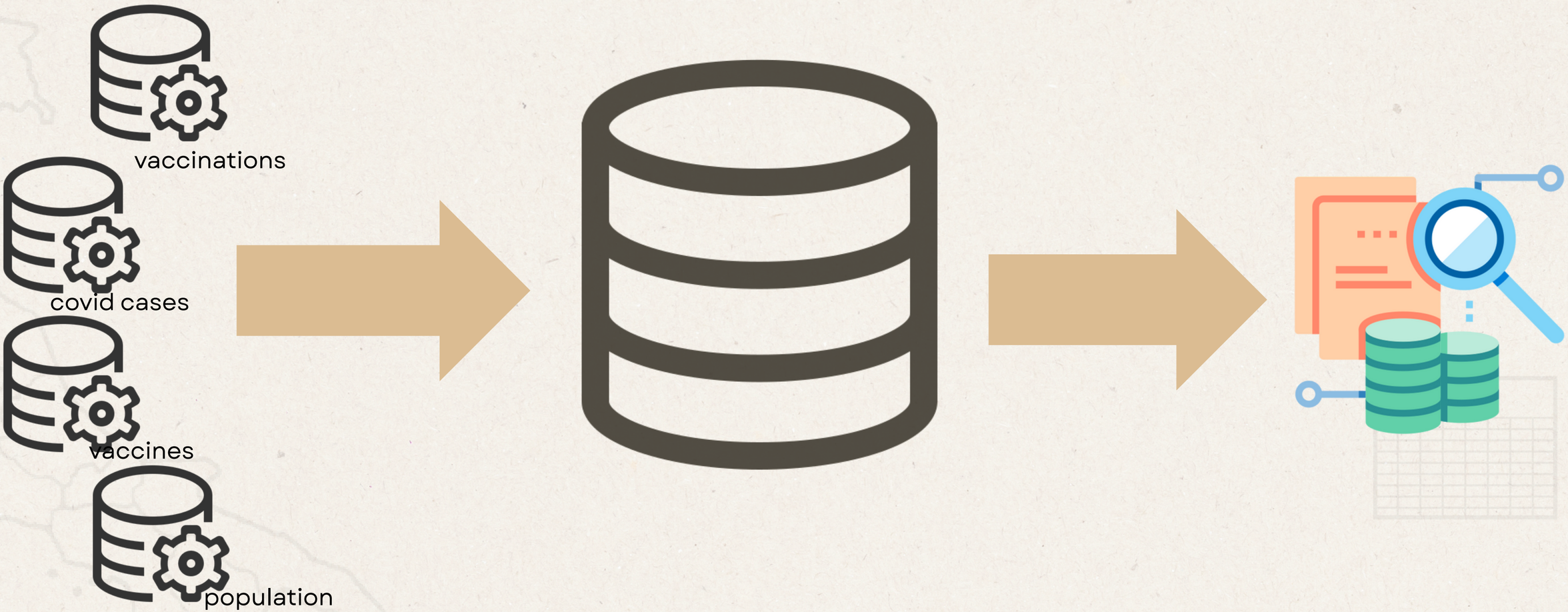
SAPIENZA
UNIVERSITÀ DI ROMA

Data Analysis of COVID-19 Vaccination and Case Data in Italy

A DataWarehouse Design

AGNESE MARIA CAPPARELLI 1794326
SAMUELE CERVO 1883147

Project Goal



What we have done

Project Timeline

Data gathering

Retrieving some data sources online.

First ETL Processes

Cleansing and filtering the original data

Conceptual Modeling

DFM Schema

Logical Modeling

Snowflake Schema

ROLAP Queries

SQL Queries

Analysis with Tableau

What we have done

Project Timeline

Data gathering

Retrieving some data sources online.

First ETL Processes

Cleansing and filtering the original data

Conceptual Modeling

DFM Schema

Logical Modeling

Snowflake Schema

ROLAP Queries

SQL Queries

Analysis with Tableau

What we have done

Project Timeline

Data gathering

Retrieving some data sources online.

First ETL Processes

Cleansing and filtering the original data

Conceptual Modeling

DFM Schema

Logical Modeling

Snowflake Schema

ROLAP Queries

SQL Queries

Analysis with Tableau

What we have done

Project Timeline

Data gathering

Retrieving some data sources online.

First ETL Processes

Cleansing and filtering the original data

Conceptual Modeling

DFM Schema

Logical Modeling

Snowflake Schema

ROLAP Queries

SQL Queries

Analysis with Tableau

What we have done

Project Timeline

Data gathering

Retrieving some data sources online.

First ETL Processes

Cleansing and filtering the original data

Conceptual Modeling

DFM Schema

Logical Modeling

Snowflake Schema

ROLAP Queries

SQL Queries

Analysis with Tableau

What we have done

Project Timeline

Data gathering

Retrieving some data sources online.

First ETL Processes

Cleansing and filtering the original data

Conceptual Modeling

DFM Schema

Logical Modeling

Snowflake Schema

ROLAP Queries

SQL Queries

Analysis with Tableau

What we have done

Project Timeline

Data gathering

Retrieving some data sources online.

First ETL Processes

Cleansing and filtering the original data

Conceptual Modeling

DFM Schema

Logical Modeling

Snowflake Schema

ROLAP Queries

SQL Queries

Analysis with Tableau

Data Gathering

Different data sources

covid_cases
DPC
---
informations about covid cases, hospedalizations, new positives, swabs and other data during covid period per regions and date.

regions
ISTAT
---
that contains regions name, region_code and region_abbreviation

population20XX
ISTAT
---
information about population per regions, divided by males, females in the year 20XX

vaccinations
OpenData
---
that contains information about vaccination during the covid period divided by age range, males, females, vaccine name, region and date

vaccines
GitHub OpenData
---
informations about every single vaccine, like supplier, name and technology

ETL Process

covid_cases (Dataset from DPC)

Cleansing and integration in phases of the original data were made through Python scripts.
Below there are detailed operations performed on each of the csv files.

- Renamed coloumns
- Replaced date from time code to yy/mm/dd
- Uniformed region_code field with a 2 characters string
- Filtered unnecessary coloumns
- Uniformed empty fields with 0 and null values.

date	country_code	**region_code**	latitude	logitude	**hospitalized**	intensitive_care	total_hospitalized	home_isolation	total_positives
variation_total_positive	**new_positives**	recovered	**deceased**	suspected_cases	screaning_cases	total_cases	intensitive_care_admission	positive_molecular_test	positive_rapid_antigen_test
molecular_swabs	rapid_antigen_swabs	nuts1_code	nuts2_code	notes	test_notes	case_notes	**swabs**	tested_cases	

ETL Process

population20XX (6 dataset from 2019 to 2024 from ISTAT)

Cleansing and integration in phases of the original data were made through Python scripts.
Below there are detailed operations performed on each of the csv files.

- Renamed coloumns
- Union the 6 dataset in one dataset grouped by year and region

year	region	sex	population

ETL Process

covid_cases (Dataset from DPC)

Cleansing and integration in phases of the original data were made through Python scripts.
Below there are detailed operations performed on each of the csv files.

- Renamed coloumns

region_code	region_name	region_abbr

ETL Process

vaccinations (Dataset from GitHub OpenData)

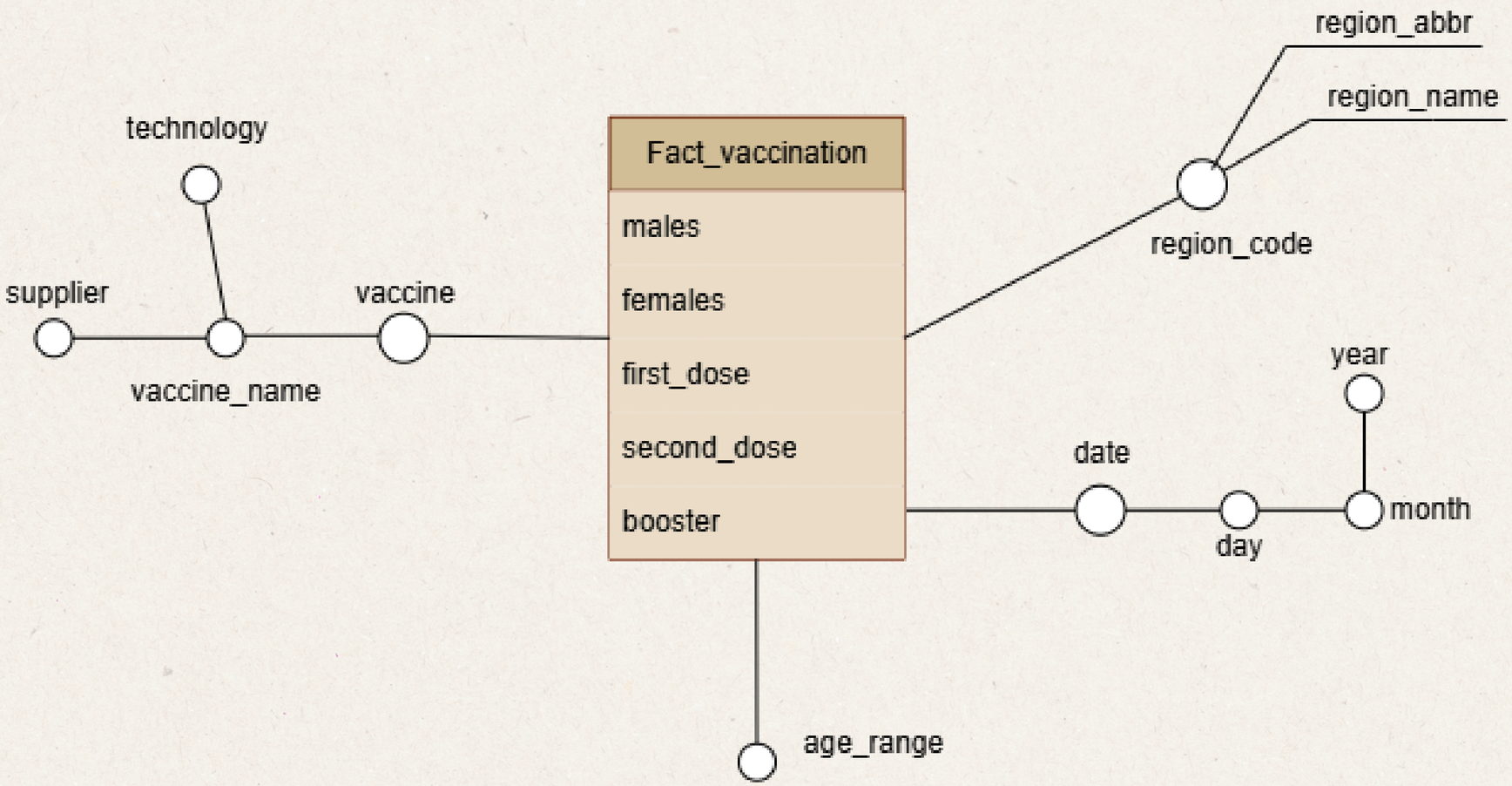
Cleansing and integration in phases of the original data were made through Python scripts.
Below there are detailed operations performed on each of the csv files.

- Renamed coloumns
- Replaced date from time code to yy/mm/dd
- Uniformed region_code field with a 2 characters string
- Filtered unnecessary coloumns
- Uniformed empty fields with 0 and null values.

data	**fornitore**	area	**eta**	**m**	**f**	**d1**	**d2**
dpi	**db1**	**db2**	**db3**	N1	N2	**ISTAT**	reg

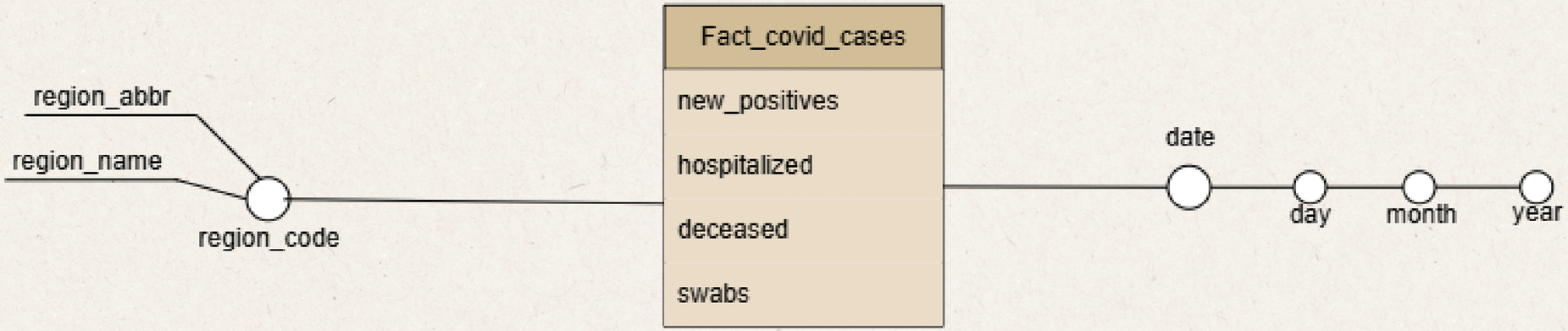
Conceptual modeling

DFM Schema



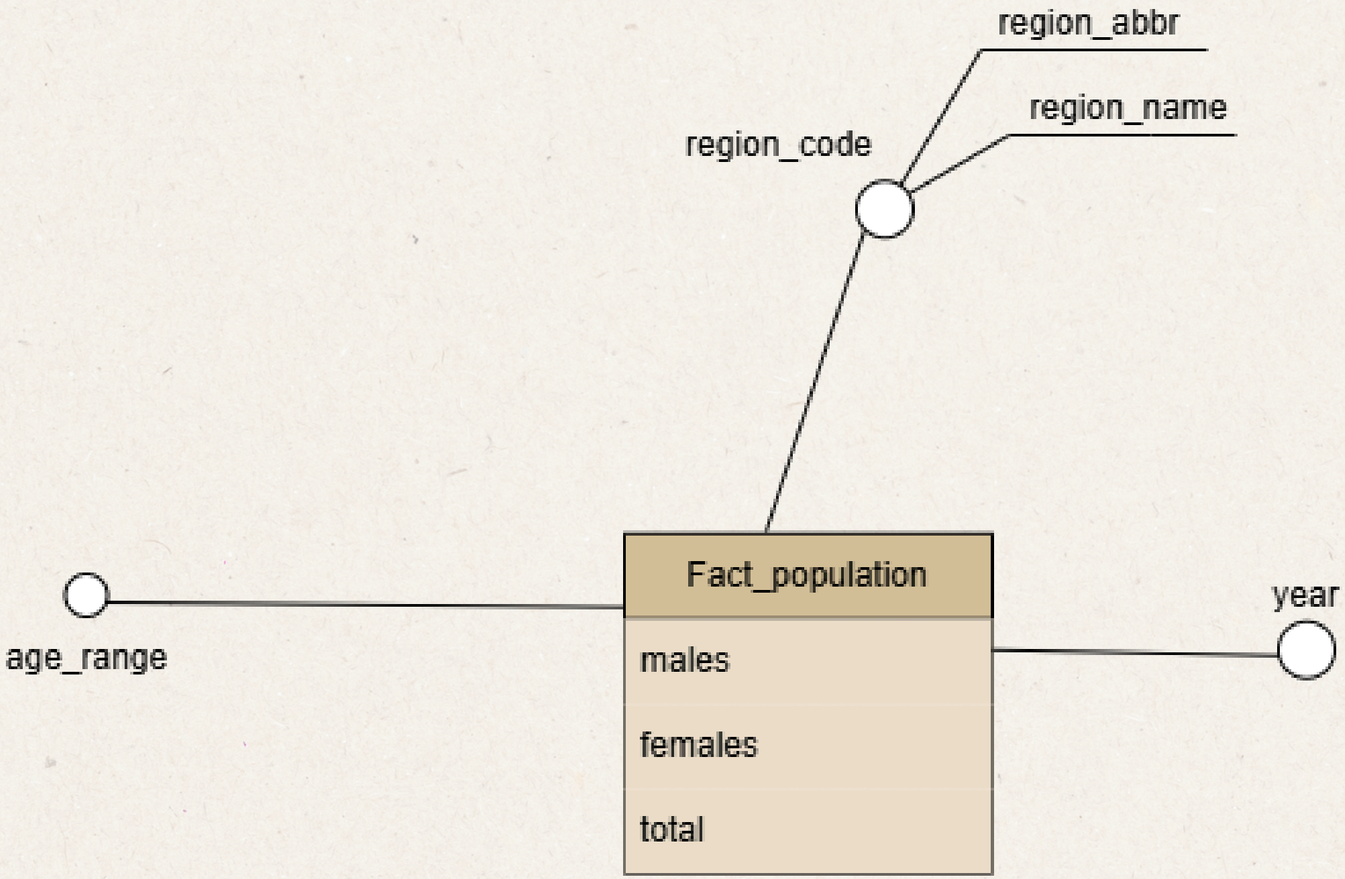
Conceptual modeling

DFM Schema



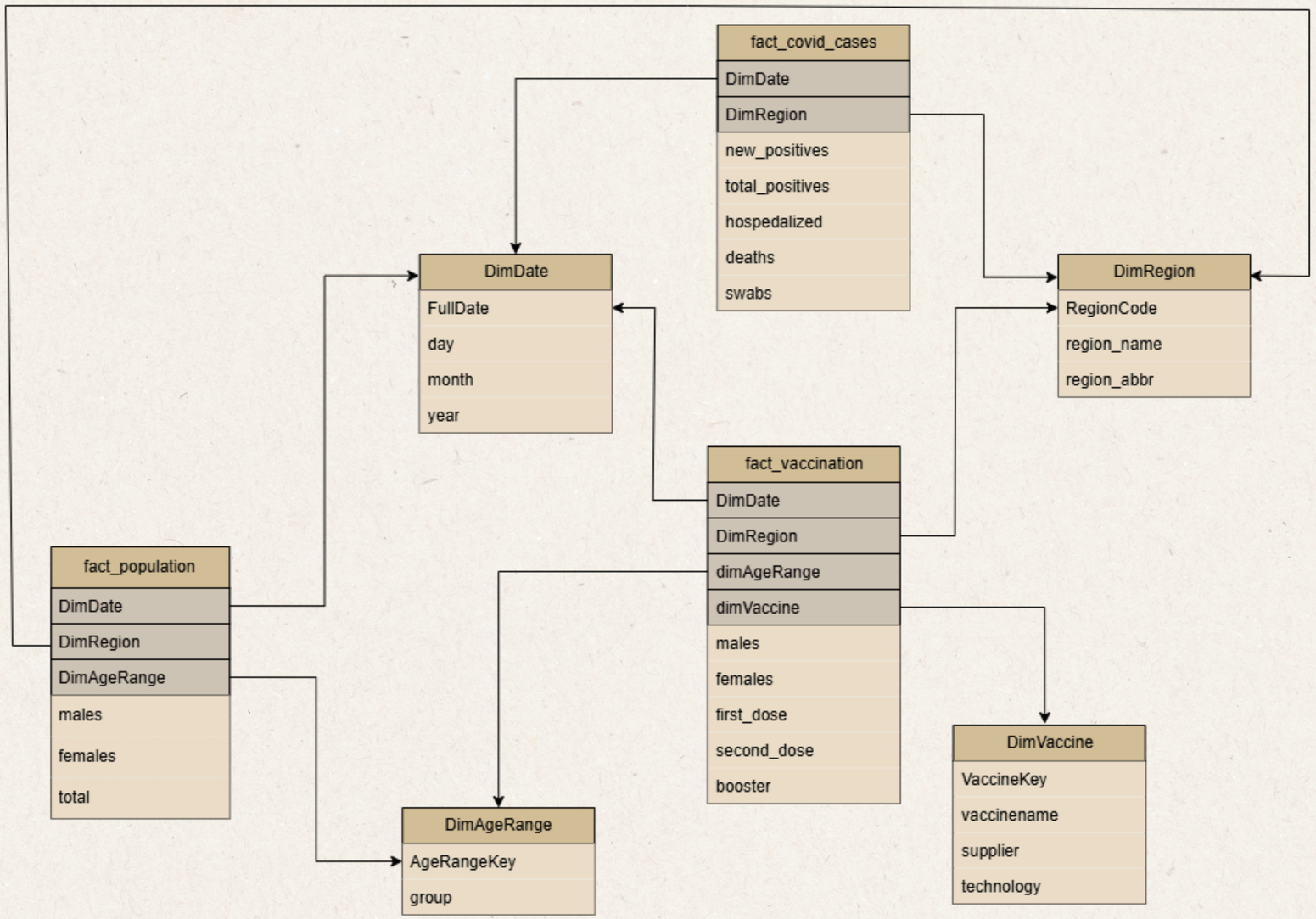
Conceptual modeling

DFM Schema



Logical modeling

SNOWFLAKE schema



Views for ROLAP Queries

```
CREATE OR REPLACE VIEW dw_layer.vw_vaccination_cumulative AS
SELECT
 date_key,
 region_key,
 age_key,
 SUM(first_dose) OVER (PARTITION BY region_key, age_key ORDER BY date_key)
AS cumulative_first_dose
FROM dw_layer.fact_vaccination;
```

Cumulative first doses

```
CREATE OR REPLACE VIEW dw_layer.vw_covid_monthly AS
SELECT
    d.year,d.month,fc.region_key,
    SUM(fc.new_positives) AS monthly_new_cases,
    SUM(fc.hospitalized) AS monthly_hospitalized
FROM dw_layer.fact_covid_cases fc
JOIN dw_layer.dimDate d
    ON fc.date_key = d.date_key
GROUP BY d.year, d.month, fc.region_key;
```

Monthly covid cases

Rolap Queries

Monthly Aggregation of New COVID-19 Cases and Cumulative First Dose Vaccinations by Region

```
SELECT r.region_name,
       d.year,
       d.month,
       AVG(c.new_positives) AS avg_new_positives,
       MAX(vc.cumulative_first_Dose) AS vaccination
FROM dw_layer.fact_covid_cases c
JOIN dw_layer.vw_vaccination_cumulative vc
    ON c.region_key = vc.region_key AND c.date_key = vc.date_key
JOIN dw_layer.dimRegion r
    ON c.region_key = r.regionKey
JOIN dw_layer.dimDate d
    ON c.date_key = d.dateKey
GROUP BY r.region_name, d.year, d.month
ORDER BY d.year, d.month, r.region_name;
```

Rolap Queries

Yearly Vaccination Coverage by Region and Age Group with Population-Based Vaccination Rate

```
SELECT
    d.year,r.region_name,a.age_group,
    MAX(vc.cumulative_first_dose) AS total_vaccinations,
    fp.total_count AS population,
    ROUND(MAX(vc.cumulative_first_dose)::NUMERIC / fp.total_count * 100, 2) AS vaccination_rate_percent
FROM dw_layer.vw_vaccination_cumulative vc
JOIN dw_layer.dimDate d
    ON vc.date_key = d.date_key
JOIN dw_layer.fact_population fp
    ON vc.region_key = fp.region_key
   AND vc.age_key = fp.age_key
   AND d.year = fp.year
JOIN dw_layer.dimRegion r
    ON vc.region_key = r.regionKey
JOIN dw_layer.dimAgeRange a
    ON vc.age_key = a.age_key
GROUP BY d.year,r.region_name,a.age_group,fp.total_count
ORDER BY d.year,r.region_name,a.age_group;
```

Rolap Queries

Vaccination Totals and Coverage Percentages by Region and Age Group for 2024

```
SELECT
    r.region_name,
    v.age_key AS age_group,
    SUM(fv.first_dose) AS total_first_dose,
    SUM(fv.second_dose) AS total_second_dose,
    SUM(fv.booster) AS total_booster,
    fp.total_count AS population_2024,
    ROUND(SUM(fv.first_dose)::numeric / NULLIF(fp.total_count,0) * 100, 2) AS pct_first_dose,
    ROUND(SUM(fv.second_dose)::numeric / NULLIF(fp.total_count,0) * 100, 2) AS pct_second_dose,
    ROUND(SUM(fv.booster)::numeric / NULLIF(fp.total_count,0) * 100, 2) AS pct_booster
FROM dw_layer.fact_vaccination fv
JOIN dw_layer.dimRegion r ON fv.region_key = r.regionKey
JOIN dw_layer.dimAgeRange v ON fv.age_key = v.age_key
LEFT JOIN dw_layer.fact_population fp
       ON fv.region_key = fp.region_key
      AND fv.age_key = fp.age_key
      AND fp.year = 2024
GROUP BY r.region_name, v.age_key, fp.total_count
ORDER BY r.region_name, v.age_key;
```

Rolap Queries

Annual COVID-19 Deaths and Population by Region (2020–2022)

```
SELECT
    d.year, r.region_name, p.population,
    MAX(fc.deceased) AS total_deaths
FROM dw_layer.fact_covid_cases fc
JOIN dw_layer.dimDate d
    ON fc.date_key = d.datekey
JOIN dw_layer.dimRegion r
    ON fc.region_key = r.regionKey
JOIN (
    SELECT year, region_key, SUM(total_count) AS population
    FROM dw_layer.fact_population
    GROUP BY year, region_key
) p
    ON p.region_key = fc.region_key
   AND p.year = d.year
WHERE d.year BETWEEN 2020 AND 2022
GROUP BY d.year, r.region_name,  p.population
ORDER BY d.year, r.region_name;
```

Rolap Queries

Annual cumulative first-dose vaccinations by vaccine technology with population and coverage percentage

```
SELECT
  d.year, v.technology AS vaccine_technology,
  SUM(fv.first_dose) AS total_vaccinations_cumulative, p.total_population,
  ROUND(SUM(fv.first_dose)::numeric / NULLIF(p.total_population,0), 4) * 100 AS
pct_vaccinated
FROM dw_layer.fact_vaccination fv
JOIN dw_layer.dimVaccine v
  ON fv.vaccine_key = v.vaccine_name
JOIN dw_layer.dimDate d
  ON fv.date_key = d.dateKey
JOIN (
  SELECT year, SUM(total_count) AS total_population
  FROM dw_layer.fact_population
  GROUP BY year
) p
  ON d.year = p.year
GROUP BY d.year, v.technology, p.total_population
ORDER BY d.year, v.technology;
```

Rolap Queries

Monthly COVID-19 Cases, Hospitalizations, and Vaccination Coverage by Age Group in Selected Regions (2021-2023)

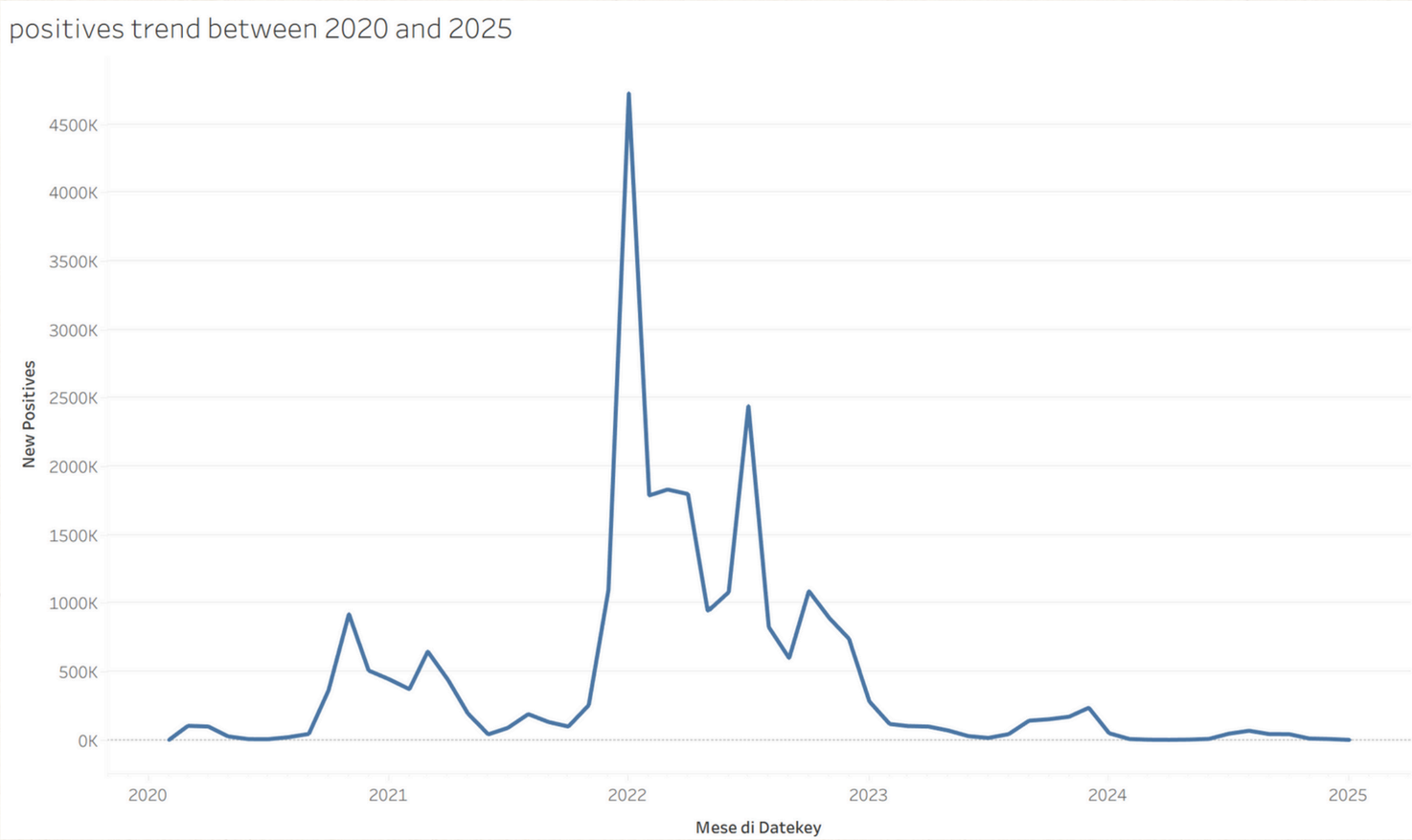
```
SELECT
    d.year, d.month, r.region_name, a.age_group,
    MAX(vc.cumulative_first_dose) AS total_vaccinations_cumulative,
    cm.monthly_new_cases AS total_new_positives, cm.monthly_hospitalized AS total_hospitalized,
    fp.total_count AS population,
    ROUND(cm.monthly_new_cases::NUMERIC / NULLIF(fp.total_count, 0) * 100000, 2) AS incidence_per_100k,
    ROUND(MAX(vc.cumulative_first_dose)::NUMERIC / NULLIF(fp.total_count, 0) * 100, 2) AS pct_vaccinated
FROM dw_layer.vw_vaccination_cumulative vc
JOIN dw_layer.dimDate d
    ON vc.date_key = d.datekey
JOIN dw_layer.vw_covid_monthly cm
    ON cm.region_key = vc.region_key AND cm.year = d.year AND cm.month = d.month
JOIN dw_layer.dimRegion r
    ON vc.region_key = r.regionKey
JOIN dw_layer.dimAgeRange a
    ON vc.age_key = a.age_key
JOIN dw_layer.fact_population fp
    ON vc.region_key = fp.region_key AND vc.age_key = fp.age_key AND d.year = fp.year
WHERE r.region_name IN ('Lombardia', 'Lazio') AND d.year BETWEEN 2021 AND 2023
GROUP BY d.year, d.month, r.region_name, a.age_group, cm.monthly_new_cases, cm.monthly_hospitalized, fp.total_count
ORDER BY d.year, d.month, r.region_name, a.age_group;
```

Tableau Analysis

doses trend per region

	Date Key (Fact Vaccination)					
	First Dose		Second Dose		Booster	
Region Name	2021	2022	2021	2022	2021	2022
Abruzzo	1.004.488	45.584	951.751	63.414	415.995	460.701
Basilicata	432.671	21.954	406.394	26.315	169.357	202.060
Calabria	1.407.728	104.449	1.296.119	137.953	508.288	631.917
Campania	4.282.757	254.265	3.965.771	360.770	1.706.662	1.816.754
Emilia-Romagna	3.516.240	143.827	3.291.597	223.720	1.503.380	1.619.072
Friuli Venezia Giulia	913.682	40.502	849.638	58.912	386.815	429.279
Lazio	4.624.738	217.291	4.164.694	315.728	2.072.743	1.949.853
Liguria	1.214.948	58.377	1.136.051	83.826	481.993	541.448
Lombardia	8.096.484	329.931	7.510.467	463.244	3.625.712	3.719.474
Marche	1.147.177	41.734	1.083.096	62.418	497.093	495.954
Molise	241.021	10.792	224.751	15.075	113.076	94.035
Piemonte	3.336.692	150.723	3.114.668	214.246	1.442.170	1.594.836
Puglia	3.172.100	157.801	2.937.443	234.990	1.468.007	1.304.921
Sardegna	1.283.894	58.253	1.220.449	78.334	496.358	573.402
Sicilia	3.605.048	245.739	3.315.696	334.623	1.137.531	1.709.437
Toscana	3.007.216	125.533	2.821.490	191.370	1.234.990	1.377.629
Trentino-Alto Adige	806.366	34.472	741.989	61.297	371.141	310.764
Umbria	694.252	28.936	646.122	46.232	304.042	302.037
Valle d'Aosta	92.937	3.812	87.031	6.182	44.273	41.535
Veneto	3.715.622	141.818	3.440.371	247.236	1.728.556	1.564.189

Tableau Analysis

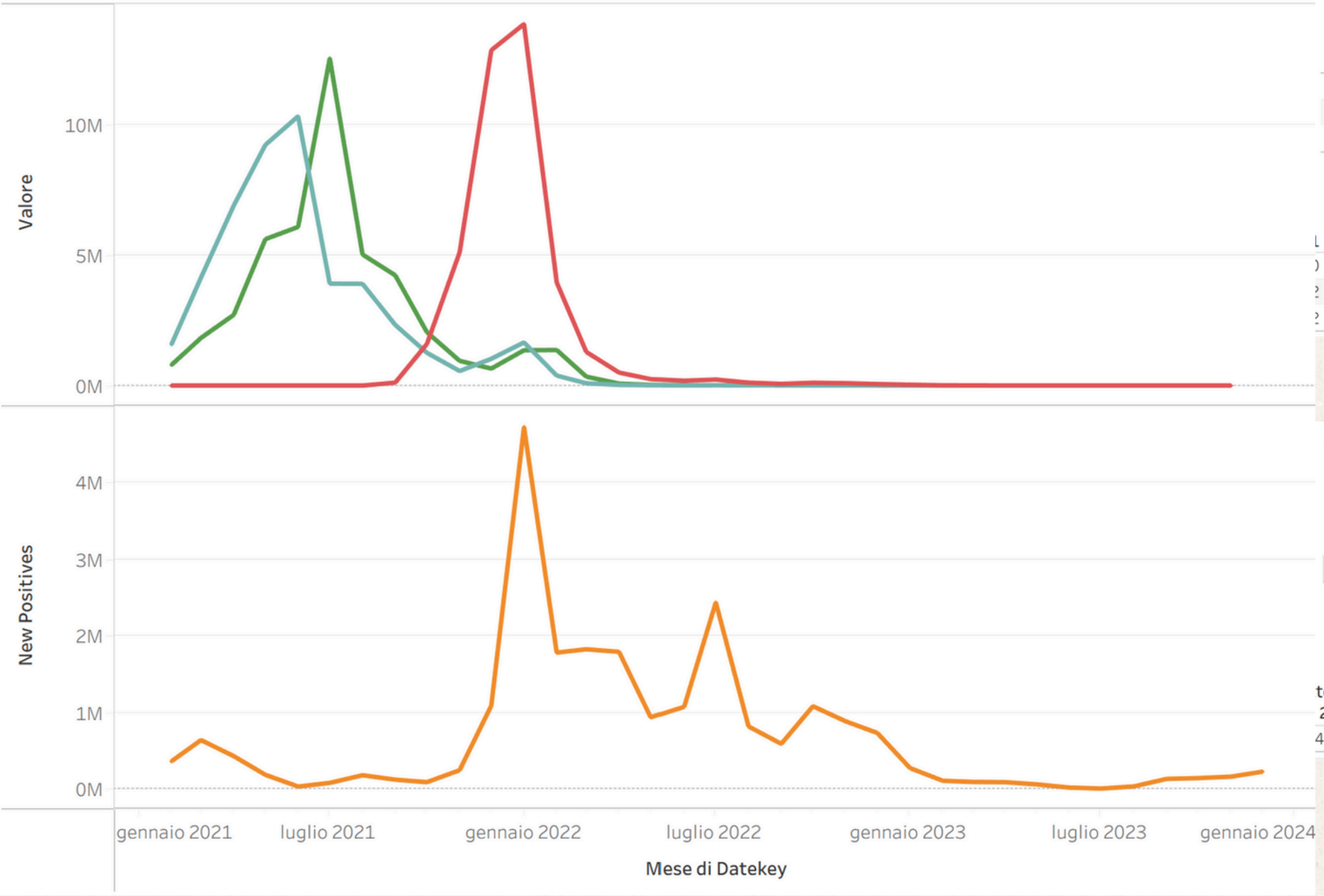


tab posotives trend

Mese di Date Key																				
febbraio 2020	marzo 2020	aprile 2020	maggio 2020	giugno 2020	luglio 2020	agosto 2020	settembre 2020	ottobre 2020	novembre 2020	dicembre 2020	gennaio 2021	febbraio 2021	marzo 2021	aprile 2021	maggio 2021	giugno 2021	luglio 2021	agosto 2021	settembre 2021	ottobre 2021
1.120	104.664	99.671	27.556	7.584	7.006	21.702	45.623	364.607	922.206	508.914	445.585	372.503	648.200	439.446	195.273	42.614	90.174	190.152	132.220	99.684

Tableau Analysis

comparison between vaccinations and new positives from 2021 to 2024



vaccination tab

Mese di Datekey

	dicembre 2020	gennaio 2021	febbraio 2021	marzo 2021	aprile 2021	maggio 2021	giugno 2021	luglio 2021
Booster	0	0	0	0	0	0	0	0
First Dose	40.428	1.354.195	1.579.627	4.150.186	6.914.053	9.249.568	10.345.464	3.916.862
Second Dose	0	644.284	792.371	1.830.980	2.716.141	5.625.297	6.094.540	12.573.862

agosto 2021	settembre 2021	ottobre 2021	novembre 2021	dicembre 2021	gennaio 2022	febbraio 2022	marzo 2022	a
0	112.480	1.598.398	5.114.177	12.883.127	13.888.970	3.960.496	1.292.599	
3.909.584	2.330.866	1.258.437	558.065	1.029.154	1.658.432	383.098	87.476	
5.036.285	4.236.632	2.052.874	949.834	652.488	1.358.185	1.364.895	343.367	

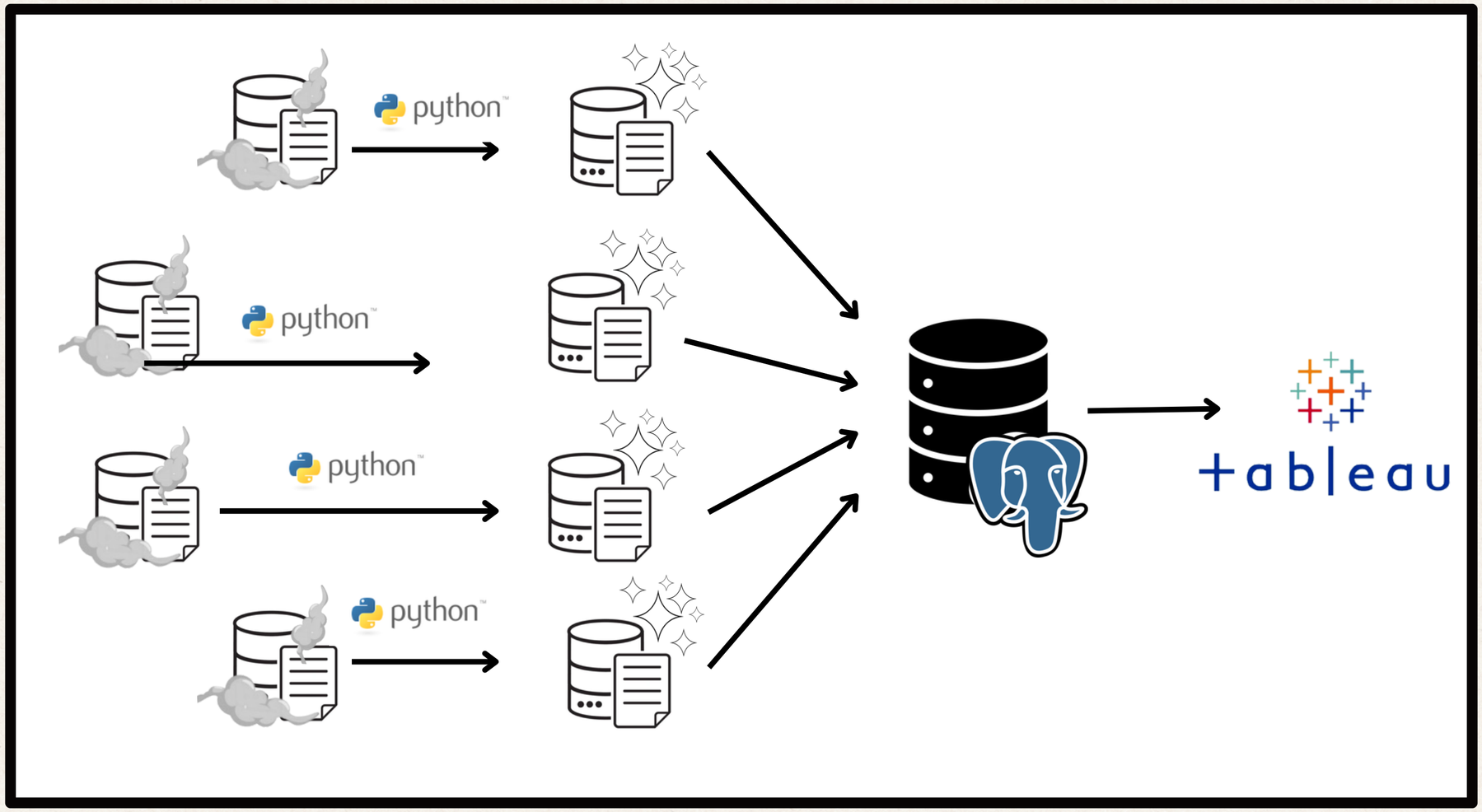
tab posotives trend

Mese di Date Key

febbraio 2020	marzo 2020	aprile 2020	maggio 2020	giugno 2020	luglio 2020	agosto 2020	settembre 2020	ottobr 202
1.120	104.664	99.671	27.556	7.584	7.006	21.702	45.623	364.60

tobre 2020	novembre 2020	dicembre 2020	gennaio 2021	febbraio 2021	marzo 2021	aprile 2021	maggio 2021	giugno 2021	
4.607	922.206	508.914	445.585	372.503	648.200	439.446	195.273	42.614	

Conclusions



Thanks for your attention!