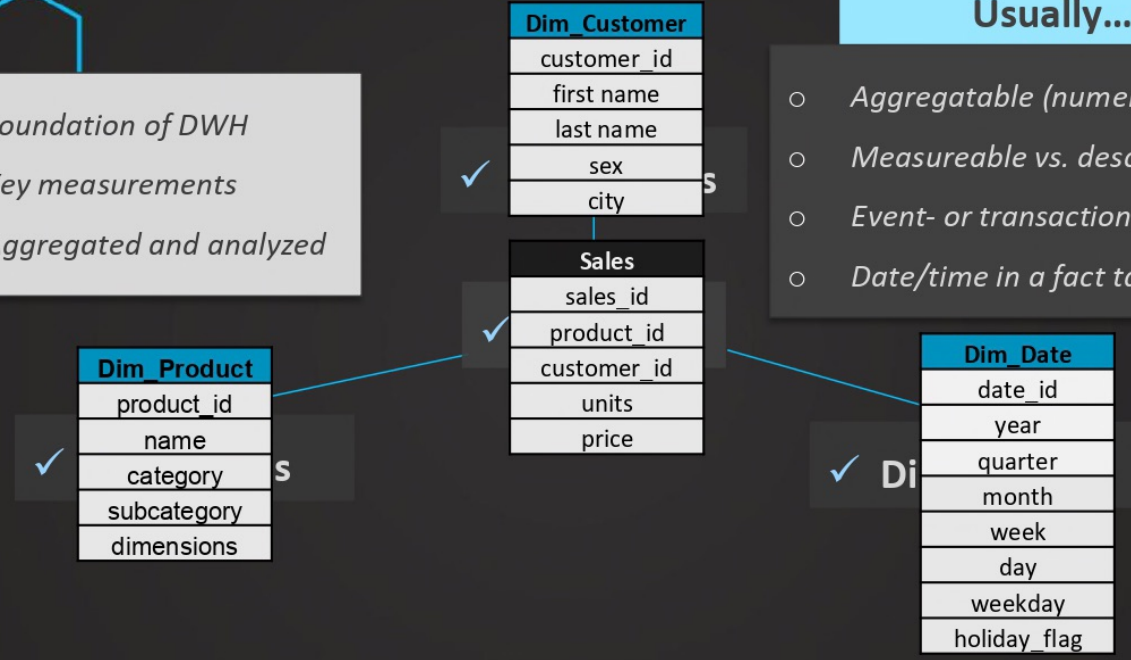


Facts

- *Foundation of DWH*
- *Key measurements*
- *Aggregated and analyzed*

Usually…

- *Aggregatable (numerical values)*
- *Measureable vs. descriptive*
- *Event- or transactional data*
- *Date/time in a fact table*



Facts

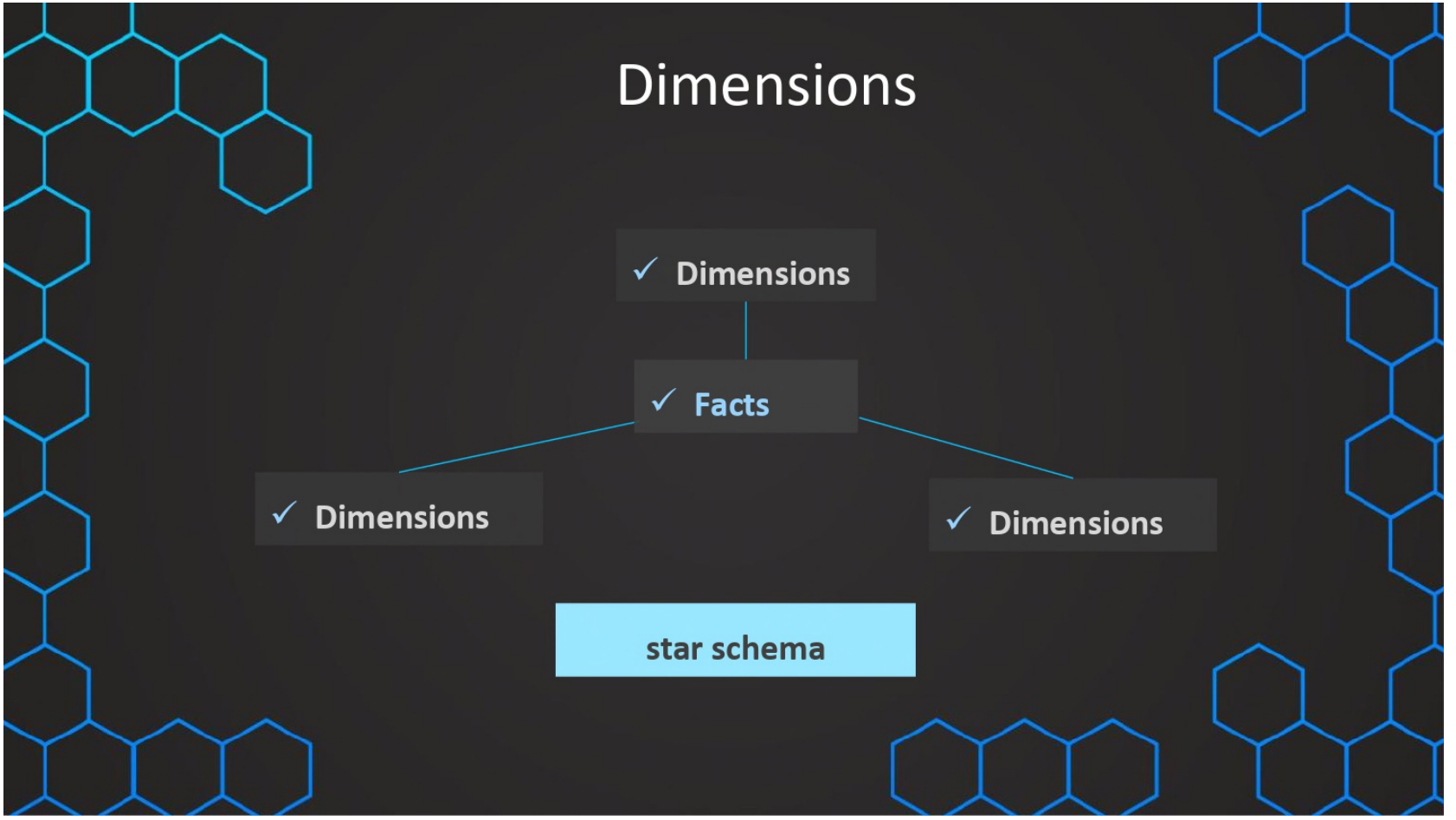
- ✓ Fact table: PK, FK & Facts
- ✓ Grain: Most atomic level facts are defined

id	date_id	region_id	profit
1	20220102	1	$23
2	20220102	2	$12
3	20220102	2	$93
4	20220102	3	$23
5	20220102	16	$21

- ✓ Different types of facts

Dimensions

Dimensions



Dimensions

- *Categorizes facts*
- *Supportive & descriptive*
- *Filtering, Grouping & Labeling*

Usually...

- *Non-Aggregatable*
- *Measureable vs. descriptive*
- *(More) static*

Dim_Customer
customer_id
first name
last name
sex
city

Sales
sales_id
product_id
customer_id
units
price

Dim_Product
product_id
name
category
subcategory
dimensions

Dim_Date
date_id
year
quarter
month
week
day
weekday
holiday_flag

Avg Price by Category, Customer Name

Dimensions

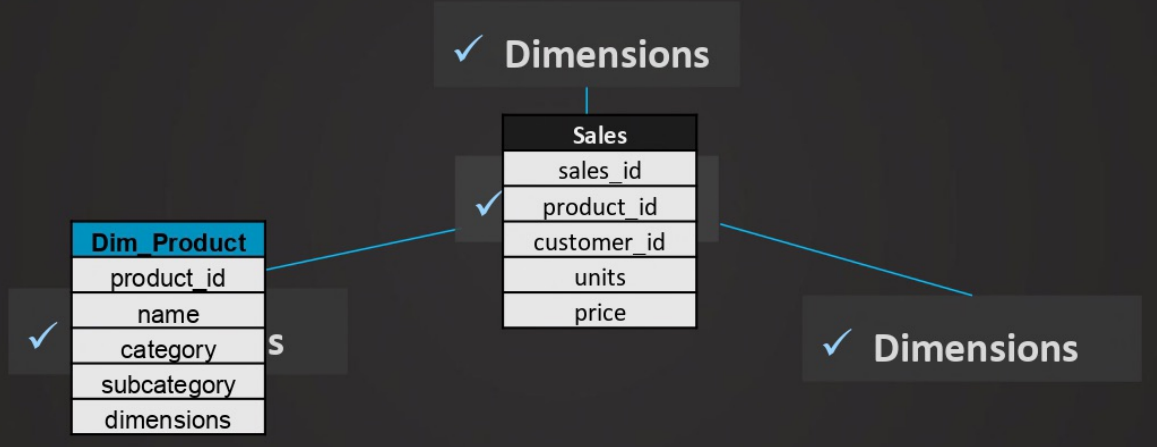
- ✓ Dimension table: PK, Dimension, (FK)
- ✓ People, products, places, time

customer_id	first_name	first_name	email
1	Mike	Miller	mike323@gmail.com
2	Sofia	Snider	snider_sof@gmail.com
2	Marco	Steadman	mstread23@gmail.com
3	Sarah	Griffith	sarah.griff@gmail.com
4	Jennifer	Lovell	jlovell@gmail.com

- ✓ Different types of dimension

Star schema

Star schema


✓ Dimensions
Sales
sales_id
product_id
customer_id
units
price
Dim_Product
product_id
name
category
subcategory
dimensions
✓ Dimensions

Star schema

Normalized

- *Technique to avoid redundancy*
- *Minimizes storage*
- *Performance (write / update)*
- *Many tables*
- *Many joins necessary*

FK

sales_id	product_id	customer_id	units	price
1	3	23	1	2.99
2	5	13	1	1.99
3	2	7	2	3.49
4	3	16	1	2.29
5	3	13	5	1.49

✓ Facts

1:n

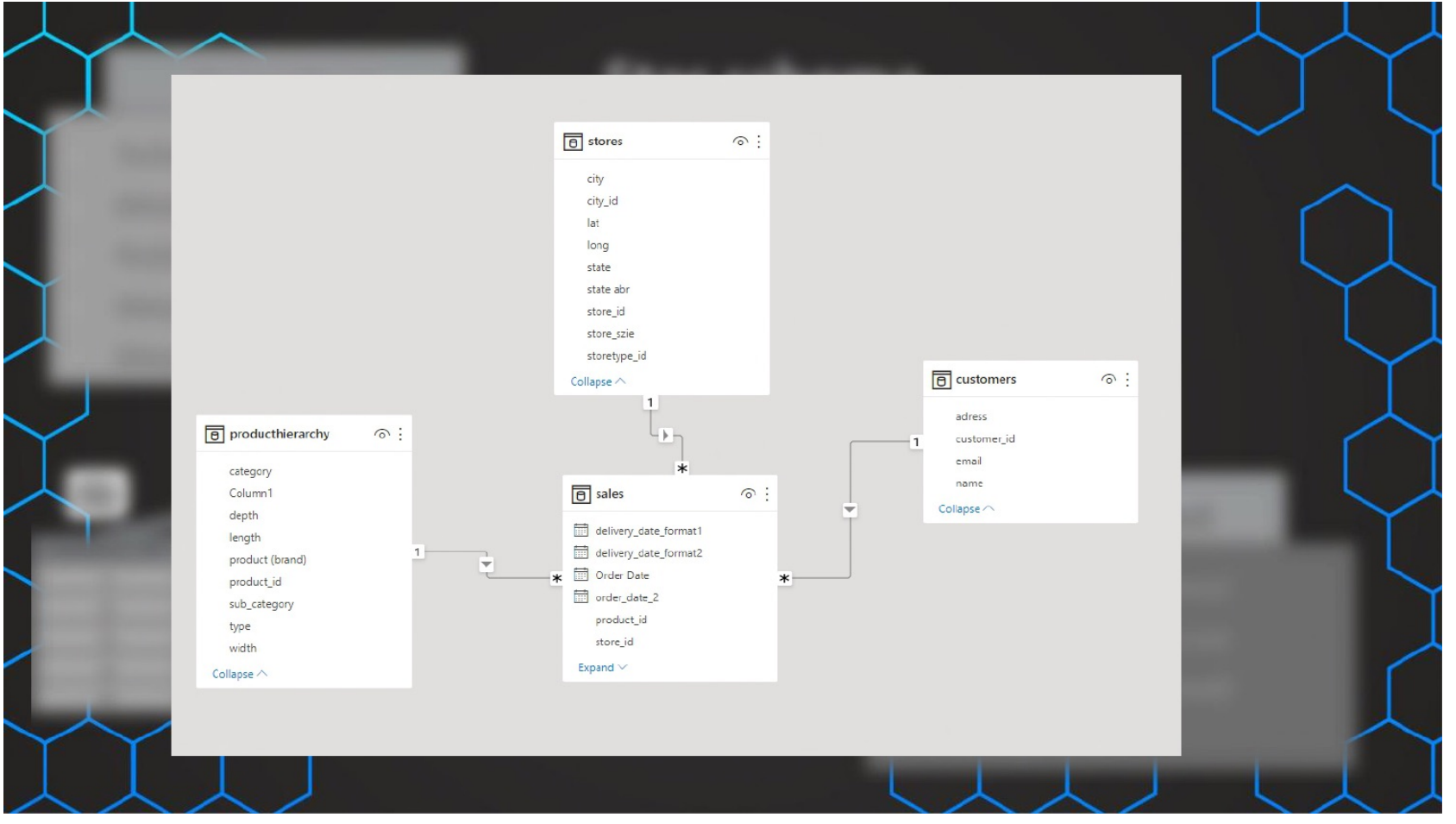
PK

✓ Dimensions

product_id	name	category	sub_category
1	Chili	Herbs	Spices
2	Garlic	Fruits & Vegetables	Vegetable
3	Banana	Fruits & Vegetables	Fruits
4	Chocolate	Sweets & Snacks	Sweets
5	Chips	Sweets & Snacks	Snacks

Denormalized

- *There is data redundancy!*
- *Optimized to get data out*
- *Query performance (read)*
- *User experience*


stores
city
city_id
lat
long
state
state abr
store_id
store_szie
storetype_id
Collapse
customers
adress
customer_id
email
name
Collapse
producthierarchy
category
Column1
depth
length
product (brand)
product_id
sub_category
type
width
Collapse
sales
delivery_date_format1
delivery_date_format2
Order Date
order_date_2
product_id
store_id
Expand
1
*

Star schema

- ✓ Most common schema in Data Mart
- ✓ Simplest form (vs. snowflake schema)
- ✓ Work best for specific needs
  (simple set of queries vs complex queries)
- ✓ Usablity + Performance for specific (read) use-case

Snowflake schema

Star schema

sales_id	product_id	customer_id	units	price
1	3	23	1	2.99
2	5	13	1	1.99
3	2	7	2	3.49
4	3	16	1	2.29
5	3	13	5	1.49

✓ Facts

✓ Dimensions

product_id	name	category	sub_category
1	Chili	Herbs	Spices
2	Garlic	Fruits & Vegetables	Vegetable
3	Banana	Fruits & Vegetables	Fruits
4	Chocolate	Sweets & Snacks	Sweets
5	Chips	Sweets & Snacks	Snacks

Snowflake schema

✓ Facts

sales_id	product_id	customer_id	units	price
1	3	23	1	2.99
2	5	13	1	1.99
3	2	7	2	3.49
4	3	16	1	2.29
5	3	13	5	1.49

product_id	name	category_id	sub_category
1	Chili	1	Spices
2	Garlic	2	Vegetable
3	Banana	2	Fruits
4	Chocolate	3	Sweets
5	Chips	3	Snacks

category_id	category
1	Herbs
2	Fruits & Vegetables
3	Sweets & Snacks

Snowflake schema

(More) normalized

Snowflake schema

Advantage

- ✓ Less space (storage cost)
- ✓ No (less) redundant data (easier to maintain/update, less risk of corrupted data)
- ✓ Solves write slow downs

Disadvantage

- ✓ More complex
- ✓ More joins (more complex SQL queries)
- ✓ Less performance Data Marts / Cubes

Types of fact tables

Transactional fact table

- ✓ 1 row = measurement of 1 event / transaction
- ✓ Taken place at a specific time
- ✓ One transaction defines the lowest grain

	FK	FK	Measure
sales_id	product_id	date_id	units
1	3	20220101	1
2	5	20220102	1
3	2	20220102	2
4	3	20220103	1
5	3	20220104	5

Sales transactions

	FK	FK	FK	Measure
call_id	emp_id	date_id	customer_id	duration
1	3	20220101	1	43
2	5	20220102	1	12
3	2	20220102	2	134
4	3	20220103	1	62
5	3	20220104	5	22

Calls

Characteristics

- ✓ Most common and very flexible
- ✓ Typically additive
- ✓ Tend to have a lot of dimensions associated
- ✓ Can be enormous in size

	FK	FK	Measure
sales_id	product_id	date_id	units
1	3	20220101	1
2	5	20220102	1
3	2	20220102	2
4	3	20220103	1
5	3	20220104	5

Sales transactions

	FK	FK	FK	Measure
call_id	emp_id	date_id	customer_id	duration
1	3	20220101	1	43
2	5	20220102	1	12
3	2	20220102	2	134
4	3	20220103	1	62
5	3	20220104	5	22

Calls

Periodic snapshot fact table

- ✓ 1 row = summarizes measure of many events / transactions
- ✓ Summarized of standard period (e.g. 1 day, 1 week etc.)
- ✓ Lowest period defines the grain

	Measure	Measure	Measure
week_id	revenue	sales	cost
1	323	123	12
2	541	322	31
3	242	108	12
4	352	212	51
5	312	198	25

Sales transactions

	Measure	Measure	Measure
day_id	no. calls	missed calles	duration
1	31	3	432
2	25	4	142
3	52	2	134
4	23	6	562
5	53	4	122

Calls

Characteristics

- ✓ Tend to be not as enormous in size
- ✓ Typically additive
- ✓ Tend to have a lot of facts and fewer dimensions associated
- ✓ No events = null or 0

	Measure	Measure	Measure
week_id	revenue	sales	cost
1	323	123	12
2	541	322	31
3	242	108	12
4	352	212	51
5	312	198	25

Sales transactions

	Measure	Measure	Measure
day_id	no. calls	missed calles	duration
1	31	3	432
2	25	4	142
3	52	2	134
4	23	6	562
5	53	4	122

Calls

Accumulation snapshot fact table

- ✓ 1 row = summarizes measure of many events / transactions
- ✓ Summarized of lifespan of 1 process (e.g. order fulfillment)
- ✓ Definite beginning & definite ending (& steps in between)

	Date FK	Measure	Date FK	Date FK	Date FK	Date FK	Date FK	Measure
order_id	Order Date FK	No. Products	Product_FK	Production Start FK	Production End FK	Inspection Date FK	Shipping Date FK	Damaged products
1	20220102	100	32	20220103	20220110	20220112	20220113	3
2	20220103	100	32	20220104	20220112	20220113	20220113	4
3	20220103	100	32	20220103	20220112	20220113	20220114	1
4	20220104	100	32	20220106	20220110	20220112	20220113	0
5	20220104	100	32	20220108	20220117	20220119	20220120	6

Order production

Characteristics

- ✓ Least common
- ✓ Workflow or process analysis
- ✓ Multiple Date/Time foreign keys (for each process step)
- ✓ Date/Time keys associated with role-playing dimension

	Date FK	Measure	Date FK	Date FK	Date FK	Date FK	Date FK	Measure
order_id	Order Date FK	No. Products	Product_FK	Production Start FK	Production End FK	Inspection Date FK	Shipping Date FK	Damaged products
1	20220102	100	32	20220103	20220110	20220112	20220113	3
2	20220103	100	32	20220104	20220112	20220113	20220113	4
3	20220103	100	32	20220103	20220112	20220113	20220114	1
4	20220104	100	32	20220106	20220110	20220112	20220113	0
5	20220104	100	32	20220108	20220117	20220119	20220120	6

Order production

Steps to create a fact table

Steps to create a fact table

Po_Id →
Or_Id →
Ord_Id →

1) Identify business process for analysis

Example:
Sales,
Order processing

sales_id	date	Sales amount
1	2022-01-01	$41
2	2022-01-02	$15
3	2022-01-02	$24
4	2022-01-03	$13
5	2022-01-04	$52

2) Declare the grain

Example: Transaction, Order, Order lines, Daily, Daily + location

3) Identify dimensions that are relevant

What, when, where, how and why

Example: Time, locations, products, customers,...

Filtering & grouping

"Soul" for analysis

4) Identify facts for measurement

Defined by the grain & not by specific use-case

Factless fact table

- ✓ Facts are usually numeric
- ✓ Sometimes only dimensionals aspects of an event are recorded
- ✓ Example new employee is registered

reg_id	Entry Date FK	dep_id	region_id	manager_id	Pos_id
1	20220102	1	2	3	10
2	20220103	3	3	4	112
3	20220103	4	6	3	202
4	20220104	4	8	6	110
5	20220104	3	4	8	17

Employee registration

Events

No metrics

Factless fact table

- ✓ How many employees have been registered last month?
- ✓ How many employees have been registered in a certain region?
- ✓ Example new employee is registered

reg_id	Entry Date FK	dep_id	region_id	manager_id	Pos_id
1	20220102	1	2	3	10
2	20220103	3	3	4	112
3	20220103	4	6	3	202
4	20220104	4	8	6	110
5	20220104	3	4	8	17

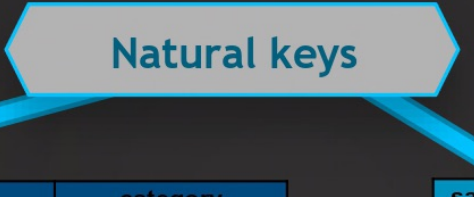
Employee registration

Events

No metrics

Natural vs. Surrogate key

Natural vs. Surrogate key



product_id	name	category
PX30	Chili	Herbs
PT32	Garlic	Fruits & Vegetables
AX42	Banana	Fruits & Vegetables
DA24	Chocolate	Sweets & Snacks
PO20	Chips	Sweets & Snacks

Products

sales_id	date	Sales amount
GXF-EFS	2022-01-01	$41
DOS-FWA	2022-01-02	$15
DSF-GWS	2022-01-02	$24
PTG-DWD	2022-01-03	$13
ERW-DWD	2022-01-04	$52

Sales

Natural vs. Surrogate key

Natural keys

- ✓ Come out of the source system

Product_PK	product_id	name	category
1	PX30	Chili	Herbs
2	PT32	Garlic	Fruits & Vegetables
3	AX42	Banana	Fruits & Vegetables
4	DA24	Chocolate	Sweets & Snacks

Surrogate key

Artificial keys

- ✓ Integer number
- ✓ _PK or _FK suffix
- ✓ Created by the database / ETL tool

Benefits

Surrogate key

- ✓ Improve performance (less storage/better joins)
- ✓ Handle dummy values (nulls / missing values) e.g. 999 or -1
- ✓ Integrate multiple source systems
- ✓ Easier administrate / update
- ✓ Sometimes there are even no natural keys available

Dimensions tables

Sales_PK	Website_FK	Customer_FK	Order_id	Order_line_ID
1001	2	312	2314	P034
1002	2	312	2314	P156
1003	2	312	2314	P643

- ✓ Always has a Primary Key (PK)

Product_ID	Name	Category
P001	Sunglases TR-7	Assecoirs
P002	Chocolate bar 70% cacao	Sweets
P003	Oat meal biscuits	Sweets

- ✓ Use surrogate key

Product_PK	Name	Category
1	Sunglases TR-7	Assecoirs
2	Chocolate bar 70% cacao	Sweets
3	Oat meal biscuits	Sweets

Product_PK	Product_ID
1	P001
2	P002
3	P003

- ✓ Lookup table

Dimensions tables

Sales_PK	Website_FK	Customer_FK	Order_id	Order_line_ID
1001	2	312	2314	P034
1002	2	312	2314	P156
1003	2	312	2314	P643

Product_ID	Name	Category
P001	Sunglases TR-7	Assecoirs
P002	Chocolate bar 70% cacao	Sweets
P003	Oat meal biscuits	Sweets

```
SELECT
      S.*,
      P.Product_PK
FROM Sales_Fact S
LEFT JOIN Product_Dim as P
ON P.Product_ID = S.Order_line_ID
```

Product_PK	Name	Category
1	Sunglases TR-7	Assecoirs
2	Chocolate bar 70% cacao	Sweets
3	Oat meal biscuits	Sweets

Product_PK	Product_ID
1	P001
2	P002
3	P003

Sales_PK	Website_FK	Customer_FK	Order_id	Order_line_ID	Product_FK	Unit_price
1001	2	312	2314	P034	34	22.99
1002	2	312	2314	P156	156	8.99
1003	2	312	2314	P643	643	16.99

Dimensions tables

- ✓ Always has a Primary Key (PK)

Product_PK	Name	Category
1	Sunglases TR-7	Assecoirs
2	Chocolate bar 70% cacao	Sweets
3	Oat meal biscuits	Sweets

Customer_id	Customer name	Order_id	Order_line_name	Order_line_FK	Quantity	Unit_price	Discounted_price
312	Franklin Miller	2345	Sunglases TR-7	1	2	23.99	23.99
312	Franklin Miller	2314	Beach towel red	156	3	8.99	8.99
312	Franklin Miller	2314	Swimsuit blue	643	1	16.99	14.99

- ✓ Relatively few rows / many columns with descriptive attributes

Date Dimension

- ✓ **One of the most common & most important dimensions**
- ✓ **Contains date related features**
  - ❖ Year, Month (name & number), Day, Quarter, Week, Weekday (name & number), …
- ✓ **Meaningful surrogate key** `YYYYMMDD`

  For example `2022-04-02` ⇔ `20220402`
- ✓ **Extra row for** `no date/null (source)` ⇔ `1900-01-01 (dim)`

Date Dimension

- ✓ Time is usually a separate dimension
- ✓ Can be populated in advance (e.g. for next 5 or 10 years)

Date features

- Numbers & Text (e.g. January, 1)
- Long & Abreviated (Jan, January – Mon, Monday)
- Combinations of attributes (Q1, 2022-Q1)
- Fiscal dates (Fiscal Year etc.)
- Flags (Weekend, company holidays etc.)

Date_PK	Date	Month	Short Month	Year-Quarter	Year	Weekday	Is_Weekend
20220101	2022-01-01	January	Jan	2022-Q1	2022	Saturday	1
20220102	2022-01-02	January	Jan	2022-Q1	2022	Sunday	1
20220103	2022-01-03	January	Jan	2022-Q1	2022	Monday	0