

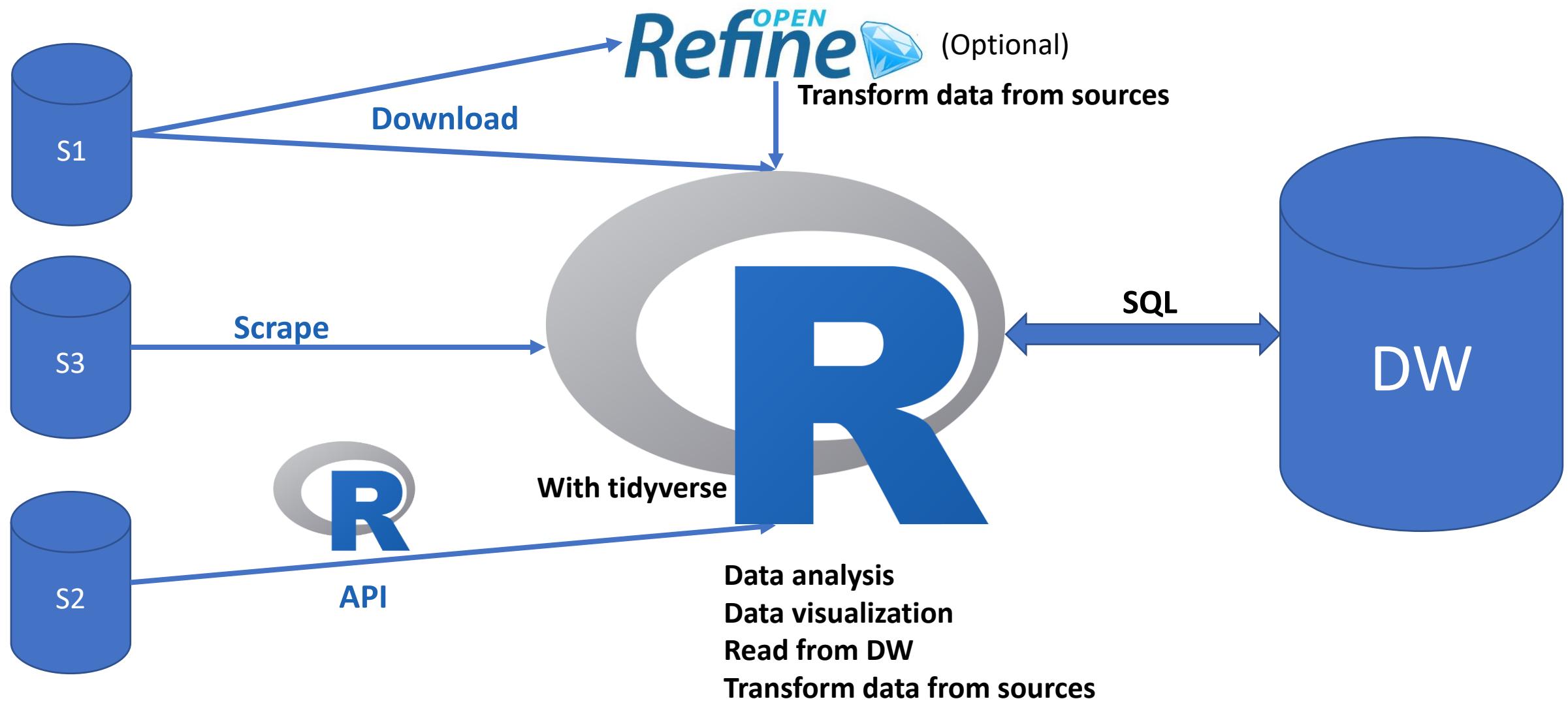
Data warehousing

Tutorial 3 – 23/05/2019

Armel Lefebvre

Business Intelligence 2018-2019

Role(s) of R in your Project



Modeling a DW

- Today we will...
 - ❖ Introduce DW modeling with star schemas
 - ❖ Give more details about the ETL pipeline

Data warehouse modeling

We will use *star schemas* (Kimball) to model our DW:

- DW is an OLAP system: Online Analytical Processing
- We use a relational database (Postgres), so we are doing ROLAP:
 - ❖ Relational Online Analytical Processing
 - ❖ OLAP with SQL
 - ❖ Facts and dimensions
- Other modeling techniques are
 - ❖ Snowflake (extension of star schema, normalization of dimensions), see ch3. p. 184 of Sharda
 - ❖ Data vault modeling, Dan Linstedt : auditability, hybrid 3NF/Star
 - ❖ Anchor modeling, Lars Rönnbäck: agile, changes by extension; not destructive
 - ❖ Data lakes: repository with schema-on-read, no (conceptual) modeling involved, see ch3. p. 193 of Sharda

Remark: Data vault and anchor modeling are just given as pointers here.

Modeling for Analytic Systems (or DW)

	Operational System	Analytic System (Data Warehouse)
Purpose	Execution of a business process	Measurement of a business process
Primary Interaction Style	Insert, Update, Query, Delete	Query
Scope of Interaction	Individual transaction	Aggregated transactions
Query Patterns	Predictable and stable	Unpredictable and changing
Temporal Focus	Current	Current and historic
Design Optimization	Update concurrency	High-performance query
Design Principle	Entity-relationship (ER) design in third normal form (3NF)	Dimensional design (Star Schema or Cube)
Also Known As	Transaction System	Data Warehouse System
	On Line Transaction Processing (OLTP) System	Data Mart
	Source System	

Facts and Dimensions: Why is that?

- 51% of Open Access Publications
- 150 registered students
- 1.67% of Gross Domestic Product

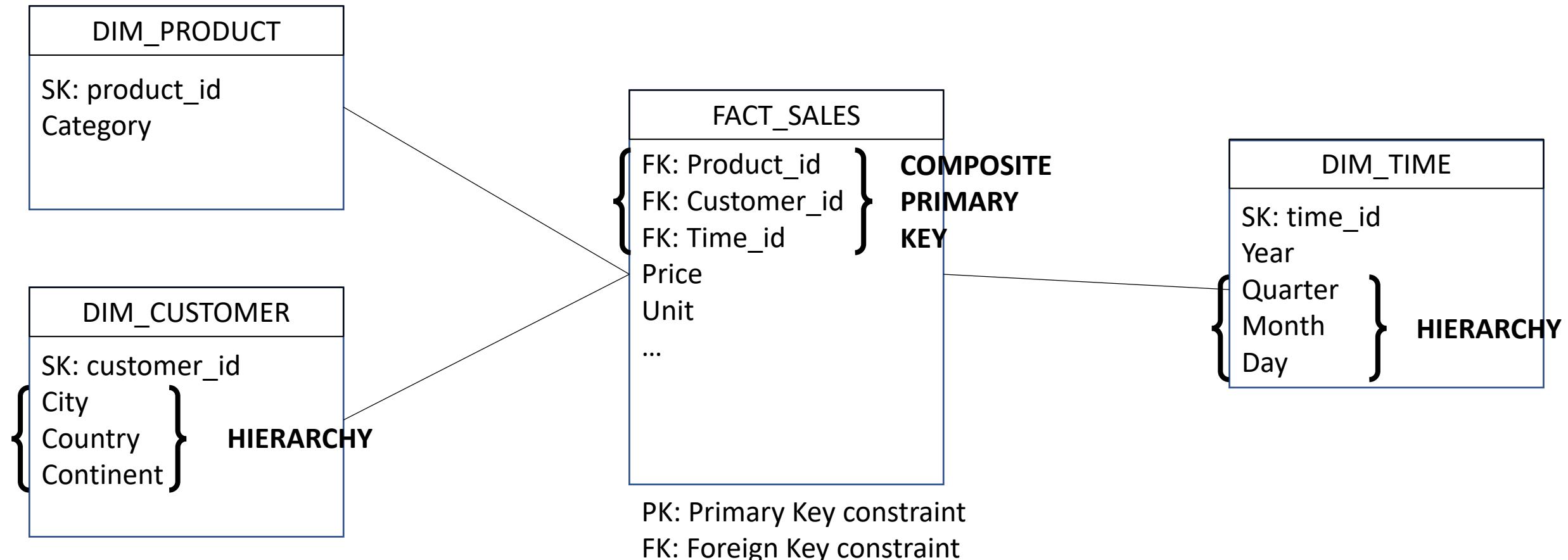
Give context/information to measures

- 51% of Open Access Publications in the International Journal on the Theory of Everything in 2017
- We predict 150 registered students for the BI course in 2021 at Utrecht University
- Total R&D expenditure in the UK in 2016 represented 1.67% of gross domestic product (GDP), unchanged from 2015, remaining below the European Union (EU-28) provisional estimate of 2.03%.
 - Actually, this one is true, see:
<https://www.ons.gov.uk/economy/governmentpublicsectorandtaxes/researchanddevelopmentexpenditure/bulletins/ukgrossdomesticexpenditureonresearchanddevelopment/2016>

Structuring measures

- Measures : Facts, e.g., 150 registered students
- Context : Dimensions, e.g., business intelligence course, Utrecht University
- Example query: Show the average *number of open access publications* **for** Brazil and France in 2018
- Fact : *number of open access publications*
- Dimensions: Country (France, Brazil), Year (2018).

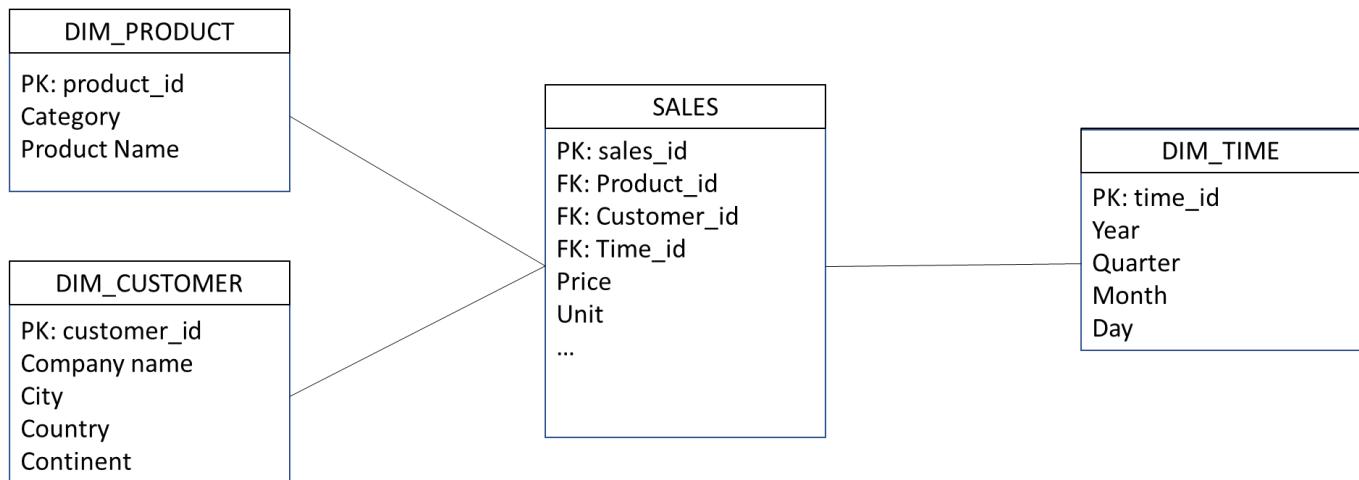
STAR schema (ROLAP)



SK means surrogate key, which is a primary key which is generated by the database

Not normalized, high redundancy!

Query Example



SQL:

```
SELECT product_name, AVG(price)  
FROM sales INNER JOIN dim_product  
ON sales.product_id = dim_product.product_id  
INNER JOIN dim_customer  
ON dim_customer.customer_id = sales.customer_id  
INNER JOIN dim_time  
ON dim_time.time_id = sales.time_id  
WHERE dim_customer.continent = 'Europe' AND  
dim_time.Year = 2017 AND product_category = 'Fruit'  
GROUP BY product_name;
```

Cube

- Cube is a technical thing
- It refers to Multidimensional database
- So, it's a data warehouse using an alternative to a SQL database
- It's an Alternative to ROLAP - > MOLAP
- At a “conceptual” level, the operations of your dashboard (e.g., slice, dice, drill-down) can be visualized as manipulating a cube
 - Even with ROLAP

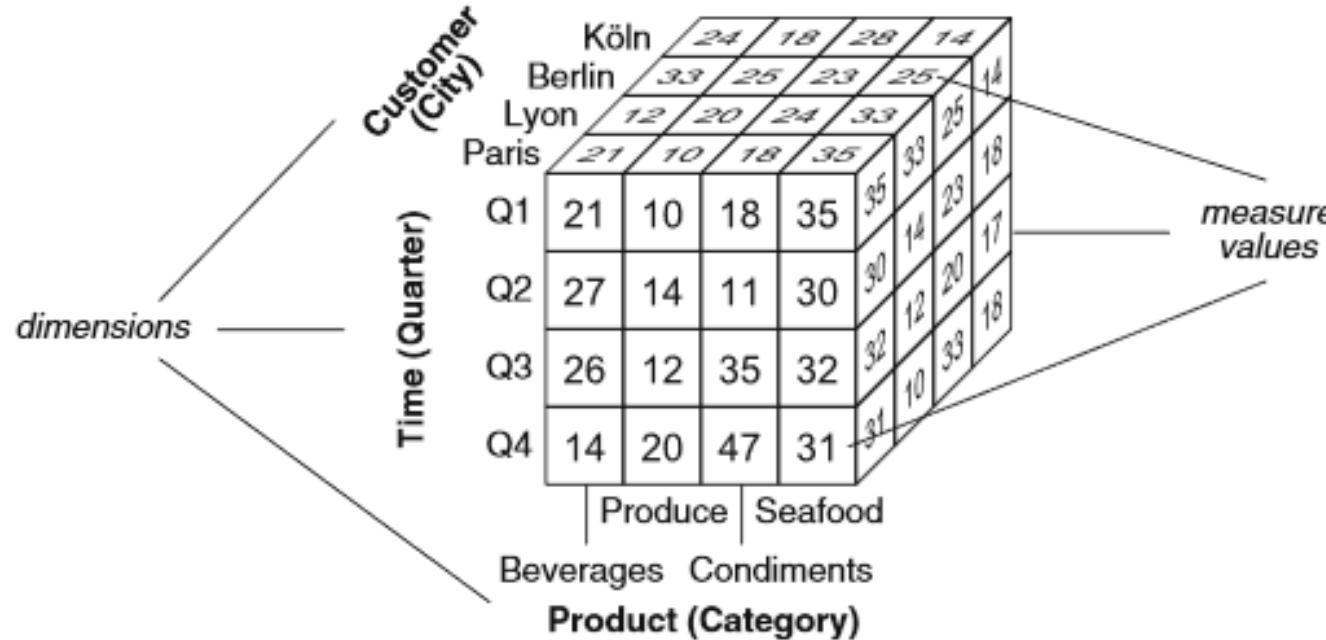


Fig. 3.1 A three-dimensional cube for sales data with dimensions Product, Time, and Customer, and a measure Quantity

Cube

- **Cube:** multidimensional representation of the data
- **A Dimension:**
 - Contains data used to slice, dice, drill-down etc.
- **A Fact:** Relates dimensions and store measures (values)
- **A Hierarchy:** Level of granularity of a dimension:
 - Example: World (All) – Continent – Country - ...
 - Roll-up, drill down

Model DW

- Use conceptual and relational data modeling
- Conceptual for communication purposes (simpler, no FK)
 - Business strategy, business processes
- Logical for implementation in relational database (with FK)
 - Implement a data warehouse in Postgres

Example

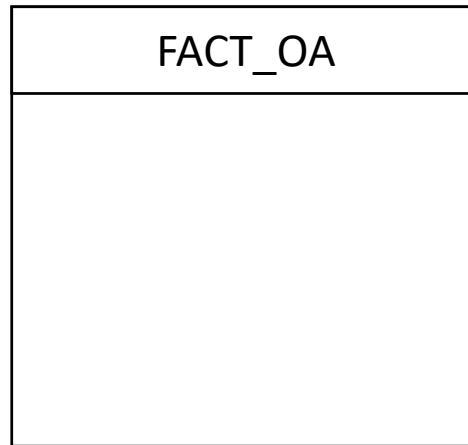
Strategy

- Identify entities/processes that you are evaluating from your CSFs
 - Open access to scientific articles for all citizens in Europe
- Open access => Measuring Open Access Attributes
- Fact : Open Access
- KPI:
 - Average number of OA pub.
 - Average APC
 - APC: Article Processing Charge

FACT_OA

Facts and Dimensions

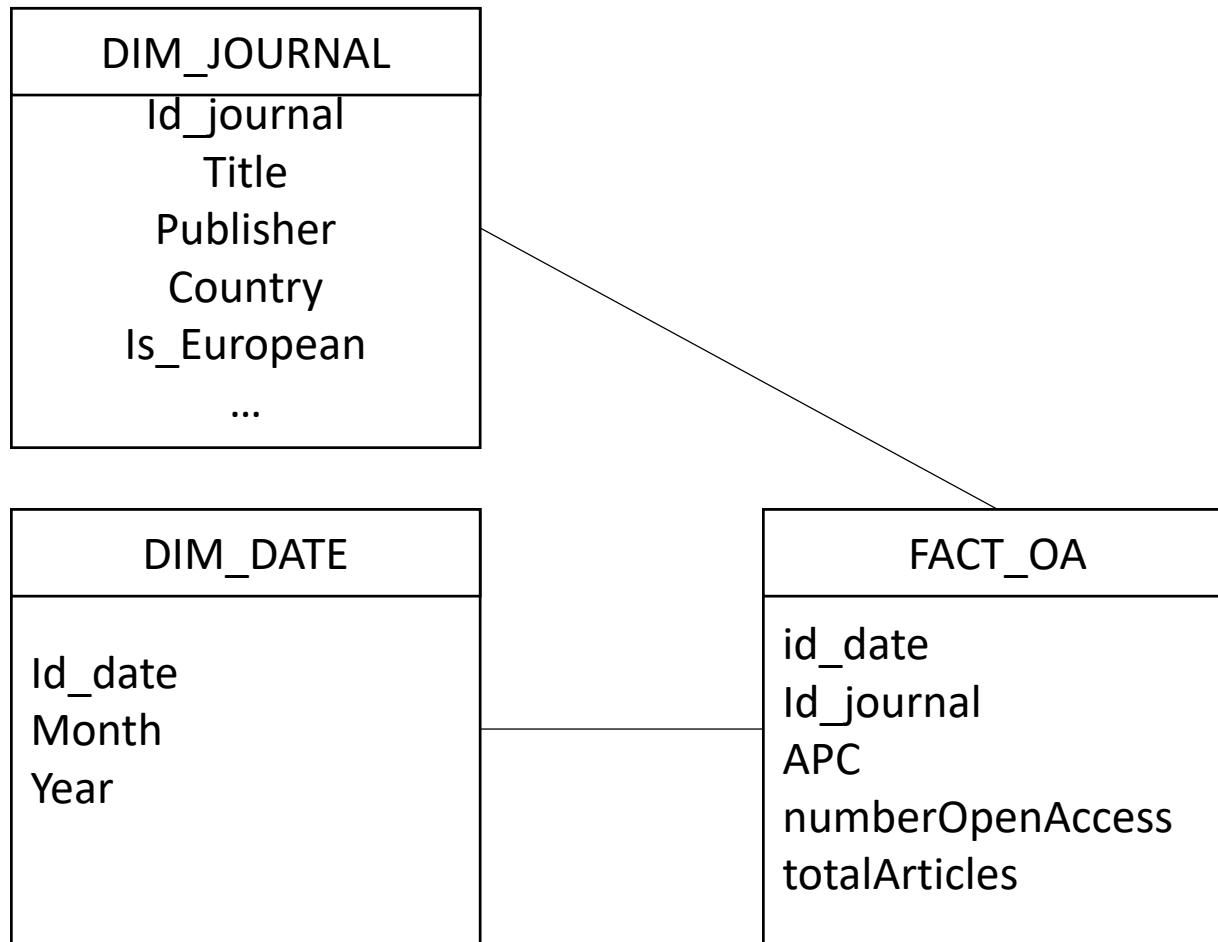
- Open access to scientific articles {for all citizens} in Europe



- Choose granularity and context: how will you define **in Europe**
 - *Origin of authors*
 - *Origin of publishers (V)*
 - ...

Historicity

- Open access to scientific articles {for all citizens} in Europe



Overall ETL pipeline

- Extract from Data sources: excel sheet, web api, csv...
- Transform: Remove columns, normalize data, deal with missing values, align data sets (time series)...
 - With Open Refine
 - With R Markdown
- Load: Structure your data in accordance with your star schema
 - With R Markdown

ETL \ Extract

- Extract sources
 - Downloads
 - API
 - Scrape (not the preferred option)

Help, too much data

- Remember your strategy
 - Select what you need
 - Justify criteria
- Example : scopus

Scopus query

Scopus

315,008 document results

TITLE-ABS-KEY ("artificial intelligence")

Edit Save Set alert Set feed

Search within results... Analyze search results Show all abstracts Sort on: Date (newest)

Refine results

Limit to Exclude

Access type ①

- Open Access (18,894) >
- Other (296,114) >

Year

- 2020 (52) >
- 2019 (7,635) >
- 2018 (27,055) >
- 2017 (23,849) >
- 2016 (22,950) >

View more

Author name

Document title	Authors	Year	Source	Cited by
1 The Impact of Artificial Intelligence on the Accounting Industry	Shi, Y.	2020	Advances in Intelligent Systems and Computing 928, pp. 971-978	0
2 The Artificial Intelligence Application in the Management of Contemporary Organization: Theoretical assumptions, current practices and research review	Jelonek, D., Mesjasz-Lech, A., Stępiński, C., Turek, T., Ziora, L.	2020	Lecture Notes in Networks and Systems 69, pp. 319-327	0
3 Research on the Development Trend of Online Education Industry Considering the Influence of Big Data and Artificial Intelligence	Fu, Y.	2020	Advances in Intelligent Systems and Computing 928, pp. 852-859	0

View abstract UBU link Related documents

View abstract UBU link Related documents

View abstract UBU link Related documents

ETL \ Transform

- The source can be messy
- Unify attributes and values before you load them in Postgres
- <http://openrefine.org/>

Load with RMarkdown

```
```{sql connection=DATABASE}

CREATE TABLE IF NOT EXISTS dim_year(
 year_id char(4) PRIMARY KEY
);

CREATE TABLE IF NOT EXISTS dim_country(
 country_id char(3) PRIMARY KEY
);

CREATE TABLE IF NOT EXISTS fact_science_hr(
 year_id char(4) REFERENCES dim_year(year_id) NOT NULL,
 country_id char(3) REFERENCES dim_country(country_id) NOT NULL,
 value DECIMAL,
 PRIMARY KEY(year_id, country_id)
);
````
```

Load with RMarkdown

```
```{r}
Load to PostgreSQL

dim_year <- as_tibble(unique(result$year)) %>% rename(year_id = value)
dim_country <- result %>% distinct(geo) %>% rename(country_id = geo)

fact_science_hr <- result %>% rename(year_id = year, country_id = geo, value = AvgSciencPop)

#We prefer to create tables ourselves to make sure the schema is right, so the table already exists and append=TRUE
#must be used
dbwriteTable(DATABASE, "dim_year", dim_year, append=TRUE, row.names = FALSE)
dbwriteTable(DATABASE, "dim_country", dim_country, append=TRUE, row.names = FALSE)
dbwriteTable(DATABASE, "fact_science_hr", fact_science_hr, append=TRUE, row.names = FALSE)
...```

```

# Cubes

# CUBE operations with SQL/Dyplr

- SLICE/DICE
  - WHERE/HAVING
  - Select one or more dimensions
- Roll-up/Drill-down
  - GROUP BY
  - Zoom in or out (granularity)
- Pivot
  - Rotate dimensions

## Operations

a

Customer (City)	Köln				Berlin				Lyon				Paris			
	24	18	28	14	33	25	23	25	12	20	24	33	21	10	18	35
Time (Quarter)	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	Produce	Seafood														
	Beverages	Condiments														
	Product (Category)															

Pivot

Customer (City)	Seafood				Condiments				Produce				Beverages			
	35	30	32	31	18	11	35	47	10	14	12	20	21	27	26	14
Time (Quarter)	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	Paris	21	27	26	14	14	20	21	10	21	17	21	21	27	26	14
	Lyon	12	14	11	13	13	28	20	13	12	11	13	13	28	20	13
	Berlin	33	28	35	32	32	19	47	32	33	28	32	32	19	47	32
	Köln	24	23	25	18	18	19	18	24	23	25	18	24	23	25	18
		Q1	Q2	Q3	Q4											
		Time (Quarter)														

## Operations

a

Customer (City)	Köln				Berlin				Lyon				Paris							
	24	18	28	14	33	25	23	25	12	20	24	33	25	14	21	10	18	35		
Time (Quarter)	Q1	Q2	Q3	Q4																
Product (Category)	Produce	Seafood	Beverages	Condiments																
Customer (City)	Köln	Berlin	Lyon	Paris																
Time (Quarter)	Q1	Q2	Q3	Q4																
Product (Category)	Produce	Seafood	Beverages	Condiments																

Roll up

Customer (Country)	Germany				France															
	57	43	51	39	33	30	42	68												
Time (Quarter)	Q1	Q2	Q3	Q4																
Product (Category)	Produce	Seafood	Beverages	Condiments																
Customer (Country)	Germany	France	Germany	France																
Time (Quarter)	Q1	Q2	Q3	Q4																
Product (Category)	Produce	Seafood	Beverages	Condiments																

## Operations

a

Customer (City)		Köln				Berlin				Lyon				Paris								
		24	18	28	14	33	25	23	25	12	20	24	33	25	14	21	10	18	35			
Time (Quarter)	Q1	21	10	18	35	35	33	25	14	23	20	33	18	17	27	14	11	30	30	12	20	18
	Q2	27	14	11	30	30	33	25	14	23	20	33	18	17	26	12	35	32	32	10	33	18
	Q3	26	12	35	32	32	33	25	14	23	20	33	18	17	21	10	18	35	35	14	23	20
	Q4	14	20	47	31	31	33	25	14	23	20	33	18	17	14	20	47	31	31	10	33	18
		Produce	Seafood	Beverages	Condiments																	

Dice: City and quarter



Customer (City)		Time (Quarter)				Product (Category)			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
Lyon	Paris	21	10	18	35	35	33	14	17
		27	14	11	30	30	33	12	18
		26	12	35	32	32	33	10	18
		14	20	47	31	31	33	10	18

Slide would be one dimension (so not a 3 dimensional figure)

Figure from: Vaisman, A., & Zimányi, E. (2014). *Data warehouse systems*. Springer, Heidelberg.