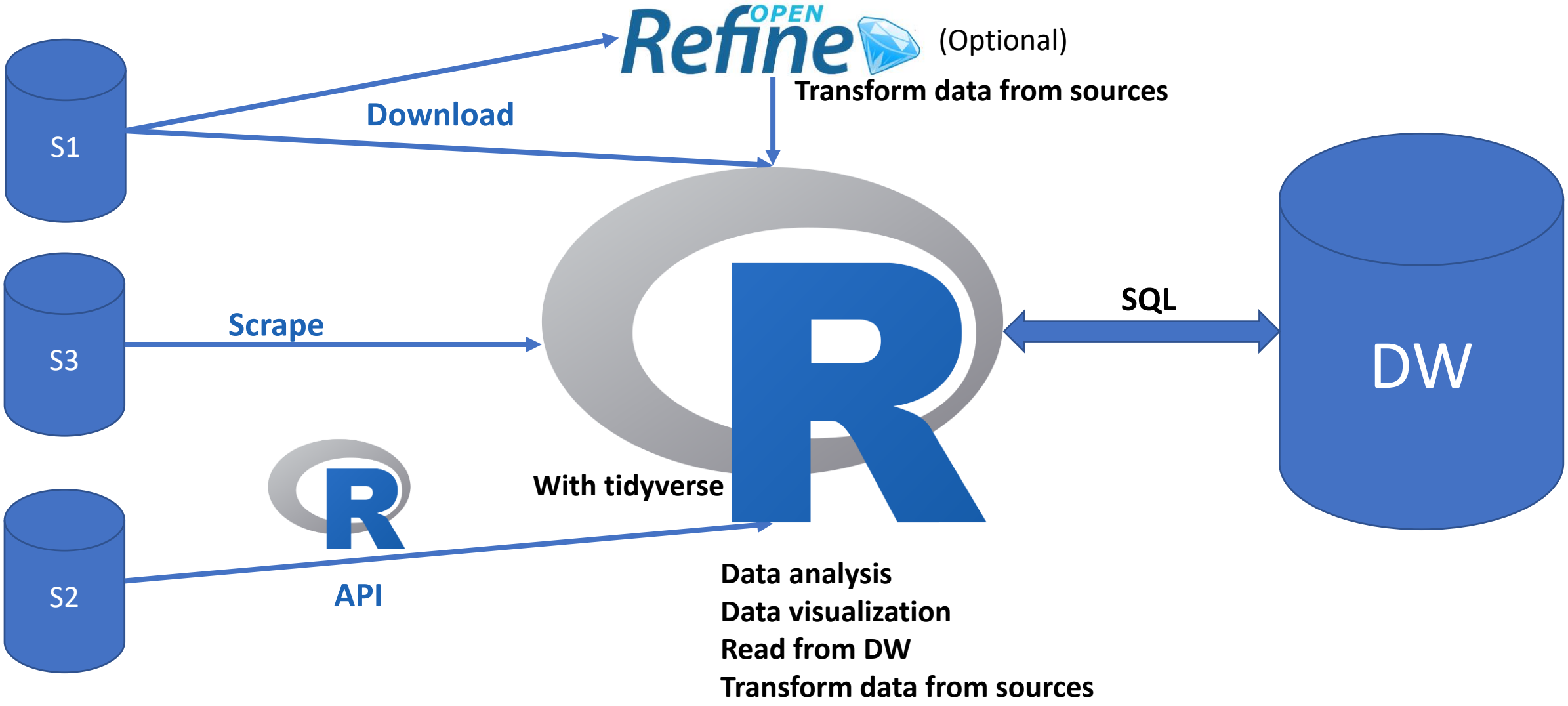
Data warehousing

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.

Pages are for last edition of the book

Sharda, R., Delen, D., Turban, E., & King, D. (2017). *Business intelligence : A managerial approach, global edition*. Retrieved from https://ebookcentral.proquest.com

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	

Star Schema: The Complete Reference Chapter 1 - Analytic Databases and Dimensional Design
McGraw-Hill/Osborne © 2010

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

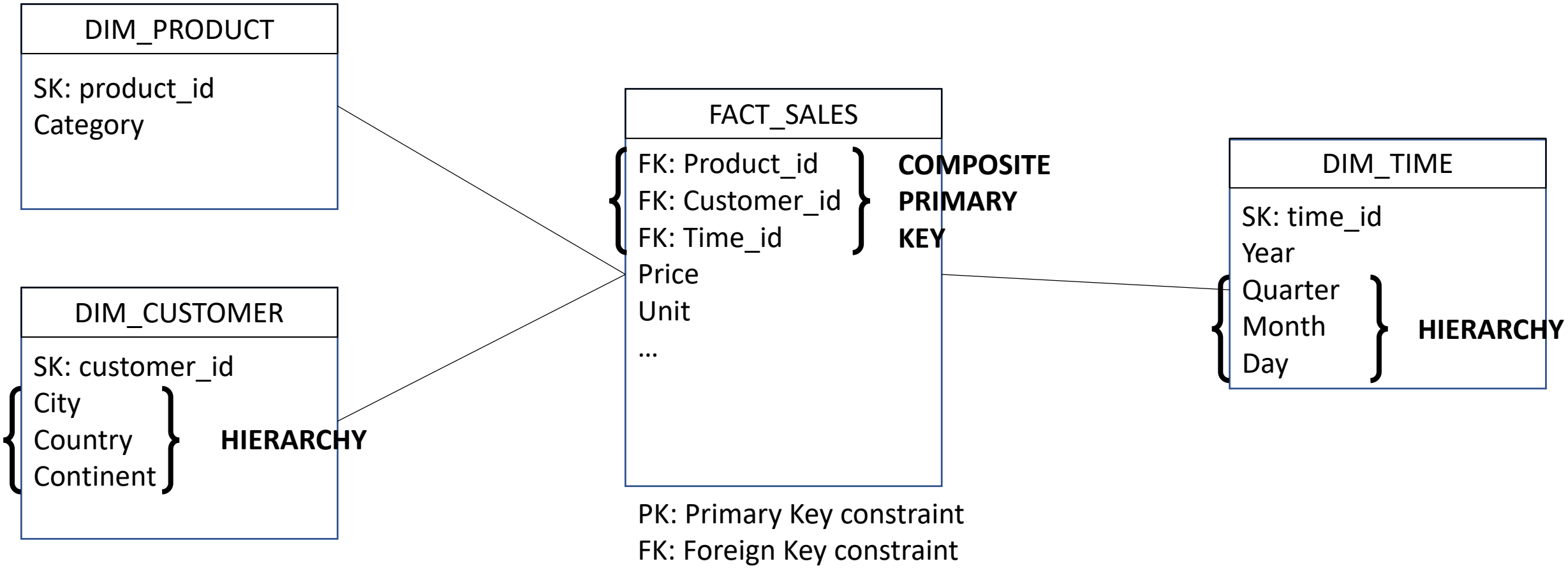
Structuring measures

- 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).

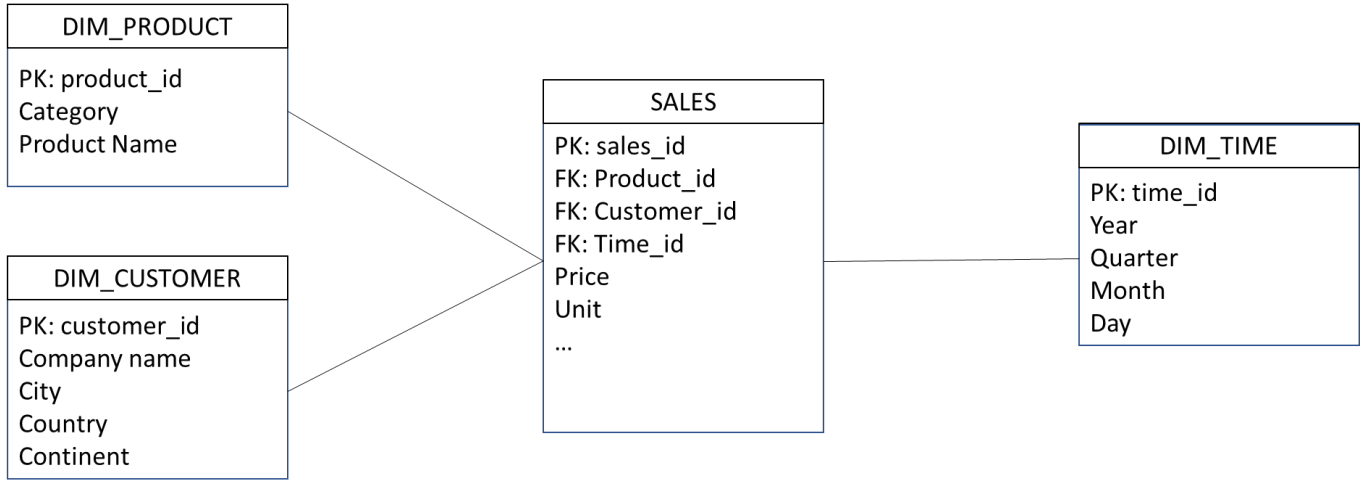
STAR schema (ROLAP)



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

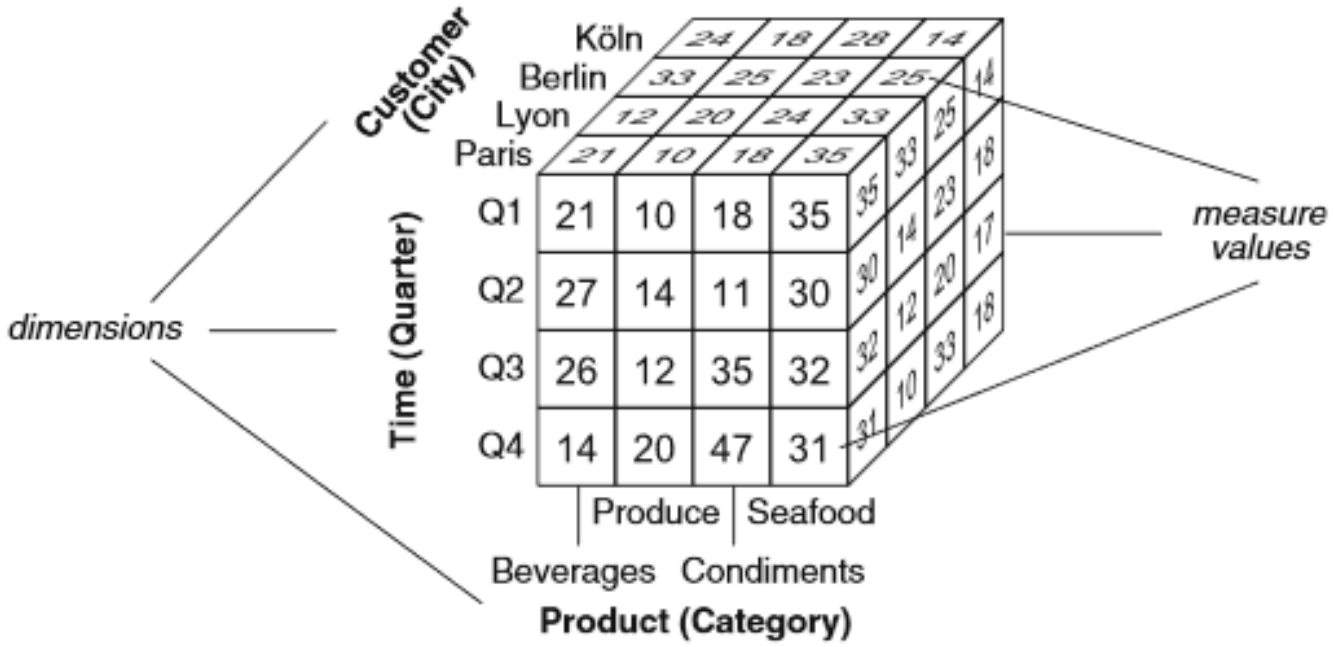


Fig. 3.1 A three-dimensional cube for sales data with dimensions Product, Time, and Customer, and a measure Quantity

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

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

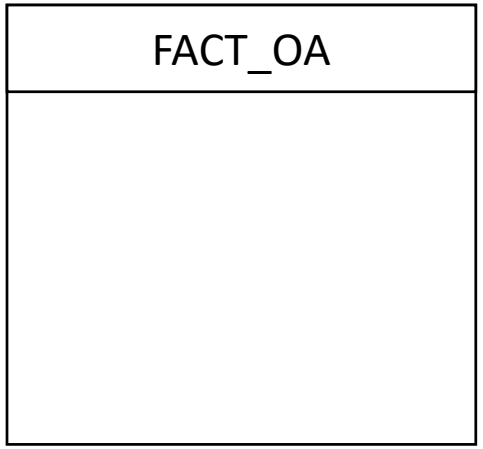
Model DW

- 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

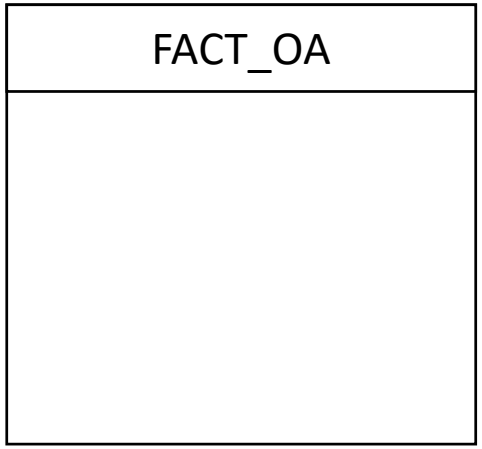
Strategy

- 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



Facts and Dimensions

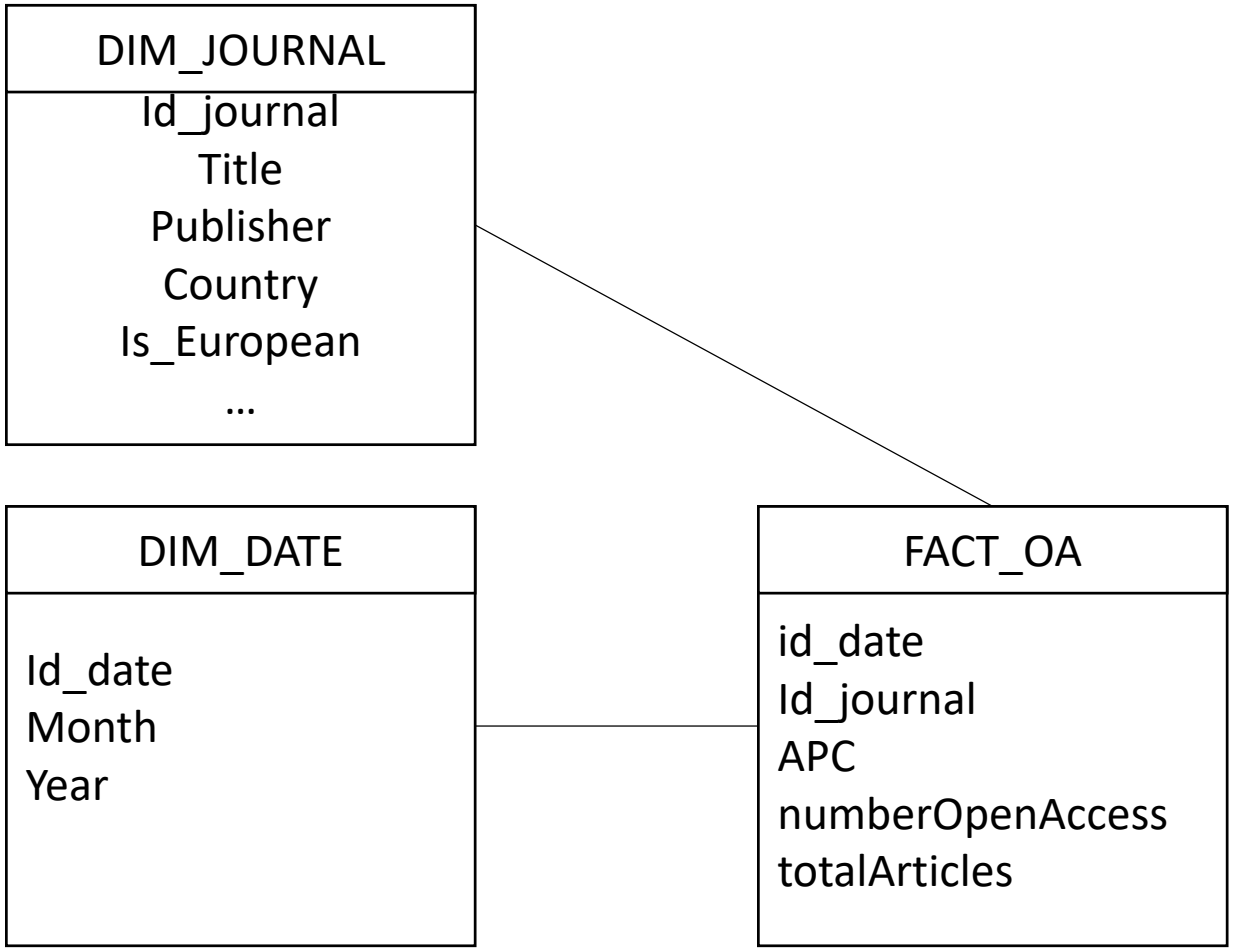
- *Open access* to scientific articles {for all citizens} **in** *Europe*



- *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

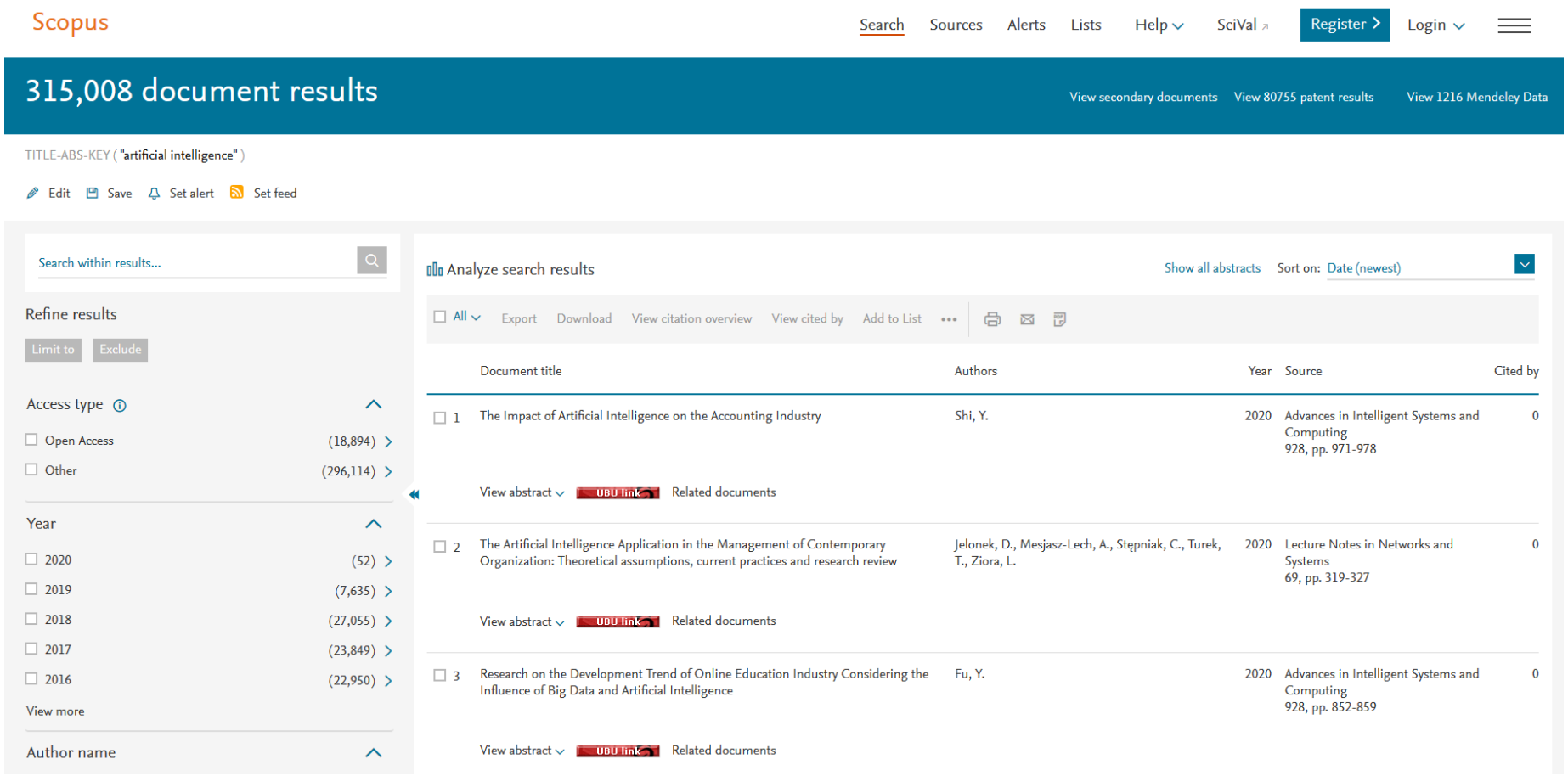
- Extract sources
  - Downloads
  - API
  - Scrape (not the preferred option)

Help, too much data

- Remember your strategy
  - Select what you need
  - Justify criteria

  - Example : scopus

Scopus query



ETL \ Transform

- The source can be messy
- Unify attributes and values before you load them in Postgres

- http://openrefine.org/

Load with RMarkdown

```sql
```{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

```
### Load to PostgreSQL

```{r}

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 = AvgScienPop)

#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

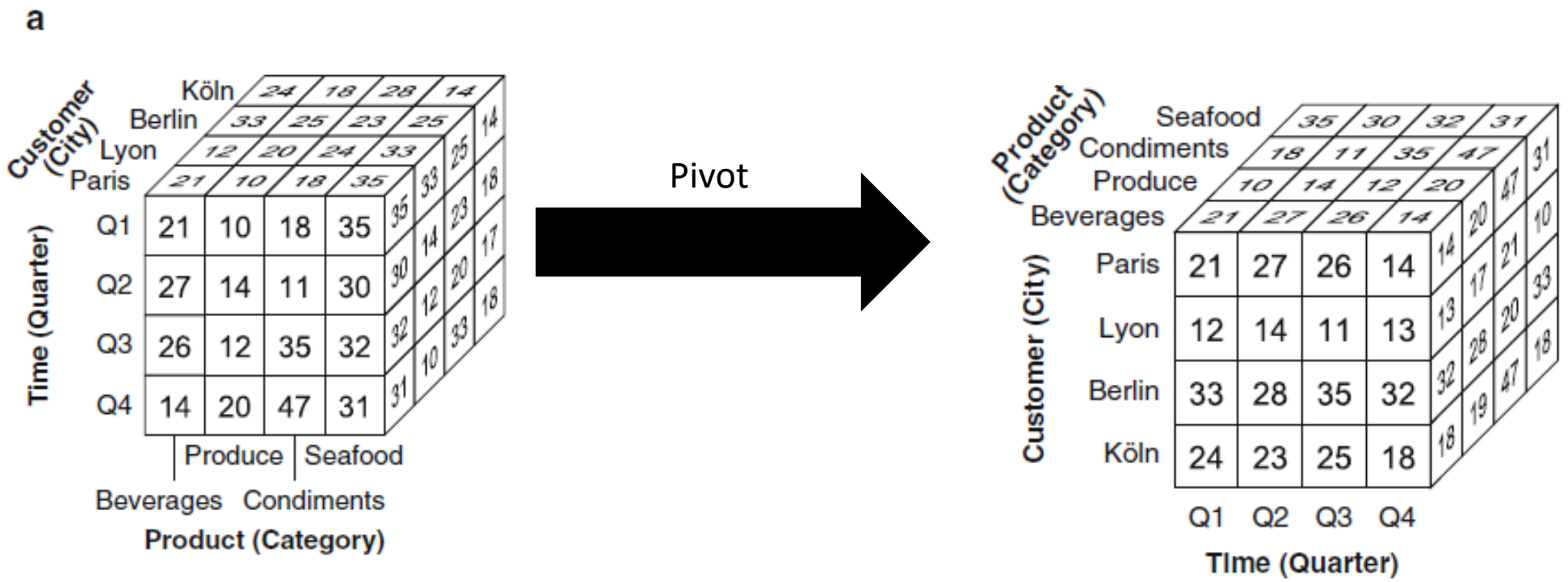


Figure from: Vaisman, A., & Zimányi, E. (2014). *Data warehouse systems*. Springer, Heidelberg.

Operations

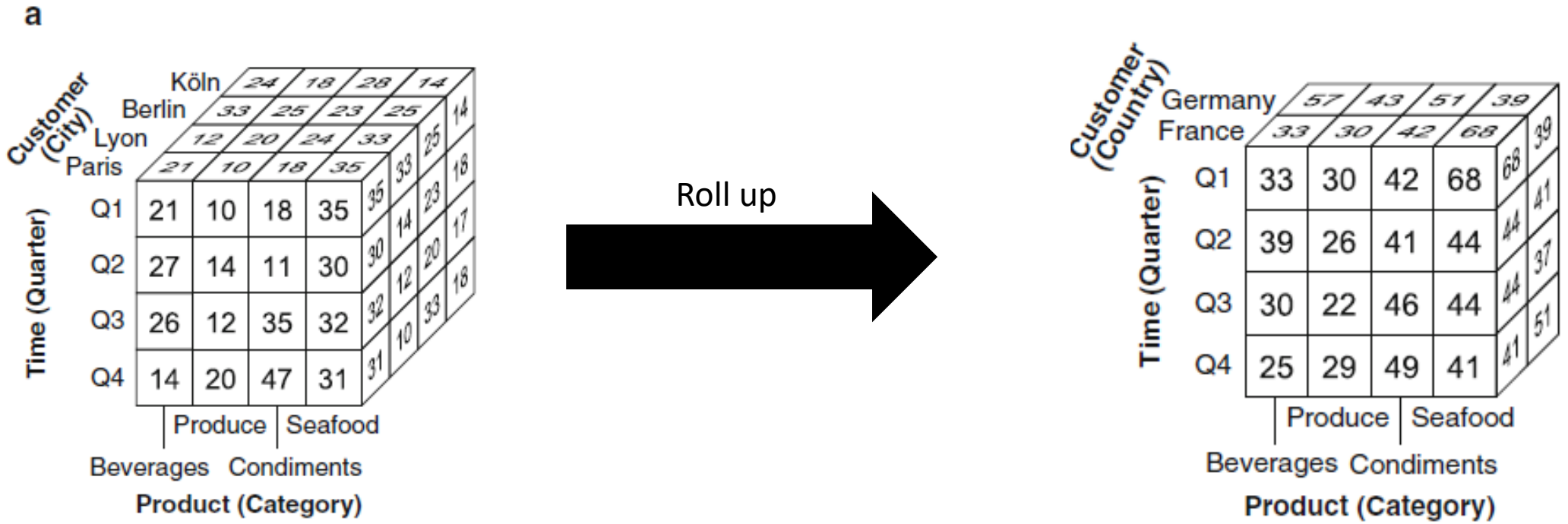


Figure from: Vaisman, A., & Zimányi, E. (2014). *Data warehouse systems*. Springer, Heidelberg.

Operations

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

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