Portada

Nombre del curso: Curso Básico de Introducción al Big Data

Título de la tarea: Proyecto Final

Nombre del docente: Prof. Ing. Richard D. Jiménez-R.

Nombre de los estudiantes: Orlando Semidei, Néstor Morel, María Luján Ibarra Benegas

Identificador de grupo: 9

Fecha de entrega: 09 de mayo de 2025

Análisis de la distribución de vacunas contra el COVID-19

1. Introducción

Este proyecto tiene como objetivo analizar la distribución y cobertura de las vacunas aplicadas a la población, utilizando un conjunto de datos público con aproximadamente 1.000.000 de registros de personas vacunadas. A través de un análisis detallado, se busca responder preguntas claves como: ¿Cuáles son los establecimientos con mayor actividad de vacunación? ¿Qué tipos de vacunas fueron administradas con mayor frecuencia? ¿Cómo ha evolucionado la tasa de vacunación a lo largo del tiempo, tanto a nivel diario como mensual? Además, se analizarán métricas clave como la cantidad de personas vacunadas por dosis, la distribución de vacunas por establecimiento, y el número de personas que han recibido refuerzos. Este análisis proporcionará una visión completa de la situación de la vacunación y permitirá identificar tendencias y áreas de oportunidad para mejorar la cobertura

2. Arquitectura del Sistema y Modelo de Datos

Para el almacenamiento y manejo de los datos, se utilizó **PostgreSQL** como sistema de gestión de base de datos. Se creó una base de datos denominada **bigdatafinal**, dentro de la cual se configuró un esquema llamado **final** para organizar las tablas y procesos relacionados.

Inicialmente, se diseñó la tabla **vacunados_temporal** con los siguientes campos: nombre, apellido, cédula, establecimiento, fecha de aplicación, dosis, descripción de la vacuna y la fecha de última actualización. A continuación, se presenta el código SQL para la creación de esta tabla:

```
CREATE TABLE vacunados_temporal (
    nombre TEXT,
    apellido TEXT,
    establecimiento TEXT,
    fecha_aplicacion DATE,
```

```
    cedula VARCHAR(25),
    dosis INTEGER,
    descripcion_vacuna TEXT,
    actualizado_al TIMESTAMP
);
```

Luego, se creó una segunda tabla denominada **vacunados_datos_limpios** que tiene una estructura similar, pero con la finalidad de almacenar los datos ya filtrados y transformados para su posterior análisis.

```
CREATE TABLE vacunados_datos_limpios (
    nombre TEXT,
    apellido TEXT,
    establecimiento TEXT,
    fecha_aplicacion DATE,
    cedula VARCHAR(25),
    dosis INTEGER,
    descripcion_vacuna TEXT,
    actualizado_al TIMESTAMP
);
```

A medida que avanzaba el proceso, se identificó una oportunidad para optimizar la estructura de las tablas y mejorar la normalización de los datos. En este sentido, se decidió dividir los datos en tablas adicionales para representar mejor las relaciones, creando las tablas **establecimientos** y **vacunas**, que se vinculan a través de claves foráneas en la tabla principal. Las nuevas tablas fueron diseñadas de la siguiente manera:

```
CREATE table establecimientos (
    id SERIAL PRIMARY KEY,
    descripcion TEXT UNIQUE
);
```

```
CREATE TABLE vacunas (
    id SERIAL PRIMARY KEY,
    descripcion TEXT UNIQUE
);
```

Finalmente, se diseñó la tabla principal **vacunados_oficial**, que incluye relaciones con las tablas de **establecimientos** y **vacunas**, reemplazando las descripciones de estos campos por sus respectivos **id** (clave foránea). Esto mejora la eficiencia y flexibilidad de la base de datos, optimizando la integridad de los datos.

```
CREATE TABLE vacunados_oficial (
    id SERIAL PRIMARY KEY,
    nombre TEXT,
    apellido TEXT,
    establecimiento_id INTEGER REFERENCES establecimientos(id),
    fecha_aplicacion DATE,
    cedula VARCHAR(25),
    dosis INTEGER,
    vacuna_id INTEGER REFERENCES vacunas(id),
    actualizado_al TIMESTAMP
);
```

3. Proceso ETL (Extracción, Transformación y Carga)

Se utilizó Pentaho Kettle para el proceso ETL, extrayendo datos desde un archivo CSV. Se aplicaron transformaciones como limpieza de datos, normalización de nombres de vacunas, conversión de fechas y eliminación de duplicados. Finalmente, los datos se cargaron en las tablas de referencia y en la tabla principal 'vacunados_oficial'.

4. Análisis y KPIs

A continuación, se presentan las principales consultas SQL utilizadas para el análisis de los datos de vacunación, junto con su propósito:

4.1. Total, de personas vacunadas (cédulas únicas)

```
SELECT COUNT(DISTINCT cedula) AS total_personas_vacunadas
FROM final.vacunados_oficial; -- Resultado: 980.697
```

Esta consulta calcula el total de personas únicas que recibieron al menos una dosis de vacuna.

4.2. Personas con más de una dosis aplicada

```
SELECT cedula,
    COUNT(*) AS total_dosis_registradas
FROM final.vacunados_oficial
GROUP BY cedula
HAVING COUNT(*) > 1
ORDER BY total_dosis_registradas DESC;
```

Identifica personas que figuran con más de una dosis registrada.

4.3. Detalle por cédula, nombre y apellido para personas con múltiples dosis

```
SELECT cedula,
    nombre,
    apellido,
    COUNT(*) AS total_dosis_registradas
FROM final.vacunados_oficial
GROUP BY cedula, nombre, apellido
HAVING COUNT(*) > 1
ORDER BY total_dosis_registradas DESC;
```

Complementa la consulta anterior con información personal para un análisis más detallado.

4.4. Dosis aplicadas por tipo de vacuna

```
SELECT vac.descripcion AS tipo_de_vacuna,
    COUNT(*) AS total_dosis_aplicadas
FROM final.vacunados_oficial vo
JOIN final.vacunas vac ON vo.vacuna_id = vac.id
GROUP BY vac.descripcion
ORDER BY total_dosis_aplicadas DESC;
```

Muestra la distribución de dosis aplicadas según el tipo de vacuna.

4.5. Dosis aplicadas por establecimiento de salud

```
SELECT est.descripcion AS establecimiento,
    COUNT(*) AS total_dosis_aplicadas
FROM final.vacunados_oficial vo
JOIN final.establecimientos est ON vo.establecimiento_id = est.id
GROUP BY est.descripcion
ORDER BY total_dosis_aplicadas DESC;
```

Indica qué establecimientos aplicaron más dosis.

4.6. Dosis aplicadas por mes y año

```
SELECT
  EXTRACT(YEAR FROM fecha_aplicacion) AS anho,
  EXTRACT(MONTH FROM fecha_aplicacion) AS mes_num,
  TO_CHAR(fecha_aplicacion, 'TMMonth') AS mes_nombre,
  COUNT(*) AS total_dosis_aplicadas
FROM final.vacunados_oficial
```

```
GROUP BY anho, mes_num, mes_nombre
ORDER BY anho, mes_num;
```

Analiza la evolución mensual de la vacunación para detectar campañas intensivas.

4.7. Distribución de cantidad de dosis por persona

```
SELECT total_dosis_registradas, COUNT(*) AS cantidad_personas
FROM (
    SELECT cedula, COUNT(*) AS total_dosis_registradas
    FROM final.vacunados_oficial
    GROUP BY cedula
) sub
GROUP BY total_dosis_registradas
ORDER BY total_dosis_registradas;
```

Muestra cuántas personas recibieron determinada cantidad de dosis (1, 2, 3, 4.).

5. Conclusiones

- Se vacunaron **980.697 personas únicas** según los registros de cédula.
- Alrededor de **200 personas recibieron más de una dosis**, lo que indica una baja adherencia a esquemas completos en el conjunto de datos analizado.
- Las vacunas más aplicadas, en orden descendente, fueron:

    1. **Pfizer** – 578.044 dosis aplicadas
    2. **AstraZeneca** – 261.643
    3. **Sputnik V** – 124.989
    4. **Moderna** – 23.553
    5. **Coronavac** – 7.488
    6. **Hayat Vax** – 2.302
    7. **Covaxin** – 1.390

8. **Sinopharm** – 392

- Se observaron **picos mensuales de vacunación**, posiblemente vinculados a campañas intensivas o aumentos en la disponibilidad de vacunas.
- Algunos establecimientos fueron responsables de un volumen significativamente mayor de dosis aplicadas, lo que podría reflejar su capacidad operativa o su ubicación estratégica.
- Este análisis podría complementarse con datos geográficos y de población para estimar mejor la **tasa de cobertura vacunal**.
- Se recomienda avanzar hacia el desarrollo de modelos predictivos que ayuden a planificar **campañas de vacunación más eficientes** en el futuro.