UNIVERSIDAD TÉCNICA PARTICULAR DE LOJA

PROYECTO INTEGRADOR

DOCENTE:

ING. DANILO JARAMILLO

POR:

BYRON ALVAREZ

ERICK TOLEDO

FABIAN CAÑAR

FECHA:

25 DE JULIO DEL 2025

ASIGNATURA: PROGRAMACIÓN AVANZADA

Contenido

Introducción

En el campo de la salud pública, el análisis de datos epidemiológicos se ha convertido en una herramienta crucial para comprender la evolución de enfermedades y tomar decisiones basadas en evidencia. La emergencia sanitaria provocada por la propagación del virus SARS-CoV-2 puso a prueba los sistemas de vigilancia epidemiológica de los países, exigiendo una respuesta ágil, coordinada y sustentada en información oportuna. En este contexto, el presente estudio se basa en un conjunto de datos extraídos del Sistema Integral de Vigilancia Epidemiológica (SIVE) del Ministerio de Salud Pública del Ecuador, el cual recoge los casos notificados de infección respiratoria aguda por COVID-19 desde febrero de 2020 hasta el 4 de marzo de 2022.

Este reporte, estructurado en formato CSV, contiene registros individuales que representan casos epidemiológicos, siendo posible que una misma persona aparezca en múltiples entradas del sistema. A través de consultas estructuradas sobre este conjunto de datos, se busca explorar diversos aspectos del impacto de la pandemia en la población ecuatoriana, como la distribución de casos por fechas, regiones, edades y clasificaciones clínicas.

El objetivo principal de esta actividad es desarrollar habilidades en el manejo, consulta y análisis de datos mediante herramientas computacionales, permitiendo interpretar tendencias y patrones relevantes en la evolución de la pandemia. Al mismo tiempo, se pretende evidenciar la importancia del procesamiento adecuado de información epidemiológica para fortalecer la respuesta del sistema de salud y promover políticas públicas informadas. Así, este ejercicio no solo tiene un enfoque técnico, sino también un valor práctico para futuras acciones de prevención y control de enfermedades.

Repositorio GitHub:

Repositorio: https://github.com/By-Alvarez06/PROG-AV-ProyectoFinalZeppelin.git

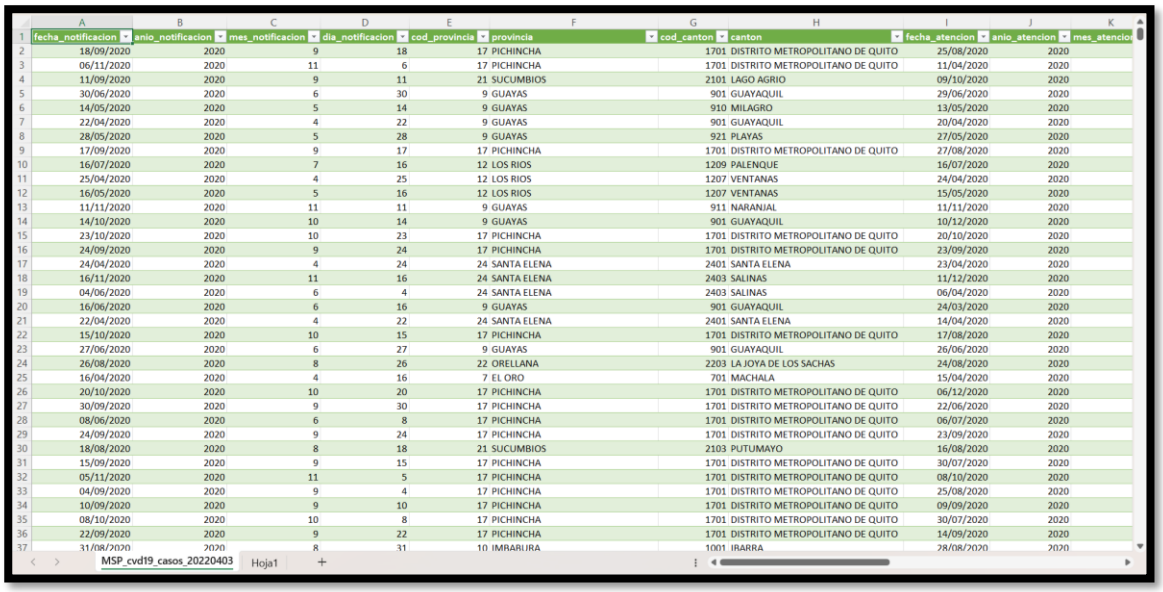
Datos Base:

Los datos base contienen información detallada sobre la infección respiratoria aguda causada por el coronavirus SARS-CoV-2 durante el periodo de febrero de 2020 al 04 de marzo de 2022. A continuación, se presenta una descripción de las variables presentes en los datos:

Nombre del campo	**Descripción del campo**
fecha_notificacion	Se refiere a la fecha en que se realizó el ingreso al aplicativo informático de Vigilancia para COVID-19 (2)
anio_notificacion	Año de la fecha de ingreso de la notificación
mes_notificacion	Mes de la fecha de ingreso de la notificación
dia_notificacion	Día de la fecha de ingreso de la notificación

cod_provincia	Código de provincia de ubicación del establecimiento de salud donde se atendió el paciente
provincia	Nombre de provincia de ubicación del establecimiento de salud donde se atendió el paciente
cod_canton	Código de cantón de ubicación del establecimiento de salud donde se atendió el paciente
canton	Nombre del cantón de ubicación del establecimiento de salud donde se atendió el paciente
fecha_atencion	Corresponde a la fecha en que, el paciente sospechoso de COVID-19, es atendido en un establecimiento de salud de la Red pública integral de salud y red complementaria. (2)
anio_atencion	Año de la fecha atención del paciente en el establecimiento de salud
mes_atencion	Mes de la fecha de atención del paciente en el establecimiento de salud
dia_atencion	Día de la fecha de atención del paciente en el establecimiento de salud
cod_provincia_residencia	Código de provincia de residencia habitual del paciente
provincia_residencia	Nombre de provincia de residencia habitual del paciente
cod_canton_residencia	Código de cantón de residencia habitual del paciente
canton_residencia	Nombre de cantón de residencia habitual del paciente
sexo_paciente	Sexo del paciente
edad_paciente	Edad declarada por el paciente
tipo_edad	Tipo de métrica usada para la edad
condicion_final	Conocido también como condición de egreso. Estado final del paciente al salir del sistema de vigilancia.
fecha_defuncion	Se refiere a la fecha en que fallece el paciente, confirmado o probable para COVID-19 (2)
anio_defuncion	Año tomado de la fecha de fallecimiento del paciente
mes_defuncion	Mes tomado de la fecha de fallecimiento del paciente
dia_defuncion	Día tomado de la fecha de fallecimiento del paciente
clasificacion_final	Estado final del evento sujeto a investigación (depende del resultado de examen de laboratorio o criterio clínico/Epidemiológico) (2)
ae_se_notificacion	En salud, se suele utilizar la división del año en semanas epidemiológicas que tienen 7 días, empezando siempre un Domingo y finalizando un Sábado; la

	primera semana del año, debe tener a los menos 4 días de enero, sino se vuelve la semana 52 o 53 del año



fecha_notificacion	anio_notificacion	mes_notificacion	dia_notificacion	cod_provincia	provincia	cod_canton	canton	fecha_atencion	anio_atencion	mes_atencio
18/09/2020	2020	9	18	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	25/08/2020	2020	
06/11/2020	2020	11	6	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	11/04/2020	2020	
11/09/2020	2020	9	11	21	SUCUMBIOS	2101	LAGO AGRIO	09/10/2020	2020	
30/06/2020	2020	6	30	9	GUAYAS	901	GUAYAQUIL	29/06/2020	2020	
14/05/2020	2020	5	14	9	GUAYAS	910	MILAGRO	13/05/2020	2020	
22/04/2020	2020	4	22	9	GUAYAS	901	GUAYAQUIL	20/04/2020	2020	
28/05/2020	2020	5	28	9	GUAYAS	921	PLAYAS	27/05/2020	2020	
17/09/2020	2020	9	17	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	27/08/2020	2020	
16/07/2020	2020	7	16	12	LOS RIOS	1209	PALENQUE	16/07/2020	2020	
25/04/2020	2020	4	25	12	LOS RIOS	1207	VENTANAS	24/04/2020	2020	
16/05/2020	2020	5	16	12	LOS RIOS	1207	VENTANAS	15/05/2020	2020	
11/11/2020	2020	11	11	9	GUAYAS	911	NARANJAL	11/11/2020	2020	
14/10/2020	2020	10	14	9	GUAYAS	901	GUAYAQUIL	10/12/2020	2020	
23/10/2020	2020	10	23	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	20/10/2020	2020	
24/09/2020	2020	9	24	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	23/09/2020	2020	
24/04/2020	2020	4	24	24	SANTA ELENA	2401	SANTA ELENA	23/04/2020	2020	
16/11/2020	2020	11	16	24	SANTA ELENA	2403	SALINAS	11/12/2020	2020	
04/06/2020	2020	6	4	24	SANTA ELENA	2403	SALINAS	06/04/2020	2020	
16/06/2020	2020	6	16	9	GUAYAS	901	GUAYAQUIL	24/03/2020	2020	
22/04/2020	2020	4	22	24	SANTA ELENA	2401	SANTA ELENA	14/04/2020	2020	
15/10/2020	2020	10	15	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	17/08/2020	2020	
27/06/2020	2020	6	27	9	GUAYAS	901	GUAYAQUIL	26/06/2020	2020	
26/08/2020	2020	8	26	22	ORELLANA	2203	LA JOYA DE LOS SACHAS	24/08/2020	2020	
16/04/2020	2020	4	16	7	EL ORO	701	MACHALA	15/04/2020	2020	
20/10/2020	2020	10	20	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	06/12/2020	2020	
30/09/2020	2020	9	30	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	22/06/2020	2020	
08/06/2020	2020	6	8	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	06/07/2020	2020	
24/09/2020	2020	9	24	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	23/09/2020	2020	
18/08/2020	2020	8	18	21	SUCUMBIOS	2103	PUTUMAYO	16/08/2020	2020	
15/09/2020	2020	9	15	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	30/07/2020	2020	
05/11/2020	2020	11	5	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	08/10/2020	2020	
04/09/2020	2020	9	4	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	25/08/2020	2020	
10/09/2020	2020	9	10	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	09/09/2020	2020	
08/10/2020	2020	10	8	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	30/07/2020	2020	
22/09/2020	2020	9	22	17	PICHINCHA	1701	DISTRITO METROPOLITANO DE QUITO	14/09/2020	2020	
31/08/2020	2020	8	31	10	IMBABURA	1001	IBARRA	28/08/2020	2020	

MSP_cvd19_casos_20220403 | Hoja1

canton_residencia	edad_paciente	tipo_edad	sexo_paciente	condicion_final	fecha_defuncion	anio_defuncion	mes_defuncion	dia_defuncion	clasificacion_final	ae_se_notificacion
DISTRITO METROPOLITANO DE QUITO	24	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202038
DISTRITO METROPOLITANO DE QUITO	39	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202045
LAGO AGRIO	25	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202037
GUAYAQUIL	26	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202027
MILAGRO	52	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202020
GUAYAQUIL	65	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202017
GUAYAQUIL	57	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202022
OTAVALO	45	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202038
PALENQUE	32	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202029
VENTANAS	61	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202017
VENTANAS	61	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202020
NARANJAL	56	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202046
GUAYAQUIL	56	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202042
DISTRITO METROPOLITANO DE QUITO	27	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202043
DISTRITO METROPOLITANO DE QUITO	38	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202039
SANTA ELENA	52	ANIOS	HOMBRE	MUERTO	2020-04-25	2020	4	25	CONFIRMADO	202017
LA LIBERTAD	52	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202047
LA LIBERTAD	52	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202023
GUAYAQUIL	51	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	PROBABLE	202025
SANTA ELENA	74	ANIOS	HOMBRE	MUERTO	2020-04-14	2020	4	14	CONFIRMADO	202017
DISTRITO METROPOLITANO DE QUITO	28	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202042
GUAYAQUIL	38	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202026
LA JOYA DE LOS SACHAS	25	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202035
MACHALA	40	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202016
DISTRITO METROPOLITANO DE QUITO	38	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202043
DISTRITO METROPOLITANO DE QUITO	43	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202040
DISTRITO METROPOLITANO DE QUITO	27	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202024
DISTRITO METROPOLITANO DE QUITO	77	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202039
PUTUMAYO	23	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202034
DISTRITO METROPOLITANO DE QUITO	30	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202038
DISTRITO METROPOLITANO DE QUITO	56	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202045
DISTRITO METROPOLITANO DE QUITO	36	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202036
DISTRITO METROPOLITANO DE QUITO	80	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202037
DISTRITO METROPOLITANO DE QUITO	17	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	PROBABLE	202041
DISTRITO METROPOLITANO DE QUITO	31	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	DESCARTADO	202039
SAN MIGUEL DE URCUQUI	76	ANIOS	HOMBRE	VIVO	NA	NA	NA	NA	CONFIRMADO	202036

MSP_cvd19_casos_20220403 | Hoja1

Estos datos, junto con los análisis que se realicen, serán almacenados en un repositorio de GitHub para facilitar su gestión, revisión y colaboración entre investigadores y especialistas interesados en el tema. El repositorio permitirá mantener un registro histórico de los cambios realizados, fomentando así la transparencia y la reproducción de los resultados obtenidos.

Datos Complementarios

Datos MSP COVID-19:

Datos obtenidos de https://www.datosabiertos.gob.ec/dataset/https-almacenamiento-msp-gob-ec-index-php-s-maihh1064vskrbl/resource/4a15d681-743f-493c-8133-4b88715f9947

Códigos de parroquias-cantón:

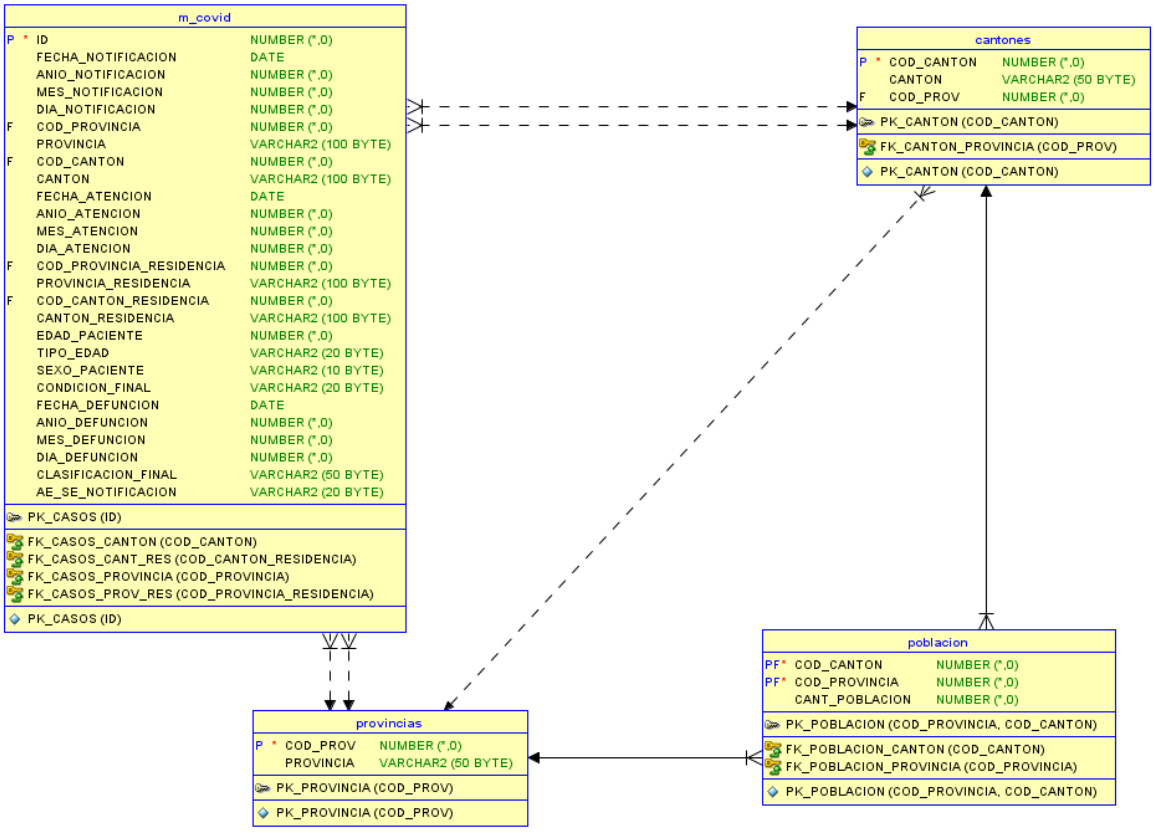
Datos obtenidos de

https://www.utmachala.edu.ec/archivos/SNIESE/Codificacion_Provincia-Canton-Parroquia.xlsx

Población por Provincias:

Datos obtenidos de https://www.romerostories.com/post/cantones-y-su-poblaci%C3%B3n-en-ecuador

Diseño lógico relacional



Diccionario de datos

ENTIDAD	ATRIBUTOS	DESCRIPCION
m_covid	• fecha_notificacion • anio_notificacion • mes_notificacion • dia_notificacion • cod_provincia • provincia • cod_canton • canton • fecha_atencion • anio_atencion • mes_atencion • dia_atencion • cod_provincia_residencia • provincia_residencia • cod_canton_residencia • canton_residencia • sexo_paciente • edad_paciente • tipo_edad • condicion_final • fecha_defuncion • anio_defuncion • mes_defuncion • dia_defuncion • clasificacion_final • ae_se_notificacion	• fecha en que se realizó el ingreso al aplicativo informático de Vigilancia • Año de la fecha de ingreso de la notificación • Mes de la fecha de ingreso de la notificación • Día de la fecha de ingreso de la notificación • Código de provincia de ubicación del establecimiento de salud donde se atendió el paciente • Nombre de provincia de ubicación del establecimiento de salud donde se atendió el paciente • Código de cantón de ubicación del establecimiento de salud donde se atendió el paciente • Nombre del cantón de ubicación del establecimiento de salud donde se atendió el paciente • Corresponde a la fecha en que, el paciente sospechoso de COVID-19, es atendido en un establecimiento de salud • Año de la fecha atención del paciente en el establecimiento de salud • Mes de la fecha de atención del paciente en el establecimiento de salud • Día de la fecha de atención del paciente en el establecimiento de salud • Código de provincia de residencia habitual del paciente • Nombre de provincia de residencia habitual del paciente • Código de cantón de residencia habitual del paciente • Nombre de cantón de residencia habitual del paciente • Sexo del paciente (Hombre/Mujer) • Edad declarada por el paciente • Tipo de métrica usada para la edad (ANIOS, MESES, DIAS)

		• Estado final del paciente al salir del sistema de vigilancia. (VIVO/MUERTO) • Se refiere a la fecha en que fallece el paciente, confirmado o probable para COVID-19 • Año tomado de la fecha de fallecimiento del paciente • Mes tomado de la fecha de fallecimiento del paciente • Día tomado de la fecha de fallecimiento del paciente • Estado final del evento sujeto a investigación (CONFIRMADO, DESCARTADO, PROBABLE, CON SOSPECHA) • División del año en semanas epidemiológicas que tienen 7 días. Codigo de primeros 4 digitos según el año, y los últimos dos de la semana epidemiológica
provincias	• cod_provincia • provincia	• Código de provincia del Ecuador • Nombre de provincia del Ecuador
cantones	• cod_canton • canton • cod_prov	• Código de cantón del Ecuador • Nombre del cantón del Ecuador • Código de provincia a la que pertenece el cantón
poblacion	• cod_canton • cod_provincia • cant_poblacion	• Código de cantón del Ecuador correspondiente a la entidad “cantones” • Código de provincia del Ecuador correspondiente a la entidad “provincias” • Cantidad de población de cada cantón

Normalización:

Tabla m_covid:

fecha_notificacion	anio_notificacion	mes_notificacion	dia_notificacion	cod_canton
18/09/2020	2020	9	18	1701
06/11/2020	2020	11	6	1701

Tabla cantones:

cod_canton	canton	cod_prov
101	CUENCA	1
108	SANTA ISABEL	1

Tabla provincias:

cod_prov	provincia
1	AZUAY
2	BOLIVAR

Tabla poblacion:

cod_canton	cod_provincia	cant_poblacion
101	1	546815
102	1	17399

En este modelo normalizado, se han creado las tablas adicionales "cantones" y “provincias” para eliminar la dependencia transitiva entre las tablas m_covid y poblacion. Cada tabla ahora contiene información coherente y las relaciones están establecidas a través de las claves primarias y foráneas correspondientes.

Script SQL

Para el Script realizamos la creación de una sola tabla a la cual le agregamos las columnas de datos que utilizamos para este proyecto y se definió el nombre de las variables con su respectivo tipo de dato, para la carga de datos se utilizó el comando LOAD de la herramienta MySQL importando los datos desde el CSV original, importando así los más de 2 millones de datos. Y para el resto de tablas, se usó de igual manera la herramienta de importación de datos de MySQL
Link Script:

https://github.com/By-Alvarez06/PROG-AV-ProyectoFinalZeppelin/blob/c251471043aa6ff2958db6cf0b7d115e2647adac/Script/Covid19_SQL.sql

Herramientas utilizadas

- Sistema operativo: Ubuntu 22 LTS o superior

- Entorno virtualizado: VirtualBox o VMware

- Lenguajes y motores: Java, Python, Scala, Spark, MySQL

- Plataforma de análisis interactivo: Apache Zeppelin

Sistema operativo Ubuntu

¿Qué es Zeppelin?

Apache Zeppelin es una herramienta de código abierto diseñada para la creación de notebooks interactivos, especialmente útil en entornos de análisis de datos, ciencia de datos y desarrollo colaborativo.

Características clave

- Ejecución de código multidisciplinar: soporta lenguajes como Python, Scala, SQL, R, entre otros, permitiendo análisis híbrido desde un mismo entorno.
- Visualización de datos: incorpora gráficos interactivos y dinámicos que se actualizan en tiempo real, facilitando la interpretación y presentación de resultados.
- Integración con motores de procesamiento: se conecta fácilmente con tecnologías como:
- Apache Spark para procesamiento distribuido en memoria.
- MySQL y otros sistemas de gestión de bases de datos para extracción y consulta de información.

Configuración del entorno

Instalación de la Máquina Virtual

Descarga e instala Oracle VirtualBox.
Crea una nueva máquina virtual con al menos 4 GB de RAM y 25 GB de disco
Monta la imagen iso de Ubuntu y sigue el asistente de instalación

Permisos de usuario de Ubuntu

- Para otorgar privilegios de superusuario a un usuario:

```
Terminal

Sudo visudo

nuevo_usuario ALL=(ALL:ALL) ALL
```

Guarda los cambios realizados en el archivo sudoers en nano. Para ello, presiona Ctrl + O, luego presiona Enter para confirmar el nombre del archivo y finalmente, presiona Ctrl + X para salir de nano.

Una vez que hayas completado estos pasos, el usuario "nuevo_usuario" tendrá privilegios de superusuario y podrá usar el comando sudo para ejecutar comandos con privilegios elevados.

Instalación de MySQL

Ejecutamos

```
Terminal mysql

CREATE USER 'zeppelin'@'localhost' IDENTIFIED
BY 'tu_clave';

GRANT ALL PRIVILEGES ON *.* TO
```

Crea usuario y base de datos si es necesario:

Podrás realizar esta configuración ejecutando este comando, desde una terminal o alguna IDE para el manejo de bases de datos como mysql workbench o datagrip

```
Terminal

sudo apt install mysql-server

sudo mysql_secure_installation
```

Instalación de Zeppelin

- Descarga desde zeppelin.apache.org.

- Extrae y configura

```
Terminal

export ZEPPELIN_HOME=/opt/zeppelin

export PATH=$PATH:$ZEPPELIN_HOME/bin
```

- Inicia Zeppelin (en una terminal dentro de la carpeta de Zeppelin):

```
$ZEPPELIN_HOME/bin/zeppelin-daemon.sh start
```

- Cierra zeppelin

```
$ZEPPELIN_HOME/bin/zeppelin-daemon.sh start
```

Instalación de Java

Ejecutar en la terminal los comandos

```
Terminal
sudo apt update
sudo apt install openjdk-11-jdk
```

y para verificar

```
java -version
```

Instalación de Python

Ejecutamos

```
Terminal
sudo apt install python3 python3-pip
```

Y para comprobar la instalación, abrimos un Shell de Pyhton con el comando:

```
python
python3
```

Instalación de Spark

- Descarga Spark desde spark.apache.org.

- Extrae y configura variables de entorno

Ejecutamos

```
Terminal
export SPARK_HOME=/opt/spark
export PATH=$PATH:$SPARK_HOME/bin
```

Instalación de Scala

Ejecutamos

```
sudo apt install scala
```

Interprete Zeppelin configuración de MySQL, Spark , Python

Configuración de MySQL en Interprete de Zepellin

- Crea intérprete %mysql con:

```
default.driver: com.mysql.cj.jdbc.Driver
default.url: jdbc:mysql://localhost:3306/DB
default.user: zeppelin
default.password: contraseña
```

Y al final configurar, en las dependencias, el driver a utilizar (en el caso de MySQL 8.0.33):

```
mysql:mysql-connector-java:8.0.33
```

Configuración de Python en Interprete de Zeppelin

Para poder realizar este procedimiento, con el zeppelin corriendo debemos

- ir a Interpreter > Create > selecciona %python.

- Configura:

```
zeppelin.python: python3
```

Análisis a realizar

Visualización de consultas

Consulta 1

Datos atípicos de los valores de edad de los pacientes, según su clasificación, encontramos registros de pacientes con edad en Años, Meses o días. Encontramos el conteo de los pacientes, el promedio de edades, desviación estándar, valor mínimo y valor máximo según el grupo de edad

```
SELECT
    tipo_edad,
    COUNT(edad_paciente) AS 'count',
    AVG(edad_paciente) AS 'mean',
    STDDEV(edad_paciente) AS 'stddev',
    MIN(edad_paciente) AS 'min',
    MAX(edad_paciente) AS 'max'
FROM m_covid
GROUP BY tipo_edad;
```

tipo_edad	Conteo	Promedio	Estandar	Minimo	Maximo
ANIOS	2852025	39.4239	17.77072743496206	1	136
MESES	7462	6.0503	3.2435771330575354	1	11
DIAS	4546	5.8713	7.045133230048723	0	30

INTERPRETACION/ANALISIS

La consulta muestra un resumen estadístico de la edad de los pacientes registrados en la base de datos, segmentada por el tipo de unidad utilizada para la edad (tipo_edad): AÑOS (ANIOS), MESES y DÍAS.

La mayoría de los registros están expresados en años, lo que refleja que la gran mayoría de pacientes fueron personas adultas. La edad promedio es de aproximadamente 39 años, con una desviación estándar alta (17.77), lo que indica gran dispersión en los datos. El valor máximo de 136 años sugiere un posible dato atípico, ya que excede significativamente la esperanza de vida promedio.

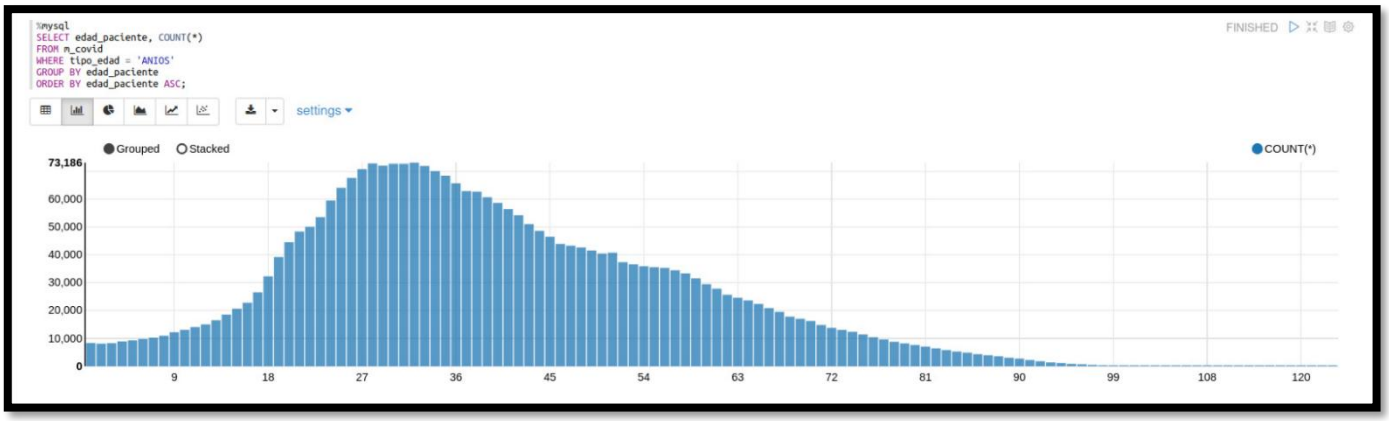
En cuanto a **meses**, este grupo representa a bebés menores de 1 año, y está correctamente acotado en el rango de 1 a 11 meses. El promedio de edad es de aproximadamente 6 meses, lo que parece coherente. No se observan valores atípicos en este grupo.

Finalmente, en cuanto a **días**, este grupo corresponde a recién nacidos, con edades entre 0 y 30 días. La media está cerca de los 6 días, y la desviación indica una cierta dispersión, pero dentro de lo esperable. El valor mínimo 0 días representa probablemente nacimientos recientes o neonatos registrados el mismo día.

Consulta 2

Gráfico estadístico en barras acerca del conteo de casos registrados de COVID-19 agrupados por edad en años ordenados de manera ascendente

```
SELECT edad_paciente, COUNT(*)
FROM m_covid
WHERE tipo_edad = 'ANIOS'
GROUP BY edad_paciente
ORDER BY edad_paciente ASC;
```



INTERPRETACION/ANALISIS

- **Forma de campana asimétrica (distribución sesgada a la derecha)**:
  - La distribución de frecuencias tiene una forma **asimétrica (positivamente sesgada)**.
  - El pico se alcanza aproximadamente entre las edades de **27 a 35 años**, con el máximo en **73,186 casos** en una sola edad (32 años).
- **Crecimiento progresivo hasta los 30 años:**
  - Desde edades tempranas (1 año) hasta aproximadamente los **30 años**, se observa un crecimiento constante en el número de casos.
  - Esto podría reflejar la mayor movilidad y exposición al virus de personas en edad productiva. Las personas entre 20 y 40 años concentran el mayor número de casos,

lo cual es esperable, ya que es el grupo con más actividad laboral, educativa y social.

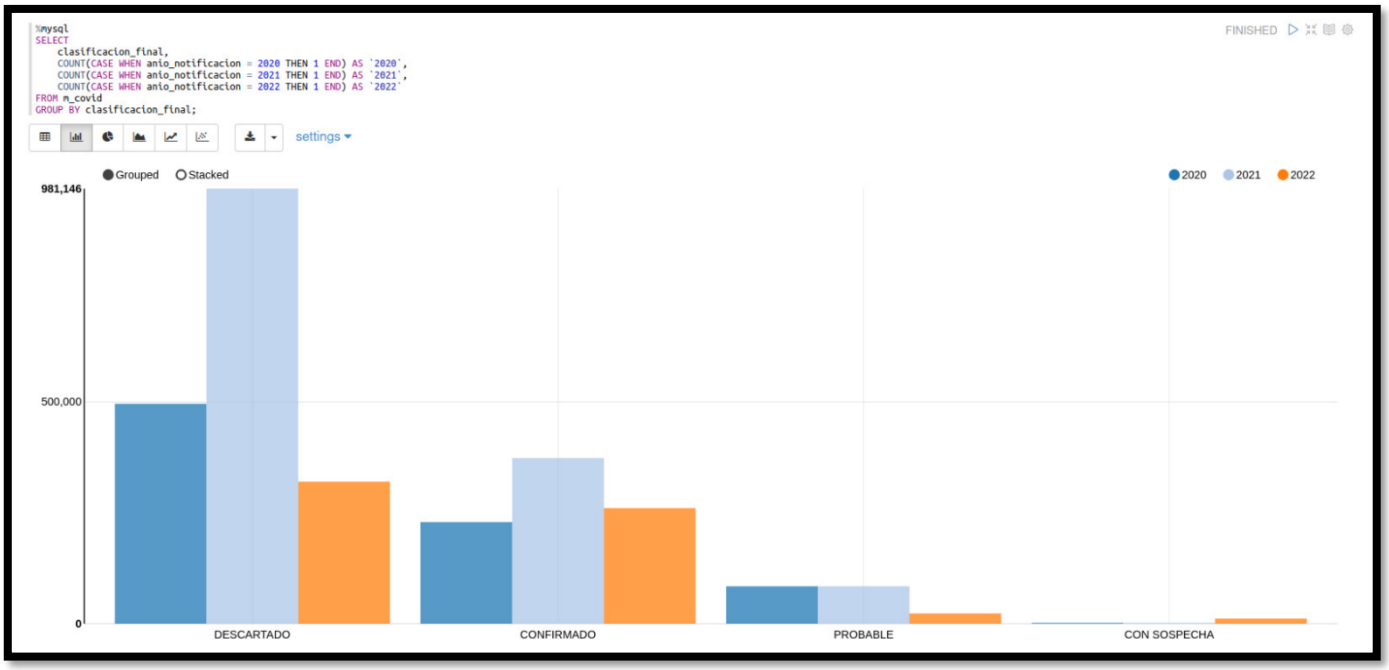
- **Disminución paulatina a partir de los 35 años:**
  - A partir de esta edad, se nota un descenso en el número de casos, aunque el número sigue siendo considerable hasta los 60 años.
  - Después de los **65 años**, el descenso es más marcado. A partir de los 60 años hay una reducción en los casos reportados, lo que podría estar relacionado con una menor movilidad, aislamiento preventivo o, en algunos casos, subregistro.
- **Presencia de datos hasta los 120 años:**
  - Aunque el conteo es muy bajo en edades superiores a los 90 años, **aún existen registros incluso hasta los 120 años**.

Consulta 3

Gráfico de barras comparativas (por grupos), donde cada grupo representa una clasificacion_final y dentro de cada uno se muestran las barras correspondientes a los años 2020, 2021 y 2022.

```
SELECT
   clasificacion_final,
   COUNT(CASE WHEN anio_notificacion = 2020 THEN 1 END) AS 2020,
   COUNT(CASE WHEN anio_notificacion = 2021 THEN 1 END) AS 2021,
   COUNT(CASE WHEN anio_notificacion = 2022 THEN 1 END) AS 2022
FROM m_covid
```

GROUP BY clasificacion_final;



INTERPRETACIÓN/ANÁLISIS

- **Clasificación "DESCARTADO"**

  - La categoría con los valores más altos, encabezando el año 2021 con un máximo de 981.146 casos descartados, ya que se realizan muchas pruebas con resultado negativo, mientras que 2020 y 2022 le siguen consecutivamente con una caída de casos descartados.

- **Clasificación "CONFIRMADO"**

  - La segunda categoría con los valores más altos, especialmente en 2021, por ser el año consecutivo al año de inicio y la expansión rápida del COVID-19. Con menores casos en el año 2022 y 2020 respectivamente

- **Clasificación "PROBABLE"**

  - Este grupo suele incluir pacientes con síntomas clínicos, pero sin confirmación por laboratorio. Los años 2020 y 2021 muestran una cantidad de casos similares con alrededor de 84.000 casos probables, por otro lado, el año 2022 fue el año con menor cantidad de casos probables.

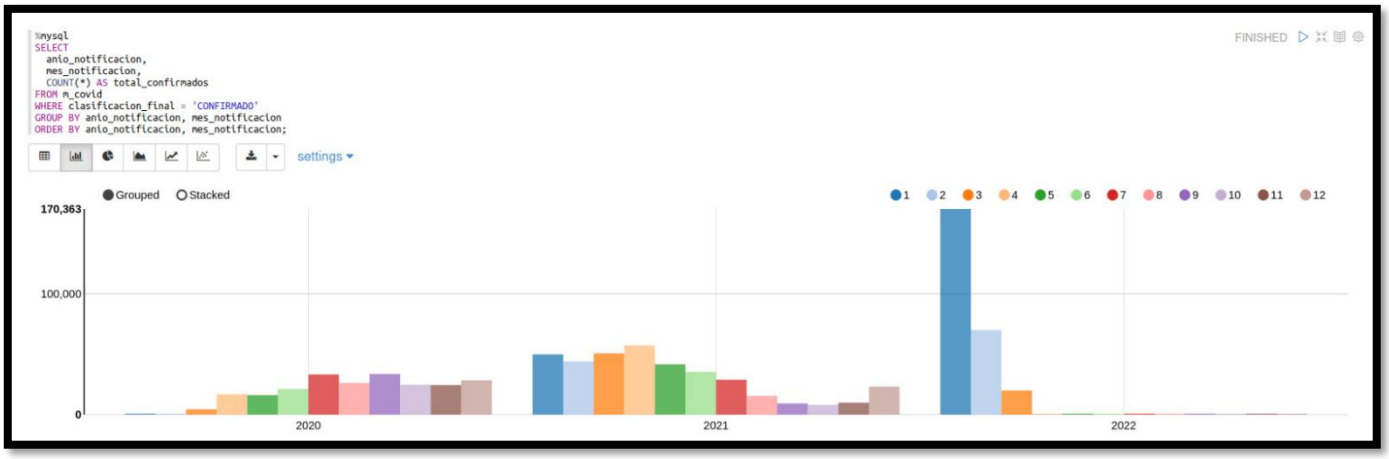
- **Clasificación "CON SOSPECHA"**

  - Similar al anterior, pero puede incluir registros en los que nunca se completó el diagnóstico. Se contienen muy pocos registros, siendo el año 2022 el único año con registros tabulados.

Consulta 4

Grafica de casos confirmados de cada mes agrupados por año

```
SELECT
  anio_notificacion,
  mes_notificacion,
  COUNT(*) AS total_confirmados
FROM m_covid
WHERE clasificacion_final = 'CONFIRMADO'
GROUP BY anio_notificacion, mes_notificacion
ORDER BY anio_notificacion, mes_notificacion;
```



INTERPRETACIÓN/ANÁLISIS

➢ **Comportamiento temporal de la pandemia**

- Esta consulta permite **identificar picos y caídas** en los contagios **mes a mes y año por año**. Los **picos altos** se esperan en meses como:
  - **Julio a agosto 2020**, inicio y expansión de la pandemia.
  - **Enero a abril 2021**, por las festividades de fin e inicio de año.
  - **Enero 2022**, repuntes por nuevas variantes como ómicron y festividades de fin de año.
- El gráfico muestra claramente las **olas de contagio**. También evidencia meses de **baja transmisión**, posiblemente por confinamientos estrictos o campañas de vacunación.

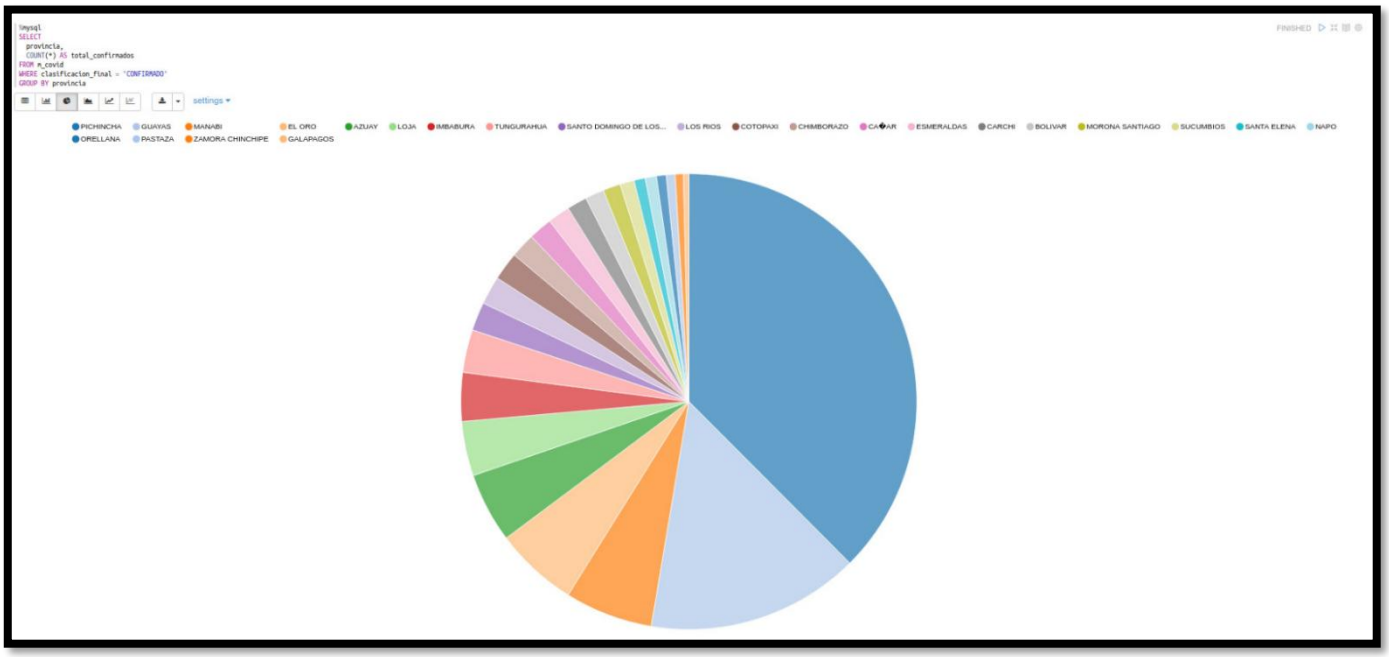
➢ **Tendencia global**

- Puede observarse una **tendencia decreciente** de casos hacia fines del 2021, posiblemente debido a que las medidas sanitarias fueron efectivas. Hay repuntes inesperados, como inicios del 2022, se podrían analizar correlaciones con eventos sociales, relajamiento de restricciones o nuevas variantes.

Consulta 5

Gráfica de casos confirmados agrupados por provincia

```
SELECT
  provincia,
  COUNT(*) AS total_confirmados
FROM m_covid
WHERE clasificacion_final = 'CONFIRMADO'
GROUP BY provincia
ORDER BY total_confirmados DESC;
```



INTERPRETACIÓN/ANÁLISIS

➢ **Distribución geográfica de contagios**

- Muestra **cuáles provincias tienen mayor carga de casos confirmados**.
- Las provincias **Pichincha, Guayas y Manabí** están en los primeros lugares, debido a:
    - Mayor densidad poblacional.
    - Más centros urbanos e interacciones sociales.
    - Mayor capacidad de diagnóstico (más pruebas realizadas).
- Las provincias con altos casos requieren **más atención sanitaria, hospitales, personal médico y campañas de prevención**. Aquellas con menos casos pueden reflejar zonas rurales o bien menor densidad poblacional con menor exposición, como el caso de Galápagos.

Conclusiones

El análisis realizado sobre la base de datos de casos de COVID-19 en el Ecuador permitió extraer información relevante tanto desde una perspectiva descriptiva como exploratoria, proporcionando una visión integral de la evolución y distribución de los contagios en el país.

En primer lugar, el estudio de los datos atípicos de edad permitió obtener estadísticas como el promedio de edad, la desviación estándar y los valores mínimo y máximo por categoría. Este análisis no solo mostró la variabilidad en las edades de los pacientes, sino que también permitió identificar valores extremos que podrían corresponder a casos excepcionales o errores en el registro. Por otra parte, el gráfico estadístico en barras del conteo de casos por edad evidenció una concentración significativa en determinados grupos de edad, lo que sugiere que ciertas edades estuvieron más expuestas al contagio o presentaron mayor detección de casos.

El análisis comparativo por clasificación y por años permitió identificar tendencias temporales y diferenciar la evolución del diagnóstico de los casos, reflejando tanto el avance de la pandemia como el impacto de las medidas sanitarias a lo largo del tiempo. Asimismo, la gráfica de casos confirmados por mes agrupados por año facilitó la identificación de los picos de contagio en las diferentes olas de la pandemia, evidenciando patrones estacionales y momentos críticos donde los contagios se dispararon.

En conjunto, las consultas y visualizaciones generadas permitieron una comprensión más profunda de los datos, revelando patrones relevantes y apoyando la toma de decisiones basadas en evidencia. Este trabajo demostró la importancia del uso de herramientas de análisis y consulta en bases de datos para transformar datos crudos en información útil y accionable.

Bibliografía

Apache Software Foundation. (2025). *Apache Zeppelin Documentation*. Recuperado de https://zeppelin.apache.org/docs/latest/

Apache Zeppelin | Cloudera. (2022, 21 noviembre). Cloudera. https://es.cloudera.com/products/open-source/apache-hadoop/apache-zeppelin.html#:~:text=Zeppelin%20es%20una%20moderna%20plataforma,de%20datos%20cada%20vez%20mayor.

Canonical Ltd. (2025). Ubuntu Documentation. Recuperado de https://ubuntu.com/server/docs

Cordero, P. (2020, 29 septiembre). *Como instalar Ubuntu en VirtualBox | Oficina de software libre*. https://osl.ugr.es/2020/09/29/como-instalar-ubuntu-en-virtual-box/

Ministerio de Salud Pública del Ecuador. (2022). Datos abiertos de casos COVID-19.

Ubuntu. (s.f.). Acerca de Ubuntu. Recuperado el 19 de julio de 2023, de https://ubuntu.com/about