Tarea Evaluación 1

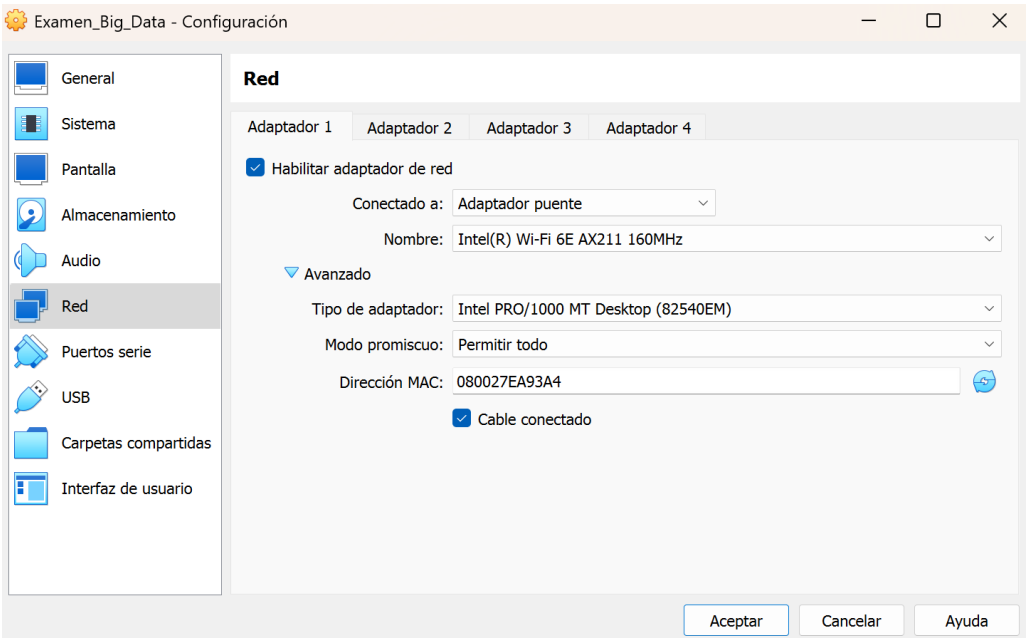
Preparación del entorno

1. Descargar máquina virtual del siguiente enlace:

Examen_Big_Data.ova

2. Importar OVA en VirtualBox

Importar OVA renovando interfaces de red. Actualizar MAC si da error.



3. Modificar fichero de hosts con la IP de la máquina virtual

Iniciar la máquina virtual con usuario hadoop/hadoop y consultar la IP.

```
ip a
```

Cambiar la IP del nodomaster por la que se ha obtenido.

```
sudo vi /etc/hosts
        10.85.10.103 nodomaster
```

Salimos guardando (:wq)

4. Arrancar Hadoop

```
start-all.sh
```

5. Arrancar Hive

```
hiveserver2 &
```

6. Comprobar que todos los servicios están en ejecución

```
jps
        Picked up _JAVA_OPTIONS: -Djava.io.tmpdir=/tmp
        2352 RunJar
        1504 DataNode
        1904 ResourceManager
        1396 NameNode
        2939 Jps
        2012 NodeManager
        1711 SecondaryNameNode
```

Si falta algún servicio, detener todo y volver a iniciarlo.

```
        kill -9 ResourceManager
        stop-all
```

7. Arrancar Hue

```
cd hue
docker-compose up
```

* Esperar hasta que complete el arranque (30s – 1min)

8. Abrir el cliente web de Hue en el navegador:

```
http://10.85.10.103:8888
usuario: hadoop
contraseña: hadoop
```

9. Abrir una nueva ventana de terminal:

```
ssh hadoop@10.85.10.103
```

Ejercicio 1. Preparación de los datos.

1. Crea un directorio “examen” en el home del usuario y sitúate en él.

2. Descarga los ficheros de las siguientes URLs (`wget enlace`):
   https://github.com/antoniojcalvo/BDA/raw/refs/heads/main/city_temperature.csv.tar.gz
   https://raw.githubusercontent.com/antoniojcalvo/BDA/refs/heads/main/olive.csv
   https://raw.githubusercontent.com/antoniojcalvo/BDA/refs/heads/main/palm.csv
   https://raw.githubusercontent.com/antoniojcalvo/BDA/refs/heads/main/sunflowerseed.csv

3. Descomprime el primer fichero.

```
tar –xzvf city_temperature.csv.tar.gz
```

4. En el sistema de archivos HDFS del clúster, crea un directorio “/examen”.

5. Copia los ficheros descargados al directorio “/examen” del clúster hdfs.

Ejercicio 2. Tablas externas en Hive

1. Crea una carpeta “/examen/oil” en el clúster hdfs y mueve el fichero “olive.csv” a dicha carpeta.

2. Crea una tabla externa sobre el directorio “oil”, con las columnas que tiene el fichero. Sustituye el texto marcado por los valores adecuados.

```
CREATE EXTERNAL TABLE oil_prod (
        country STRING,
        year INT,
        Beginning_Stocks DOUBLE,
        Domestic_Consumption DOUBLE,
        Ending_Stocks DOUBLE,
        Exports DOUBLE,
        Feed_Waste DOUBLE,
        Food_Use DOUBLE,
        Imports DOUBLE,
        Industrial DOUBLE,
        Production DOUBLE,
        Total_Distribution DOUBLE,
        Total_Supply DOUBLE)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY 'separador'
STORED AS formato_fichero
LOCATION '/ruta/a/ficheros'
TBLPROPERTIES ("skip.header.line.count"="1");
```

3. Realiza las siguientes consultas y captura pantalla de los resultados de cada una, incluyéndolas en el fichero de respuesta:
   a. Muestra las 10 primeras filas de la tabla
   b. Cuenta el número total de registros
   c. Obtén país y año de máxima producción (pista: 2 sentencias sql)

4. Mueve los ficheros “palm.csv” y “sunflowerseed.csv” al directorio “/examen/oil” en el clúster hdfs, y repite las consultas del punto 3, guardando las capturas de pantalla.

Ejercicio 3. Tablas temporales, formato parquet y particiones

1. Crea una tabla temporal, añade propiedad para que se salte la primera línea de cabecera y carga los datos del fichero “city_temperature.csv”.

```
CREATE TEMPORARY TABLE staging_temperatures (
        Region STRING,
        Country STRING,
        State STRING,
        City STRING,
        Month INT,
        Day INT,
        Year INT,
        AvgTemperature FLOAT)
ROW FORMAT DELIMITED
FIELDS TERMINATED BY 'separador'
STORED AS formato_fichero;

ALTER TABLE staging_temperatures
SET TBLPROPERTIES ("skip.header.line.count"="1");

LOAD DATA INPATH '/ruta/fichero.csv' INTO TABLE staging_temperatures;
```

2. Realiza dos consultas para comprobar los datos:
    a. 100 primeras filas.
    b. 10 primeras filas con valores de AvgTemperature negativos.

3. Comprueba los ficheros almacenados en el warehouse de Hive. Captura pantalla e inclúyela en el fichero a entregar.

    Ruta en cluster hdfs: `/user/hive/warehouse/`

4. Crea una tabla en formato parquet sobre los datos de la tabla temporal, filtrando aquellos que tengan valores incorrectos (temperatura <> -99)

```
CREATE TABLE temperatures
STORED AS formato_fichero
AS SELECT * FROM staging_temperatures where AvgTemperature <> -99;
```

5. Para la región ‘Europe’ y el año 2015, muestra las temperaturas mayores de 85, incluyendo ciudad, mes, día y temperatura.
   Obtén el plan de ejecución de la consulta y captura la pantalla de resultado donde se muestra el número de registros leídos antes de filtrar, e inclúyela en el fichero a entregar.

```
EXPLAIN SELECT … FROM … WHERE …
```

6. Crea una tabla en formato parquet, particionada por country y year. Activa antes las particiones.

```
SET hive.exec.dynamic.partition=true;
SET hive.exec.dynamic.partition.mode=nonstrict;

CREATE TABLE temperaturesbyRegionYear (
        Country STRING,
        State STRING,
        City STRING,
        Month INT,
        Day INT,
        AvgTemperature FLOAT)
PARTITIONED BY (Region STRING, Year INT)
STORED AS formato_fichero;
```

7. Inserta los datos de la tabla temperaturas a partir de 2010 (por la limitación de memoria de nuestra máquina virtual).

```
INSERT INTO TABLE temperaturesbyRegionYear
        PARTITION (Campo1,Campo2)
SELECT Country, State, City, Month, Day, AvgTemperature, Region, Year
        FROM temperatures
        WHERE Region IS NOT NULL and Year IS NOT NULL and Year > 2010;
```

8. Repite el paso 5.

9. Repite el paso 3

   Nota: el comando -ls admite la opción -R para mostrar los directorios de forma recursiva. Amplía la ventana y/o reduce el tamaño de letra para que se muestren las líneas enteras.