DATA MODELING & POWER QUERY

1 Column Formatting and Minor Transformations

Upon importing the datasets into Power BI, several column types were **misinterpreted** (e.g., excess_mortality being detected as text instead of numeric). These were adjusted **directly** in **Power BI** for simplicity, including:

- **Setting proper data types** for numeric metrics
- **Formatting percentage-based columns** (e.g., handwashing_facilities)
- **Limiting decimals to enhance readability** in charts and cards

Additionally, in the **VAERS_Symptoms table**, all **symptom columns** (SYMPTOM1 to SYMPTOM5) were **unpivoted**. This transformation was essential to analyze symptom **frequency regardless** of **position**, allowing for dynamic symptom **ranking** and **correlation** with outcomes like death or hospitalization.

2 Dimensional Modeling: Star Schemas

The project uses a **dual-star schema model** built around two **fact tables**:

OWID_Compact → contains **global** pandemic data by **country** and **day**.

VAERS_Data → contains **individual**-level **reports** of **adverse events**.

These are **surrounded** by smaller **dimension tables** (calendar, symptoms, vaccines) and filtered using slicers and disconnected fields.

To enable **time-based analysis**, a **custom calendar table** (Dim_Calendar) was **created** with the following DAX code (written in Spanhis):

```
Dim_Calendar = ADDCOLUMNS (
    CALENDAR ( DATE(2020,1,1), DATE(2025,12,31) ),
```

```
    "FechaSK", FORMAT([Date], "YYYYMMDD"),
    "#Año", YEAR([Date]),
    "#Trimestre", QUARTER([Date]),
    "#Mes", MONTH([Date]),
    "#Día", DAY([Date]),
    "Trimestre", "T" & FORMAT([Date], "Q"),
    "Mes", FORMAT([Date], "MMMM"),
    "MesCorto", FORMAT([Date], "MMM"),
    "#DíaSemana", WEEKDAY([Date], 2),
    "#SemanaAño", WEEKNUM([Date], 2),
    "CierreSemana", [Date] + 7 - WEEKDAY([Date], 2),
    "Día", FORMAT([Date], "DDDD"),
    "DíaCorto", FORMAT([Date], "DDD"),
    "AñoTrimestre", FORMAT([Date], "YYYY") & "/T" & FORMAT([Date], "Q"),
    "Año#Mes", FORMAT([Date], "YYYY/MM"),
    "AñoMesCorto", FORMAT([Date], "YYYY/mmm"),
    "InicioMes", EOMONTH([Date], -1) + 1,
    "FinMes", EOMONTH([Date], 0)
)
```

This calendar provides **granular date hierarchies** (year, month, quarter, weekday, etc.) and was marked as the **official Date Table** using **[Date]** as the primary field.

Relationship Decisions:

Key relationships in the model include:

1) **Dim_Calendar[Date] → OWID_Compact[date] (1:*), active**
2) **Dim_Calendar[Date] → VAERS_Data[ONSET_DATE] (1:*), active**
3) **OWID_Compact[country] →** OWID_Vacc[country], **inactive relationship** (used selectively with USERELATIONSHIP)
4) **VAERS_Data[VAERS_ID] →** VAERS_Symptoms[VAERS_ID] and VAERS_VAX[VAERS_ID] **(1:*), active**

All surrogate key fields used for joining were **hidden** in the report view, ensuring **clean** and **centralized filter control.** Only one instance

of each key (e.g., VAERS_ID, country, date) is visible to users, maintaining **best practices** in **report design**.

Note: For VAERS_Data, the date was linked to ONSET_DATE (symptom onset), instead of VAX_DATE, since the analysis **focused** on **when symptoms began**. However, both **dates** are **available** for **deeper time comparisons** in the dashboards.

Additional Notes:

1) **20 total CSV files** were used, spread across **5 structural formats**.
2) Many countries (notably in Africa and Asia) were **excluded** due to **low data quality** or **missing values**, especially regarding **deaths** and **vaccinations**. For example, WHO estimates that **some African countries underreport COVID deaths by a factor of 8–10**.
3) **Relationships** and **filtering** logic were built to preserve **data granularity**, while enabling **normalized comparisons** across time, geography, and clinical severity.

A visual diagram of the final model and connections is available here: