Examen BDA1 - Tarea práctica

Ejercicio 1. Preparación de los datos

* Crea un directorio “examen” en el home del usuario y sitúate en él.
* Descarga los ficheros de las siguientes URLs (wget enlace)
* Descomprime el primer fichero



* En el sistema de archivos HDFS del clúster, crea un directorio “/examen”
* Copia los ficheros descargados al directorio “/examen” del clúster hdfs

Texto

Descripción generada automáticamente Texto

Descripción generada automáticamente

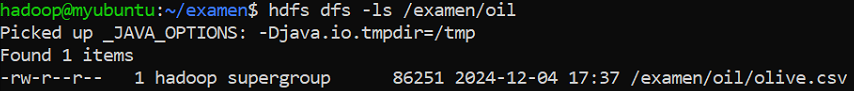
Ejercicio 2. Tablas externas en Hive

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

Texto

Descripción generada automáticamente





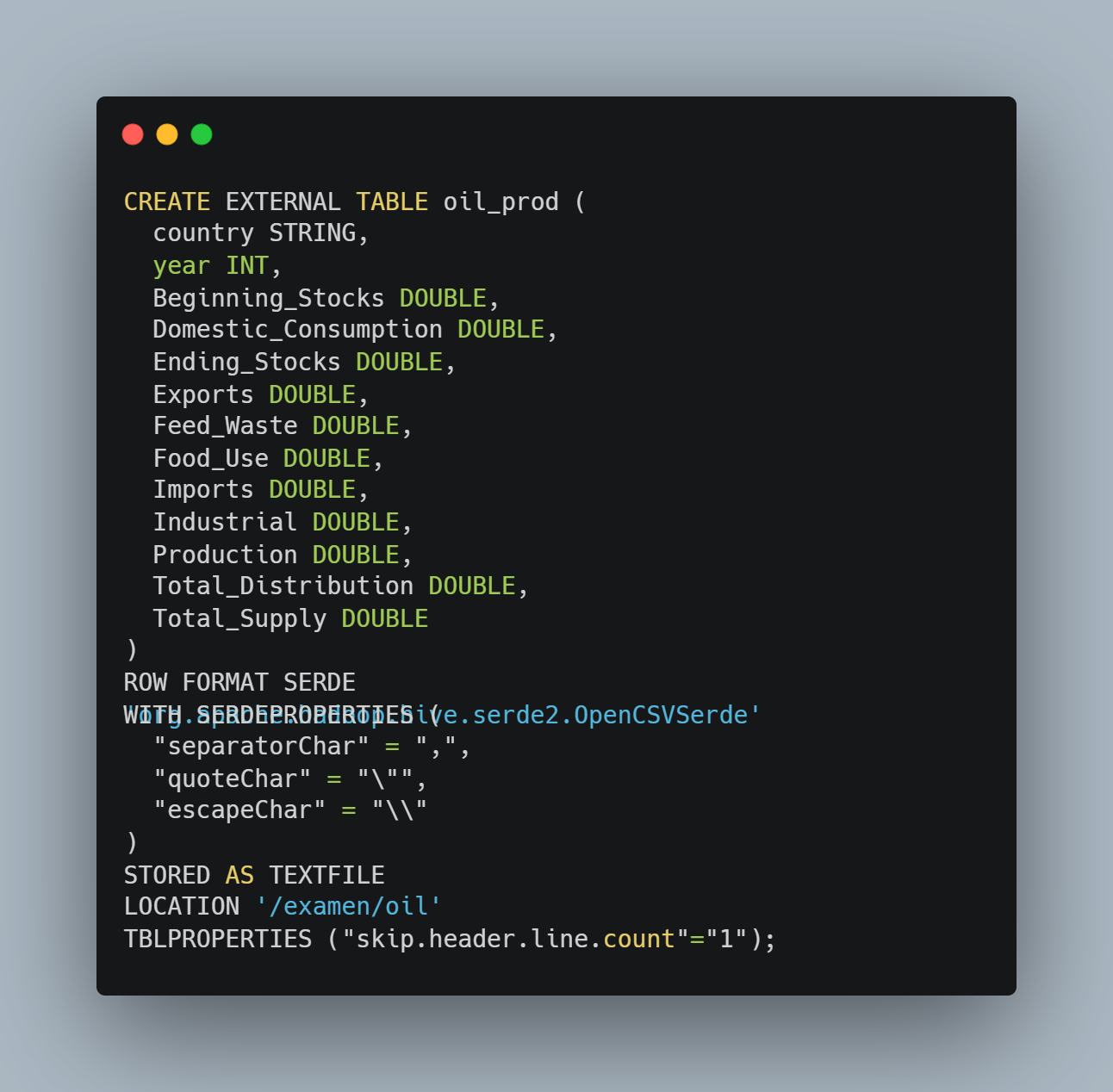
* Crea una tabla externa sobre el directorio “oil”, con las columnas que tiene el fichero. Sustituye el texto marcado por los valores adecuados.

Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente



* Realiza las siguientes consultas y captura pantalla de los resultados de cada una, incluyéndolas en el fichero de respuesta:
  + Muestra las 10 primeras filas de la tabla



Pantalla de computadora con letras

Descripción generada automáticamente con confianza media

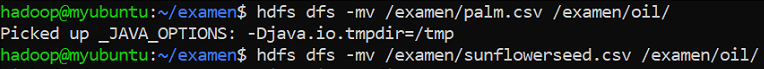
* + Cuenta el número total de registros

No he sido capaz de hacerlo, la consulta alcanzaba los 8 minutos incluso con LIMIT 50.

* + Obtén país y año de máxima producción (pista: 2 sentencias sql)

No he sido capaz de hacerlo, la consulta alcanzaba los 8 minutos.

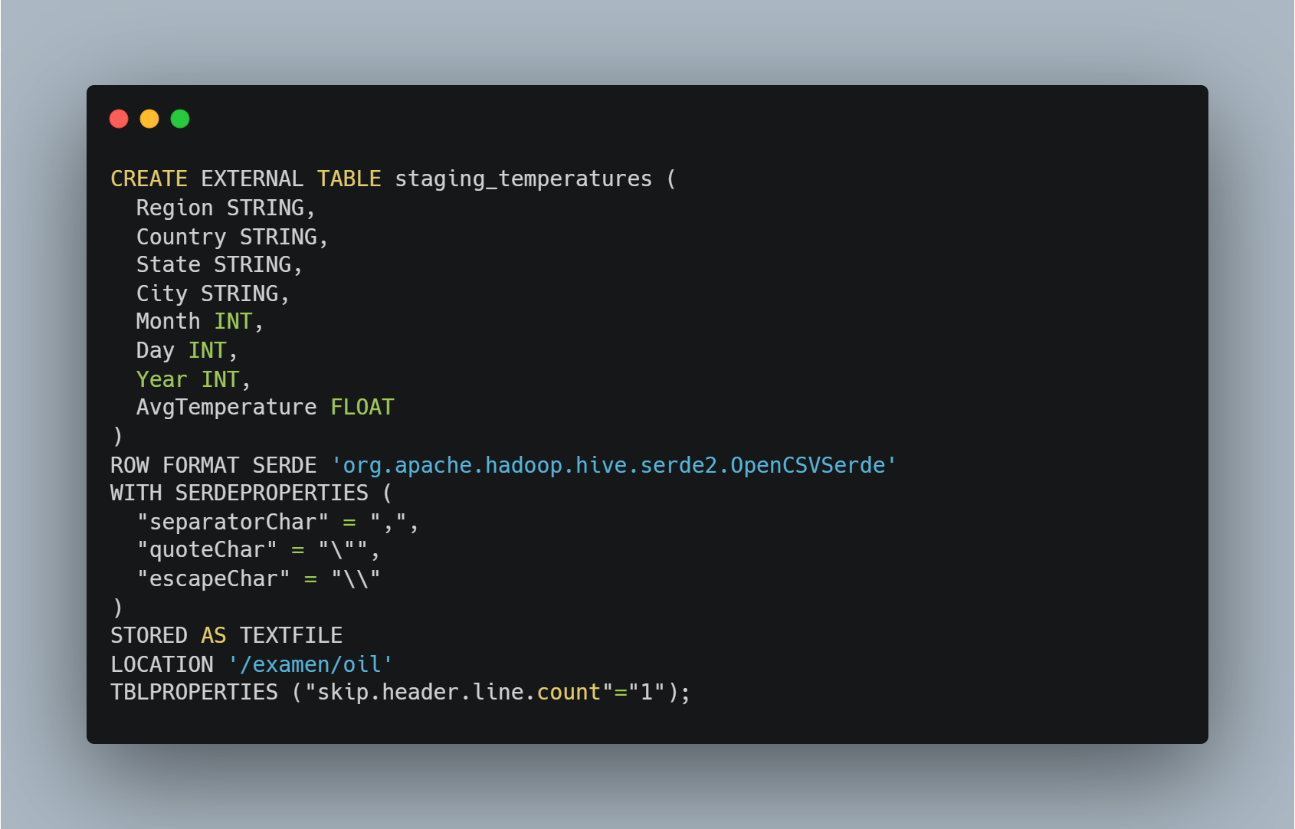
* Mueve los ficheros “palm.csv” y “sunflowerseed.csv” al directorio “/examen/oil” en el clúster hdfs, y repite las consultas del punto 3, guardando las capturas de pantalla.

 Texto

Descripción generada automáticamente

Ejercicio 3. Tablas temporales, formato parquet y particiones

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



* Realiza dos consultas para comprobar los datos:
  + 100 primeras filas.



Pantalla de un video juego

Descripción generada automáticamente con confianza media

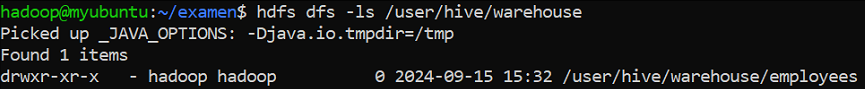


* + 10 primeras filas con valores de AvgTemperature negativos.



No he sido capaz de hacerlo, la consulta me sale vacía

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



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



No he sido capaz de hacerlo, la consulta alcanzaba los 8 minutos.

* 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.

Texto

Descripción generada automáticamente con confianza baja

Texto

Descripción generada automáticamente con confianza baja

Captura de pantalla de computadora

Descripción generada automáticamente



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

Imagen que contiene naranja, oscuro, firmar, gente

Descripción generada automáticamente

Imagen que contiene Texto

Descripción generada automáticamente

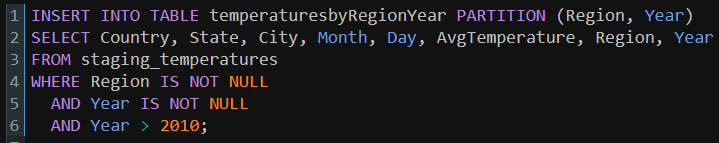
Texto

Descripción generada automáticamente

Texto

Descripción generada automáticamente

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



No he sido capaz de hacerlo, la consulta alcanzaba los 8 minutos.