

# AUFBAU EINER DATENARCHITEKTUR AUF BASIS DES ZENSUS 2022 ZU BILDUNG UND ERWERBSTÄTIGKEIT

STUDIENARBEIT DATENINFRASTRUKTUR, MAREIKE LASS-HENNEMANN

JULI/AUGUST 2024

# Daten

- ▶ Ursprung: Zensus 2022
- ▶ Excel Arbeitsmappe, 10 aufbereitete und 10 einfache Arbeitsblätter
- ▶ Begleitende Informationen wie Erläuterungen, Methodik, Impressum
- ▶ Variablen: berichtszeitpunkt, \_rs (Gemeindeschlüssel), name, regionalebene, jeweils spezifische kombinierte Typen+Geschlecht-Attribute mit Zählwerten
- ▶ Beispiel:

Berichtszeitpunkt	RS	Name	Regionalebene	SCHUELER_KLSS_STP	SCHUELER_KLSS_STP_M	SCHUELER_KLSS_STP_W	SCHUELER_KLSS_STP_1
20220515	00	Deutschland	Bund	8903780	4536040	4367740	3048570
20220515	01	Schleswig-Holstein	Land	320060	162880	157180	109240
20220515	01001	Flensburg, Stadt	Stadtkreis/kreisfreie Stadt/Landkreis	10380	5010	5370	3530
20220515	010010000000	Flensburg, Stadt	Gemeinde	10380	5010	5370	3530
20220515	01002	Kiel, Landeshauptstadt	Stadtkreis/kreisfreie Stadt/Landkreis	24340	12270	12070	8770
20220515	010020000000	Kiel, Landeshauptstadt	Gemeinde	24340	12270	12070	8770



# TOOLS

- ▶ PostgreSQL als relationale Datenbank
- ▶ PG Admin zur Verwaltung
- ▶ VS Code als Entwicklungsumgebung und verbindendes Tool
- ▶ Erweiterungen in VS Code: Jupyter Notebooks/Python Kernel für ETL und Visualisierung, PostgreSQL Explorer für Normalisierung und Sternschema

# ETL – worauf musste bei der Transformation geachtet werden:

- ▶ Breites pivot-ähnliches Format anstatt langes Format
- ▶ Viele aggregierte Werte (insgesamt, zusammen, Bund)
- ▶ Aggregierte Werte stimmen z.T. nicht mit anderen Ebenen überein: Zahlen auf Ebene Gemeinde und Gemeindeverband sowie Regierungsbezirk

Zeilenbeschriftungen	Summe von Insgesamt	Summe von Männlich	Summe von Weiblich
Bund	8903780	4536040	4367740
Gemeinde	6626570	3379180	3247890
Gemeindeverband	1309460	667400	641970
Land	8903780	4536030	4367750
Regierungsbezirk	7013880	3573710	3440120
Stadtkreis/kreisfreie Stadt/Landkreis	8903770	4536180	4367720

- ▶ \_RS liegt als string vor und enthält führende Nullen
- ▶ Zwei mal thematisch Berufsabschlüsse vorhanden mit teilweise gleichen Attributen (Benamung „et\_“)



# Normalisierung

raw_klassenstufe	
PK	<u>rs</u>
PK	subtyp
PK	geschlecht
	regionalebene
	einheit
	berichtszeitpunkt
	name
	anzahl

raw_schulform	
PK	<u>rs</u>
PK	subtyp
PK	geschlecht
	regionalebene
	einheit
	berichtszeitpunkt
	name
	anzahl

außerdem noch in gleicher art:  
 raw\_schulabschluss  
 raw\_berufsabschluss  
 raw\_alter  
 raw\_et\_berufsabschl  
 raw\_berufsgruppe  
 raw\_stellung  
 raw\_wirtschaft  
 raw\_erwerbsstatus

← 1NF

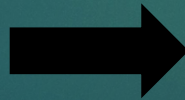
land	
PK	<u>land_id</u>
	land_name

kreis	
PK	<u>kreis_id</u>
	kreis_name
FK	land_id

geschlecht	
PK	<u>geschlecht_id</u>
	geschlecht_typ

zeitpunkt	
PK	<u>zeitpunkt_id</u>
	berichtszeitpunkt

2NF



schulform	
PK	<u>haupt_id</u>
	schulform_anzahl
FK	schulform_id
FK	geschlecht_id
FK	kreis_id
FK	zeitpunkt_id

schulform_typ	
PK	<u>schulform_id</u>
	schulform_typ
	einheit

klassenstufe	
PK	<u>haupt_id</u>
	klassenstufe_anzahl
FK	klassenstufe_id
FK	geschlecht_id
FK	kreis_id
FK	zeitpunkt_id

klassenstufe_typ	
PK	<u>klassenstufe_id</u>
	klassenstufe_typ
	einheit

schulabschluss	
PK	<u>haupt_id</u>
	schulabschluss_anzahl
FK	schulabschluss_id
FK	geschlecht_id
FK	kreis_id
FK	zeitpunkt_id

schulabschluss_typ	
PK	<u>schulabschluss_id</u>
	schulabschluss_typ
	einheit

berufsabschluss	
PK	<u>haupt_id</u>
	berufsabschluss_anzahl
FK	berufsabschluss_id
FK	geschlecht_id
FK	kreis_id
FK	zeitpunkt_id

berufsabschluss_typ	
PK	<u>berufsabschluss_id</u>
	berufsabschluss_typ
	einheit

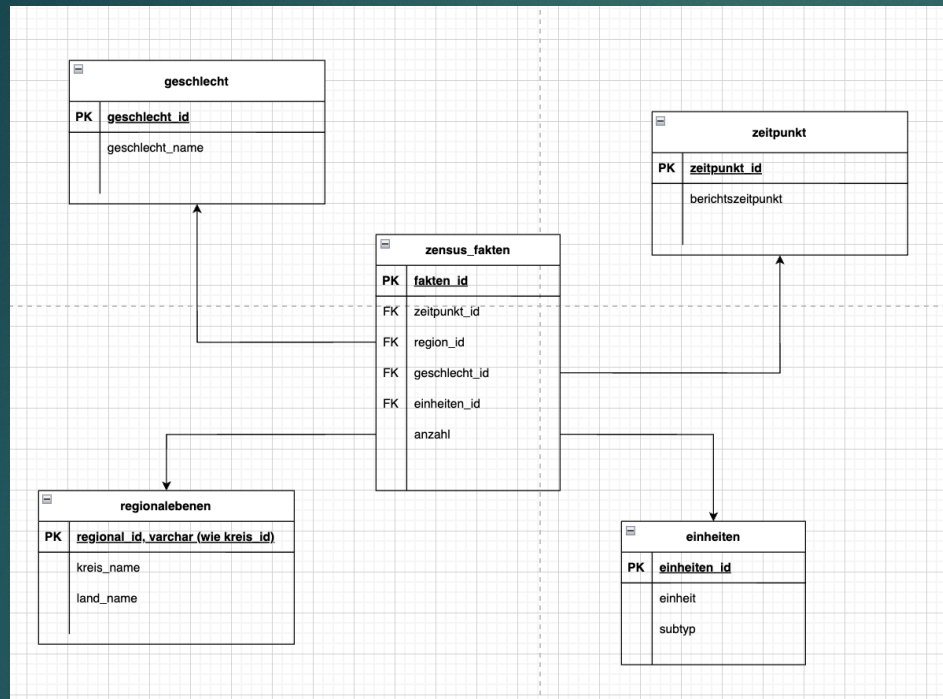
alter	
PK	<u>haupt_id</u>
	alter_anzahl
FK	alter_id
FK	geschlecht_id
FK	kreis_id
FK	zeitpunkt_id

alter_typ	
PK	<u>alter_id</u>
	alter_typ
	einheit

außerdem noch in gleicher art:  
 et\_berufsabschl  
 berufsgruppe  
 stellung  
 wirtschaft  
 sowie:  
 erwerbsstatus

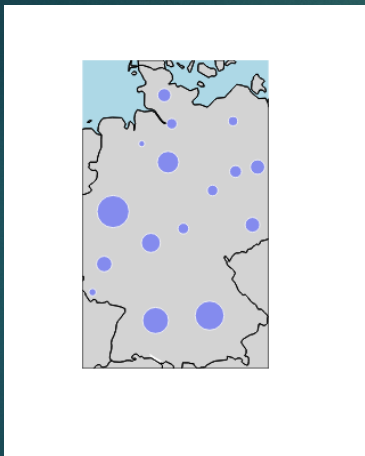
außerdem noch in gleicher Art:  
 et\_berufsabschluss\_typ,  
 berufsgruppe\_typ,  
 stellung\_typ,  
 wirtschaft\_typ  
 sowie erwerbsstatus\_typ

# Sternschema



# Visualisierungen, Anwendung des Sternschemas

Anzahl erwerbstätiger Personen  
in den Bundesländern



Prozentualer Anteil Berufsabschlüsse erwerbstätiger Personen

subtyp	bachelor, master, diplom, promotion (et)	fachschulabschluss (et)	lehre, duales system (et)	ohne beruflichen bildungsabschluss (et)
land				
baden-württemberg	26.075775	14.379530	36.710136	22.834560
bayern	25.967154	10.027345	46.216294	17.789207
berlin	40.696769	10.779379	26.467402	22.056450
brandenburg	21.458894	18.815011	45.197567	14.528528
bremen	28.565200	9.473979	34.034071	27.926750
hamburg	39.359101	6.413202	33.960674	20.267024
hessen	28.386813	12.212306	35.076636	24.324245
mecklenburg- vorpommern	18.687441	17.256539	49.874191	14.181828
niedersachsen	21.034161	14.653751	42.228671	22.083417
nordrhein-westfalen	24.177123	13.589736	38.448173	23.784967
rheinland-pfalz	21.685135	15.114391	40.278882	22.921592
saarland	21.002148	14.221910	42.280780	22.495162
sachsen	23.465295	18.039888	44.783283	13.711534
sachsen-anhalt	16.912839	18.905186	49.803542	14.378433
schleswig-holstein	21.717854	9.679491	50.272386	18.330269
thüringen	19.919061	18.953865	46.262647	14.864427



# Visualisierungen, Anwendung des Sternschemas

