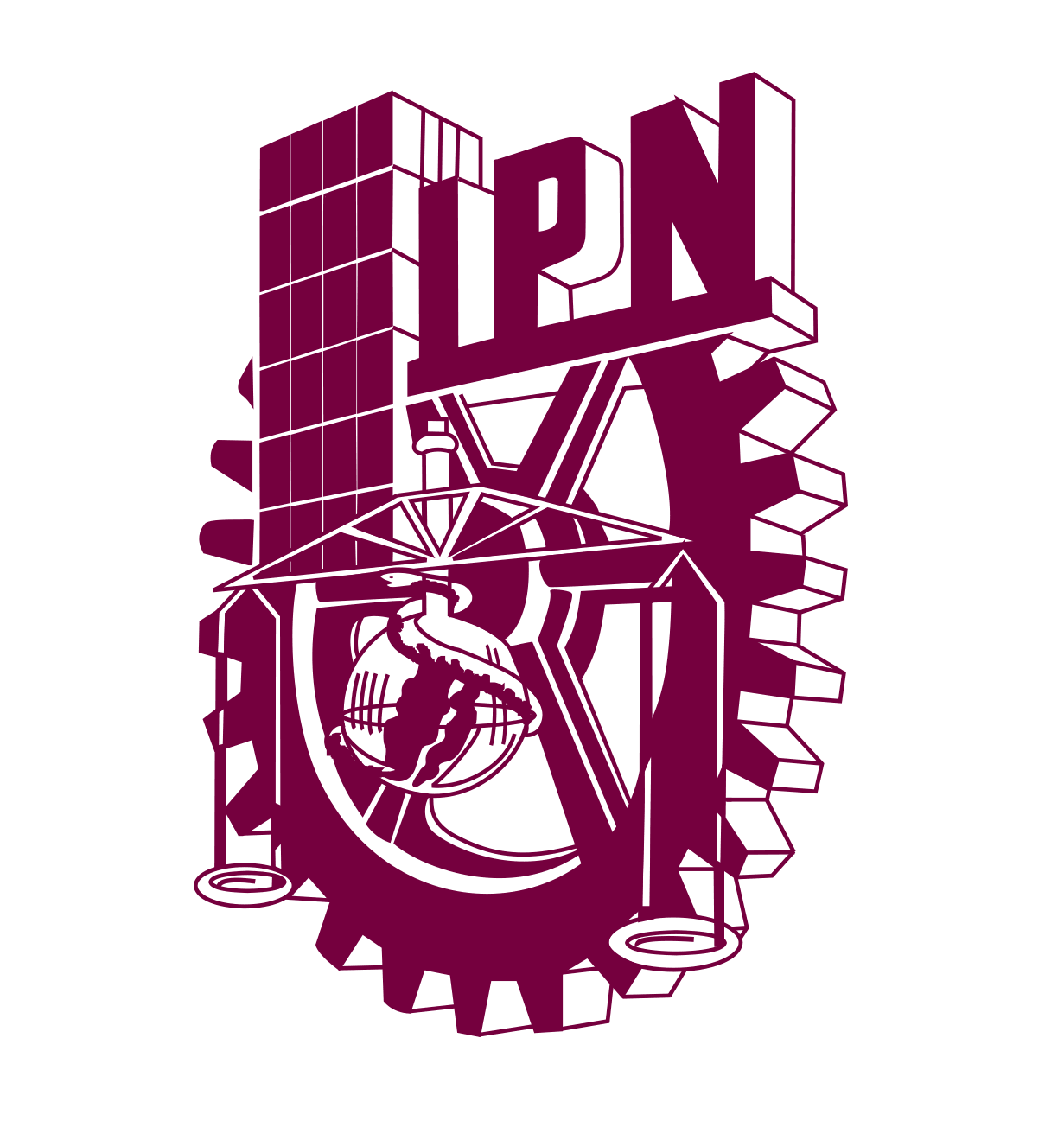
**INSTITUTO POLITÉCNICO NACIONAL**

**ESCUELA SUPERIOR DE COMPUTO**

**SISTEMAS DISTRIBUIDOS**

[**Práctica 4. Procesamiento de tablas agregadas utilizando MapReduce**](https://m4gm.com/moodle/mod/assign/view.php?id=533)

**7CV2**

**Ingeniería en Sistemas Computacionales**

**PROFESOR:**  **Pineda Guerrero Carlos**

**ALUMNO: Dominguez Olvera Leonardo Daniel**

**BOLETA: 2022630452**

**FECHA DE ENTREGA: 22 de junio 2025**

# DESARROLLO

**Creación y carga de la base de datos**

1. Crear una máquina virtual con Ubuntu **en** Azure, 4 GB de memoria RAM, 2 CPU virtuales y disco SSD estándar.

El nombre de la máquina virtual **deberá** ser: "P5-MAP-" concatenando el número de boleta del alumno o alumna; por ejemplo, si el número de boleta es 12345678, entonces la máquina virtual deberá llamarse: P3-MAP-12345678.

Para este paso, solo incluir en el documento PDF la captura de la última pantalla, correspondiente a la creación de la máquina virtual.

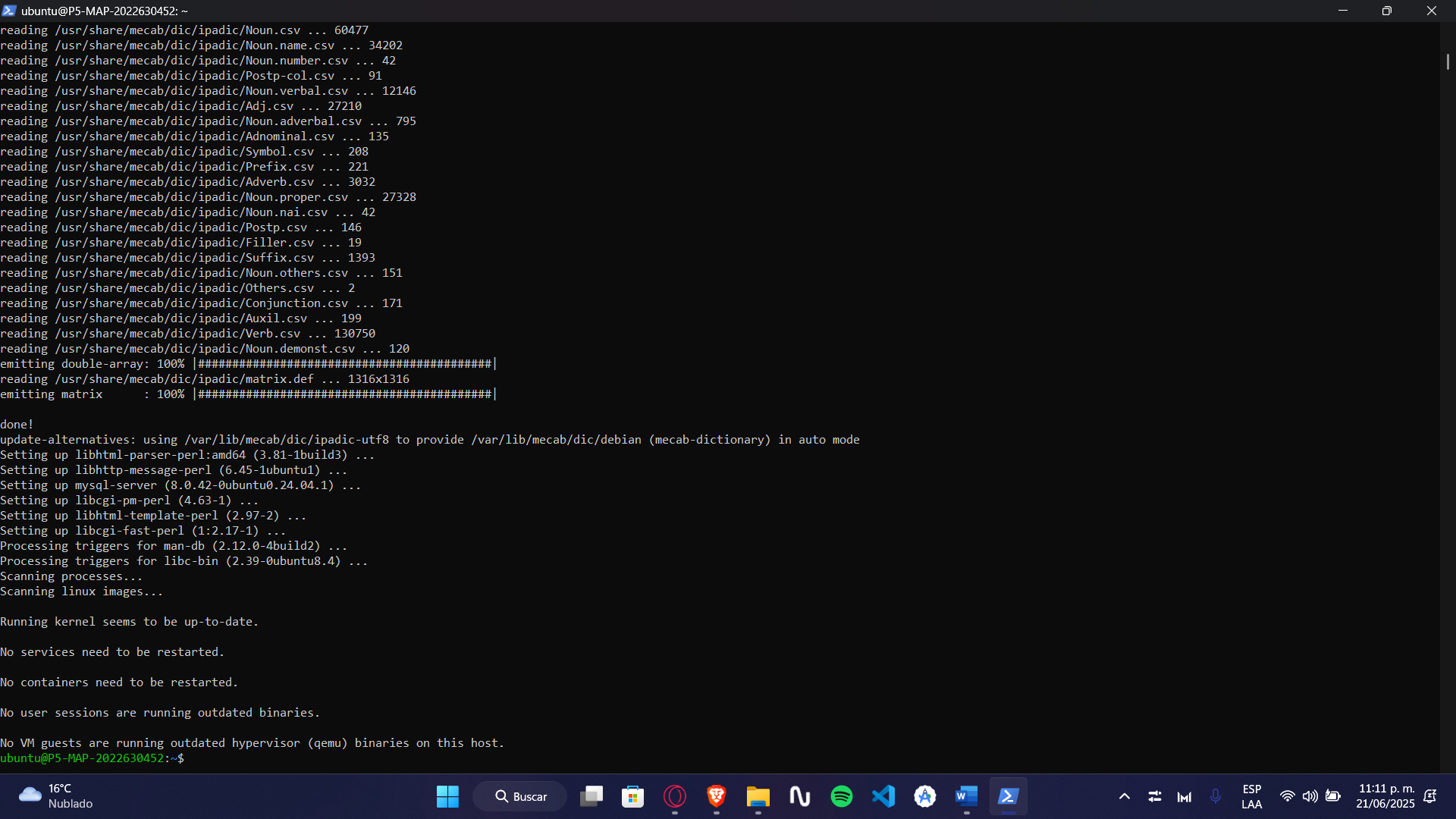
Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

2. Instalar el servidor MySQL en la máquina virtual.



3. Crear la base de datos "practica\_olap":

mysql -u root -p  
create database practica\_olap;

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

4. Descargar el archivo "practica\_olap.sql" de la práctica 4 en la plataforma Moodle.

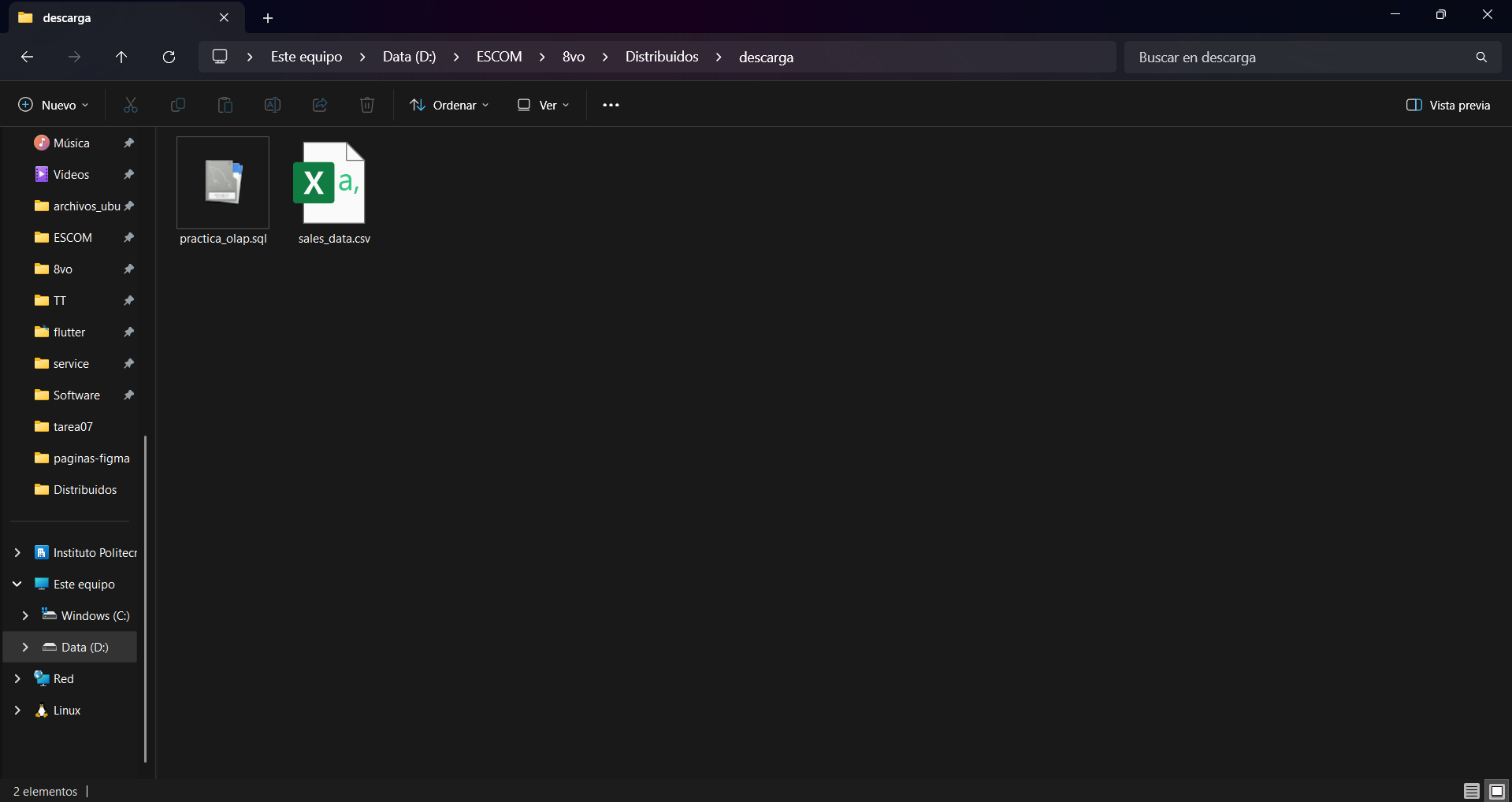
Captura de pantalla de un celular

El contenido generado por IA puede ser incorrecto.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

5. Descargar el archivo "sales\_data.csv" de la práctica 4 en la plataforma Moodle.



Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

6. Crear las tablas ejecutando el script "practica\_olap.sql" en la línea de comandos:

mysql -u root -p practica\_olap < practica\_olap.sql

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

7. Ejecutar el monitor de mysql incluyendo la opción "local-infile" para habilitar la carga de datos:

mysql --local-infile=1 -u root -p practica\_olap

Captura de pantalla de computadora

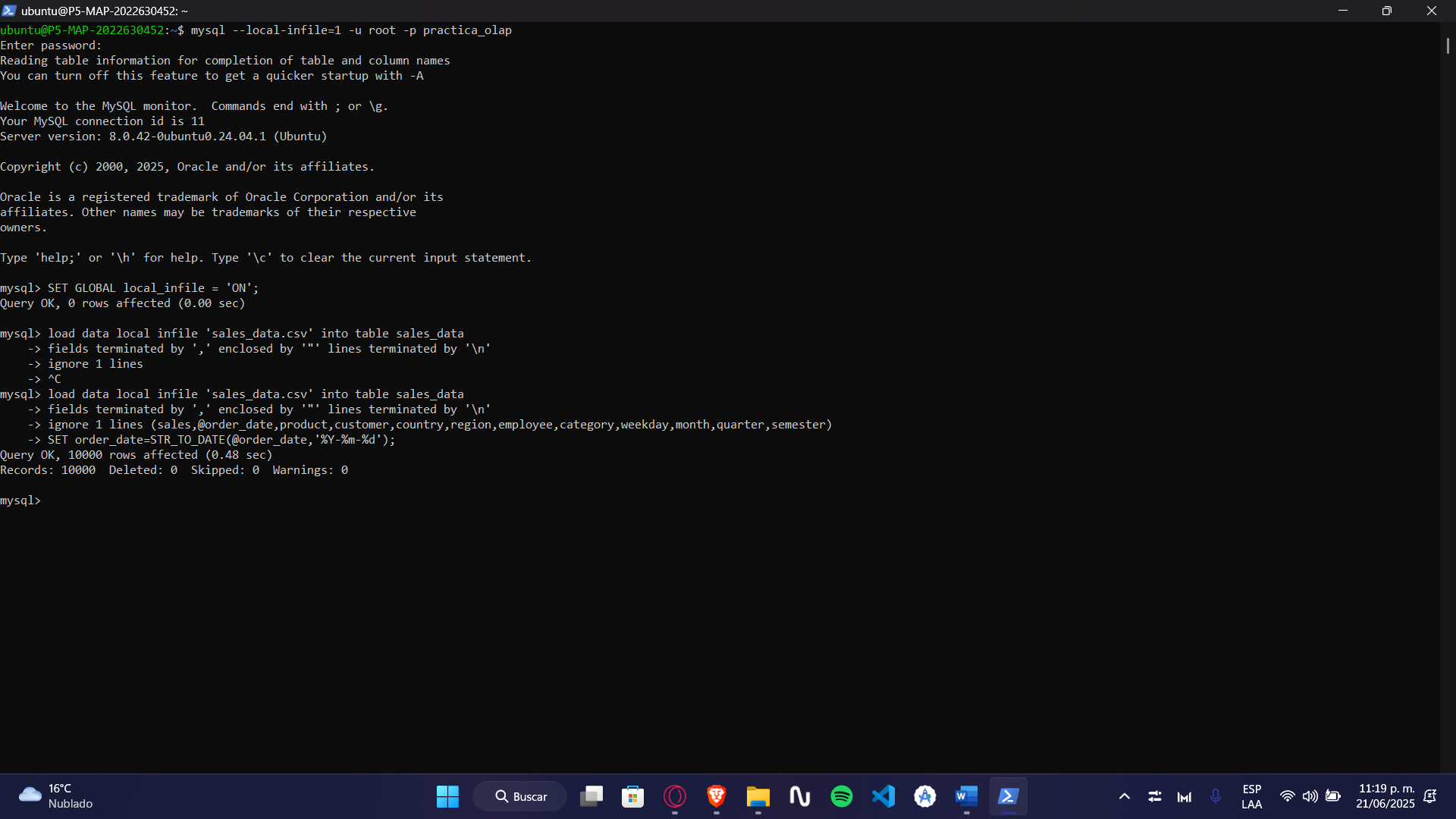
El contenido generado por IA puede ser incorrecto.

8. Cargar los datos a la tabla "sales\_data", ejecutando los siguientes comandos en el monitor de mysql:

SET GLOBAL local\_infile = 'ON';

load data local infile 'sales\_data.csv' into table sales\_data   
fields terminated by ',' enclosed by '"' lines terminated by '\n'   
ignore 1 lines (sales,@order\_date,product,customer,country,region,employee,category,weekday,month,quarter,semester)  
SET order\_date=STR\_TO\_DATE(@order\_date,'%Y-%m-%d');

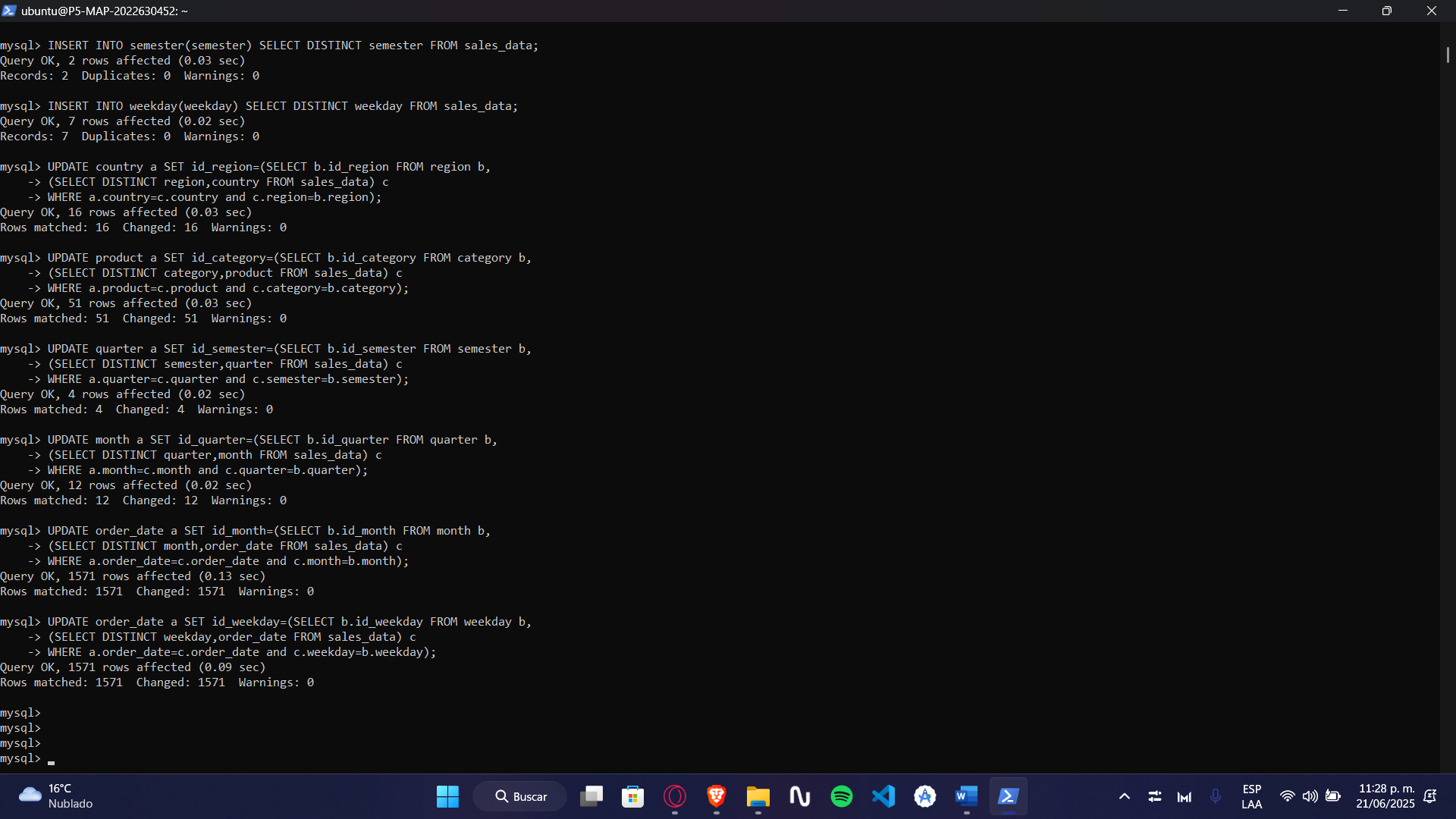
Donde "ruta-archivo-sales\_data.csv" es la ruta absoluta del archivo "sales\_data.csv", descargado de la plataforma Moodle.



10. Cargar las tablas de dimensiones y la fact table, ver las "Actividades individuales a realizar" de la clase del día 26/05/2025.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.



Pantalla de computadora con fondo negro

El contenido generado por IA puede ser incorrecto.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

11. Ejecutar el monitor de mysql:

mysql -u root -p practica\_olap

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

12. Ejecutar la siguiente instrucción SELECT para generar el archivo CSV country\_category\_product.csv con las columnas id\_contry, id\_category, id\_product y sales:

SELECT a.id\_country,b.id\_category,a.id\_product,a.sales   
INTO OUTFILE '/tmp/**country\_category\_product.csv**'  
FIELDS TERMINATED BY ',' OPTIONALLY ENCLOSED BY '"'  
LINES TERMINATED BY '\n'  
FROM fact\_table a,product b   
WHERE a.id\_product=b.id\_product;

**Se cambio el commando a**  
SELECT a.id\_country, b.id\_category, a.id\_product, a.sales

INTO OUTFILE '/var/lib/mysql-files/country\_category\_product.csv'

FIELDS TERMINATED BY ',' OPTIONALLY ENCLOSED BY '"'

LINES TERMINATED BY '\n'

FROM fact\_table a, product b

WHERE a.id\_product = b.id\_product;

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

**Instalación y ejecución de Apache Hadoop**

14. Ahora vamos a descargar e instalar Apache Hadoop. Ejecutar la siguiente instrucción en la máquina virtual:

wget <https://dlcdn.apache.org/hadoop/common/hadoop-3.4.0/hadoop-3.4.0-aarch64.tar.gz>

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

15. Descomprimir y desempacar el archivo descargado:

gunzip hadoop-3.4.0-aarch64.tar.gz  
tar xvf hadoop-3.4.0-aarch64.tar

Pantalla de computadora con fondo negro

El contenido generado por IA puede ser incorrecto.

16. Apache Hadoop requiere Java, entonces necesitamos instalar el JDK en la máquina virtual:

sudo apt-get update  
sudo apt install openjdk-16-jdk-headless

Pantalla de computadora con fondo negro

El contenido generado por IA puede ser incorrecto.

17. Ahora vamos a editar el archivo "hadoop-3.4.0/etc/hadoop/hadoop-env.sh":

nano hadoop-3.4.0/etc/hadoop/hadoop-env.sh

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

18. Quitar el comentario a "export JAVA\_HOME" y asignarle "/usr" (el directorio que contiene el directorio "bin" que a su vez contiene los programas "java" y "javac"):

export JAVA\_HOME=/usr

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

19. Ahora vamos a crear una aplicación Hadoop que obtenga una tabla agregada a partir del archivo "country\_category\_product.csv".

19.1 Definimos las siguientes variables de entorno:

export JAVA\_HOME=/usr  
export HADOOP\_HOME=/home/***usuario***/hadoop-3.4.0  
export PATH=$PATH:$JAVA\_HOME/bin:$HADOOP\_HOME/bin

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

Donde ***usuario***es el usuario actual de la máquina virtual

19.2 Creamos un directorio "prueba", y dentro del directorio creamos las clases "AggregationMapper.java", "AggregationReducer.java" y "AggregationDriver.java" que vimos en clase.

mkdir prueba

Pantalla de computadora con fondo negro

El contenido generado por IA puede ser incorrecto.

nano prueba/AggregationMapper.java

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

nano prueba/AggregationReducer.java

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.  
nano prueba/AggregationDriver.java

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

19.3 Compilamos las clases y creamos un archivo .jar (aplicación Hadoop):

javac -classpath `$HADOOP\_HOME/bin/hadoop classpath` -d prueba prueba/\*.java  
jar -cvf Aggregation.jar -C prueba .

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

19.4 Creamos un directorio donde colocaremos todos los archivos de entrada para nuestra aplicación Hadoop:

mkdir prueba/input

Pantalla de computadora con fondo negro

El contenido generado por IA puede ser incorrecto.

19.5 Copiamos el archivo "country\_category\_product.csv" (obtenido en el paso 13) al directorio "prueba/input".

sudo cp /var/lib/mysql-files/country\_category\_product.csv ~/prueba/input/

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

19.6 Ejecutamos el job de Hadoop (si el directorio prueba/output ya existe, es necesario eliminarlo antes de ejecutar el job de Hadoop):

hadoop jar Aggregation.jar AggregationDriver prueba/input prueba/output

Entonces podemos ver el resultado del proceso en el directorio "prueba/output", en este caso el resultado es la tabla agregada.

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

20. Ahora se cargará el archivo resultante del proceso MapReduce, a la tabla agregada "country\_category\_product" de la base de datos "practica\_olap":

20.1 Podemos ver que el archivo resultante se llama "part-r-00000" y se encuentra en el directorio "prueba/output".

Si desplegamos este archivo veremos que consta de parejas clave-valor separadas por tabulador, por tanto, para poder cargar el archivo a la tabla "country\_category\_product" debemos cambiar el tabulador por coma (para que sea un archivo CSV); para cambiar el tabulador por una coma ejecutamos el siguiente comando en la línea de comandos:

sed -i 's/\t/,/g' prueba/output/part-r-00000

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

20.2 Ejecutar el monitor de MySQL incluyendo la opción "local-infile" para habilitar la carga de datos:

mysql --local-infile=1 -u root -p practica\_olap

Captura de pantalla de computadora

El contenido generado por IA puede ser incorrecto.

20.3 Cargar el archivo "part-r-00000", ejecutando los siguientes comandos en el monitor de mysql:

load data local infile 'ruta-absoluta-del-archivo-part-r-00000' into table country\_category\_product   
fields terminated by ',' enclosed by '"' lines terminated by '\n'   
(id\_country,id\_category,id\_product,sales);

Donde "ruta-absoluta-del-archivo-part-r-00000" es la ruta absoluta del archivo part-r-00000.

Pantalla de computadora con fondo negro

El contenido generado por IA puede ser incorrecto.

# CONCLUSIÓN

Se integraron herramientas como Apache Hadoop con una base de datos relacional como MySQL, demostrando un flujo completo de extracción, transformación y carga (ETL) de datos. A partir de un archivo CSV, se utilizó un programa MapReduce para realizar una agregación por país, categoría y producto, obteniendo como resultado una tabla resumida de ventas.

Durante el proceso, se evidenció la importancia de tener los datos correctamente estructurados y formateados, ya que cualquier error en la delimitación o permisos del archivo puede impedir la ejecución del job o la carga en MySQL. La combinación de Hadoop y MySQL permitió no solo procesar un volumen considerable de datos, sino también almacenarlos de forma estructurada para su análisis posterior.

Y se consolidó el modelo estrella mediante la creación de una tabla de hechos y una tabla agregada, lo que representa una base sólida para sistemas de inteligencia de negocios y análisis de datos.